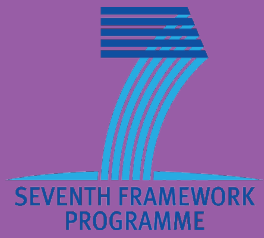
PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Deliverable 6.3
Report on the outcomes of the third year evaluation activities

Version 1.0, 14 August 2013



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

Document Information

Deliverable number:	D6.3
Deliverable title:	Report on the outcomes of the third year evaluation activities
Delivery date:	31/08/2013
Lead contractor for this deliverable	HES-SO
Author(s):	Alba G. Seco de Herrera, Henning Müller, Maria Gäde, Florina Piroi, Theodora Tsikrika
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	25/09/12	Draft	Alba G. Seco de Herrera (HES-SO), Henning Müller (HES-SO) and Maria Gäde (UBER)	First draft
0.2	27/05/13	First version	All authors (several partners)	Inclusion of work of the various partners
1.0	14/08/13	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 the outcomes of the evaluation activities in the third year of PROMISE. PROMISE organizes experimental evaluation activities for multilingual and multimedia information access systems at an international level and on an annual basis. The Cross-Language Evaluation Forum (CLEF) promotes R&D in multilingual information access by developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts. In addition, CLEF creates test-suites of reusable data which can be employed by system developers for benchmarking purposes. Since 2010 CLEF is running in the context of the PROMISE Network of Excellence. Therefore, this report presents the outcomes of the CLEF 2012 conference and labs, with emphases on the three main PROMISE use cases participating (unlocking culture, search for innovation and visual clinical decision support). In addition, an outlook on CLEF 2013 organization is presented.

Table of Contents

Document Information 2
Abstract 3
Table of Contents 4
Executive Summary 6
1 Introduction 8
2 Overview of the third year evaluation activities 9
2.1 CLEF 2012 Conference and Labs 9
2.1.1 CLEF 2012 Conference 9
2.1.2 Participation in the CLEF 2012 Labs 14
2.2 Main advancements 14
2.3 Main trends and experimental outcomes 15
2.4 Main problems 17
3 Outcomes of evaluation activities: CLEF 2012 lab test collections 18
3.1 Collections 18
3.2 Topics 18
3.3 Ground truth 18
4 Outcomes of the evaluation activities for the “Visual Clinical Decision Support” Use Case 20
4.1 Medical Modality Classification Task 21
4.2 Medical Image Retrieval Task 24
4.3 Medical Case Retrieval Task 27
4.4 Summary of the outcomes of the “Visual Clinical Decision Support” Use Case 29
5 Outcomes of the evaluation activities for the “Search for Innovation” Use Case 31
5.1 Passage Retrieval Starting from Patent Claims Task 32
5.2 Flowchart Recognition Task 34
5.3 Chemical Recognition Task 36
5.4 Other Activities in the ‘Search for Innovation’ Use Case 37
5.5 Outcomes Summary in the ‘Search for Innovation’ Use Case 38
6 Outcomes of the evaluation activities for the “Unlocking Culture” Use Case 39
7 Impact Analysis for the CLEF evaluation campaign (2000–2009) 41
7.1 Bibliometric Analysis 42
8 Outlook on future evaluation activities: CLEF 2013 45
9 References 50
Appendix I: Questionnaires sent to CLEF 2012 Labs organizers 53

Appendix II: Participation in the CLEF 2012 labs.....53
Appendix III: Main outcomes of the CLEF 2012 Labs 58
Appendix IV: CLEF 2012 Labs Test Collections 73

Executive Summary

The main outcomes of the evaluation activities in the third year of PROMISE are presented in this deliverable. Mainly it focuses the on the outcomes of the CLEF 2012 conference and labs, with an emphasis on the three main PROMISE use case domains. Finally, an outlook on CLEF 2013 organization is presented.

Evaluation activities in CLEF 2012: Conference and Labs

Conference and Labs of the Evaluation Forum (CLEF) has been an activity of the PROMISE Network of Excellence since 2010. CLEF 2012 consisted of an independent peer-reviewed conference and a set of labs and a workshop which test different aspects of information retrieval systems. This deliverable is presented in the same format as previous deliverables ((Tsikrika, y otros, 2011)and (Piori, y otros, 2012)). First a short overview of the **CLEF 2012 conference** with a short description of the **CLEF 2012 labs** and the participation to them are introduced. The questionnaires sent to the CLEF 2012 lab organizers report the outcomes of the CLEF 2012 labs and form a point of reference for monitoring the evolution and progress of the CLEF labs over the coming years. These results can be summarized as follows:

1. **Tasks**: A total of 25 tasks were investigated in the CLEF 2012 labs: twelve (ad-hoc) information retrieval tasks, four classification tasks, five question answering tasks, while the rest cover of a variety of task, namely annotation, image recognition, expert search and summarization.
2. **Main advancements:** The increasing tendency in the number of benchmarking labs and in the number of their tasks continued in 2012. In 2012 more types of tasks were introduced such as annotation, image recognition and summarization tasks.
3. **Main trends in the participants' approaches**: An analysis over each task is presented due to the high heterogeneity of the tasks. Main trends and experimental outcomes are compared between previous CLEF organized by PROMISE and CLEF 2012.
4. **Main problems**: The main problems were the low participation rates for most of the new tasks and the time needed to generate the data.
5. **Test collections generated by the CLEF 2012 labs**
    a. **Collections:** The CLEF 2012 Labs employed in total 19 collections, one more than in 2011. Most of the collections were used for the first time due to the introduction of new labs and tasks. Indeed only three collections were reused from last years although other four collections used parts of previous years. Hence, PROMISE has distributed an increasing amount of data. Opposite to previous years' tendencies, only eight of the collections were multilingual.
    b. **Topics**: New topics are constantly being created to address new evaluation problems. The number of topics among the tasks varies between 1 and 30,000.
    c. **Ground truth**: Continuing the tendency of 2011, most of the tasks employed human assessors. Ground truth development was still tedious and time-consuming but PROMISE helped to provide the resources to build ground truth data sets. As well, some tasks reused ground truth information and

extended previous work.

Evaluation activities for PROMISE use cases

Three main use cases have been deeply investigated during PROMISE project. Therefore, a detailed description of the outcomes of the evaluation activities carried out by them is presented in this deliverable.

1. **"Visual Clinical decision Support" Use case** (Medical task at ImageCLEF lab)
   a. *In its 9th edition this task reminds one of the most popular tasks in CLEF.* The largest number of runs was submitted for the Image-based Retrieval subtask. However the number of submitted runs at the Modality Classification task increased compared with respect to last year.
   b. *More research is necessary.* There are still different situations as to whether visual, textual or combined techniques perform better depending on the task. Also many groups explored the same or similar descriptors obtaining often quite differing results.
2. **"Search for Innovation" Use Case** (CLEF-IP lab)
   a. *Passage retrieval Starting from Claims task is well formulated*. Patent experts pointed that this task correctly reflects the patentability searches of patent examiner.
   b. *Very good scores obtained in Flow-chart Recognition task.* It shows that treating flow-chart is technically possible to reach a digitalization of flow-charts.
   c. *Not many researchers in the area seem to have interest in this evaluation campaign.*
3. **"Unlocking Culture" Use Case** (CHiC lab)
   a. Six research groups participated in the CHiC pilot lab. In its first year as a lab, 126 runs were submitted.
   b. More work is necessary to improve the quality of the pilot tasks. It is planned to integrate collections in more languages as well as complete Europeana collection. The existing tasks will be updated and two more tasks will be defined.

1 Introduction

PROMISE has advanced in the participative research and experimentation providing a virtual laboratory to carry out, advance and bring automation into the evaluation and benchmarking of complex information systems. PROMISE promotes collaboration and re-use over the acquired knowledge-based and stimulates knowledge transfer and uptake.

The Conference and Labs of the Evaluation Forum (CLEF) conducts these evaluation activities. As in 2010 and 2011, CLEF 2012 and 2013 have been organize in the framework of PROMISE. Therefore, this deliverable reports on the outcomes of the evaluation activities that have taken place during the third year of PROMISE, based 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 past CLEF which were organized by PROMISE and CLEF 2012 are performed based on the material in PROMISE Deliverable 6.1 (Tsikrika, y otros, 2011) and 6.2 (Piori, y otros, 2012). CLEF 2013 will take place after the end of the PROMISE project but it has been organized by PROMISE Network. Consequently, the state of the activities and labs for CLEF 2013 will also be mentioned in this report.

This deliverable is structured following the same format as previous deliverables ((Tsikrika, y otros, 2011)and (Piori, y otros, 2012)). Section 2 aims to provide an overview of the third year evaluation activities by discussing the main outcomes of and the lessons learned from the CLEF 2012 conference and labs. Section 3 emphases the outcomes of these experimental evaluation activities, the CLEF 2012 lab test collections. Sections 4, 5 and 6 focus into the outcomes of the evaluation activities for the three PROMISE Use Cases. Section 7 gives a brief description of the main results of the impact analysis for the CLEF initiative. Section 8 concludes with an outlook on the current status of the CLEF 2013 conference and labs.

2 Overview of the third year evaluation activities

2.1 CLEF 2012 Conference and Labs

CLEF 2012 was the third CLEF conference continuing the popular CLEF campaigns which have run since 2000 contributing to the systematic evaluation of information access systems, primarily through experimentation on shared tasks.

Building on the format first introduced in 2010, CLEF 2012 consisted of an independent peer-reviewed conference on a broad range of issues in the fields of multilingual and multimodal information access evaluation, and a set of labs and workshops designed to test different aspects of mono and cross-language Information retrieval systems. Together, the conference and the lab series maintained and expanded upon the CLEF tradition of community-based evaluation and discussion on evaluation issues.

CLEF 2012 was an activity of the PROMISE Network of Excellence. CLEF 2012 was hosted by the University "La Sapienza" in Rome, Italy, from 17th to 20th September 2012. For further information about CLEF 2012 conference, see (Forner, PROMISE Deliverable 7.9: Third PROMISE Annual Conference and Proceedings, 2012).

2.1.1 CLEF 2012 Conference

While preserving CLEF's traditional core business and goals, namely the benchmarking activities carried in various Tracks, these were complemented with a peer-reviewed conference component on experimental evaluation.

The CLEF 2012 conference aimed at advancing research in the evaluation of complex information systems for cross-language tasks and scenarios, in order to support individuals, organizations, and communities who design, develop, employ, and improve such systems. Experimental evaluation - both laboratory and interactive - is a key to fostering the

development of multilingual and multimodal information systems that address increasingly complex information needs.

Conference proceedings are now published in Springer LNCS Series. The online version is available at http://link.springer.com/book/10.1007/978-3-642-33247-0/page/1.
Table 1 Table 1 lists the CLEF 2012 labs and the tasks organised within each of them. Compared to 2011:

- four benchmarking labs (CLEF-IP, ImageCLEF, PAN and QA4MRE[1]) returned;
- a workshop-style lab (CHiC) became a benchmarking lab;
- two new benchmarking lab (INEX and RepLab) were introduced;
- a new workshop-style (CLEFeHealth) was introduced;
- two CLEF 2011 labs did not return (LogCLEF and MusicCLEF).

Table 1: CLEF 2012 benchmarking and workshop-style labs.

Benchmarking labs, their tasks, and subtasks		
CHiC	Ad-hod Retrieval	
	Variability	
	Semantic Enrichment	
CLEF-IP	Chemical Image Extraction and Recognition	
	Flowchart Recognition	
	Passage Retrieval Starting from Claims	
ImageCLEF	Flickr Photo Annotation and Retrieval	Concept Annotation
		Concept Retrieval
	Medical Image Classification and Retrieval	Ad-hoc image-based retrieval
		Case-based Retrieval
		Modality Classification
	Plant Identification*	
	Pilot Task on Personal Photo Retrieval	Retrieval of Visual Concepts
		Retrieval of Events
	Robot Vision	Task 1
		Task 2
	Scalable Image Annotation Using General Web data	Improving Performance in Flickr Concept Annotation task
		Scalable Concept Image Annotation
INEX	Linked Data	Ad-hoc retrieval
		Faceted Search

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

		Jeopardy
	Relevance Feedback	
	Snippet Retrieval	
	Social Book Search	
	Tweet Contextualization	
PAN	Plagiarism Detection	Candidate Document Retrieval
		Detailed Comparison
	Quality Flaw Prediction in Wikipedia	
	Traditional Authorship Attribution	Traditional Authorship Attribution
		Sexual Predator Identification
QA4MRE	Machine Reading of Biomedical Texts about Alzheimer	
	Processing Modality and Negation	
	Question Answering	
RepLab	Monitoring	
	Profiling	
Workshop-style Lab		
CLEFeHealth	Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis	

Here is a brief description of the CLEF 2012 benchmarking labs:

1 **CHiC (Cultural Heritage in CLEF):** a benchmarking activity to investigate systematic and large-scale evaluation of cultural heritage digital libraries and information access systems (Petras, et al., 2012). In 2012, three sub-tasks were offered: the *Ad-hoc Retrieval* task for measuring information retrieval effectiveness with respect to user input in the form of queries; the *Variability* task which required systems to present a list of objects, which are relevant to the query and should present a particular good overview over the different object types and categories targeted towards a casual user; and the *Semantic Enrichment* task which required systems to present a ranked list of related concepts for a query to semantically enrich the query and / or guess the user's information need or original query intent.

2 **CLEF- IP: a** benchmarking activity to investigate IR techniques in the patent domain, running since 2009 (Piroi, Lupu, Hanbury, Sexton, Magdy, & Filippov, 2012). Three tasks were offered in 2012: the *Chemical Image Extraction and Recognition* task for identifying the location of the chemical structures depicted on patent pages and, for each of them, return the corresponding structure in a MOL file (a chemical structure file format); the *Flowchart Recognition* task for extracting the information in patent images representing flow-charts and returning it in a predefined textual format; and the *Passage Retrieval Starting from Claims* task for retrieving relevant documents in the collection and mark out the relevant passages in patent application documents.

3 **ImageCLEF:** a benchmarking activity on the experimental evaluation of image classification and retrieval, focusing on the combination of textual and visual

evidence, running at CLEF since 2004. Six tasks were offered in 2012: the *Flicker Photo Annotation and Retrieval* task (Thomee & Popescu, 2012) for analyzing a collection of Flickr photos in terms of their visual and/or textual features in order to detect the presence of one or more concepts and then for automatically annotating the images or for retrieving the best matching images to a given concept-oriented query; the *Medical Image Classification and Retrieval* task (Müller, García Seco de Herrera, Kalpathy-Cramer, Demmer Fushman, Antani, & Eggel, 2012) 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 *Plant Identification task* (Goëau H. , y otros, 2012) for tree species identification based on leaf images; the *Pilot task on Personal Photo Retrieval* (Zellhoefer, 2012) for providing a test bed for QBE-based retrieval scenarios in the scope of personal information retrieval based on a collection of 5,555 personal images plus rich metadata; the *Robot Vision* task (Martinez-Gomez, Garcia-Varea, & Caputo, 2012) for visual place classification, this time with the use of images acquired with the Kinect depth sensor; and the task (Villegas & Paredes, 2012) for scalable image annotation using as training a collection of automatically obtained Web images.

4 **INEX:** a benchmarking activity on the evaluation of XML retrieval. Five tasks were offered in 2012: the *Linked Data* task (Wang, y otros, 2012) which aims to close the gap between IR-style keyword search and Semantic-Web-style reasoning techniques; the *Relevance Feedback* task (Chappell & Geva, 2012) which covers the use of focused feedback, a relevance feedback model wherein users specify segments of the document considered relevant to the search topic; the *Snippet Retrieval* task (Trappett, Geva, Trotman, Scholer, & Sanderson, 2012) for generating informative snippets for search results; the *Social Book Search* task (Koolen, Kazai, Kamps, Preminger, Doucet, & Landoni, 2012) for supporting users in searching and navigating the full texts of digitized books and complementary social media as well as providing a forum for the exchange of research ideas and contributions; and the *Tweet Contextualization* task (Sanjuan, Moriceau, Tannier, Bellot, & Mothe, 2012) for providing some context about the subject of a given tweet, in order to help the reader to understand it.

5 **PAN:** a benchmarking activity on uncovering plagiarism, authorship and social software misuse, running at CLEF since 2010. Three tasks were offered in 2012: the *Plagiarism Detection* task (Potthast, y otros, 2012) for external plagiarism detection; the *Quality Flaw Prediction in Wikipedia* task (Anderka & Stein, 2012) for deciding whether an untagged Wikipedia article suffers from a particular quality flaw; the *Traditional Author Identification* task (Juola, 2012) for determining the authorship of anonymous documents based on internal evidence.

6 **QA4MRE:** a benchmarking activity on the evaluation of Machine Reading systems through Question Answering and Reading Comprehension Tests. Three tasks were offered in 2012: the *Machine Reading of Biomedical Texts about Alzheimer* task (Morante, Krallinger, Valencia, & Daelemans, 2012) exploring the ability of a machine reading system to answer questions about a scientific topic, namely Alzheimer's disease; the *Processing Modality and Negation* task (Morante & Daelemans,

Annotating Modality and Negation for a Machine Reading Evaluation, 2012) for determining whether an event mentioned in a text is presented as negated, modalized (i.e. affected by an expression of modality), or both; and the *Question Answering* task (Peñas, y otros, Overview of QA4MRE at CLEF 2012: Question Answering for Machine Reading Evaluation, 2012) for answering multiple-choice questions on documents concerned with four different topics.

7 **RepLab:** a benchmarking activity on reputation management technologies (Amigó, Corujo, Gonzalo, Meij, & de Rijke, 2012). Two tasks were offered in 2012: the *Monitoring* task for clustering the most recent tweets thematically, and assigning relative priorities to the clusters; and the *Profiling* task for annotation of the ambiguity and the polarity for reputation on tweets.

The CLEF 2012 benchmarking labs consisted of 25 tasks (see Table 1), 8 more than CLEF 2011. In 2012 the types of the tasks were:

- 9 (Ad-hoc) Information Retrieval tasks
- 3 Classification tasks
- 3 Question Answering tasks
- 3 Annotation tasks
- 2 Image Recognition tasks
- 1 Expert Search
- 2 both Classification and Retrieval tasks
- 1 both Question Answering, Retrieval and Summarization task
- 1 both Question Answering and Retrieval task.

The increasing tendency in the number of benchmarking labs and in the number of their tasks continued in 2012 (see Table 2). Compared to 2011, log analysis tasks were not offered in 2012. In 2012 other types of tasks were introduced such as annotation, image recognition and summarization tasks. Classification, information retrieval, question answering and expert search continued to be offered in CLEF2012.

Table 2: Evolution of the number of CLEF Labs and tasks

	2010	2011	2012
No. Labs	5	6	7
No. Tasks	11	17	25

The following workshop-style lab was also held at CLEF 2012:

- **CLEFeHealth:** workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis. It was organized as speaking and discussion session to explore issues of evaluation methodology, metrics, and processes in information access and closely related fields.

2.1.2 Participation in the CLEF 2012 Labs

The participation in the CLEF 2012 is described in more the detail in (Forner, PROMISE Deliverable 7.9: Third PROMISE Annual Conference and Proceedings, 2012). Overall, CLEF 2012 was attended by almost 200 people from different academic and industrial institutions; the attendance has increased with respect to last year confirming the positive trend started since 2010. More than 280 groups initially registered in the Labs showing interest in the benchmarking activities proposed this year at CLEF 2012. This proves, once again that CLEF has achieved a high visibility.

ImageCLEF was still the most popular Lab able to attract many participants not only from Europe but also from the United States and a few from other continents. In 2012, also the PAN lab experienced great popularity and success and was able to reach wide research communities growing considerably with respect to last year. Also the Question Answering for Machine Reading Evaluation attracted a considerable number of registrations and participations. The CLEFeHealth workshop, newly proposed this year, managed to draw the attention of a considerable number of researches.

About 180 groups, both researchers and system developers, submitted runs to the different Labs, despite the number of registered participants to the benchmarking activities was much higher than that. It is anyway a very good result also if we compare it to 2011 the number of participants taking part to the labs has almost doubled (95 institutions participated in 2011).

Return participations from the previous year are on average around 70%, an increase compared to previous years (50% in 2011 and 40% in 2010), indicating that a large number of researchers rely year after year on the resources created in the context of the CLEF evaluation activities. All the tasks that ran in previous CLEF editions had return participants. As in 2011, in QA4MRE lab the 100% of the participants also participated in previous years.

The number of submissions varies greatly per task, with an average of 43 a bit less than in 2011 (49).

ImageCLEF stand out with 519 submissions. A notable case is CHiC with 126 submissions, high number for a new lab. Two groups used DIRECT (CHiC and one task of CLEF IP) although only CHiC used it for the entire evaluation procedure, including experiments submissions, relevance assessments and metric computation.

2.2 Main advancements

As in 2011, CLEF 2012 introduced a considerable number of new tasks. Although 4 CLEF 2011 labs continued in CLEF 2012, only 6 tasks remained the same. Table 18 in Appendix III: Main outcomes of the CLEF 2012 Labs, presents the main differences between CLEF 2011 and 2012 as pointed out by the task organizers.

Again, many of the tasks employed larger collections, either by updating existing collections, adding new elements or creating new ones (Chemical Image Extraction and Recognition CLEF-IP task, Flicker Photo Annotation and Retrieval, Medical Image Classification and Retrieval and Plant Identification ImageCLEF tasks, Plagiarism Detection PAN task, Question Answering QA4MRE task). In contrast, the Traditional Authorship Attribution task, from PAN lab, used a much smaller corpus in 2012 since in 2011 the large

corpus was considered to create impracticalities for many participants.

The efforts to make the task more realistic continued in 2012, not only improving the collections, but also modifying topics (Flicker Photo Annotation and Retrieval ImageCLEF task, Plant Identification ImageCLEF task, Tweet Contextualization INEX task, Question Answering QA4MRE task,...) and even modifying the hierarchical classification (Medical Image Classification and retrieval ImageCLEF task).

Plagiarism Detection PAN task used a new experimentation platform for software submission. Also a new performance measurement was applied in order to further push the limits of evaluating plagiarism detectors.

2.3 Main trends and experimental outcomes

Table 19 in Appendix III: Main outcomes of the CLEF 2012 Labs presents the main trends among the participants' approaches, as well as the main experimental outcomes based on the participants' results. As in previous deliverables, we will present the analysis over each task:

1. *CHiC:* Most groups focused on monolingual tasks. For the Ad-hoc Retrieval task, most groups used open information retrieval systems. Translations for bilingual and multilingual tasks were produced with open source solutions like Google Translate, Wikipedia entries (with associated translations) and Microsoft's translation service. Semantic enrichments were mainly derived from Wikipedia.
2. *CLEF-IP Chemical Image Extraction and Recognition*: A set of heuristics was the key element to solve the disambiguating between the elements present in the images, and the one with the better heuristics obtained the best results.
3. *CLEF-IP Flowchart recognition*: The three participants used diverse methods but generally they agree in the order of identifying different components (first graphs and then textual contents). Overall participants did a surprisingly good job, representing over 80% of the original graphs. Next year the flowcharts will be more complicated in the test collection.
4. *CLEF-IP Passage retrieval starting from claims*: The solutions chosen by the submitting participants range from two-step retrieval approaches, namely a document level retrieval in the first step and a passage level retrieval in the second step to using Natural Language Processing techniques and trigrams to extract relevant passages. All participants have used translation tools on the generated queries.
5. *ImageCLEF Flickr Photo Annotation and Retrieval:* Bag of textual/visual words as well as SVMs were still very popular but did not guarantee good performances. On the other hand, Fisher vectors, soft coding, optimal fusion, semantic contextualization of tags have led to good results.
6. *ImageCLEF Medical Image Classification and Retrieval:* The main trend was the use of Lucene, concept-based approaches and multiple visual features. Visual, textual or mixed runs perform differently based on the subtasks so it is difficult to conclude which method is better. Although it is clear that the expansion of the training set (introduced in CLEF 2011) and the use of multiple visual features were successful.
7. *ImageCLEF Plant Identification:* Most of the participants focussed on interactive

segmentation, shape boundary features or more generic approaches (SVMs, sparse coding...). Fully automatic identification from unconstrained photographs was still very challenging. Performances on leaf scans were correct although the increased number of species already shows the limit of a leaf-based only system.

8. *ImageCLEF Pilot Task on Personal Photo Retrieval:* Only one group decided to exploit the browsing data instead of the provided metadata. Surprisingly, there was no interest in solving the so-called user-centered initiative of the subtasks.
9. *ImageCLEF Robot Vision:* The main trend was the used of Fisher Vectors and SVM-based. As a result, frame by frame recognition worked reasonably well.
10. *ImageCLEF Scalable Image Annotation using General Web Data:* Most participants relied on the use of online learning methods that scale well to large datasets and are able to handle noisy data. For some concepts, the annotation systems based on automatic data had a comparable performance than systems based on manually labelled data.
11. *INEX Linked Data:* Most participants used traditional IR approaches. DB approaches employed by the participants performed much worse.
12. *INEX Relevance Feedback:* There was relatively little, if any, advantage or disadvantage from using the full document collection rather than the pool. Although more work needs to be done, participants were still able to find the relevant documents in the full collection.
13. *INEX Snippet Retrieval:* The participation was too little to determinate any output.
14. *INEX Social Book Search:* The most effective systems incorporated the full topic statement, which included the title of the topic thread, the name of the discussion group, and the full first message that elaborates on the request.
15. *INEX Tweet Contextualization:* Participants that combined simple Indri query language features with state of the art Part of Speech tagging and summarization tools clearly out-performed pure single approaches based on advanced focused IR or fast summarization algorithms for large data. Many participants used Anaphora resolution but it did not help readability. Although sentence reordering based on anaphora detection did. Tweet reformulation based on local LDA also improved results.
16. *PAN Plagiarism Detection:* For candidate retrieval, an analysis of the participants' notebooks reveals a number of building blocks that were commonly used to build candidate retrieval algorithms: chunking, keyphrase extraction, query formulation, search control and download filtering. For text alignment, comparison algorithm was commonly used such as seeding, match merging and extraction filtering.
17. *PAN Traditional Authorship Attribution:* Most of the participants used ensemble methods; they used multiple classifiers and average them. Participants can be really good at this task with enough training data.
18. *PAN Quality Flaw Prediction in Wikipedia:* Three quality flaw classifiers have been developed, which employ a total of 105 features to quantify the ten most important quality flaws in the English Wikipedia. Two classifiers achieve promising performance for particular flaws.
19. *QA4MRE Machine Reading of Biomedical Texts about Alzheimer:* Most teams applied text similarity methods. The background collection was used by most teams

and it seems to be necessary. Index expansion techniques work well for the task and, on the other hand, simple text similarity techniques do not suffice to perform the task.

20. *QA4MRE Processing modality and negation:* Rule-based systems were built based on linguistic knowledge and they obtained good results in general.
21. *QA4MRE Question Answering:* Participants prefer to perform ranking of answers and selection of the most promising one instead of validation of them, which is the purpose of the current setting.
22. *RepLab Monitoring:* Systems are not substantially contributing to solve the problem yet.
23. *RepLab Profiling: Most of the participants focus on s*entiment polarity detection software adapted to the reputation scenario and/or the textual source (tweets). The profiling task is far from being solved automatically.

2.4 Main problems

Table 18 in Appendix III: Main outcomes of the CLEF 2012 Labs presents the main problems from the organizational point of view. For most of the new tasks, the main issue was the low participation rates. Also the Medical Image Classification and retrieval ImageCLEF task has a low participation rate compared to the number of registrations. Finally, the Robot Vision ImageCLEF task finds difficulties attracting participants to submit Working Note papers. The PROMISE framework can help to increase the overall participation promoting evaluation tasks.

The other main significant problem is the time needed to generated data. Some labs and tasks, such as CHiC, Flowchart Recognition CLEF-IP or Flicker Photo Annotation and Retrieval ImageCLEF, also found difficulties identifying appropriates evaluation measures.

3 Outcomes of evaluation activities: CLEF 2012 lab test collections

3.1 Collections

The CLEF 2012 Labs employed in total 19 collections, one more than in 2011. A description of each collection and some statistics are presented in Appendix IV: CLEF 2012 Labs Test Collections.

Only 4 out of the 19 collections (~21%) were not created primarily for the labs. In 2012, most of the collections (~84%) were used for the first time, a big increase from last years (53% in 2010 and 66% in 2011). Only three collections were reused from last years although other four collections used parts of previous years. The MIRFLICKR collection is the oldest collection in CLEF; it has been used for 4 years in Flickr Photo Annotation and Retrieval ImageCLEF task.

Differently to previous years' tendencies, only eight of the collections were multilingual (~42% in 2012 vs. 72% in 2011). One task (Plant Identification ImageCLEF task) was totally language independent and all of the ImageCLEF tasks on multimedia retrieval were language independent by nature.

As in 2011, the size of the collections and the number of documents they contain vary widely, with the size ranging between 2MB and 132GB and the number of documents between 8 and ~23 million.

3.2 Topics

A further description of the topics can be found in Table 21 in Appendix IV: CLEF 2012 Labs Test Collections.

The number of topics varies between 1 and 30,000. Most of the tasks provided less than one hundred topics and just a few tasks provided more than 1,000. In general, the languages of the topic are consistent with the language of the dataset, although the Medical Image Classification and Retrieval ImageCLEF task provided the topics in English, Spanish, French and German while the corpus is mainly in English.

3.3 Ground truth

Table 22 in Appendix III: Main outcomes of the CLEF 2012 Labs briefly presents the process for the ground truth generation followed in each of the CLEF 2012 tasks and also provides estimates on the applied human effort.

Ground truth generation is tedious and time-consuming. Some tasks reused ground truth information and extended previous work. As in 2011, most of the tasks employed human assessors, mostly volunteers such as task organizers, students or participants. Quality Flaw Prediction in Wikipedia PAN task made used of Wikipedia users. Some tasks also used external expert assessors, e.g., physician medical doctors for the Medical Image Classification and Retrieval ImageCLEF task, members of the Telebotanica social network for the Plant Identification ImageCLEF task or reputation management experts for the

RepLab.

Finally, some other tasks employed crowd workers such as Amazon's Mechanical Turk.

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

The Medical Image Classification and Retrieval task in 2012 is the evaluation activity for the Visual Clinical Decision Support PROMISE use case. This ImageCLEF tasks is supported by the project. Further information about this task can be found at (Müller, García Seco de Herrera, Kalpathy-Cramer, Demmer Fushman, Antani, & Eggel, 2012).This task covered image modality classification and image retrieval with visual, semantic and mixed topics in several languages using a data collection from the biomedical literature. In 2012, there were three types of tasks in the medical image classification and retrieval task:

- Modality Classification
- Image-based Retrieval
- Case-based Retrieval

In ImageCLEFmed 2012, a larger database than the 2011 one was provided using the same types of images and the same journals. The database contains over 300,000 images of 75'000 articles of the biomedical open access literature that allow free redistribution of the data. The ImageCLEF database is a subset of the PubMedCentral[2] database containing in total over 1.5 million images. PubMedCentral contains all articles in PubMed that are open access but the exact copyright policy for redistribution varies among the journals.

In total over 60 groups registered for the medical tasks and obtained access to the data sets. ImageCLEF in total had over 200 registrations in 2012, with a bit more than 30% of the groups submitting results. 17 of the registered groups submitted results to the medical tasks, the same number as in previous years.

Table 3 shows the groups which submitted at least one run.

Table 3: ImageCLEFmed participants in 2012 (participants marked with a star had not participated in the medical task in 2011)

Research Group	Country
Bioingenium, National University of Colombia*	Colombia
BUAA AUDR, BeiHang University	China
DEMIR, Dokuz Eylul University	Turkey
ETFBL, Faculty of Electrical Engineering Banja Luka	Bosnia and Herzegovina
FINKI, University in Skopje*	Macedonia
GEIAL, General Electric Industrial Automation Limited*	United States
IBM Multimedia Analytics*	United States
IPL, Athens University of Economics and Business	Greece

[2] http://www.ncbi.nlm.nih.gov/pmc/

ITI, Image and Text Integration Project, NLM*	United States
LABERINTO, Universidad de Huelva	Spain
Lambdasfsu, San Francisco State University*	United States
medGIFT, University of Applied Sciences Western Switzerland	Switzerland
IRACL, Higher Institute of Computer Science and Multimedia of Sfax*	Tunisia
MRIM, Laboratoire d'Informatique de Grenoble	France
ReDCAD (National School of Engineering of Sfax*	Tunisia
UESTC, University of Electronic Science and Technology	China
UNED–UV, Universidad Nacional de Educación a Distancia and Universitat de València	Spain

A total of 202 valid runs were submitted, 43 of which were submitted for modality detection, 122 for the image-based topics and 37 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 studies have shown that imaging modality is an important information on the image for medical retrieval. In user-studies, clinicians have indicated that modality is one of the most important filters through which they would like to re-fine their search. Many image retrieval websites (Goldminer, Yottalook) allow users to limit the search results to a particular modality. Using the modality information, the retrieval results can often be improved significantly. An improved ad–hoc hierarchy with 31 classes in the sections compound or multipane images, diagnostic images and generic biomedical illustrations was created based on the existing data set (Figure 1).
The class codes with descriptions are the following ([Class code] Description):

- [COMP] Compound or multipane images (1 category)
- [Dxxx] Diagnostic images:
  - [DRxx] Radiology (7 categories):
    - [DRUS] Ultrasound
    - [DRMR] Magnetic Resonance
    - [DRCT] Computerized Tomography
    - [DRXR] X–Ray, 2D Radiography
    - [DRAN] Angiography
    - [DRP E] PET
    - [DRCO] Combined modalities in one image
  - [DV xx] Visible light photography (3 categories):
    - [DV DM] Dermatology, skin
    - [DV EN] Endoscopy
    - [DV OR] Other organs
  - [DSxx] Printed signals, waves (3 categories):
    - [DSEE] Electroencephalography

    - [DSEC] Electrocardiography
    - [DSEM] Electromyography
  - [DMxx] Microscopy (4 categories):
    - [DMLI] Light microscopy
    - [DMEL] Electron microscopy
    - [DMT R] Transmission microscopy
    - [DMF L] Fluorescence microscopy
  - [D3DR] 3D reconstructions (1 category)
  - [Gxxx] Generic biomedical illustrations (12 categories):
    - [GT AB] Tables and forms
    - [GP LI] Program listing
    - [GF IG] Statistical figures, graphs, charts
    - [GSCR] Screenshots
    - [GF LO] Flowcharts
    - [GSY S] System overviews
    - [GGEN] Gene sequence
    - [GGEL] Chromatography, Gel
    - [GCHE] Chemical structure
    - [GMAT] Mathematics, formulae
    - [GNCP] Non–clinical photos
    - [GHDR] Hand–drawn sketches

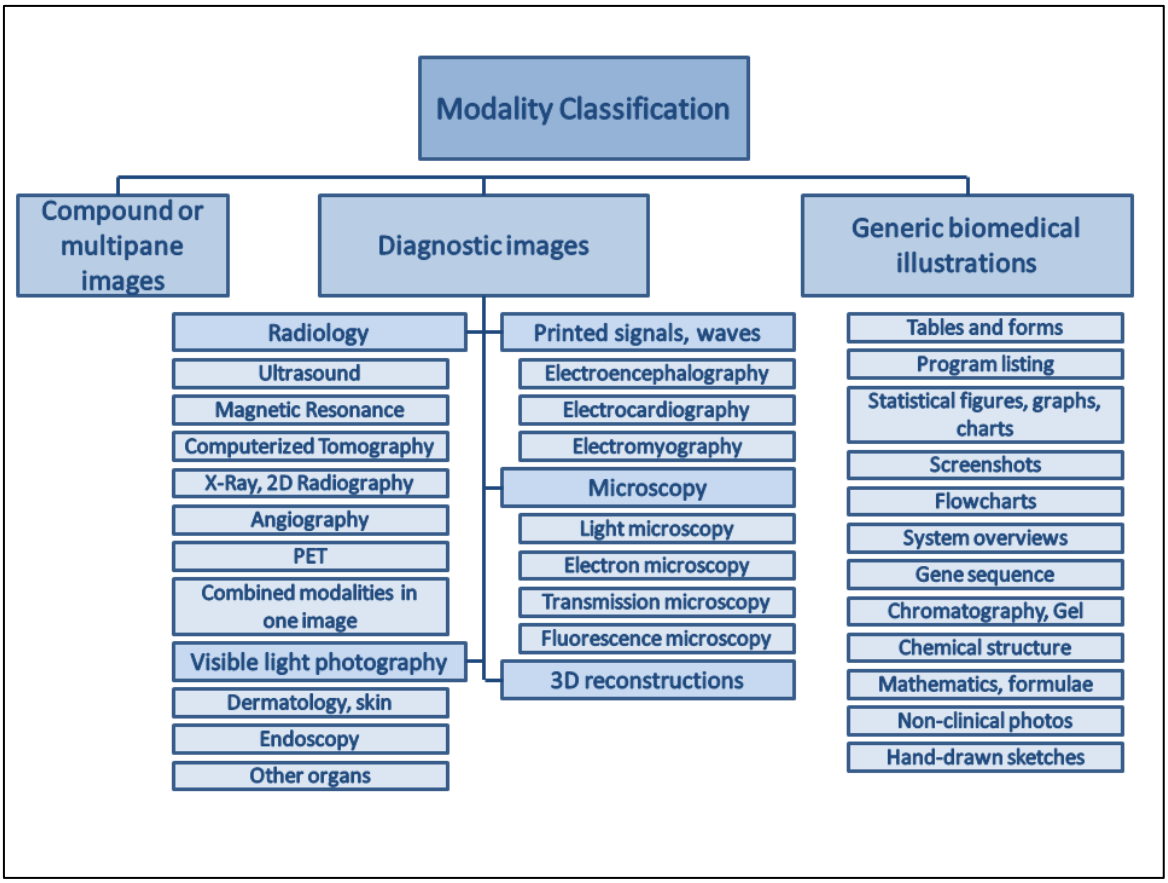


Figure 1: Modality categories of the ImageCLEF 2012 Medical Modality Classification task.

For this hierarchy 1,000 training images and 1,000 test images were provided to the participants. Labels for the training images were known whereas labels for the test images were distributed after the results submission, only.

The results of the modality classification task were compared using classification accuracy. With a higher number of classes, this task was more complex than in previous years. As seen in Table 4, the best result was obtained by the IBM Multimedia Analytics [13] group using visual methods (69.6%). In previous years combining visual and textual methods most often provided the best results. The best run using visual methods had a slightly better accuracy than the best run using mixed methods (66.2%) by the medGIFT group [14]. Only a single group submitted text–based results that performed worse than the average of all runs. The best run using textual methods alone obtained a much lower accuracy (41.3%).

Table 4: Top-10 results per run type for the 2012 ImageCLEF Medical Modality Classification task.

Run Id	Group	Run Type	Classification Accuracy
medgift–nb–mixed–reci–14–mc	medGIFT	Mixed	66.2

medgift–orig–mixed–reci–7–mc	medGIFT	Mixed	64.6
medgift–nb–mixed–reci–7–mc	medGIFT	Mixed	63.6
Visual Text Hierarchy w Postprocessing 4 Illustration	ITI	Mixed	63.2
Visual Text Flat w Postprocessing 4 Illustration	ITI	Mixed	61.7
Visual Text Hierarchy	ITI	Mixed	60.1
Visual Text Flat	ITI	Mixed	59.1
medgift–b–mixed–reci–7–mc	medGIFT	Mixed	58.8
Image Text Hierarchy Entire set	ITI	Mixed	44.2
IPL MODALITY SVM LSA BHIST 324segs 50k WithTextV	IPL	Mixed	23.8
Text only Hierarchy	ITI	Textual	41.3
Text only Flat	ITI	Textual	39.4
preds Mic Combo100Early MAX extended100	IBM	Visual	69.6
LL fusion nfea 20 rescale	IBM	Visual	61.8
preds Mic comboEarly regular	IBM	Visual	57.9
UESTC–MKL3	UESTC	Visual	59.8
UESTC–MKL2	UESTC	Visual	56.6
UESTC–MKL5	UESTC	Visual	55.9
UESTC–MKL6	UESTC	Visual	55.9
NCFC ORIG 2 EXTERNAL SUBMIT	IBM	Visual	52.7
UESTC–SIFT	UESTC	Visual	52.7
Visual only Hierarchy	ITI	Visual	51.6

4.2 Medical Image Retrieval Task

The image-based retrieval task is the classic medical retrieval task, similar to the tasks organized from 2004 to 2011 where the query targets are single images. Participants were given a set of 22 textual queries (in English, Spanish, French and German) with 1-7 sample images for each query. The queries were classified into textual, mixed and semantic queries, based on the methods that are expected to yield the best results. The topics for the image-based retrieval task were based on a selection of queries from search logs of the Goldminer radiology image search system. Only queries occurring 10 times or more (about 200 queries) were considered as candidate topics for this task. A radiologist assessed the importance of the candidate topics, resulting in 50 candidate topics that were checked for at least occurring a few times in the database. The resulting 22 queries were then distributed among the participants and example query images were selected from a past collection of ImageCLEF.

13 teams submitted 36 visual, 54 textual and 32 mixed runs for the image-based retrieval

task (Table 5, Table 6 and Table 7). The best result in terms of mean average precision (MAP) was obtained by ITI using multimodal methods. The second best run was a purely textual run submitted by Bioingenium. As in previous years, visual approaches achieved much lower results than the textual and multimodal techniques.

Table 5: Top-10 results of the multimodal runs for the 2012 ImageCLEF Medical Image Retrieval task.

Run Id	Group	MAP	GM-MAP	bPref
nlm-se	ITI	0.2377	0.0665	0.2542
Merge_RankToScore_weighted	ITI	0.2166	0.0616	0.2198
mixedsum(CEDD,FCTH,CLD)+1.7TFIDFmax2012	DEMIR	0.2111	0.0645	0.2241
mixedFCTH+1.7TFIDFsum2012	MedGIFT	0.2085	0.0621	0.2204
medgift-ef-mixed-mnz-ib	DEMIR	0.2005	0.0917	0.1947
mixedCEDD+1.7TFIDFsum2012	ITI	0.1954	0.0566	0.2096
nlm-lc	ITI	0.1941	0.0584	0.1871
nlm-lc-cw-mf	ITI	0.1938	0.0413	0.1924
nlm-lc-scw-mf	ITI	0.1927	0.0395	0.194
nlm-se-scw-mf	ITI	0.1914	0.0206	0.2062

Table 6: Top-10 results of the textual runs for the 2012 ImageCLEF Medical Image Retrieval task.

Run Id	Group	MAP	GM-MAP	bPref
UNAL	Bioingenium	0.2182	0.082	0.2173
AUDR_TFIDF_CAPTION[QE2]_AND_ARTICLE	BUAA AUDR	0.2081	0.0776	0.2134
AUDR_TFIDF_CAPTION[QE2]_AND_ARTICLE	BUAA AUDR	0.2016	0.0601	0.2049
IPL_A1T113C335M1	IPL	0.2001	0.0752	0.1944
IPL_A10T10C60M2	IPL	0.1999	0.0714	0.1954
TF_IDF	DEMIR	0.1905	0.0531	0.1822
AUDR_TFIDF_CAPTION_AND_ARTICLE	BUAA AUDR	0.1891	0.0508	0.1975
IPL_T10C60M2	IPL	0.188	0.0694	0.1957
AUDR_TFIDF_CAPTION[QE2]	BUAA AUDR	0.1877	0.0519	0.1997
TF_IDF	DEMIR	0.1865	0.0502	0.1981

Table 7: Top-10 results of the visual runs for the 2012 ImageCLEF Medical Image Retrieval task.

Run Id	Group	MAP	GM-MAP	bPref
RFBr23+91qsum(CEDD,FCTH,CLD)max2012	DEMIR	0,0101	0,0004	0,0193

IntgeretedCombsum(CEDD,FCTH,CLD)max	DEMIR	0,0092	0,0005	0,019
unal	Bioingenium	0,0073	0,0003	0,0134
FOmixedsum(CEDD,FCTH,CLD)max2012	DEMIR	0,0066	0,0003	0,0141
edCEDD&FCTH&CLDmax2012	DEMIR	0,0064	0,0003	0,0154
medgift-lf-boc-bovw-mnz-ib	MedGIFT	0,0049	0,0003	0,0138
Combined_LateFusion_Fileterd_Merge	ITI	0,0046	0,0003	0,0107
FilterOutEDFCTHsum2012	DEMIR	0,0042	0,0004	0,0109
finki	FINKI	0,0041	0,0003	0,0105
EDCEDDSUMmed2012	DEMIR	0,004	0,0003	0,0091

4.3 Medical Case Retrieval Task

The case–based retrieval 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, 30 case descriptions with patient demographics, limited symptoms and test results including imaging studies were provided (but not the final diagnosis). The goal was to retrieve cases including images that a physician would judge as relevant for differential diagnosis. Unlike the ad–hoc task, the unit of retrieval here was a case, not an image. The topics were created form an existing medical case database. Topics included a narrative text and several images.

In 2012, 37 runs were submitted in the case–based retrieval task (**Table 8**,

Table 9 and Table 10). As in previous years most of them were textual runs. Only the medGIFT team submitted visual and multimodal case–based retrieval runs. Although textual runs achieved the best results, a mixed approach performs better than the average of all

submitted runs in this task. Visual runs do not perform as well as most of the textual retrieval runs.

Table 8: Results of the multimodal runs for the 2012 ImageCLEF Medical Case Retrieval task.

Run Id	Group	MAP	GM-MAP	bPref
medgift-ef-mixed-mnz-cb	medGIFT	0,1017	0,0175	0,0857
medgift-ef-mixed-reci-cb	medGIFT	0,0514	0,009	0,0395

Table 9: Top-10 results of the textual runs for the 2012 ImageCLEF Medical Case Retrieval task.

Run Id	Group	MAP	GM-MAP	bPref
HES-SO-VS_FULLTEXT_LUCENE	Bioingenium	0,169	0,0374	0,1499
LIG_MRIM_CB_FUSION_DIR_W_TA_TB_C	BUAA AUDR	0,1508	0,0322	0,1279
LIG_MRIM_CB_FUSION_JM07_W_TA_TB_C	BUAA AUDR	0,1384	0,0288	0,11
UESTC_case_f	IPL	0,1288	0,025	0,1092
UESTC-case-fm	IPL	0,1269	0,0257	0,1117
LIG_MRIM_CB_TFIDF_W_DintQ	DEMIR	0,1036	0,0167	0,077
nlm-lc-total-sum	BUAA AUDR	0,1035	0,0137	0,1053

nlm-lc-total-max	IPL	0,1027	0,0125	0,1055
nlm-se-sum	BUAA AUDR	0,0929	0,013	0,0738
nlm-se-max	DEMIR	0,0914	0,0128	0,0736

Table 10: Results of the visual runs for the 2012 ImageCLEF Medical Case Retrieval task.

Run Id	Group	MAP	GM-MAP	bPref
medgift-lf-boc-bovw-reci-IMAGES-cb	medGIFT	0,0366	0,0014	0,0347
medgift-lf-boc-bovw-mnz-IMAGES-cb	medGIFT	0,0302	0,001	0,0293
baseline-sift-early-fusion-cb	medGIFT	0,0016	0	0,0032
baseline_sift_late_fusion_cb	medGIFT	0,0008	0	0
medgift-ef-boc-bovw-reci-IMAGES-cb	medGIFT	0,0008	0,0001	0,0007
medgift-ef-boc-bovw-mnz-IMAGES-cb	medGIFT	0,0007	0	0

4.4 Summary of the outcomes of the "Visual Clinical Decision

Support" Use Case

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

1. As in previous years, the largest numbers of runs were submitted for the Image–based Retrieval task. However, in 2012 there were 122 runs in this task, eight less than in 2011.
2. For the Case–based Retrieval task the number of runs also decreased to 37 (43 in 2011).
3. The number of submitted runs at the Modality Classification task increased to 43 (34 in 2011).
4. There are still different situations as to whether visual, textual or combined techniques perform better depending on the task.
   i. For the Modality Classification, a visual run achieved the best accuracy using training data extension.
   ii. For the Image–based Retrieval task, multimodal runs obtained best results.
   iii. For the Case–based Retrieval task textual runs obtained the best results.
5. In 2012 some groups reused techniques applied by the best group in 2011
6. Many groups explored the same or similar descriptors obtaining often quite differing results.

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

To encourage the disclosure of ideas and inventions to the society, governmental and intergovernmental organizations (patent offices) ensure that the inventors are given and can make use of exclusivity rights for a given period of time. These exclusive rights are commonly known as 'patents'. Granted patents may have important economic consequences, therefore specific searches during the examination of patent applications are very thorough. The "Search for Innovation" use case we benchmark various aspects of the specific searches performed by experts in intellectual property.

In 2012 CLEF-IP organized 3 tasks:

- Passage retrieval starting from claims (Claims to passages, CLM)
- Flowchart recognition (FC)
- Chemical structure recognition (CS)

The flowchart and the chemical structure recognition tasks were organized with support from the ImageCLEF lab organizers. Besides them, the institutions involved in organizing the CLEF-IP tasks in 2012 were:

- Institute of Software Technology and Interactive Systems, Vienna University of Technology, Austria
- Qatar Computing Research Institute, Qatar
- School of Computer Science, University of Birmingham, UK
- Chemical Biology Laboratory, SAIC-Frederick Inc. USA

A number of 28 participants registered to participate in the lab, with 13 teams (see Table 11) actually submitting a total of 51 experiments.

Table 11: CLEF-IP registered participants in 2012.

Institution	Country
University of Hildesheim, Information Science	Germany
Radboud University Nijmegen	Netherlands
University of Lugano	Switzerland
University of Birmingham	UK
Inria	France
Humbold University, Dept. of German Language and Linguistics	Germany
Joanneum Research Forschungsgesellschaft mbH, Institute of Information and Communication Technologies	Austria
Computer Vision Center, Universitat Autonoma de Barcelona	Spain
University of Applied Sciences, Information Studies, Geneva	Switzerland
Chemnitz University of Technology, Department of Computer	Germany

Science	
University of Wolverhampton, School of Technology	UK
Vienna University of Technology, Inst. f. Software Technology and Interactive Systems	Austria
Univ. of Macedonia, Department of Applied Informatics, Thessaloniki	Greece
Chemical Biology Laboratory, SAIC-Frederick Inc.	US

The following sections present the details and outcomes for each of these tasks.

5.1 Passage Retrieval Starting from Patent Claims Task

In patent documents the claim section has an important role in defining the extent of the protection rights which an inventor aims for when submitting a patent application. Patent examiner decisions refer to claims in the application documents and also give a list of prior publications relevant to the given application, publications in which often passages are underlined for being particularly relevant to the application claims.

The 'Claims to passages' task in 2012 investigated how an IR system may support a patent expert in finding the passages of interest for a set of claims. The topics in this task were based on the claims in patent application documents. Given a set of claims the participants were asked to retrieve relevant documents in the collection and mark out the relevant passages in these documents. Participants were provided also the whole patent application document where the claims of the topic occurred and were allowed to use the content of the document for query generation.

The collection of patent documents that were to be used in this task are mainly patent documents published by the European Patent Office (EPO) and have content in at least of the three languages officially accepted by the EPO: English, German, French.

A training set of 51 topics was first made available. The topic test set contained 105 topics with 35 in each of the three languages mentioned above. When compared to the previous CLEF-IP labs (with at least 2,000 topics in the textual retrieval tasks), the number of topics is low, which, in 2012, is due to the fact that the relevance assessments for the topics were to be created by humans in a non-trivial process. The topics were selected out of a pool of patent application documents – candidate documents – which had to fulfil several requirements. The candidate documents had to:

- be published after 2001 and not be part of the CLEF-IP document corpus;
- have at least two highly relevant citations in the search report attached to the candidate document;
- have textual content besides the document's bibliographic part;
- be shorter than 300 thousand words.

Creating a pool of candidate topic documents with the restrictions above has been done automatically by document meta-data inspections. The next step, extracting the set of claims and their relevant passages, had to be done manually by inspecting each search report in detail. The most tedious and time-consuming phase was matching the passage indications in the search reports with the files of the CLEF-IP corpus. To aid our work we have used an in-house developed system (Piroi, Lupu, Hanbury, Sexton, Magdy, & Filippov, 2012).

The submissions to this task were textual files where each line contained, especially, a topic identifier, the identifier of a document considered relevant and the XPath to the textual content identified as relevant. A limit of 100 relevant documents per topic was accepted, the number of marked passages in a relevant document was not limited.

The evaluation of the submissions was done on two levels. The first one, the document level evaluations, measured the quality of the retrieval results considering only the relevant documents and not the passages. Here we used Precision, Recall, MAP and PRES measures. Since no established IR evaluation measure was directly usable to evaluate the quality of the submitted results at passage level, we have adapted to our needs the Mean Average Precision and Precision measures, MAP (D) and Precision (D). More details about these measures can be found in (Piroi, Lupu, Hanbury, Sexton, Magdy, & Filippov, 2012).

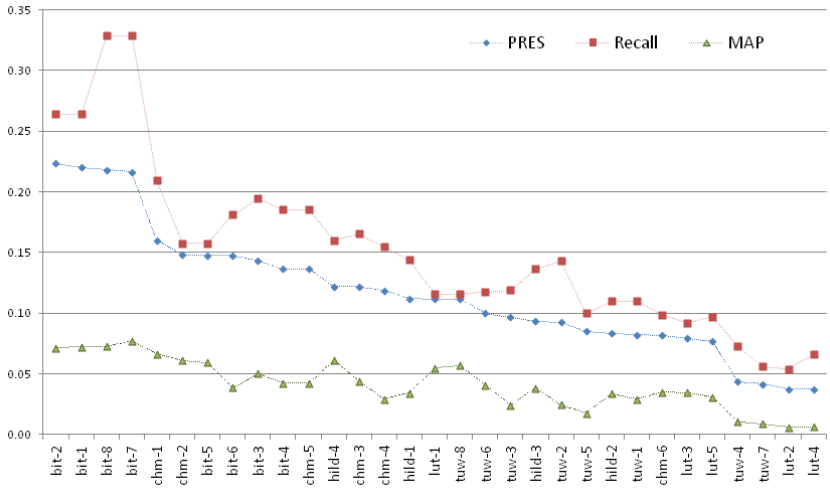


Figure 2: Evaluation at the document level.

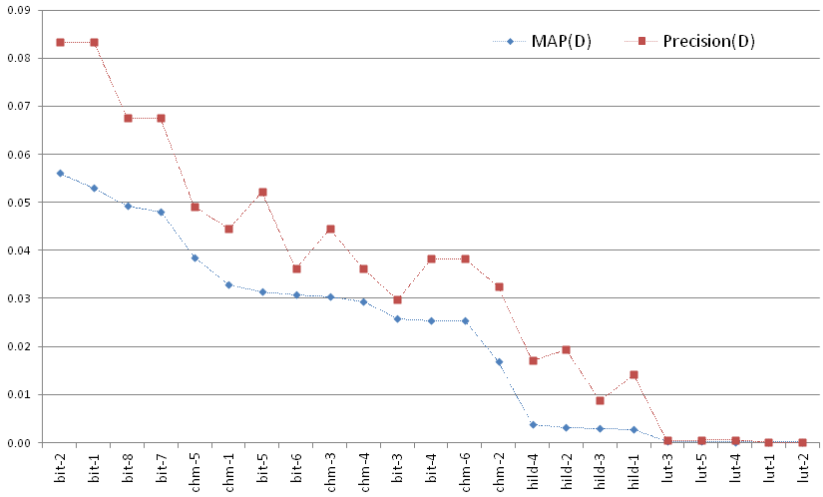


Figure 3: Measures at relevant passage level.

The retrieval solutions chosen by the participants to this task ranged from using natural language processing techniques, two-step retrieval approaches (retrieve document and then mark passage in the document) to using distributed IR and trigram-based searches. One positive fact regarding the multilingual aspect of the task: all participants have used translation tools on the generated queries in order to reach out and find relevant documents in languages other than the language of the topic. Plots of the measure values at the document and passage levels are presented in Figure 2 and Figure 3.

5.2 Flowchart Recognition Task

Images in patents are used both for concept and idea clarifications but also, during an examination process, are used to rapidly filter out documents not relevant to an examined application. The objective of the Flowchart Recognition task is to make the content of the patent images searchable and comparable.

The topics of this task were black and white patent images representing flow-charts. The participants had to extract the information in these images and store it into a predefined textual format. The textual format was defined by the task organizers and basically is a form of textual representation: the nodes of the flow-chart are listed, together with any textual or type information attached to them, then the list of connections between nodes is given, again, together with any type or textual information attached to them (Figure 4) (Piroi, Lupu, Hanbury, Sexton, Magdy, & Filippov, 2012).

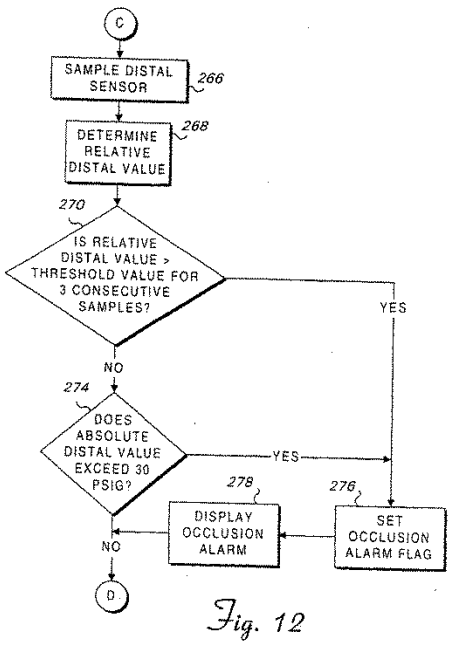


Fig. 12

```
MT Title "FIG. 12"
MT NO 16
MT DE 9
MT UE 8
CO === Nodes ===
NO 1 oval "C"
NO 2 rectangle "SAMPLE DISTAL SENSOR"
NO 3 rectangle "DETERMINE RELATIVE DISTAL VALUE"
NO 4 diamond "IS RELATIVE DISTAL VALUE > THRESHOLD VALUE FOR 3
NO 5 diamond "DOES ABSOLUTE DISTAL VALUE EXCEED 30 PSIG?"
NO 6 point ""
NO 7 circle "D"
NO 8 point ""
NO 9 rectangle "SET OCCLUSION ALARM FLAG"
NO 10 rectangle "DISPLAY OCCLUSION ALARM"
NO 11 no-box "266"
NO 12 no-box "268"
NO 13 no-box "270"
NO 14 no-box "274"
NO 15 no-box "278"
NO 16 no-box "276"
CO === Edges ===
DE 1 2 plain ""
DE 2 3 plain ""
DE 3 4 plain ""
DE 4 5 plain "NO"
UE 5 6 plain ""
DE 6 7 plain "NO"
UE 4 8 plain "YES"
DE 8 9 plain ""
```

Figure 4: Example of a flow-chart image and its textual representation.

50 flow-charts given, together with their textual representation (qrel), as a training set; the test set contained 100 flow-charts. All relevance judgements for these flow-charts were done manually.

The evaluation of the submissions was done by measuring the distance between the textually represented topic graphs and the ones in the submissions. The distance between graphs was computed using the maximal common subgraph using an implementation of the McGregor algorithm (McGregor, 1982). This is however a problem with a high time complexity depending on the factorial of the number of nodes, therefore several practice-based optimizations had to be done (Lupu, Piroi, & Hanbury, 2013).

The results for the 13 submissions are shown in Figure 5, where the higher values reflect better content recognition results.

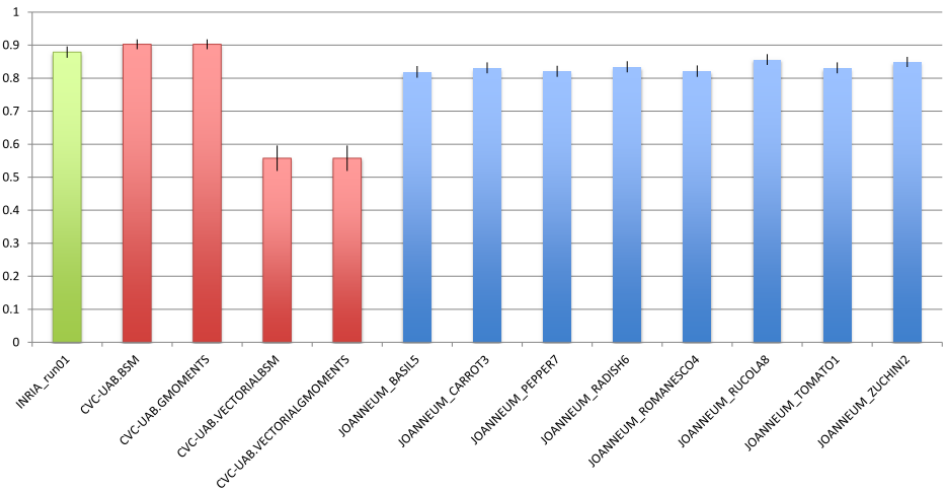


Figure 5: Measurements using the Most Commons Subgraph.

5.3 Chemical Recognition Task

The third task organized in the CLEF-IP Lab is in spirit similar to the Flow-chart Recognition task, and deals with chemical molecular diagrams in patents. Extracting molecular diagrams from patent documents and making them comparable in an automatic way is a potentially powerful approach to identify relevant documents and claims.

The task had two parts: a segmentation part, where patents rendered as monochrome multipage TIFF images with chemical diagrams were distributed with the requirement to clip out the molecular diagrams, and a recognition part, which required that – given a set of diagrams – analyze them to some recognized textual format.

The segmentation part of the task used 30 patents rendered as mentioned above; the participants had to submit the coordinates of the bounding boxes for the molecular diagrams occurring in them. The format of the submission was a comma separated value file.

To evaluate the degree of bounding box matches between the ground truth and the submissions a tool was written that automatically compared the participants' result files with the ground truth file. A result is considered a match if every side of the bounding box in the participant's file is within the tolerance number of pixels of the corresponding ground truth bounding box. The evaluations were done, then, at a range of tolerance levels, starting from 0 (perfect match) to 55 pixels (approximately 0.5 cm) (Piroi, Lupu, Hanbury, Sexton, Magdy, & Filippov, 2012).

Only one submission (by saic) was done in this sub-task, the evaluation results are shown in Table 12.

Table 12: Segmentation evaluation results.

Tolerance	**Precision**	**Recall**	**F_1**
0	0.70803	0.68622	0.69696

10	0.79311	0.76868	0.78070
20	0.82071	0.79543	0.80787
40	0.86696	0.84025	0.85340
55	0.88694	0.85962	0.87307

In the Molecular Diagram recognition task the format which the participants had to use to store their results was MOL, which is currently the most complete standard format for chemical diagrams. However, patent documents often describe a whole family of molecules with diagrams that extend the standard molecule diagrams with graphical representations of varying structures (called Markush structures). These structures cannot be represented in MOL files without some abuse of notation. For this reason there were two sets of topics, one containing 865 diagrams (the automatic set) fully representable as MOL files, and one containing 95 diagrams (the manual set) containing some amount of variability in their structure.

The evaluations for the automatic topic set were done by automatic comparisons of the submissions with the ground truth using the OpenBabel open source chemistry toolbox. The evaluations for the manual set was done by using the MarvinView tool to render the textual representation graphically and then visually comparing the topic image with the rendering of the submission. This visual comparison was done by humans. The results of these evaluations are shown in Table 13

Table 13: Diagram recognition evaluation results.

Participant ID	**Automatic set structures recalled**	**Manual set structures recalled**	**Total structures recalled**
saic	88%	40%	83%
uob-1	96%	46%	91%
uob-2	95%	59%	91%
uob-3	95%	46%	90%
uob-4	96%	57%	92%

5.4 Other Activities in the 'Search for Innovation' Use Case

Efforts were done in 2010 and 2011 to collaborate with other campaigns using patent data. Following the TREC-CHEM campaign and the PatOlympics effort[3] experts that helped organizing the Chemical Recognition Task were attracted.

In 2012 the PatOlympics effort was organized not as a competition, but as a study of how search systems specialized on patent data are used. PatOlympics 2012 was organized as a

[3] The 3rd PatOlympics. http://www.ir-facility.org/patolympics-and-demos

demo session at IRFC 2012[4]. There 7 systems doing patent search were demoed, conference and demo-session participants were allowed to use the systems. This permitted us to log user-system interactions and to do screen-casts.

There is ongoing work within PROMISE to evaluate the knowledge accumulated in the PatOlympics laboratory and adopt the lessons learned into a more general competitive demonstration, in order to test its generality and adaptability to different use-cases.

5.5 Outcomes Summary in the 'Search for Innovation' Use Case

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

1. Following the discussions with participants during the CLEF-IP 2012 workshop, where patent experts were present as well, we concluded the following:
   i. The 'Passage Retrieval Starting From Claims' correctly reflects the patentability searches of a patent examiner.
   ii. It was pointed out that examiners have access and currently use the previously published patent documents that are family members of the application they examine. In 2013 we plan to follow this recommendation and provide the respective documents as well.
   iii. It was stated that this kind of passage retrieval is difficult and that the evaluation results, even though low in absolute values, mark a degree of success which reflect the fact that participants did think of how to best tackle this problem.
2. The results in the Flow-chart Recognition task proved to be surprisingly good for the chosen set of topics. Considering it as having been warming-up round, the next CLEF-IP will include more complicated and interesting flow-charts in the topic set. It is worth mentioning, though, that treating flow-charts is less a research problem and much more an engineering problem, which shows that technically it is possible to reach a digitization of flow-chart to a quality level that can be accepted by IP experts.
3. The results obtained in the evaluation of the manual topic set of the Chemical Recognition task strongly show a need to develop some alternative to or an extension to the MOL file representation which can accommodate the Markush structures and provide tools for manipulating, comparing, etc. such structures and their visual renderings.
4. We remark that researchers in the area seem to have low interest in evaluation campaigns like CLEF-IP.

[4] 5th IRF Conference. http://www.ir-facility.org/irf-conference-2012

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

The CHiC 2012 pilot evaluation lab aimed at moving towards a systematic and large-scale evaluation of cultural heritage digital libraries and information access systems. The lab's goal is to increase our understanding on how to integrate examples from the cultural heritage community into a CLEF-style evaluation framework and how results can be fed back into the CHiC community. CHiC has cooperated with Europeana[5], Europe's largest digital library, museum and archive for cultural heritage objects to provide a realistic environment for experiments.

Five institutions were involved in the preparation and organization of this year's lab (Table 14).

Table 14: CHiC 2012 organizers.

Research Group	Country
Berlin School of Library and Information Science, Humboldt-Universität zu Berlin	Germany
Department of Information Engineering, University of Padova	Italy
Royal School of Library and Information Science, Copenhagen	Denmark
The Information School, University of Sheffield	United Kingdom
Europeana	Netherlands

At the CLEF 2011 conference, a first workshop on information retrieval evaluation was put on by the organizers of the lab to discuss information needs, search practices and appropriate information retrieval tasks for this domain. The outcome of this workshop was a pilot lab proposal for the CLEF conference series suggesting three tasks relevant for cultural heritage information systems. In its first year, CHiC offered three tasks:

- ad-hoc, which measured retrieval effectiveness according to relevance of the ranked retrieval results (standard 1000 document TREC output),
- variability, which required participants to present a list of 12 records that represent diverse information contexts and
- semantic enrichment, which asked participants to provide a list of 10 semantically related concepts to the one in the query to be used in query expansion experiments. All tasks were offered in monolingual, bilingual and multilingual modes.

In total 48 conference participants indicated their interest for CHiC from which roughly 35 attended the lab. From the 21 groups that registered for participation in CHiC 6 (Table 15) research groups submitted 126 different experiments in total.

[5] http://www.europeana.eu

Table 15: CHiC 2012 participating groups.

Institution	Country
Chemnitz University of Technology, Dept. of Computer Science	Germany
GESIS – Leibniz Institute for the Social Sciences	Germany
Unit for Natural Language Processing, Digital Enterprise Research Institute, National University of Ireland	Ireland
University of the Basque Country, UPV/EHU & University of Sheffield	Spain / UK
School of Information at the University of California, Berkeley.	USA
Computer Science Department, University of Neuchatel	Switzerland

The complete Europeana data index (March 2012) was downloaded for collection preparation. The Europeana index was used in Europeana's Solr search portal which contained 23,300,932 documents with a size of 132 GB.

Europeana data consists of metadata records describing digital representations of cultural heritage objects. Roughly 62% of the metadata records describe images, 35% text, 2% audio and 1% video recordings. The metadata contains title and description data, media type and chronological data as well as provider information.

For 2012, three of the 14 language subcollections (Table 16) were used for each task:

- English collection: all Europeana documents with English metadata records.
- French collection: all Europeana documents with French metadata records.
- German collection: all Europeana documents with German metadata records.

Table 16: CHiC 2012 Subcollections by Language and Media Type.

Language	Sound	Text	Image	Video	Total
German	23,370	664,816	3,169,122	8,372	**3,865,680**
French	13,051	1,080,176	2,439,767	102,394	**3,635,388**
English	5,169	45,821	1,049,622	6,564	**1,107,176**
Total	**455,162**	**8,371,581**	**14,304,289**	**169,899**	**23,300,932**

Original user queries were extracted from Europeana query logs from which 50 queries were selected that covered a wide range of topics and represented a distribution of query categories that was found in a previous study [9]. For later relevance assessments, descriptions of the underlying information need were added, but were not admissible for information retrieval. All 50 queries were then translated into French and German. For the variability and semantic enrichment tasks, only the first 25 topics were used for the experiments.

For 2013, it is planned to integrate collections in more languages as well as the complete Europeana collection. In addition, the existing tasks will be improved and two more tasks

focusing on interaction and multilingual issues prepared.

7 Impact Analysis for the CLEF evaluation campaign (2000–2009)

Measuring the impact of benchmarking activities, such as CLEF, is crucial for assessing which of their aspects have been successful, and thus obtain guidance for the development of improved evaluation methodologies and information access systems. Given that their contribution to the field is mainly indicated by the research that would otherwise not have been possible, it is reasonable to consider that their success can be measured, to some extent, by the scholarly impact of the research they foster. Recent investigations have reported on the scholarly impact of TRECVid (Thornley, Johnson, Smeaton, & Lee, 2011) and ImageCLEF (Tsikrika, García Seco de Herrera, & Müller, Assessing the scholarly impact of ImageCLEF, 2011). Building on this work, we summarize the first results of a study that assesses the scholarly impact of the first ten years of CLEF activities by performing a citation analysis on a dataset of publications obtained from the CLEF proceedings. Further results and more detailed analysis are presented in the PROMISE Deliverable 6.4 – Report on the impact analysis for the CLEF Initiative.

CLEF's annual evaluation cycle culminates in a workshop where participants of all labs present and discuss their findings with other researchers. This event is accompanied by the **CLEF working notes**, where research groups publish, separately for each lab and task, participant notebook papers that describe their techniques and results. In addition, the organizers of each lab (and/or each task) publish overview papers that present the evaluation resources used, summarize the approaches employed by the participating groups, and provide an analysis of the main evaluation results. Moreover, evaluation papers reflecting on evaluation issues, presenting other evaluation initiatives, or describing and analyzing evaluation resources and experimental data may also be included. These (non-refereed) CLEF working notes papers are available online on the CLEF website[6]. From 2000 to 2009, participants were invited to publish after each workshop more detailed descriptions of their approaches and more in–depth analyses of the results of their participation, together with further experimentation, if possible, to the **CLEF proceedings**. These papers went through a reviewing process and the accepted ones, together with updated versions of the overview papers, were published in a volume of the Springer Lecture Notes in Computer Science series in the year following the workshop and the CLEF evaluation campaign. Moreover, CLEF participants and organizers may extend their work and publish in journals, conferences, and workshops. The same applies for research groups from academia and industry that, while not official participants of the CLEF activities, may decide at a later stage to use CLEF resources to evaluate their approaches. These **CLEF–derived** publications are a good indication of the impact of CLEF beyond the environment of the evaluation campaign. This analysis focusses on the CLEF proceedings publications;

[6] http://www.clef-initiative.eu/

analysis of the working notes and CLEF-derived publications is described in PROMISE Deliverable 6.4.

7.1 Bibliometric Analysis

The list of CLEF 2000–2009 proceedings publications consists of 873 papers and was obtained through DBLP. All publications were semi-automatically annotated with their *type* (i.e., evaluation, participant or overview) and the *lab(s)* and/or *tasks(s)* they refer to. Their citations were obtained in an 24-hour period in April 2013 using the following citation data sources: (i) Scopus and (ii) Google Scholar. Scopus provides citation analysis tools to calculate various metrics of scholarly impact, such as the h-index [49]. Google Scholar, on the other hand, does not offer such capabilities for arbitrary publication sets; citation analysis using its data can though be performed by systems such as the Online Citation Service (OCS)[7] and Publish or Perish (PoP)[8].

Main results: The results of the bibliometric analysis of the citation data found by the three sources (OCS, PoP and Scopus) for the 873 CLEF proceedings publications are presented in Table 17. Over the years, there is a steady increase in the number of publications, in line with the continuous increase in the number of offered labs (with the exception of 2007). The coverage of publications varies significantly between Scopus and Google Scholar, with the former indexing a subset that does not include the entire 2000 and 2001 CLEF proceedings and another four individual publications, and thus contains 92% of all publications, while the latter does not index 22 (0.02%) of all publications. The number of citations varies greatly between Scopus and Google Scholar, with the latter finding around ten times more citations than Scopus. When examining the distributions over the years, OCS and PoP reach their peak in terms of number of citations and h-index values in 2006, while Scopus does so in 2009. The average number of citations per publication peaks much earlier though, indicating that the publications of the early CLEF years have on average much more impact than the more recent ones.

Overall, the total number of citations over the 873 CLEF proceedings publications are 9,137 and 8,878 as found by OCS and PoP respectively, resulting in 10.47 and 10.17 average cites per paper, respectively. This is slightly higher, but in essence comparable to the findings of the studies on the scholarly impact of TRECVid (Thornley, Johnson, Smeaton, & Lee, 2011) and ImageCLEF (Tsikrika, García Seco de Herrera, & Müller, Assessing the scholarly impact of ImageCLEF, 2011), with the difference that the former considers a much larger dataset (2,073 publications with 15,828 citations) that also includes TREC-derived papers, while the latter a much smaller one (249 publications with 2,147 citations). Finally, since OCS achieves a slightly higher recall, OCS data will be used for the analysis below.

[7] http://dbs.uni-leipzig.de/ocs/

[8] http://www.harzing.com/pop.htm/

Table 17: The citations, average number of citations per publication, and h-index of the CLEF proceedings publications as found by the three sources.

	#labs	#publ.	OCS			PoP			Scopus		
			# cit.	avg	h-index	# cit.	avg	h-index	# cit.	avg	h-index
2000	3	27	501	18.56	15	507	18.78	15	-	-	-
2001	2	37	904	24.43	17	901	24.35	17	-	-	-
2002	4	44	636	14.45	14	634	14.41	14	74	1.68	4
2003	6	65	787	12.11	15	776	11.94	15	87	1.34	5
2004	6	81	989	12.21	17	942	11.63	16	137	1.69	5
2005	8	112	1231	10.99	18	1207	10.78	17	133	1.19	5
2006	8	127	1278	10.06	18	1250	9.84	18	133	1.05	5
2007	7	116	1028	8.86	16	902	7.78	15	119	1.03	5
2008	10	131	1002	7.65	16	989	7.55	16	78	0.60	3
2009	10	133	781	5.87	12	770	5.79	12	144	1.08	5
Total	14	873	9,137	10.47	41	8,878	10.17	41	905	1.04	10

CLEF publication types: Figure 6 compares the relative number of publications of the three types (evaluation, overview and participant) with their relative citation frequency. The participants' publications account for a substantial share of all publications, namely 86%, but only receive 64% of all citations. On the other hand, overview and evaluation publications receive three times or twice the percentage of citations compared to their publications' percentage. This indicates the significant impact of these two types.

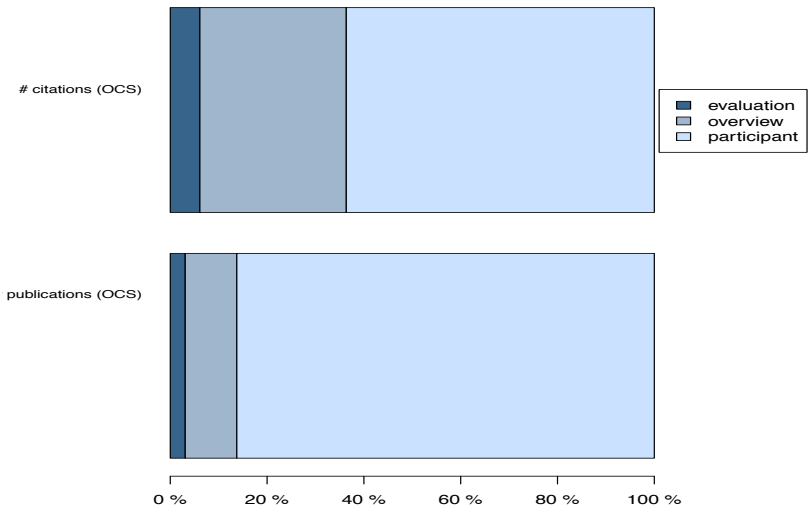


Figure 6: Relative impact of different types of CLEF proceedings publications.

CLEF labs and tasks:Figure 7 depicts the number of citations for the 14 CLEF labs and their tasks organized by CLEF during its first 10 years. Two more "pseudo-labs", CLEF and Other are also listed; these are used for classifying the evaluation type publications not assigned to specific labs, but rather pertaining to evaluation issues related to CLEF or other evaluation campaigns, respectively. Three labs, Adhoc, ImageCLEF and QA@CLEF, clearly dominate; they account for 67% of all publications and for 72% of all citations. Regarding the tasks, the Medical Retrieval and Medical Annotation ImageCLEF tasks have had the greatest impact, closely followed by the main QA task and the main Cross/Monolingual ad-hoc task. This also indicates a bias towards older, most established labs and tasks. Finally, although it is difficult to identify trends over all labs and tasks, in many cases there appears to be a peak in their second or third year of operation, followed by a decline. Exceptions include the Photo Annotation ImageCLEF task, which attracted significant interest in its fourth year when it employed a new collection and adopted new evaluation methodologies, and also the Cross–Language Speech Retrieval (CL–SR) lab that increased its impact in 2005 following a move from broadcast news to conversational speech. Such novel aspects result in renewed interest in labs and tasks, and also appear to strengthen their impact.

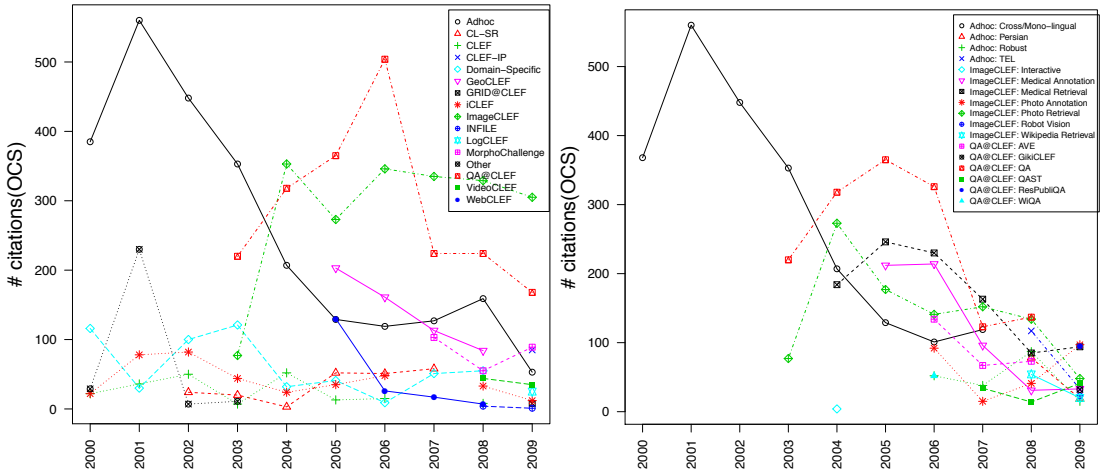


Figure 7: The impact of CLEF labs (left) and tasks (right) over the years.

In summary, this bibliometric analysis of the CLEF 2000–2009 proceedings has shown the considerable impact of CLEF during its first ten years in several diverse multi-disciplinary research fields. The high impact of the overview publications further indicates the significant interest in the created resources and the developed evaluation methodologies, typically described in such papers. Further, more detailed analysis can be found in the PROMISE Deliverable 6.4 – Report on the impact analysis for the CLEF Initiative.

8 Outlook on future evaluation activities: CLEF 2013

This section provides an outlook on the upcoming evaluation activities. CLEF 2013 conference takes place just after the end of PROMISE but its organization is part of the project. This section provides a brief summary of the steps taken towards the organization of the CLEF 2013 conference and its current status and by listing the selected labs.

CLEF 2013 conference[9]: The CLEF 2013 is the fourth CLEF conference continuing the popular CLEF campaigns which have run since 2000 contributing to the systematic evaluation of information access systems, primarily through experimentation on shared tasks. Since 2011, CLEF 2013 is an activity organizes by PROMISE and in 2013 it will take place in Valencia, Spain, on September 23-26, 2013.

Building on the format first introduced in 2010, CLEF 2013 consists of an independent peer-reviewed conference on a broad range of issues in the fields of multilingual and multimodal information access evaluation, and a set of labs and workshops designed to test different 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.

539 research groups were initially registered to CLEF 2013, almost the double of registration than in 2012. Finally 197 groups submitted runs and 183 groups submitted working notes describing their experiments.

CLEF 2013 labs: As in previous years 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.

There were 12 lab submissions although only ten labs were finally accepted. Nine labs will follow a "campaign-style" evaluation practice for specific information access problems in the tradition of past CLEF campaign tracks:

1. *CHiC - Cultural Heritage in CLEF*[10] aims at moving towards a systematic and large scale evaluation of cultural heritage digital libraries and information access systems. After a pilot lab in 2012, where a standard ad-hoc information retrieval scenario was tested together with two use-case-based scenarios (diversity task and semantic enrichment task), the 2013 lab strives to diversify more of the tasks and become more realistic in its tasks organization. The 2013 CHiC lab will organizethree task:
   i. *Multilingual Task*: is a continuation of the 2012 CHiC lab, using similar task scenarios, but requiring multilingual retrieval and results. This task has a predefined and fine-tuned set of requirements from last year's lab, but will

[9] http://clef2013.org/

[10] http://www.promise-noe.eu/chic-2013/home/

assess all runs against the multilingual collection.

ii. *Polish Task*: is a continuation of the 2012 CHiC monolingual lab, namely using topic descriptions written in the Polish language to retrieve cultural object descriptions also written in Polish. The main objective of this task is to have a better understanding of information retrieval for complex languages such as the Polish one.

iii. *Interactive Task*: is a new task whose purpose is to gauge user experience by observing user activity with Europeana under controlled and simulated conditions, aiming for as much "real-life" experiences intruding into the experimentation.

2. *CLEFeHealth - CLEF eHealth Evaluation Lab*[11] is a benchmarking activity aiming at developing processing methods and resources to enrich difficult-to-understand health text as well as their evaluation setting. Three task are organized:
   i. *Task 1*: consists in the identification of disorders from clinical reports and mapping of the SNOMED CT disorders to UMLS codes
   ii. *Task 2*: releases on mapping abbreviations and acronyms in clinical reports to UMLS codes
   iii. *Task 3*: focuses on information retrieval to address questions patients may have when reading clinical reports
3. *CLEF-IP - Retrieval in the Intellectual Property Domain*[12] is a benchmarking activity to investigate IR techniques in the patent domain. Three challenging tasks are foreseen:
   i. *Passage retrieval starting from claims (patentability or novelty search)*: The topics in this task are based on the claims in patent application documents. Given a claim, the participants are asked to retrieve relevant documents in the collection and mark out the relevant passages in these documents.
   ii. *Text to image/image to text*: Given a patent application document - as an XML file - and the set of images occurring in the application, the participants extract the links between the image labels and the text pointing to the object of the image label.
   iii. *Structure Recognition Task*: The topics in this third task are patent images representing flow-charts. Participants in this task are asked to extract the information in these images and return it. The task is similar to the one organized in 2012. This year images with more challenging cases of flow-charts, and - as resources permit - images representing electrical block schemes are added.
4. *ImageCLEF - Cross Language Image Annotation and Retrieval*[13] is a benchmarking activity on the experimental evaluation of image classification and retrieval, focusing on the combination of textual and visual evidence. ImageCLEF 2013 organizes three main tasks, plus one task that is associated with the AMIA 2013 conference[14]:
   i. *Photo Annotation and Retrieval*: aims to advance the state of the art in multimedia research by providing a challenging benchmark for visual concept

[11] http://nicta.com.au/business/health/events/clefehealth_2013/

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

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

[14] http://www.amia.org/amia2013/

detection, annotation and retrieval in the context of diverse collections of photos. The benchmark consists of two subtasks:
   a) *Scalable Concept Image Annotation*: purposes to accurately detect a wide range of semantic concepts for the purpose of scalable automatic image annotation on a large collection of web images.
   b) *Personal Photo Retrieval*: focus on correctly retrieve relevant images from personal photo collections based on typical scenarios in which a user wants to find some of their own photos according to certain criteria.

ii. *Plant identification*: is a new challenge dedicated to botanical data. This year, the task will be focused on tree and herb species identification, based on different types of images.

iii. *Robot Vision*: address the problem of semantic place classification using visual and depth information. This time, the task also addresses the challenge of object recognition.

iv. *AMIA-Medical task*: will for the first time organize a workshop outside of Europe; the ImageCLEF meeting is planned at the annual AMIA meeting in the form of a workshop. There are four types of tasks in 2013:
   a) *Modality Classification*: is organized in the same format as in 2012. In 2013, a larger number of compound figures will be present making the task significantly harder but corresponding much more to the reality of biomedical journals.
   b) *Compound figure separation*: aims to detect compound figures and then separate them into sub figures that can subsequently be classified into modalities and made available for research.
   c) *Ad-hoc image-based retrieval*: is the classic medical retrieval task, similar to those organized in 2005-2012.
   d) *Case-based retrieval*: is a more complex task first introduced in 2009, but one that we believe is closer to the clinical workflow.

5. *INEX - INitiative for the Evaluation of XML retrieval*[15] builds evaluation benchmarks for search with rich structure - such as document structure, semantic metadata, entities, or genre/topical structure - as of increasing importance on the web and in professional search. In 2013, INEX is running four types of task:
   i. *Social Book Search*: investigates techniques to support users in searching and navigating the full texts of digitized books and complementary social media as well as providing a forum for the exchange of research ideas and contributions. Towards this goal the track is building appropriate evaluation benchmarks complete with test collections for focused, social and semantic search tasks. The track touches on a range of fields, including information retrieval (IR), information science (IS), human computer interaction (HCI), digital libraries (DL), and eBooks. The Social Book Search Track runs two search subtasks:
      a) *Social Book Search*: aims to evaluate book search and to investigate the relative value of traditional book metadata and user-generated content for book search.

[15] https://inex.mmci.uni-saarland.de/

b) *Prove It!*: proposes participants to devise a retrieval system that, in response to a claim, returns lists of pages, in descending order of likelihood that they confirm / refute the claim, and at the same time have the authority to do so (and therefore can be trusted).

ii. *Linked Data*: investigates retrieval techniques over a combination of textual and highly structured data, where RDF properties carry additional key information about semantic relations among data objects that cannot be captured by keywords alone. For INEX 2013, two different retrieval tasks are explored that continue from INEX 2012:
   a) *Ad-hoc Retrieval*: investigates informational queries to be answered mainly by the textual contents of the Wikipedia articles.
   b) *Jeopardy*: employs natural-language Jeopardy clues which are manually translated into a semi-structured query format based on SPARQL with keyword conditions.

iii. *Tweet Contextualization*: is running since 2010. In 2013, the goal of the task and the evaluation metrics remain unchanged but tweet diversity has been improved. More specially, a significant part of tweets with hashtags have been included in the tweet set. Hashtags are authors' annotation on key terms of their tweets. In the two past years, hashtags have been underused, unless they are core components of tweets.

iv. *Snippet Retrieval*: determines how best to generate informative snippets for search results. Such snippets should provide sufficient information to allow the user to determine the relevance of each document, without needing to view the document itself.

6. *PAN - Uncovering Plagiarism, Authorship, and Social Software Misuse*[16] is the 9th evaluation lab on uncovering plagiarism, authorship, and social software misuse. In 2013, PAN offers three tasks:
   i. *Plagiarism Detection*: is divided into two sub-tasks:
      a) *Source Retrieval*: aims to retrieve all plagiarized sources while minimizing retrieval, given a suspicious document and a web search API.
      b) *Text Alignment*: propose to identify all contiguous maximal-length passages of reused text between them, given a pair of documents.
   ii. *Author Identification*: focuses on authorship verification and methods to answer the question whether two given documents have the same author or no. This question accurately emulates the real-world problem that most forensic linguists face every day.
   iii. *Author Profiling*: is concerned with predicting an author's demographics from her writing. Besides being personally identifiable, an author's style may also reveal her age and gender.
7. *QA4MRE - Question Answering for Machine Reading Evaluation*[17] is a benchmarking activity on the evaluation of machine reading systems through question answering and reading comprehension tests. Beside the Main Task, also two pilot tasks are offered in 2013:

[16] http://pan.webis.de/

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

i. *Machine Reading*: addresses the problem of building a bridge between knowledge encoded as natural text and the formal reasoning systems that need such knowledge.
ii. *Machine Reading of Biomedical Texts about Alzheimer's Disease*: is aimed at setting questions in the Biomedical domain with a special focus on one disease, namely Alzheimer. This pilot task will explore the ability of a system to answer questions using scientific language.
iii. *Entrance Exams*: aims at evaluating systems under the same conditions humans are evaluated to enter the University. In this first campaign we will reduce the challenge to Reading Comprehension exercises contained in the English exams. More types of exercises will be included in subsequent campaigns (2014–2016) in coordination with the "Entrance Exams" task at NTCIR.

8. *QALD-3 - Question Answering over Linked Data*[18] is the third in a series of evaluation campaigns on question answering over linked data, this time with a strong emphasis on multilinguality. It offers two open tasks:
   i. *Multilingual question answering*: provides a benchmark for comparing different approaches and systems that mediate between a user, expressing his or her information need in natural language, and semantic data.
   ii. *Ontology lexicalization*: consists in finding English lexicalizations of a set of classes and properties from the DBpedia ontology in a Wikipedia corpus.
9. *RepLab 2013* is the second CLEF lab on Online Reputation Management. In 2013, RepLab focus on the task of monitoring the reputation of entities (companies, organizations, celebrities,) on Twitter. The monitoring task for analysts consists of searching the stream of tweets for potential mentions to the entity, filtering those that do refer to the entity, detecting topics (i.e., clustering tweets by subject) and ranking them based on the degree to which they signal reputation alerts (i.e., issues that may have a substantial impact on the reputation of the entity).

One lab will be run as a workshop organized as speaking and discussion session to explore issues of evaluation methodology, metrics, and processes in information access and closely related fields:

1. *CLEF-ER - Entity Recognition @ CLEF*[19] is a workshop on multilingual annotation of named entities and terminology resources acquisition.

[18] http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/

[19] http://www.limosine-project.eu/events/replab2013

9 References

McGregor, J. (1982). Backtrack search algorithms and the maximal common subgraph problem. *Software Practice and Experience , 12* (1), 23-34.

Amigó, E., Corujo, A., Gonzalo, J., Meij, E., & de Rijke, M. (2012). Overview of RepLab 2012: Evaluating Online Reputation Management Systems. *CLEF 2012 working notes.* Rome.

Anderka, M., & Stein, B. (2012). Overview of the 1th International Competition on Quality Flaw Prediction in Wikipedia. *CLEF 2012 working notes.* Rome.

Argamon, S., & Juola, P. (2011). Overview of the International Authorship Identification Competition at PAN-2011. *CLEF.* Amsterdam.

Chappell, T., & Geva, S. (2012). Overview of the INEX 2012 Relevance Feedback Track. *CLEF 2012 working notes.* Rome.

Di Nunzio, G. M., Leveling, J., & Mandl, T. (2011). LogCLEF 2011 Multilingual Log File Analysis:Language identification, query classification, and success of a query. *CLEF.* Amsterdam.

Forner, P. (2011). PROMISE Deliverable 7.5: Second PROMISE Annual Conference and Proceedings.

Forner, P. (2012). *PROMISE Deliverable 7.9: Third PROMISE Annual Conference and Proceedings.*

Gäde, M., Ferro, N., & Lestari Paramita, M. (2011). CHiC 2011 – Cultural Heritage in CLEF: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage. *CLEF.* Amsterdam.

Goëau, H., Bonnet, P. J., Boujemaa, N., Barthelemy, D., Molino, J.-F., Birnbaum, P., et al. (2011). The CLEF 2011 Plant Images Classification Task. *CLEF.* Amsterdam.

Goëau, H., Bonnet, P., Joly, A., Yahiaoui, I., Barthelemy, D., Boujemaa, N., et al. (2012). The ImageCLEF 2012 Plant Identification Task. *CLEF 2012 working notes.* Rome.

Hersh, W., Müller, H., & Kalpathy-Cramer, J. (2009). The ImageCLEFmed Medical Image Retrieval Task Test Collection. *Digital Imaging , 22* (6), 648-645.

Juola, P. (2012). An Overview of the Traditional Authorship Attribution Subtask. *CLEF 2012 working notes.* Rome.

Kalpathy–Cramer, J., Müller, H., Bedricks, S., Eggel, I., G. Seco de Herrera, A., & Tsikrika, T. (2011). Overview of the CLEF 2011 Medical Image Classification and Retrieval Tasks. *CLEF.* Amsterdam.

Koolen, M., Kazai, G., Kamps, J., Preminger, M., Doucet, A., & Landoni, M. (2012). Overview of the INEX 2012 Social Book Search Track. *CLEF 2012 working notes.* Rome.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision , 60* (2), 91-110.

Lupu, M., Piroi, F., & Hanbury, A. (2013). Evaluating Flowchart Recognition for Patent Retrieval. *The Fifth International Workshop on Evaluating Information (EVIA).* Tokyo, Japan.

Müller, H., Despont-Gros, C., Hersh, W., Jensen, J., Lovis, C., & Geissbuhler, A. (2006). Medical image analysis and retrieval, User testing and task analysis. *Proceedings of the*

Medical Informatics Europe Conference (MIE 2006). Maastricht.

Müller, H., García Seco de Herrera, A., Kalpathy-Cramer, J., Demmer Fushman, D., Antani, S., & Eggel, I. (2012). Overview of the ImageCLEF 2012 Medical Image Retrieval and Classification Tasks. *CLEF 2012 working notes.* Rome.

Martinez-Gomez, J., Garcia-Varea, I., & Caputo, B. (2012). Overview of the ImageCLEF 2012 Robot Vision Task. *CLEF 2012 working notes.* Rome.

Morante, R., & Daelemans, W. (2012). Annotating Modality and Negation for a Machine Reading Evaluation. *CLEF 2012 working notes.* Rome.

Morante, R., & Daelemans, W. (2011). Overview of the QA4MRE Pilot Task: Annotating Modality and Negation for a Machine Reading Evaluation. *CLEF.* Amnasterdam.

Morante, R., Krallinger, M., Valencia, A., & Daelemans, W. (2012). Machine Reading of Biomedical Texts about Alzheimer's Disease. *CLEF 2012 working notes.* Rome.

Nowak, S., Nagel, K., & Liebetrau, J. (2011). The CLEF 2011 Photo Annotation and Concept-based Retrieval Tasks, Notebook Papers. *CLEF.* Amsterdam.

Orio, N., & Rizo, D. (2011). Overview of MusiCLEF 2011. *CLEF.* Amsterdam.

Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Forascu, C., et al. (2011). Overview of QA4MRE at CLEF 2011: Question Answering for Machine Reading Evaluation. *CLEF.* Amsterdam.

Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Sporleder, C., et al. (2012). Overview of QA4MRE at CLEF 2012: Question Answering for Machine Reading Evaluation. *CLEF 2012 working notes.* Rome.

Petras, V., & Clogh, P. (2011). Introduction to the CLEF 2011 Labs. *CLEF.* Amsterdam.

Petras, V., Ferro, N., Gäde, M., Isaac, A., Kleineberg, M., Masiero, I., et al. (2012). Cultural Heritage in CLEF (CHiC) Overview 2012. *CLEF 2012 working notes.* Rome.

Petras, V., Forner, P., & Clough, P. D. (2011). CLEF 2011 Labs ans Workshop. *CLEF.* Amsterdam.

Piori, F., Lupu, M., Hanbury, A., & Zenz, V. (2011). Retrieval in the Intellectual property Domain. *CLEF.* Amsterdam.

Piori, F., Petras, V., Gäde, M., Larsen, B., Tsikrika, T., García Seco de Herrera, A., et al. (2012). *PROMISE Deliverable 6.2 - Report of the second year evaluation activities.*

Piroi, F., Lupu, M., Hanbury, A., Sexton, A., Magdy, W., & Filippov, I. (2012). CLEF-IP 2012: Retrieval Experiments in the Intellectual Property Domain. *CLEF 2012 working notes.* Rome.

Potthast, M., & Holfeld, T. (2011). Overview of the 2nd International Competition on Wikipedia Vandalism Detection. *CLEF.* Amsterdam.

Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Overview of the 3rd International Competition on Plagiarism Detection. *CLEF.* Amsterdam.

Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., Oberländer, A., et al. (2012). Overview of the 4th International Competition on Plagiarism Detection. *CLEF 2012 working notes.* Rome.

Sanjuan, E., Moriceau, V., Tannier, X., Bellot, P., & Mothe, J. (2012). Overview of the INEX 2012 Tweet Contextualization Track. *CLEF 2012 working notes.* Rome.

Thomee, B., & Popescu, A. (2012). Overview of the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task. *CLEF 2012 working notes.* Rome.

Thornley, C. V., Johnson, A. C., Smeaton, A. F., & Lee, H. (2011). The Scholarly Impact of TRECVid (2003-2009). *Journal of the American Society for Information Science and Technology , 62* (4), 613–627.

Trappett, M., Geva, S., Trotman, A., Scholer, F., & Sanderson, M. (2012). Overview of the INEX 2012 Snippet Retrieval Track. *CLEF 2012 working notes.* Rome.

Tsikrika, T., García Seco de Herrera, A., & Müller, H. (2011). Assessing the scholarly impact of ImageCLEF. *Proceedings of the 2nd CLEF conference.* Amsterdam, The Netherland.

Tsikrika, T., Müller, H., Forner, P., Frieseke, M., Piori, F., Agosti, M., et al. (2011). *PROMISE Deliverable 6.1: Report on the outcomes of the first year evaluation activities.*

Tsikrika, T., Popecu, A., & Kludas, J. (2011). Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011. *CLEF.* Amsterdam.

Villegas, M., & Paredes, R. (2012). Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task. *CLEF 2012 working notes.* Rome.

Wang, Q., Kamps, J., Ramirez Camps, G., Marx, M., Schuth, A., Theobald, M., et al. (2012). Overview of the INEX 2012 Linked Data Track. *CLEF 2012 working notes.* Rome.

Zellhoefer, D. (2012). Overview of the Personal Photo Retrieval Pilot Task at ImageCLEF 2012. *CLEF 2012 working notes.* Rome.

Appendix I: Questionnaires sent to CLEF 2012 Labs organizers

- CLEF 2012 Labs
  https://docs.google.com/spreadsheet/ccc?key=0At0_0UCBbb7ldFFEbV94U2dnQnFxU3NJXzVqbVlDaEE#gid=0
- CLEF 2012 Labs: collections
  https://docs.google.com/spreadsheet/ccc?key=0At0_0UCBbb7ldEY1dUxyUzREMXVkbFhSaEtjdVRqQVE#gid=0

Appendix II: Participation in the CLEF 2012 labs

Table 17: Participation to the CLEF 2012 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
CHiC	Ad-hoc Retrieval Variability Semantic Enrichment	2	21	6	n/a	n/a	6 groups / 126 runs	DIRECT
CLEF-IP	Chemical Image Extraction and Recognition	1	11	2	2	10	5	Sent by email
	Flowchart Recognition	1	16	3		10	13	Sent by email
	Passage Retrieval Starting from Claims	1	20	5	4	8	16	DIRECT
	Flickr Photo Annotation and	10	108	annotation: 18, retrieval: 7	unknown	annotation: 5, retrieval	annotation: 80, retrieval: 47	ImageCLEF submission

Lab	Task(s)	Number of years the task is part of CLEF	Registrations	Participations	Return participations	Submissions allowed per participant	Total submissions	Submission system
	Retrieval							system
	Plant Identification	2	68	11	5	67	40	ImageCLEF submission system
	Pilot Task on Personal Photo Retrieval	1	62	3	Not applicable	5	15	ImageCLEF submission system
	Robot Vision	3	43	8	3	10	100+	ImageCLEF submission system
	Scalable Image Annotation using General Web Data	1	47	3	Not applicable	20, 10 per subtask	35	ImageCLEF submission system
INEX	Linked Data Track	1	73	7	9	3	25	Upload to server
	Relevance Feedback	1	52	2	1	Unrestricted	15	Upload to server
	Snippet Retrieval	1	71	3	-	Unrestricted		Upload to server
	Social Book Search	1	55	5	-	6	28	Upload to server

Lab	Task(s)	Number of years the task is part of CLEF	Registrations	Participations	Return participations	Submissions allowed per participant	Total submissions	Submission system
	Tweet Contextualization	1	62	15	13	3	33	Upload to server
PAN	Plagiarism Detection	3	39	11	6	1 per subtask	17	Sent by mail + logging at ChatNoir (candidate retrieval), software submissions (text alignment)
	Quality Flaw Prediction in Wikipedia	1	21	3	Not applicable	Unrestricted	4	Sent by email
	Traditional Authorship Attribution	4	-	12	-	Unrestricted	25	Sent by email
QA4MRE	Machine Reading of Biomedical	1	22	7	7	10	43	Upload to server

Lab	Task(s)	Number of years the task is part of CLEF	Registrations	Participations	Return participations	Submissions allowed per participant	Total submissions	Submission system
	Texts about Alzheimer							
	Processing modality and negation	2	12	3	3	3	6	Sent by email
RepLab	Monitoring	1	39	3	Not applicable	5	11	Sent by email
	Profiling	1		13	Not applicable	5	31	Sent by email

Appendix III: Main outcomes of the CLEF 2012 Labs

Table 18: Main advancements and main problems in the CLEF 2012 Labs

Lab	Task(s)	Task type	Main differences/advancements from 2011	Main problems
CHiC	Ad-hoc Retrieval Variability Semantic Enrichment	Retrieval	Not applicable	Alternative tasks require alternative measures that need to be considered in tools provided for the participants such as DIRECT. Cultural heritage information systems are looking to incorporate more user interactions into their systems. The information retrieval evaluation field has often been criticized for viewing the viewer as outside of the scope of study. This domain and the available system (Europeana) enable us to combine and collaborate on information retrieval and information interaction research. CHiC is attempting to move towards this direction.
CLEF-IP	Chemical Image Extraction and Recognition	Image recognition	The first step: extracting the chemical images from the patent PDFs. The addition of chemical images containing elements not representable as InCHI formulas.	Attracting participants

Lab	Task(s)	Task type	Main differences/advancements from 2011	Main problems
	Flowchart Recognition	Image recognition	Not applicable	Creating the QRELs identifying an appropriate similarity measure
	Passage Retrieval Starting from Claims	Retrieval	The requirement to identify paragraphs rather than documents	Generating QRELs: had to create a dedicated tool to semi-automate the process
ImageCLEF	Flickr Photo Annotation and Retrieval	Retrieval, Classification	Bigger emphasis on ground truth collection. Concepts and queries slightly modified to take feedback from last year into account as well as adding/removing/updating them based on what people actually search for on the web as determined through an inspection of the query logs of Yahoo! Image Search.	Ground truth annotation takes too much time and costs too much money. This will have to change otherwise I cannot justify to my company to organize again. Participants' results are disappointing; no improvement has been made since last year. While concepts and queries were more difficult, advances in science should have been able to keep up with this. Even the detection quality of simple concepts like sunset dropped significantly, which is a worrying trend. Evaluation measures need an update, we've been using the same ones as previous years to not change too much, but the current MAP etc. are not ideal. Yet at the same time, for some teams there is a specific focus on optimizing their techniques for one particular variant of the evaluation measure, which is an attitude that needs to change: tweaking performance to improve 0.01 on some evaluation measure while overall performance

Lab	Task(s)	Task type	Main differences/advancements from 2011	Main problems
				is disappointingly low means they have their priorities set incorrectly; at the same time these participants attack us (somewhat aggressively) via email regarding our choices of evaluating the results. Such participants do not encourage organizing the task next year. Also, lack of feedback, enthusiasm and gratitude from participants is worrying. I noticed this last year already when I was present as an observer to take over the task Stefanie organized, and this year once again. It does not make for a very simulating experience. (that being said, the atmosphere at the conference amongst the organizers was excellent, which made up for a lot)
	Medical Image Classification and Retrieval	Retrieval, Classification	Larger dataset. Improved hierarchy in the classification task	Many groups do not submit runs.
	Pilot Task on Personal Photo Retrieval	Retrieval	Not applicable	The generation of user-centered tasks and the acquisition of the accompanying data take much more time than expected. Low participation.
	Plant Identification	Retrieval	More data. More topics.	Many work during summertime for both organizers and participants. Some of them did complain about that.
	Robot Vision	Classification	Competition did not run in 2011; wrt previous edition, introduction of multi modal data as opposed to vision only.	Only 4 of the 8 groups that submitted results submitted a WN paper.

Lab	Task(s)	Task type	Main differences/advancements from 2011	Main problems
	Scalable Image Annotation using General Web Data	Annotation	Not applicable	Since there was another image annotation task, most of the participants preferred the other one (that has run for many years and is an easier challenge), so we had a low participation.
INEX	Linked Data	Retrieval, Question Answering	The data collection switched from IMDB to Wikipedia+DBpedia. There was a new task: Jeopardy! task.	Due to the complexity of Wikipedia markup, the preprocessed data collection was not comprehensive enough.
	Relevance Feedback	Retrieval	The entire Wikipedia collection was used, rather than a substantially smaller submission pool. Instead of submitting a module, participants would link their module to the evaluation platform and use it to generate and upload submissions. Participants would be able to create their module in any programming or scripting language that supports standard input and output.	Low participation
	Snippet Retrieval	Retrieval	Full document-based assessment was used in addition to snippet-based assessment, so that both the snippet and the full document were assessed by the same assessor. To keep the assessment load manageable that document-based assessment was included, the number of topics and snippets was reduced: The document title was shown alongside each snippet in the assessment software – it was	Low participation

Lab	Task(s)	Task type	Main differences/advancements from 2011	Main problems
			not necessary to include the document title in the snippet itself (although this was not forbidden). Snippets were limited to 180 characters (down from 300 in 2011). There was a baseline run included in the evaluation, consisting of the first 180 characters of each document in the reference run.	
	Social Book Search	Expert Search	The choice to focus on topics with post-catalogued suggestions (PCS topics) resulted in a topic set that is slightly different from the topics used last year, where the personal catalogued of the topic creator was ignored and all topics that have a book request were considered, a descriptive title and at least one suggestion.	Low participation
	Tweet Contextualization	Retrieval, Question Answering, Summarization	Using a large amount of real tweets in Json format from twitter as topics.	We succeeded in providing a cleaned preprocessed xml Wikipedia dump as corpus and in sticking with a tight schedule. We shall focus on getting more participants in next edition.
PAN	Plagiarism Detection	Retrieval	Complete redevelopment of the evaluation framework. For candidate retrieval: • New plagiarism corpus crowdsourced using oDesk workers. • Each plagiarism case was written manually.	The candidate retrieval task required us to set up the search infrastructure for the participants. Keeping the software running was a challenge. Participants started to develop their candidate retrieval algorithms only just before the final deadline for submitting runs: Hence, we had to

Lab	Task(s)	Task type	Main differences/advancements from 2011	Main problems
			• The sources were retrieved manually from the ClueWeb using the ChatNoir search engine. • New performance measures for candidate retrieval. For text alignment: • New experimentation platform: software submissions. • New performance measurement. • Introduction of real plagiarism cases.	redo the test phase for candidate retrieval. The software submissions for text alignment were sometimes difficult to be run at our side which caused a lot of back and forth between participants and us.
	Quality Flaw Prediction in Wikipedia	Classification	Not applicable	Low participation.
	Traditional Authorship Attribution	Classification	Different and much smaller corpus.	Organization needed to be tighter in terms of both timing and quality control.
QA4MRE	Machine Reading of Biomedical Texts about Alzheimer	Question Answering	Not applicable	Compiling the background collection took a lot of time. Converting pdf files into text format
	Processing Modality and Negation	Question Answering	The 2011 task had a more fine-grained set of labels, which made it more difficult.	We were not able to provide training data.

Lab	Task(s)	Task type	Main differences/advancements from 2011	Main problems
RepLab	Monitoring	Annotation	Not applicable	According to the current Twitter Terms of Service, the organization cannot distribute the tweets themselves, but rather the link to the tweets, so that participants have to download the tweets themselves. But the set of available tweets changes over time: users cancel their accounts, change their privacy settings or remove specificc tweets. That means that, over time, the RepLab 2012 test collection will be continuously shrinking in size.
	Profiling	Annotation	Not applicable	

Table 19: Main trends in the approaches employed by the participants to the CLEF 2012 Labs and the main experimental outcomes.

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
CHiC	Ad-hoc Retrieval	Most groups concentrated on the monolingual tasks. For the ad-hoc task, most groups used open information retrieval systems. For translations in the bilingual and multilingual tasks, Google Translate, Wikipedia entries (with associated translations) and Microsoft's translation service were used. For the semantic enrichment task, the most often used external source for terms was Wikipedia at different levels of detail	Bilingual and multilingual experiments also seem to perform better than the monolingual experiments on average
	Variability		
	Semantic Enrichment		

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
CLEF-IP	Chemical Image Extraction and Recognition	The biggest problem is disambiguating the elements present in the image, elements which have a particular significance in chemistry. Overall, a set of heuristics was the key element, and the one with the better heuristics obtained the best results. It was not necessarily the case that chemistry knowledge was an asset.	There is a huge gap between "basic" chemistry (whatever can be represented as InCHI formulas) and everything else. The set of 95 structures evaluated manually because they had no InCHI representation obtained a maximum of 59% recall, while the automatic set reached 96%.
	Flowchart Recognition	The common parts were generally the order of identifying different components: first the graph parts (nodes, edges) and then only the textual contents. Other than that, despite having only 3 participants, the methods were quite diverse. Shapes were identified both by eliminating first the empty space around them or by filling in the space inside them. The discussions during the workshop showed how complementary the methods were and how much the participants learned from each other. They also used different	Overall participants did a surprisingly good job. 11 of 13 runs returned graphs representing over 80% of the original graphs. This lead us to conclude that we underestimated their capacity to solve the problem and the suggestion for the next year was to let the flowcharts be more complicated in the test collection.

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
		OCR tools or methods.	
	Passage Retrieval Starting from Claims	The solutions chosen by the submitting participants range from two-step retrieval approaches, namely a document level retrieval in the first step and a passage level retrieval in the second step to using Natural Language Processing techniques and trigrams to extract relevant passages. Another approach was to simulate a distributed IR system by splitting the CLEF-IP collection by IPC codes. All participants have used translation tools	The experiment results measured at the passage level have low scores, we believe that the Precision (D) and MAP (D) measures are not fit enough to measure the success of this task.

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
		on the generated queries.	
ImageCLEF	Flickr Photo Annotation and Retrieval	Fisher vectors, soft coding, optimal fusion and semantic contextualization of tags have led to good results. Bag of textual/visual words as well as SVMs still very popular but no guarantee for good performance.	Fisher vectors, soft coding, optimal fusion and semantic contextualization of tags have led to good results. Bag of textual/visual words as well as SVMs still very popular but no guarantee for good performance.
	Medical Image Classification and Retrieval	Lucene, concept-based approaches & used of multiple visual features	Visual, textual or mixed runs perform differently based on the subtasks. Same or similar descriptors differ on results. Expansion of the training set and the used of multiple visual features were successful.
	Plant Identification	Interactive segmentation + shape boundary features OR Generic approaches (SVMs, sparse coding).	Fully automatic identification from unconstrained photographs is still very challenging. Performances on leaf scans are correct but the increased number of species already shows the limit of a leaf-based only system. We plan to extend the task to more organs next year.
	Pilot Task on Personal Photo Retrieval	Only the group REGIM decided to exploit the browsing data instead of the provided metadata	Interestingly, there was no interest in solving the so-called user-centered initiative of the subtasks

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
	Robot Vision	Fisher Vectors, SVM-based classifiers	frame by frame recognition works reasonably well, still challenging how to exploit the temporal continuity
	Scalable Image Annotation using General Web Data	The use of online learning methods that scale well to large datasets and are able to handle noisy data.	For some concepts, the annotation systems based on automatic data have a comparable performance than systems based on manually labelled data.
INEX	Linked Data	Due to the tight schedule and it was the first year of this new track, most participants used traditional IR approaches to do the tasks, except that Max-Plank Institute employed the DB+IR approach to evaluate the SPARQL FullText queries.	The combination approaches used by Renmin University of China performed best, which combined retrievals over textual documents and RDF data. It was followed by the groups that employed traditional IR approaches. For many reasons, DB approaches employed by the participants performed much worse.
	Relevance Feedback	Too little participation to determine	Relatively little if any advantage or disadvantage from using the full document collection rather than the pool. Although more work needs to be done by participants with the track in this form, participants are still able to find the relevant documents in the full collection
	Snippet Retrieval	Too little participation to determine	Too little participation to determine

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
	Social Book Search	Too little participation to determine	The most effective systems incorporate the full topic statement, which includes the title of the topic thread, the name of the discussion group, and the full first message that elaborates on the request.
	Tweet Contextualization	Combining Natural Language Processing tools with Information Retrieval tools in an effective way. Tweet reformulation using semantic expansion based on terminology analysis or latent probabilistic models. Sentence extraction based on PoS tagging, scoring based on similarity measures and re-ordering to improve readability. Anaphora detection and resolution. Pure summarization approaches on a large number of documents. Pure focused passage information retrieval	Participants that combined simple Indri query language features with state of the art Part of Speech tagging and summarization tools clearly out-performed pure single approaches based on advanced focused IR or fast summarization algorithms for large data. Anaphora resolution did not help readability but sentence reordering based on anaphora detection did. Tweet reformulation based on local LDA also improved results.
PAN	Plagiarism Detection	For candidate retrieval: • An analysis of the participants' notebooks reveals a number of building blocks that were commonly used to build candidate retrieval algorithms: (1) chunking, (2) keyphrase extraction, (3)	

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
		query formulation, (4) search control, and (5) download filtering. In what follows, we describe them in detail. They are described in detail in the overview paper. For text alignment: • An analysis of the participants' notebooks reveals a number of building blocks that were commonly used to build detailed comparison algorithms: (1) seeding, (2) match merging, and (3) extraction filtering. They are described in detail in the overview paper.	
	Traditional Authorship Attribution	Ensemble methods -- use multiple classifiers and average them. The primary features used were low-level character n-grams.	People can be really good at this with enough training data.
	Quality Flaw Prediction in Wikipedia	Too little participation to determine	Three quality flaw classifiers have been developed, which employ a total of 105 features to quantify the ten most important quality flaws in the English Wikipedia. Two classifiers achieve promising performance for particular flaws.

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
QA4MRE	Machine Reading of Biomedical Texts about Alzheimer	Most teams applied text similarity methods. The background collection was used by most teams.	Index expansion techniques work well for the task. The use of the background collection is necessary. Simple text similarity techniques do not suffice to perform the task. The task is realistic and the questions were well defined.
	Processing Modality and Negation	Rule-based systems were built based on linguistic knowledge.	Rule-based systems can obtain good results. The task is realistic and the degree of difficulty is correct.
RepLab	Monitoring	Too little participation to determine	Systems are not substantially contributing to solve the problem yet.
	Profiling	Sentiment polarity detection software adapted to the reputation scenario and/or the textual source (tweets)	The proling task is far from being solved automatically.

Appendix IV: CLEF 2012 Labs Test Collections

List of collections in the CLEF 2012 labs:

1. **Alzheimer's Disease Literature Corpus (ADLC corpus):** The collection contains scientific publications about Alzheimer's disease in several formats (pdf, text, html, annotated). The texts have been carefully selected to be as specific as possible for this topic and the corpus should constitute a comprehensive resource for this task in particular and for text mining efforts tailored to the Alzheimer's disease field in general.
2. **Amazon/LibraryThing:** This corpus was crawled by the University of Duisburg-Essen. The collection consists of 2.8 million book records from Amazon, extended with social metadata from LibraryThing.
3. **Background Collections - Main Task:** It is a carefully constructed background corpus to allow systems to acquire the background knowledge needed for answering the tests. The background collections should cover completely the corresponding topic. This is feasible sometimes and unrealistic at others.
4. **CLEF-IP 2012:** A collection extending the CLEF-IP 2011 collection to which two tasks (chemical and flowchart) added additional images. This collection contains patent documents.
5. **Europeana[20]:** This collection was used for all 3 tasks in CHiC, a large digital library, museum and archive, which provides access to over 20 million cultural heritage objects. The documents in the Europeana collection are metadata records consisting of brief descriptions of the object (title, keywords, description, date, and provider) and occur in multiple languages.
   For experimental purposes, the Europeana collection was divided into 3 subcollections according to metadata languages (i.e. language of metadata object provider), so that some control over the language of documents for the relevance assessments can be asserted.
6. **INEX Wikipedia collection (2009):** Wikipedia articles in XML format annotated with YAGO.
7. **MIRFLICKR:** Two subsets of images obtained from the MIRFLICKR were used. The Flickr photos are collected based on interestingness rating, including Flickr user tags and EXIF tags for most of the photos.
8. **PAN-PC-12:** This collection contains manually written plagiarized documents whose sources have been retrieved also manually from the ClueWeb using the ChatNoir search engine. The collection also contains automatically generated plagiarism cases with various degrees and kinds of obfuscation. These cases are constructed similarly to the previous PAN plagiarism corpora.
9. **PAN Wikipedia quality flaws corpus 2012 (PAN-WQF-12):** This corpus is based on the English Wikipedia snapshot from January 4, 2012. The corpus contains for each of the ten quality flaws Wikipedia articles that are exclusively tagged with the

[20] http://www.europeana.eu/

respective cleanup tag. The corpus contains also untagged articles, which have not been tagged with any cleanup tag.

10. **Plant Leaves II:** This database is a collaborative botanical dataset built during the first two years of the Pl@ntNet Project[21]. The database focuses on leaves of 126 plant species, mainly trees from French Mediterranean area. It contains 11,572 pictures of leaves subdivided into 3 different categories of pictures: 6,630 scans, 2,726 scan-like (photos of a leaf with a white uniform background), and 2,216 free natural unconstrained photos of leaves on the tree. Each picture is associated with a xml file containing various metadata (GPS, locality, full APGIII taxon, author, date...). The collaborative context induces a great diversity (morphological variation for a same species, framing and shooting conditions, places, growing stages of plants...)
11. **PubMedCentral:** A collection of medical images obtained from PubMed Central extending the subset used in 2011. 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.
12. **Pythia:** This collection consists of personal photos that have been contributed by 19 photographers. The documents for the collection have been picked randomly. To ensure a variance in photographic motifs and style, the contributors have been chosen from different demographic groups, e.g. ranging from year of birth 1,944 to 1,985.
    The presented collection has been annotated manually with a sophisticated graded relevance scale and provides rich metadata such as demographic and event information.
13. **QA4MRE 2012:** A multilingual collection of reading comprehension tests of given documents. Each test consists of one single document (Test Document) with 10 questions and 5 candidate answers.
14. **RepLab 2012:** This collection consist of tweets crawled per company name, for six companies (Apple, Lufthansa, Alcatel, Armani, Marriott, Barclays) using the company name as query, in English and Spanish. For each company's timeline, 300 tweets have been manually annotated by reputation management experts. The rest is the "background" dataset and has not been annotated.
15. **Traditional authorship attribution:** A collection of documents for authorship attribution. The corpus was collected from the free fiction collection published by Feedbooks.com, including both classic fiction that is now out-of-copyright as well as fiction, represented by the Feedbooks.com site).
16. **Tweet Contextualization 2012 Document collection:** This document collection has been built based on a recent dump of the English Wikipedia from November 2011. A plain XML corpus was target for an easy extraction of plain text answers. All notes and bibliographic references that are difficult to handle were removed and non empty Wikipedia pages (pages having at least on section) were only kept.
17. **VIDA:** This collection contains image sequences of indoor rooms (in the IDIAP research building) acquired from a mobile robot using visual and 3D point cloud data, at different times of the day with varying illumination conditions.

[21] http://www.plantnet-project.org/

18. **WEBUPV250k:** This collection is composed of 250,000 images for training, 1,000 for development and 2,000 for test. For each image, there were 7 visual feature types and 4 textual feature types. The development and test set images have been manually labelled for 115 concepts for evaluation.
19. **Wikipedia-LOD (v1.2):** The core data collection consists of Wikipedia articles and RDF properties from DBpedia 3.7 and YAGO2. Each Wikipedia article corresponds to an entity/resource in DBpedia and YAGO2. The Wikipedia articles in the collection were based on the MediaWiki-formatted dump dated on July 22, 2011, which corresponds to the version 3.7 of DBpedia. To facilitate participants process the data collection, a fused collection of XML files were built by first converting each raw Wikipedia article (originally in MediaWiki markup) into a customized XML-formatted file, and then appending the RDF triples imported from both DBpedia and YAGO2 that contain the article entity as subject or object to the article's XML file. Unfortunately, due to the complexity of MediaWiki markup, the parser employed failed in parsing all articles successfully, and thus resulted in a subset of Wikipedia articles in the fused collection.

Table 20: Collections used in the tasks of the CLEF 2012 Labs.

Lab	Task(s)	Collection	Number of documents	Size	Languages	Collection created for the lab	Number of years collection used in lab
CHiC	Ad-hoc Retrieval	Europeana	23,300,932	132 GB	EN, DE, FR	Yes	1
	Variability						
	Semantic Enrichment						
CLEF-IP	Chemical Image Extraction and Recognition	CLEF-IP 2012	3.5 million XML documents + 1,304 images in the image recognition tasks	13.5GB	EN, DE, FR	Yes	1
	Flowchart Recognition						
	Passage Retrieval Starting from Claims						
ImageCLEF	Flickr Photo Annotation and Retrieval	MIRFLICKR	annotation: 25,000, retrieval: 200,000	1,000,000	Any language	No	4

Lab	Task(s)	Collection	Number of documents	Size	Languages	Collection created for the lab	Number of years collection used in lab
	Medical Image Classification and Retrieval	PubMedCentral	Over 300,000 images of 75,000 articles	18GB	Mainly EN	Yes	1
	Plant Identification	PlantLeaves II	11,572 x 2 (jpg + xml)	23,144 files, 1.38 Go	Not applicable	Yes	2
	Pilot Task on Personal Photo Retrieval	Pythia	5,555	555	EN	No	1
	Robot Vision	VIDA	8 sequences	10MB	EN	No	1
	Scalable Image Annotation using General Web	WEBUPV250k	253,000		European languages, mainly EN.	Yes	1

Lab	Task(s)	Collection	Number of documents	Size	Languages	Collection created for the lab	Number of years collection used in lab
	Data						
INEX	Linked Data	Wikipedia-LOD (v1.2)	3,164,041	61GB	EN	Yes	1
	Relevance Feedback Snippet Retrieval	INEX Wikipedia collection (2009)	2,666,190	50.7GB	EN	No	1
	Social Book Search	Amazon/LibraryThing	2.8 million		EN	Yes	1
	Tweet Contextualization	Tweet Contextualization 2012 Document collection	3 ,691,092	7.8 Go	EN	Yes	1
PAN	Plagiarism Detection	PAN-PC-12	Candidate Retrieval: 40 Text Alignment: 8,033	1 GB	EN	Yes	1

Lab	Task(s)	Collection	Number of documents	Size	Languages	Collection created for the lab	Number of years collection used in lab
	Quality Flaw Prediction in Wikipedia	PAN Wikipedia quality flaw corpus 2012 (PAN-WQF-12)	1,592,226		EN	Yes	1
	Traditional Authorship Attribution	Traditional authorship attribution	about 70	~2MB	EN	Yes	1
QA4MRE	Machine Reading of Biomedical Texts about Alzheimer's Disease	The Alzheimer's Disease Literature Corpus (ADLC corpus)	75,994	33G	EN	Yes	1
	Question Answering	QA4MRE 2012	16	2 MB	EN, ES, DE, IT, RO, AR, BG	Yes	1
	Question Answering	Background Collections - Main Task	301,488	37 MB	EN, ES, DE, IT, RO, AR, BG	Yes	2

Lab	Task(s)	Collection	Number of documents	Size	Languages	Collection created for the lab	Number of years collection used in lab
RepLab	Monitoring Profiling	RepLab2012	~60,000		EN,ES	Yes	1

Table 21: Topics used in the tasks of the CLEF 2012 Labs.

Lab	Task	What constitutes a topic for this task?	Topics	Languages
CHiC	Ad-hoc Retrieval	Topics are taken from real-life Europeana query topics and consist of a mixture of topical and named-entity queries. Navigational queries are rarely seen in Europeana; however queries for people, places and works (named entities) occur very often. The short topics in title-format only (e.g. "Eiffel tower") reflect real expressed user needs and are distributed according to query category statistics (mostly named entities, some topical queries etc.) in a cultural heritage digital library researched previously. All 50 queries were then translated into French and German. For the variability and semantic enrichment tasks, only the first 25 topics were used for the experiments.	50 / 25	EN, FR, DE
	Variability			
	Semantic Enrichment			
CLEF-IP	Chemical Image Extraction and Recognition	A patent image, a chemical structure image.	960	Not applicable
	Flowchart Recognition	A bitmap file depicting a flowchart	150	Not applicable
	Passage Retrieval Starting from Claims	A set of claims of a patent application. However, the full patent application is available to the participants.	156	EN , FR, DE

Lab	Task	What constitutes a topic for this task?	Topics	Languages
ImageCLEF	Flickr Photo Annotation and Retrieval	For the annotation subtask a topic is a concept, whereas for the retrieval subtask it is a query.	Concepts: 94, Queries: 42	Whereas the topics themselves are specified in English, the textual metadata can be supplied in any language.
	Medical Image Classification and Retrieval	An information need in four languages and images	30 image-based topics and 10 case-based topics	EN, ES, FR, DE
	Plant Identification	A taxon (a family, a genus, a species, ...)	126	EN
	Pilot Task on Personal Photo Retrieval	Visual concepts and events	39	EN
	Robot Vision	Visual images and range images	2,445	EN

Lab	Task	What constitutes a topic for this task?	Topics	Languages
	Scalable Image Annotation using General Web Data	Concepts present in images.	115 concepts	Concepts are language independent. The textual part of the training data contains documents of most European languages, however the majority is EN.
INEX	Linked Data	Keyword queries for the ad hoc, faceted, and Jeopardy! tasks, and SPARQL FullText queries for the Jeopardy! task.	140	EN
	Relevance Feedback	A topic is a search query, typically only a few words long	50	EN
	Snippet Retrieval	A short content only (CO) query, a phrase title, a one line description of the search request and a narrative with the explanation of the information need, the context and the information need, and a description of what makes a document relevant or not.	35	EN
	Social Book Search	Each topic has a title and is associated with a group on the discussion forums	300	EN

Lab	Task	What constitutes a topic for this task?	Topics	Languages
	Tweet Contextualization	Tweets from twitter	1,000	EN
PAN	Plagiarism Detection	We have reused TREC topic descriptions to create a corpus for candidate retrieval. For text alignment we have resorted to the corpora used in PAN 2009-2011, where the notion of a topic was not present; in these cases, each suspicious document was considered a topic.	40 (candidate retrieval), 3,033 (text alignment)	EN
	Quality Flaw Prediction in Wikipedia	A set of Wikipedia articles	19,010	EN
	Traditional Authorship Attribution	An individual author	8	EN
QA4MRE	Machine Reading of Biomedical Texts about Alzheimer	Alzheimer's Disease	1	EN
	Processing Modality and Negation	A subject or area of interest as for example AIDS	4	EN, ES, DE, IT, RO, AR, BG
	Question Answering			

Lab	Task	What constitutes a topic for this task?	Topics	Languages
RepLab	Monitoring	A stream of tweets containing the name of an entity	~30,000	EN, ES
	Profiling	Tweets containing a company name, for several companies	~30,000	EN, ES

Table 22: Ground truth generation for the tasks in the CLEF 2012 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?
Cultural Heritage in CLEF (CHiC)	Ad-hoc Retrieval Variability Semantic Enrichment	80,367	6	4 internal (3 PhD students and one professor from Humboldt-Universität zu Berlin) and 2 external assessors	~200 hours
CLEF-IP	Chemical Image Extraction and Recognition	95	2	Igor Filippov, filippovi@mail.nih.gov, Chemical Biology Laboratory at NCI-Frederick Alan P. Sexton, A.P.Sexton@cs.bham.ac.uk, University of Birmingham	3 hours
	Flowchart Recognition	150	3	Mihai Lupu Florina Piroi Allan Hanbury	30 hours
	Passage Retrieval Starting from Claims	156	4	Mihai Lupu Florina Piroi Allan Hanbury Linda Andersson (Vienna University of Technology	30 hours

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	Flickr Photo Annotation and Retrieval	annotation: 2,000,000, retrieval: 100,000	Crowdsourcing, too many to tell	For the annotation task we employed crowdworkers active on Mechanical turk through the intermediary CrowdFlower, who for most of the 94 concepts had to evaluate all 25,000 images in the collection. For some concepts I could reuse many of last year's annotations. The gold standard - for filtering out badly performing assessors - I manually specified myself, roughly 125 images per concept, whereby I limited the number of images a single assessor could do to prevent them from seeing the same gold standard image twice. For the retrieval task I	Crowdsourcing, difficult to tell

Lab	Task(s)	How many documents were assessed?	How many assessors were employed?	Who were the assessors?	How much time did the assessors spend?
				had access to professional editors of NIST who performed the initial gold standard collection from the pool of aggregated results in the submitted runs, where roughly 300 images per query were assessed. These images then acted as the gold standard in the actual relevance assessment of the pools, where I once again solicited the help of crowdworkers through CrowdFlower.	
	Medical Image Classification and Retrieval	Classification: 2,000 images	Classification: 18 Retrieval:11	Classification: Researchers in the medical imaging. Retrieval: physicians Medical doctors at OHSU	Classification: 96 hours Retrieval: 235 hours

Lab	Task(s)	How many documents were assessed?	How many assessors were employed?	Who were the assessors?	How much time did the assessors spend?
	Plant Identification	11,000	Social network, 56 contributors	Members of Telabotanica social network	3 minutes per image (they really took pictures of plant)
	Pilot Task on Personal Photo Retrieval	5,555	42	Most of the assessors were students with a background in economic, the second largest group has a background in computer sciences and information technology	2-3 hours per topic
	Robot Vision		1	One of the organizer	3 hours
	Scalable Image Annotation using General Web Data	3,000	4	Members of our lab.	6 hours
INEX	Linked Data	11,000	Unknown	Amazon Mechanical Turk	10 hours

Lab	Task(s)	How many documents were assessed?	How many assessors were employed?	Who were the assessors?	How much time did the assessors spend?
	Relevance Feedback	89,756	Unknown	Unknown (Assessments were reused and combined from previous iterations of the INEX Ad Hoc track)	Unknown
	Snippet Retrieval	20 per assessor	Unknown	Unknown	Unknown
	Social Book Search	60,000	Unknown	Unknown	Unknown
	Tweet Contextualization	94,500	21	Informatively assessed by 5 organizers Readability of summaries checked by 16 participants	16 hours per assessor
PAN	Plagiarism Detection	all	Not applicable	Not applicable	Not applicable
	Traditional Authorship Attribution	~75	2	Myself and two graduate students	1 man month
	Quality Flaw Prediction in Wikipedia	208,228	Unknown	Wikipedia users	Unknown

Lab	Task(s)	How many documents were assessed?	How many assessors were employed?	Who were the assessors?	How much time did the assessors spend?
QA4MRE	Machine Reading of Biomedical Texts about Alzheimer	4	2	2 organizers. The evaluation was performed automatically	5 days
	Processing Modality and Negation	8	2	1 organiser and 1 PhD student. The evaluation was performed automatically	3 days

Lab	Task(s)	How many documents were assessed?	How many assessors were employed?	Who were the assessors?	How much time did the assessors spend?
	Question Answering	16	7	Pamela Forner, Center for the Evaluation of Language and Communication Technologies, forner@celct.it Alvaro Rodrigo, IR&NLP Group at UNED, alvarory@lsi.uned.es Richard Sutcliffe, School of CSEE (University of Essex), rsutcl@essex.ac.uk Caroline Sporleder, Saarland University, csporled@coli.uni-sb.de Corina Forascu, Al. I. Cuza University of Iasi, corinfor@info.uaic.ro Yassine Benajiba, Philips Research North America, Yassine.Benajiba@philips.com Petya Osenova, Bulgarian Academy of Sciences,	2 months

Lab	Task(s)	How many documents were assessed?	How many assessors were employed?	Who were the assessors?	How much time did the assessors spend?
				petya@bultreebank.org	
RepLab	Monitoring Profiling	1,800	Unknown	Reputation management experts	Unknown