PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Deliverable 7.12
Inventory of evaluation resources and evaluation packages

Version 1.0, September 2013

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

Document Information

Deliverable number:	7.12
Deliverable title:	Inventory of evaluation resources and evaluation packages
Delivery date:	30/09/2013
Lead contractor for this deliverable	ELDA
Author(s):	Priscille Schneller, Khalid Choukri, ELDA
Participant(s):	ELDA
Workpackage:	WP7
Workpackage title:	Dissemination, IPR and Resources
Workpackage leader:	ELDA
Dissemination Level:	PU – Public
Version:	1.0
Keywords:	Language Resources, Evaluation resources, Evaluation packages, CLEF Labs, Sustainability

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
0.1	2/03/2013	Draft	ELDA	
1.0	30/09/2013	Final	ELDA	

Abstract

The purpose of the Deliverable D7.12 is to report on the evaluation resources used and produced within the PROMISE project between September 2010 and August 2013. This includes all the data sets used and created during the evaluation tasks conducted in the framework of the Cross-Language Evaluation Forum (CLEF). In addition, this report contains the list of "Evaluation Packages" that have been made available in order to make such test collections reusable for benchmarking purposes.

Additional input on the work carried out since the establishment of the CLEF initiative is given in appendix for reference but also to shed more light on the community-built assets.

Table of Contents

Executive Summary

The purpose of the Deliverable D7.12 is to report on the evaluation resources used and produced within the PROMISE project between September 2010 and August 2013. Additional input regarding data sets compiled and packaged within previous projects built on the CLEF initiative (CLEFs, Treble-CLEF) are given as an appendix to emphasize the importance of the assets produced so far and on which the community is capitalizing.

The aim of PROMISE was to provide a virtual and open laboratory for conducting participative research and experimentation in order to carry out, advance and bring automation into the evaluation and benchmarking of complex multimedia and multilingual information systems.

In such a context, one of the activities of PROMISE was the organization of evaluation activities. A set of Evaluation Labs were jointly organized with the CLEF peer-reviewed Conference on a yearly basis. This report provides an inventory of all the data sets used during the CLEF Evaluation Labs between 2010 and 2013.

In order to make this huge amount of data reusable for benchmarking purposes, "Evaluation Packages" have been produced, containing all the data sets, topics, guidelines, relevance assessments, results and papers from the evaluation campaigns. This report provides the list of Evaluation Packages that have been compiled to be made available.

1 Introduction

The Deliverable D7.12 reports on the evaluation resources used and produced within the PROMISE project between September 2010 and August 2013.

One of the main goals of PROMISE has been to provide a virtual and open laboratory for conducting participative research and experimentation in order to carry out, advance and bring automation into the evaluation and benchmarking of complex multimedia and multilingual information systems. In order to reach this goal PROMISE planned to facilitate management and offer access, curation, preservation, re-use, analysis, visualisation, and mining of the collected experimental data.

The pillars of the PROMISE virtual institute (see Figure 1) are evaluation activities organised on a yearly basis and the realistic use cases and evaluation tasks designed for compelling user and industrial needs.

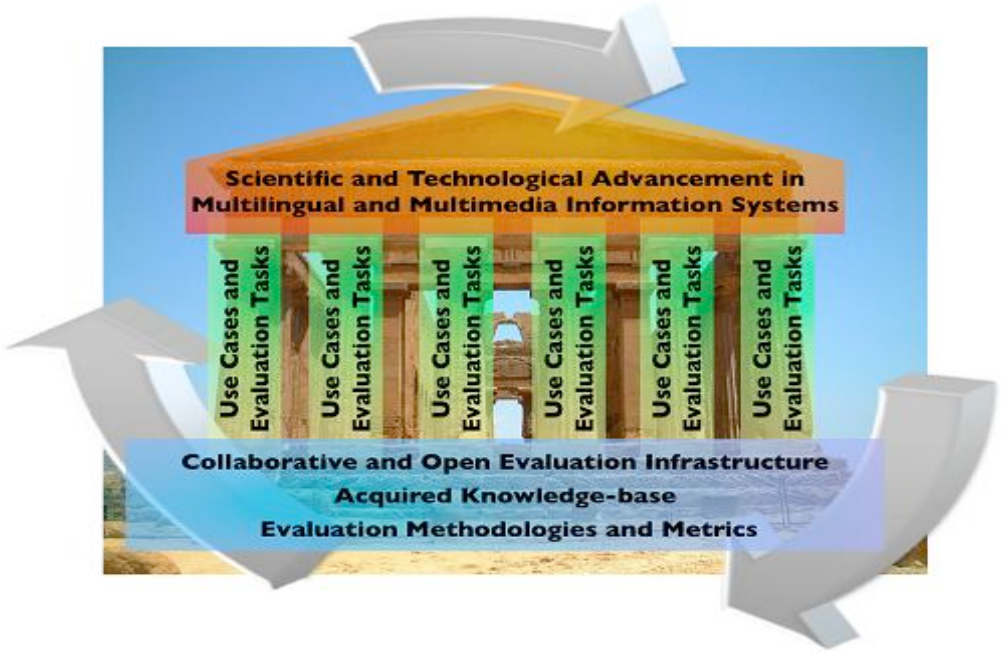

Figure 1. Representation of the PROMISE virtual institute

These evaluation tasks have been carried out through the CLEF evaluation framework, which over the past decade has been running with the support of major players in Europe, many of which are part of the PROMISE consortium.

The Cross-Language Evaluation Forum (CLEF) is a consortium-based partnership that takes in charge the organisation of an annual conference plus a set of focused benchmarking activities, called **Evaluation Labs**, i.e. laboratories to conduct evaluation of information access systems on particular topics and/or specialized domains, and associated workshops to discuss and pilot innovative evaluation activities.

Each Evaluation Lab provided to its participants data sets for training and/or testing their system. An inventory of the data sets used and/or created between 2010 and 2013 is presented in Section 2, while a summary of CLEF achievement, back to its genesis in 2000, is appended in Appendix 2.

In order to achieve one of the objectives of CLEF, which is to "create reusable test collections for benchmarking", each Lab's organizer has been requested (and highly encouraged) to create an "Evaluation Package", containing all the data sets, topics, guidelines, relevance assessments provided to the participants, along with the results and the papers from the evaluation campaign. ELRA supported such a creation and was involved in some of the legal tasks required to clear the Language Resources IPRs so that such packages could be made available to the community. The list of the Evaluation Packages made available is provided in Section 3.

An overview of the evaluation Labs conducted between 2010 and 2013 is presented below:

Task	2010	2011	2012	2013
CLEF-IP: Retrieval in the Intellectual Property Domain	●	●	●	●
ImageCLEF: Image Retrieval in CLEF	●	●	●	●
LogCLEF: Log File Analysis	●	●		
PAN: Uncovering plagiarism, authorship and social software misuse	●	●	●	●
QA@CLEF / QA4MRE: Question Answering for Machine Reading Evaluation	●	●	●	●
CriES: Cross-lingual Expert Search	●			
WePS: Searching Information about Entities in the Web	●			
MusiCLEF: Music Information Retrieval		●		
CHiC: Cultural Heritage in CLEF			●	●
INEX: Evaluation of XML retrieval			●	●
RepLab: Evaluation of Online Reputation Management Systems			●	●
CLEFeHealth: Cross-Language Evaluation for eHealth Document Analysis				●
QALD-3: Question Answering over Linked Data				●

2 Evaluation resources

Data sets used in the different Evaluation Labs between 2010 and 2013 are listed in section 2.1. Some data sets have been re-used in different Labs, which are described in section 2.2.

In conjunction with data sets creation/settlement, ground truth has been generated from collections. A list of data produced in the framework of the PROMISE project is given in section 2.3.

2.1 List of data sets used in CLEF Labs

2.1.1 CLEF 2010

Below is a table of data sets used in the 2010 CLEF Labs:

Lab / Task	Data set	Description	Language	Size
CLEF-IP2010 / Classification and Prior Art Search Tasks	CLEF-IP 2010 (extract of the MAREC database)	Database containing over 2.6 million patent documents, corresponding to approximately 1.3 million individual patents published until 2001 from the European Patent Office.	EN, DE, FR	Corpus: 9.5GB (archived) Topics: 42MB (archived)
CriEs2010	CriES Yahoo! Answers Collection	Subset of an official crawl released by Yahoo! Contains a total of 780,193 documents.	EN, DE, FR, ES	~600MB
ImageCLEF2010 / Medical Retrieval Task	Collection of RSNA images	74,902 documents. Subset of the Radiological Society of North America (RSNA) journals' image collections.	EN	16GB
ImageCLEF2010 / Visual Concept Detection and Annotation Task	MIR Flickr 25.000 images dataset	18,000 images from the MIR Flickr collection including EXIF data and Flickr user tags.	EN	approx 2.9GB
ImageCLEF2010 / Robot vision Task	COLD-Stockholm	9,592 image sequences acquired using a robot platform equipped with a stereo camera system.	Language-independent	4.5GB
ImageCLEF2010 / Wikipedia Retrieval Task	ImageCLEF2010 Wikipedia collection	237,434 Wikipedia images with user-provided annotations and Wiki articles that contains these images. Based on a September 2009 dump.	EN, DE, FR	25GB

Lab / Task	Data set	Description	Language	Size
LogCLEF2010	The European Library (TEL) logs	A large log of activities from The European Library (TEL). Total of 760,000 records from January to December 2009.	Several languages	~2GB
	Deutscher Bildungsserver (DBS) logs	The "Deutscher Bildungsserver" is a quality controlled internet directory for educational resources. A raw server log representing three months of activities on the portal is made available (09/2009 to 11/2009)	mostly DE, some EN	5GB
PAN2010 / Wikipedia Vandalism Detection Task	PAN-WVC-10 (PAN Wikipedia Vandalism Corpus 2010)	Contains meta information about edits on Wikipedia articles along with the edited articles as wiki text articles.	EN	447MB (compressed)
QA@CLEF2010	ResPubliQA 2010	Contains a subset of the JRC-Acquis multilingual parallel corpus and a small portion of the EUROPARL collection. Total of 10,700 documents.	EN, DE, FR, IT, PT, RO, ES, BG, NL	670MB
WePS2010	WePS-3	Consists of the top 200 search results from Yahoo! API for 300 different ambiguous parson names, with human assessments of the correct way to group these documents. Total of 60,000 documents.	EN	520MB
	WePS-3 ORM	Set of Twitter entries containing an ambiguous company name. Each organization is associated with the company name and its homepage. Total of 20,000 documents.	EN	81.6MB

2.1.2 CLEF 2011

Below is a table of data sets used in the 2011 CLEF Labs:

Lab / Task	Data set	Description	Language	Size
CLEF-IP2011 / Classification and Prior Art Search Tasks	CLEF-IP 2011	A set of 3000 topic documents comprising 1000 English, German and French language documents each, from the MAREC database containing over 2.6 million patent documents, corresponding to approximately 1.3 million individual patents published until 2001 from the European Patent Office (EPO) and the corresponding patent documents published by the WIPO (more than 400,000 documents).	EN, DE, FR	Corpus: 13.5GB (archived)
CLEF-IP2011 / Image-based Retrieval Task	Subset of CLEF-IP 2011 + images	The target data set contains all images for patent documents in three IPC sub-classes that have an application date previous to 2002 (291566 images grouped by patent document)	EN, DE, FR	Corpus: 338MB + 4.6GB images (archived)
CLEF-IP 2011 / Image-based Classification Task	Separate set of images	Training data with between 300 and 6,000 training images for each class (9 classes). Test data consisted of 1,000 unclassified images.	Language-independent	Training data: 337MB (archived)
ImageCLEF2011/ Medical Retrieval Task	ImageCLEFmed 2011 Collection	231,000 figures from PubMed Central articles (includes figures from BioMed Central journals)	EN	20GB
ImageCLEF2011/ Visual Concept Detection and Annotation Task	MIRFLICKR-1M Image Collection	1 million images from the MIR Flickr collection including EXIF data and semantic tags from flickr.com.	EN	Approx. 16.5GB
ImageCLEF2011/ Wikipedia Retrieval Task	ImageCLEF 2010 Wikipedia collection	Same data set as in ImageCLEF2010. See description in section 2.1.1.	EN, DE, FR	25GB

Lab / Task	Data set	Description	Language	Size
ImageCLEF2011/ Plant Identification Task	Pl@ntLeaves collection	Contains around 5436 pictures subdivided into 3 different kinds of pictures: scans (3070), scan-like photos (897) and free natural photos (2469).	mainly FR	
LogCLEF2011	The European Library (TEL) logs	A large log of activities from The European Library (TEL). Total of 950,000 records from January to December 2010,	Several languages	~2GB
	Deutscher Bildungsserver (DBS) logs	Same data set as in LogCLEF2010. See description in section 2.1.1.	mostly DE, some EN	5GB
	Sogou dataset	The Sogou query logs contain queries to the Chinese Sogou search engine.	ZH	
PAN2011 / Plagiarism Detection Task	PAN-PC-10 (PAN Plagiarism Corpus 2010) and PAN-PC-11 (PAN Plagiarism Corpus 2011)	These corpora contain books downloaded from the project Gutenberg in the form of text documents. In total it is based on 22,000 English books, 520 German books, and 210 Spanish books.	EN, DE, ES	1.7GB each (archived)
PAN2011 / Wikipedia Vandalism Detection Task	PAN-WVC-10 and PAN-WVC-11	Contains meta information about edits on Wikipedia articles and edited articles as wiki text articles.	EN	PAN-WVC-11: 370.8MB (archived)
PAN2011 / Author Identification	PAN 11 Authorship identification corpus	Subset of the Enron Email Corpus, including five training collections of real-world texts (often short and messy), one with 26 different authors, one with 72 different authors, and three each with a single author (for author verification).	EN	Approx. 3MB

Lab / Task	Data set	Description	Language	Size
QA4MRE 2011	Background collection	The Background Collections are comparable (but not identical) topic-related collections created in all the different languages of the task. Texts are drawn from many sources: newspapers, newswire, web pages, blogs and Wikipedia entries. Total of 10,700 documents.	EN, DE, IT, RO, ES	2,18MB
MusiCLEF2011	Audio collection	Large collection (more than 200,000 files) of audio files in MP3 format, manually annotated with information regarding authorship, title, relationship between recordings and compositions, physical support, label, editorship, and so on.	Language-independent	

2.1.3 CLEF 2012

Below is a table of data sets used in the 2012 CLEF Labs:

Lab / Task	Data set	Description	Language	Size
CHiC2012	French Europeana collection	Contains all the Europeana documents with French metadata records.	FR	395MB (archived)
	English Europeana collection	Contains all the Europeana documents with English metadata records.	EN	144MB (archived)
	German Europeana collection	Contains all the Europeana documents with German metadata records.	DE	528MB (archived)
CLEF-IP2012 / Claims to Passage	CLEF-IP 2012	Same as the CLEF-IP 2011 corpus. See description in Section 2.1.2	EN, DE, FR	Corpus: 13.5GB (archived)
CLEF-IP2012 / Flowchart recognition task	Separate set of images	Training data: set of 50 images containing flowcharts, and the corresponding text files Test data: 100 black and white images	Language-independent	1.7MB (archived)
CLEF-IP2012 / Chemical Structure Recognition task	Separate set of images	1) Bounding box extraction: - training data: set of 30 patents and manually extracted image clips - test data: set of patent files for which you need to extract the bounding boxes of all chemicals 2) Structure recognition: subset of the images extracted above, with the corresponding MOL files	Language-independent	209MB (archived)
ImageCLEF2012/ Medical Retrieval Task	ImageCLEFmed 2011 Collection	Same data set as in ImageCLEF2011. See description in section 2.1.2	EN	

Lab / Task	Data set	Description	Language	Size
ImageCLEF2012/ Photo Annotation and Retrieval Task (Task 1: Visual Concept Detection and Annotation)	MIRFLICKR-1M Image Collection	Same data set as in ImageCLEF2011. See description in section 2.1.2	EN	Approx. 16.5GB
ImageCLEF2012/ Photo Annotation and Retrieval Task (Task 2: Scalable Image annotation)	ImageCLEF 2012 webupv250k Image Annotation Dataset	Subset of 250,000 images extracted from a database of millions of images downloaded from the Internet. URLs of the images obtained by querying popular image search engines (namely Google, Bing and Yahoo) when searching for words from an English dictionary.	EN	21MB
ImageCLEF2012/Personal Photo Retrieval Task	Pythia Image Collection v1	Consists of 5,555 images plus rich metadata as they have been found on hard disks of 19 contributors ranging from year of birth 1944 to 1985.	EN	
ImageCLEF2012/ Plant Identification Task	Pl@ntLeaves II dataset	Contains 1572 pictures subdivided into 3 different kinds of pictures: scans (57%), scan-like photos (24%) and free natural photos (19%).	mainly FR	
ImageCLEF2012/ Robot vision Task	Set of the Robot Vision VIDA database	Depth images, acquired with the kinect device	Language-independent	
PAN2012 / Plagiarism Detection Task (Candidate Document Retrieval task)	PAN12 Plagiarism Candidate Retrieval Corpora	Set of suspicious documents, each of which about a specific topic and plagiarized from web pages on that topic found in the ClueWeb09 corpus. The training corpus contains annotations that reveal the plagiarism whereas the test corpus is without annotations.	EN	392KB

Lab / Task	Data set	Description	Language	Size
PAN2012 / Plagiarism Detection Task (Detailed Comparison task)	PAN12 Detailed Comparison Corpus	A set of pairs of suspicious document and potential source document. The suspicious document may contain passages of text plagiarized from the source document. The corpus contains automatically and manually generated plagiarism, including annotations that reveal where they are.	EN	500MB
PAN2012 / Author Identification (Traditional Authorship Attribution task)	PAN12 Authorship Attribution Corpora	Data collected from the free fiction collection published by Feedbooks.com	EN	Approx 13.8MB
PAN2012 / Author Identification (Sexual Predator Identification task)	PAN12 Sexual Predator Identification Corpora	Data collected from websites where logs of online conversations between convicted sexual predators and volunteers posing as underage are available.	EN	Approx 110.6MB
PAN2012 / Quality flaw detection in Wikipedia	PAN12 Wikipedia Quality Flaw	The training corpus contains 154,116 tagged articles for each quality flaw plus untagged articles with labels given. The test corpus contains 19,010 articles, including a balanced number of tagged articles and untagged articles for each flaw. No labels are given. All the articles were extracted from the English Wikipedia snapshot from 04/01/2012.	EN	388.2MB

Lab / Task	Data set	Description	Language	Size
QA4MRE 2012	QA4MRE 2012 Background Collections	The 2012 background collections are based on but not identical to the 2011 collections (see description in section 2.1.2), with the addition of 1,000 new documents for each topic in all languages.	AR, BG, DE, EN, ES, IT, RO	
	Alzheimer's Disease Literature Corpus (ADLC corpus)	Consists of abstracts and full text articles about Alzheimer's Disease. The following sets of documents are provided: (1) 66,222 abstracts from PubMed. (2) 8,249 Open Access full articles from PubMed Central in .pdf format. (3) 379 Full articles from Elsevier and 103 abstracts. (4) 1,041 full text articles from PubMed Central in html and txt format.	EN	
RepLab2012	Twitter data in English and Spanish	Trial data: 30,000 tweets crawled per company name, for six companies (Apple, Lufthansa, Alcatel, Armani, Marriott, Barclays) using the company name as query, in English and Spanish. Test data is identical to trial data, for a different set of 25 companies.	EN, ES	
INEX 2012 / Books and Social Search track (Social Book Search subtask)	Amazon/Library Thing collection	Contains metadata for 2.8 million books crawled from the online book store of Amazon and the social cataloging web site of LibraryThing in February and March 2009 by the University of Duisburg-Essen. The data set is in XML.	EN	

Lab / Task	Data set	Description	Language	Size
INEX 2012 / Books and Social Search track (ProveIt subtask)	Digitized Book Corpus	Consists of over 50,000 digitized, out-of-copyright books, provided by Microsoft Live Book Search and the Internet Archive (for non-commercial purposes only). The full text of the books is in an XML format, referred to as BookML.	EN	400GB
INEX 2012 / Linked Data track	Wikipedia-LOD (v1.1)	Contains an overall amount of 3.1 Million individual XML articles. Each Wikipedia-LOD article consists of a mixture of XML tags, attributes, and CDATA sections, containing infobox attributes, free-text contents, describing the entity or category that the article captures, and a section with both DBpedia and YAGO2 properties that are related to the article's entity.	EN	
INEX 2012 / Tweet Contextualization track	Set of 1000 tweets and document collection from Wikipedia	A set of about 1000 tweets in English selected among informative accounts and collected by the track organizers from Twitter Search API. Document collection built based on a dump of the English Wikipedia from November 2011.	EN	
INEX 2012 / Snippet Retrieval track and Relevance Feedback track	INEX 2009 Wikipedia collection	An XML version of the English Wikipedia, based on a dump from October 2008 and semantically annotated. Total of 2,666,190 documents.	EN	50.7GB

2.1.4 CLEF 2013

Below is a table of data sets used in the 2013 CLEF Labs:

Lab / Task	Data set	Description	Language	Size
CHiC 2013	CHiC Europeana collection	Combination of the 13 language-based sub-collections of Europeana (each collection containing more than 100,000 documents in one language). Downloaded in March 2012.	DE, FF, SV, IT, ES, NO, NL, EN, PL, FI, SL, EL, HU	
CLEF-IP2013 /	CLEF-IP 2012 collection	Same as the CLEF-IP 2011 corpus. See description in Section 2.1.2	EN, DE, FR	Corpus: 13.5GB (archived)
CLEF-IP2013 / Flowchart recognition task	Separate set of images	Same set of images than in CLEF-IP2012 with an additional set of 50 images containing flowcharts and their corresponding text files.	Language-independent	- 2012 data set: 1.7MB - 2013: 13MB (archived)
ImageCLEF2013/ Medical Retrieval Task	ImageCLEFmed 2011 Collection	Same data set as in ImageCLEF2011 and 2012 Medical Retrieval Task. See description in section 2.1.2	EN	20GB
ImageCLEF2013/ Photo Annotation and Retrieval Task (Task 1: Scalable Image annotation)	ImageCLEF 2012 webupv250k Image Annotation Dataset	Same data as in ImageCLEF2012 Scalable Image Annotation Task. See description in section 2.1.3	EN	21MB
ImageCLEF2013/ Photo Annotation and Retrieval Task (Task 2: Personal Photo Retrieval)	Pythia Image Collection v1	Same data as in ImageCLEF2012 Personal Photo Retrieval Task. See description in section 2.1.3	EN	

Lab / Task	Data set	Description	Language	Size
ImageCLEF2013/ Plant Identification Task	Pl@ntView dataset	Focuses on 250 herb and tree species from France area. It contains 26077 pictures. The training data results in 20985 images, and the test data in 5092 images.	mainly FR	
ImageCLEF2013/ Robot vision Task	O-VIDA Robot Vision dataset	Consists of different training, validation and test sequences of depth and visual images acquired within an indoor environment. From this dataset two different labelled sequences were selected for training, one labelled sequence for validation, and one unlabelled sequence for testing.	Language-independent	
PAN2013 / Plagiarism Detection Task (Source Retrieval task)	PAN'12 Training Corpus for Plagiarism Source Retrieval	Same data as in PAN2012 Plagiarism Detection Task. See description in section 2.1.3	EN	1.09MB (archived)
PAN2013 / Plagiarism Detection Task (Text Alignment task)	PAN'13 Training corpus for the Plagiarism Detection, Text Alignment task	The corpus comprises 1.827 suspicious documents as plain text and 3230 source documents as plain text. Furthermore, the corpus contains 5.000 XML files which each report for a pair of suspicious and source document the exact locations of the plagiarized passages.	EN	12MB (archived)
PAN2013 / Author Identification	PAN'13 Training Corpus for Author Identification Task	Consists of a number of separate problems, each of them having a number of "known" documents, all written by a single person and one "unknown" document. Contains 10 problems in English, 5 in Spanish and 20 in Greek.	EN, ES, EL	2.5MB

Lab / Task	Data set	Description	Language	Size
PAN2013 / Author Profiling	PAN'13 Training Corpus for Author Profiling Task	Consists of documents written in both English and Spanish. Contains posts of three age classes: 10s (13-17), 20s (23-27), and 30s (33-47). Moreover, documents from authors who pretend to be minors are included.	EN, ES	674MB
QA4MRE 2013	QA4MRE 2013 Background Collections	The 2013 background collections are based on but not identical to the 2012 collections (see description in section 2.1.3). Texts from many sources: newspapers, newswire, web pages, blogs and Wikipedia entries.	AR, BG, EN, RO, ES	3.6MB
	Alzheimer's Disease Literature Corpus (ADLC corpus)	Same data as in QA4MRE2012. See description in section 2.1.3	EN	
	"Entrance Exams" 2013 test set	Composed of reading comprehension tests taken from the Japanese Center Test, which is a nation-wide achievement test for Japanese university admissions. Data provided in the XML format	JA	
RepLab2013	Twitter data in English and Spanish	Collection of tweets referring to a selected set of 61 entities from four domains: automotive, banking, universities and music/artists. Crawling period: 1/06/2012 to 31/12/2012. For each entity, at least 2,200 tweets were collected (700 used as training set and the rest as test set).	EN, ES	

Lab / Task	Data set	Description	Language	Size
INEX2013 / Books and Social Search track (Social Book Search task)	Amazon/Library Thing collection	Same data as in INEX2013 Social Book Search Task. See description in section 2.1.3	EN	
INEX 2013 / Books and Social Search track (ProveIt task)	Digitized Book Corpus	Same data as in INEX2013 ProveIt Task. See description in section 2.1.3	EN	400GB
INEX 2013 / Linked Data track	INEX 2013 Linked Data Data set	Subset of DBpedia and YAGO2s together with a dump of Wikipedia core articles from June 2012.	EN	
	Wikipedia-LOD v2.0 (supplementary resource)	Similar to Wikipedia-LOD v1.1. (see description in section 2.1.3)	EN	
	Set of Wikipedia article texts as RDF Triples (supplementary resource)	Contains the full text (without XML markup) of each Wikipedia article in the 2012 dump.	EN	
INEX 2013 / Tweet Contextualization track	Set of 598 tweets	598 tweets in English have been collected by the organizers from Twitter®, selected among informative accounts in order to avoid purely personal tweets that could not be contextualized.	EN	
INEX 2013 / Tweet Contextualization track and Snippet Retrieval track	INEX 2013 Wikipedia dump	Document collection rebuilt based on a recent dump of the English Wikipedia from November 2012. Plain XML corpus.	EN	
CLEF eHealth / (Task 1 and 2)	MIMIC II database, v. 2.5	Deidentified clinical free-text notes	EN	

Lab / Task	Data set	Description	Language	Size
CLEF eHealth / (Task 3)	Khresmoi set of medical-related documents	Set of medical-related documents, provided by the Khresmoi project. Contains documents covering a broad set of medical topics, and does not contain any patient information.	European languages	
QALD-3	QALD-3 Dataset	Includes data from the English DBpedia (v.3.8) with multilingual labels, Spanish Dbpedia and an RDF export of MusicBrainz (English)	EN, ES	

2.2 List of data sets re-used on several years

Some data sets have been re-used within several campaigns. Below is a summary of the data sets re-used, associated with the Labs and years of use.

Data set	2010	2011	2012	2013
Extracts of the MAREC database	CLEF-IP	CLEF-IP	CLEF-IP	CLEF-IP
Europeana			CHiC	CHiC
ImageCLEF 2010 Wikipedia collection	ImageCLEF	ImageCLEF		
ImageCLEFmed 2011 collection		ImageCLEF	ImageCLEF	ImageCLEF
MIRFLICKR-1M Image Collection		ImageCLEF	ImageCLEF	
Pythia Image Collection v1			ImageCLEF	ImageCLEF
WEBUPV dataset			ImageCLEF	ImageCLEF
Amazon/Library Thing collection			INEX	INEX
Digitized Book Corpus			INEX	INEX
The European Library (TEL) logs	LogCLEF	LogCLEF		
Deutscher Bildungsserver (DBS) logs	LogCLEF	LogCLEF		
PAN Plagiarism Corpus 2010	PAN	PAN		
PAN Wikipedia Vandalism Corpus 2010	PAN	PAN		
PAN'12 Corpus for Plagiarism			PAN	PAN
Alzheimer's Disease Literature Corpus			QA	QA

2.3 List of annotation and relevance assessments produced

In the framework of the PROMISE project, annotation and relevance assessments were produced for some of the Evaluation tasks linked to the following PROMISE Use Cases:

- **Visual clinical decision support:**
  - Annotation of images for ImageCLEF 2011
  - Relevance Judgments for ImageCLEF2012
  - Relevance Judgments for ImageCLEF2013
- **Unlocking culture:**
  - Relevance assessment in German and French for CHiC2012 Lab
  - Relevance assessments in 12 European languages for CHiC2013 Lab

3 Evaluation packages

One of ELRA's roles within PROMISE (and beyond that within CLEF and similar initiatives) is to ensure that all resources used within an evaluation experiment is packaged and made publicly available so as to allow for replication of the experiments by any interested party. Each Lab's organizer has been asked to create an "Evaluation Package", containing all the data provided to the participants, along with the results and the papers from the evaluation campaign. ELRA supported such a creation and was involved in some of the legal tasks required to clear the Language Resources IPRs (Intellectual Property Rights) so that such packages could be made available to the community.

3.1 Description

An Evaluation package contains all the data used during an evaluation challenge, plus results and papers:

- datasets used for training and testing
- topics used for these evaluations
- guidelines provided to the participants
- the corresponding relevance assessments
- the official results obtained by the participants
- working notes of the CLEF campaigns

3.2 Distribution policy

As indicated above, the distribution policy should ensure that all packaged are available for evaluation purposes. In many cases, the data used within the evaluation is supplied by right holders that may (or may not) impose some legal constraints on its distribution. For instance some right holders require the data to be used exclusively for evaluation tasks and not for the development of technologies or for other purposes. Some of them even require the data to be deleted by the participants at the end of the evaluation campaign.

In order to comply with these constraints, ELRA drafted a specific license that allows using the data sets for "Evaluation Purposes Only". Such legal framework made it easy to convince the data owners to donate their resources to the project consortium and its labs.

In addition to the clarification of the purposes of use, the licence (referred to as "End-User Evaluation agreement") contains guidelines constraining the dissemination and publication of evaluation results to ensure that no comparisons of achieved results are used for marketing purposes or for endorsing products/companies (see Appendix 1).

Some data sets cannot be re-distributed. In such a case ELRA distributes an Evaluation Suite (without the dataset) and adds a pointer to data set location so that the users can access to it directly.

3.3 List of the evaluation packages made available

Evaluation package in the ELRA catalogue:

- CLEF Test Suite for the CLEF 2000-2003 Campaigns, ELRA-E0008
- CLEF AdHoc-News Test Suites (2004-2008), ELRA-E0036
- CLEF Domain Specific Test Suites (2004-2008), ELRA-E0037
- CLEF Question Answering Suites (2003-2008), ELRA-E0038
- CLEF QAST 2007-2009, ELRA-E0039

Evaluation packages to be finalized before their distribution in the ELRA catalogue:

- CLEF Question Answering Evaluation Package (2009-2013)
- CLEF-IP Evaluation Package (2009-2013)
- CLEF PAN Evaluation Package (2010-2011)
- CLEF eHealth Task 3 - Evaluation Package (2013)
- ImageCLEF – Medical image retrieval task Evaluation Package (2011-2013)
- ImageCLEF – Scalable image annotation task Evaluation Package (2011-2013)
- LogCLEF Evaluation Package(2009-2011)

Evaluation suites (package without the dataset) to be finalized before their distribution in the ELRA catalogue:

- ImageCLEF Photo Annotation task Evaluation Suite (2009-2010)
- CLEF WePS Evaluation Suite (2010)
- CLEF CriES Evaluation Suite (2010)
- CLEF eHealth Task 1 and 2 Evaluation Suite (2013)
- CLEF RepLab Evaluation Suite (2012-2013)

4 Conclusion

The inventory presented in this deliverable shows that a large amount of data has been produced for benchmarking purposes during the last four years of CLEF, either newly-created corpora or customized data sets from existing data. Most of the work was carried out with the critical support of PROMISE project.

A big effort has been devoted to making most of these packages available for future exploitation. Data produced in research projects are often lost a few years after the end of the project, especially when they are stored on simple web pages that disappear. The PROMISE project ensured that both legal, technical, practical/logistic aspects are properly addressed by the consortium to ensure such availability. A large part of the data described in this report will be re-distributed for evaluation purposes and made available through the ELRA catalogue of Language Resources (http://catalog.elra.info).

We do expect potential developers to be attracted by such packages and encouraged to share their own experiences with the community, even after the end of the official evaluations supported by PROMISE. The sharing paradigm is of paramount importance to the community to avoid duplication of efforts even if the number of copies of the CLEF evaluation packages distributed so far **is about 60 copies** (through the ELRA catalogue).

Appendix 1: Extracts of the "Evaluation License"

[…]

EXHIBIT D: Guidelines constraining the dissemination and publication of evaluation results are:

1. **SCIENTIFIC OR TECHNICAL PUBLICATIONS:** Scientific or technical publications, including newsletters from universities or research laboratories, should adhere to community standards for fairness and objectivity and should accurately and clearly state the limitations of the testing conditions and other factors which might influence scores. The experimental nature of the tasks, data and evaluation procedures should also be stated. The full evaluation packages should always be referenced.

2. **ADVERTISEMENTS:** No advertisements using the evaluation results can be placed in magazines, journals, newspapers, or other publications.

3. **PRESS RELEASES:** Press releases about the evaluation results to organizations with national/international coverage are also prohibited.

4. **MARKETING LITERATURE, LOCAL NEWSLETTERS:** Although it is recognized that extensive evaluation discussions are not appropriate in this type of literature, it is expected that any claims made on the basis of evaluation results are accurate, that the evaluation measures used to substantiate these claims are stated, and that a reference is made to the evaluation packages. Where promotional material is subject to prepublication revision by the media, the author should make every effort to see that the revision does not cause a violation of the guidelines.

5. **CROSS-SYSTEM COMPARISONS**: Cross-system comparisons may not be made with other named teams listed in the evaluation package documentation for individual tests, and may only be made when they are supported by accepted methods of statistical significance testing. Comparisons must be accompanied by the results of those tests and should reference the publication of those tests. Informal, qualitative comparisons with recognized baselines or benchmarks, and with general levels or trends in performance, must be clearly stated to be such and thus open to statistical reassessment.

Appendix 2: CLEF Resources since 2000

Task	Sub-task	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
AdHoc	CLIR on News	•	•	•	•	•	CLEF 2004-2008 Package (ELRA-E0036)			•	•				
AdHoc	Robust							•	•	•	•				
AdHoc	TEL@CLEF	CLEF 2000-2003 Package (ELRA-E0008)								•	•				
Domain spec	Scientific data retrieval (GIRT)	•	•	•	•	•	CLEF 2004-2008 Package (ELRA-E0037)			•					
Domain spec	Scientific data retrieval (Amaryllis)			•											
iCLEF	Interactive CLIR		•	•	•	•	•	•		•	•				
CL-SDR	Cross-Language Spoken Document Retrieval				•	•									
CL-SR	Cross-Language Speech Retrieval						•	•	•						
QA@CLEF	Main QA				•	CLEF 2003-2008 QA Package (ELRA-E0038)				•					
QA@CLEF	WSD (Word Sense Disambiguation)									•					
QA@CLEF	WWW (2005) and WiQA (2006)						•	•							
QA@CLEF	AVE (Answer Validation Exercise)							•	•	•					
QA@CLEF	TCE (Time-Constrained Exercise)							•							
QA@CLEF	QAST (Speech Transcripts)								•	CLEF QAST 2007-2009 (ELRA-E0039)					
QA@CLEF	GikiCLEF (*cf. GeoCLEF-GikiP*)										•				
QA@CLEF	QA (ResPubliQA)										•	•			
QA4MRE	QA												•	•	•
ImageCLEF	Photo Retrieval				•	•	•	•	•	•	•				
ImageCLEF	Interactive				•	•	•	•							
ImageCLEF	Medical Image Retrieval					•	•	•	•	•	•	•	•	•	•
ImageCLEF	Medical Image Annotation						•	•	•	•	•				
ImageCLEF	Visual Concept Detection and Annotation (photo annotation)							•	•	•	•	•	•	•	•
ImageCLEF	Wikipedia MM									•	•	•	•		
ImageCLEF	Robot Vision										•	•		•	•
ImageCLEF	Plant identification												•	•	•
ImageCLEF	Image annotation using general Web data													•	
ImageCLEF	Personal Photo Retrieval													•	•
WebCLEF	WebCLEF						•	•	•	•					
GeoCLEF	Main						•	•	•	•					
GeoCLEF	Query Parsing								•						
GeoCLEF	GikiP									•					
VideoCLEF	VideoCLEF									•	•				
INFILE@CLEF	INFILE@CLEF									•	•				
LogCLEF	LogCLEF										•	•	•		
CLEF-IP	Intellectual Property (Prior Art Candidate Search Task)										•	•	•		
CLEF-IP	Patent Classification											•	•		
CLEF-IP	Image-based Patent Retrieval												•		
CLEF-IP	Image-based Classification												•		
CLEF-IP	Claims to passage													•	•
CLEF-IP	Flowchart recognition task													•	•
CLEF-IP	Chemical Structure Recognition task													•	
CLEF-IP	Text to image/image to text														•
Grid@CLEF	Grid@CLEF										•				
PAN	Plagiarism detection											•	•	•	•
PAN	Wikipedia Vandalism Detection											•	•		
PAN	Authorship Identification												•	•	•
PAN	Quality Flaw Prediction in Wikipedia													•	•
WePS	document filtering (Online Reputation Management)											•			
WePS	Document Clustering Information Extraction											•			
CriES	expert search											•			
MusiCLEF	Content and Context-based Music Retrieval												•		
MusiCLEF	Music identification												•		
CHiC	Ad-hoc Retrieval Task													•	
CHiC	Variability Task													•	
CHiC	Semantic Enrichment Task													•	
CHiC	Multilingual Ad-hoc Retrieval Task														•
CHiC	Multilingual Semantic Enrichment Task														•
CHiC	Polish Task (automatic or manual)														•
CHiC	Interactive Task														•
RepLab	Profiling task													•	•
RepLab	Monitoring task													•	•
INEX	Social book search													•	•
INEX	Linked Data													•	•
INEX	Snippet Retrieval													•	•
INEX	Relevance Feedback													•	•
INEX	Tweet Contextualization													•	•
eHealth	Task 1 (identification of disorders from clinical reports and mapping of the SNOMED CT disorders to UMLS codes)														•
eHealth	Task 2 (mapping abbreviations and acronyms in clinical reports to UMLS codes)														•
eHealth	Task 3 (IR to address questions patients may have when reading clinical reports)														•
QALD-3															•