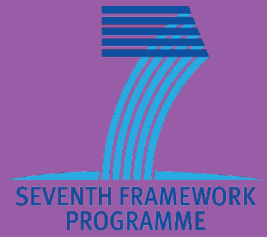
PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Deliverable 3.1
Initial prototype of the evaluation infrastructure

Version 1.0, 28th February 2011



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

Document Information

Deliverable number:	3.1
Deliverable title:	Initial prototype of the evaluation infrastructure
Delivery date:	28/02/2011
Lead contractor for this deliverable	UNIPD
Author(s):	Maristella Agosti, Giorgio Maria Di Nunzio, and Ncola Ferro - UNIPD
Participant(s):	UNIPD
Workpackage:	3
Workpackage title:	Evaluation Infrastructure
Workpackage leader:	UNIPD
Dissemination Level:	PU – Public
Version:	0.4
Keywords:	evaluation infrastructure, prototype, DIRECT, server cluster, CLEF 2011

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
0.1	22/02/2011	Draft	Giorgio Maria Di Nunzio, UNIPD	Circulated to UNIPD members
0.2	24/02/2011	Draft	Maristella Agosti, UNIPD	Circulated to UNIPD members
0.3	25/02/2011	Draft	Giorgio Maria Di Nunzio, UNIPD	Circulated to UNIPD members
0.4	26/02/2011	Final Draft	Maristella Agosti, UNIPD	Circulated to PROMISE reviewers
1.0	28/02/2001	Final	Nicola Ferro, UNIPD	Final version revised after partners' comments

Abstract

This deliverable describes the setup of the initial prototype of the evaluation infrastructure of PROMISE. The set-up has been prepared taking into account the support that the infrastructure has to give to the Cross-Language Evaluation Forum (CLEF) in 2011. In particular, this deliverable describes the migration of the DIRECT system from the original server to the new server cluster of PROMISE, and the creation of the infrastructure for the evaluation of two PROMISE use cases: search for innovation (CLEF-IP), and visual clinical decision support (ImageCLEF). They both represent exemplar cases when multilingual and multimedia information accesses are mixed together and need to be evaluated, which are one of the major goals of PROMISE.

Table of Contents

Executive Summary

Work package 3 ("Evaluation Infrastructure") is responsible for designing, developing and delivering the evaluation infrastructure at the core of the PROMISE activities. Such infrastructure serves to carry out the evaluation activities identified in WP2 and WP6, in applying the evaluation methodologies and metrics proposed by WP4, in supporting the collaboration and knowledge sharing as proposed by WP5.

This deliverable presents the state of the art of the development of the evaluation infrastructure in relation to the new hardware that underlines and facilitates it together with the new features that have been implemented to support the two PROMISE use cases, that are:

- Search for innovation (CLEF-IP);
- Visual clinical decision support (ImageCLEF).

These two evaluation activities represent a very good example where multilingual and multimedia information access techniques mix and play a central role in the PROMISE activities.

To present the status of the new evaluation infrastructure, the document reports on:

- the DIRECT system, which has been developed by University of Padua since 2005 to support large-scale evaluation campaigns, and is used as back-bone for the PROMISE evaluation infrastructure;
- the migration of DIRECT, which describes the transfer of the initial DIRECT database and the initial DIRECT Web application to the new server cluster that has been specifically acquired;
- the preparation of CLEF 2011 Lab use cases presents the work done to prepare the datasets of the use cases which use the new prototype of the evaluation infrastructure;
- the status of development of the interface of the prototype also showing examples of the functionalities currently available.

The overall goal of this initial prototype of the evaluation infrastructure is to have the possibility to test it in the context of the CLEF 2011 Labs with actual multilingual and multimedia datasets and with real users in order to gather information, additional requirements, and feedback which will serve as input for deliverable D3.2 "Specification of the evaluation infrastructure based on user requirements" due at month 12. This, in turn, will lead to a re-engineering, modularization, and extension of DIRECT in order to produce a new prototype of the evaluation infrastructure, as planned with deliverable D3.3 "Prototype of the evaluation infrastructure" due at month 18.

Please note that, due to the actual scheduling of the CLEF evaluation campaign, at the time of writing this deliverable we are still quite ahead with respect to the preparation of some of

the data, such as topics, that will be used in the CLEF 2011 activities. As a consequence, we will show interfaces and functionalities using sample data with similar features to the actual ones that will be used in CLEF 2011.

On the other hand, having the prototype ready ahead of time gives ImageCLEF and CLEF-IP organizers the possibility of familiarizing with it and to ask for adjustment that can be ready on time for the CLEF 2011 campaign.

Finally, a demo video explaining the functionalities of DIRECT was uploaded on YouTube in the PROMISE channel at:
http://www.youtube.com/watch?v=fDsXDCUPkiM

1 Introduction

Evaluation campaigns promote and stimulate the research and development of Information Retrieval and Information Access Systems, for example by creating an evaluation infrastructure and organizing regular evaluation campaigns for system testing.

One of the aims of PROMISE is to provide a virtual laboratory for conducting participative research and experimentation to carry out, advance and bring automation into the evaluation and benchmarking of complex access and retrieval systems, by facilitating management and offering access, curation, preservation, re-use, analysis, visualization, and mining of the collected experimental data.

This deliverable describes the setup of the initial prototype of the evaluation infrastructure for the CLEF 2011 campaign[1] which is based on the Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)[2] system, developed by University of Padua, and the creation of the infrastructure for the evaluation of two use cases of PROMISE:

- Search for innovation (CLEF-IP)[3];
- Visual clinical decision support (ImageCLEF)[4].

These two evaluation activities represent a very good example where multilingual and multimedia information access techniques mix together and play a central role in the PROMISE activities [Ferro, 2011].

The overall goal of this initial prototype of the evaluation infrastructure is to have the possibility to test it in the context of the CLEF 2011 Labs with actual multilingual and multimedia datasets and with real users in order to gather information, additional requirements, and feedback which will serve as input for deliverable D3.2 “Specification of the evaluation infrastructure based on user requirements” due at month 12. This, in turn, will lead to a re-engineering, modularization, and extension of DIRECT in order to produce a new prototype of the evaluation infrastructure, as planned with deliverable D3.3 “Prototype of the evaluation infrastructure” due at month 18.

The document is divided into the following sections:

- “The DIRECT System” provides an overall description of the functionalities and the architecture of the system;
- “DIRECT System Migration” describes the transfer of the DIRECT database and the DIRECT Web application to the new server cluster;

[1] http://clef2011.org/

[2] http://direct.dei.unipd.it/

[3] http://www.promise-noe.eu/search-for-innovation/

[4] http://www.promise-noe.eu/visual-clinical-decision-support/

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

- "Preparation of CLEF 2011 Labs Use Cases" presents the work done to prepare the datasets of the use cases which use the initial prototype of the evaluation infrastructure;
- "Interface of the Initial Prototype" shows examples of the functionalities currently implemented in the prototype.
  Due to the actual scheduling of the CLEF evaluation campaign, at the time of writing the present deliverable we are still ahead with respect to the preparation of some of the data, like topics, that will be used in the CLEF 2011 activities. As a consequence, we show interfaces and functionalities using sample data with similar features to the actual ones that will be used in CLEF 2011.

Finally, a demo video explaining the functionalities of DIRECT was uploaded on YouTube in the PROMISE channel at:
http://www.youtube.com/watch?v=fDsXDCUPkiM

2 The DIRECT System

Since the first prototype was developed by UNIPD for CLEF 2005 [Agosti and Ferro, 2009; Di Nunzio and Ferro 2005; Dussin and Ferro, 2009c], the DIRECT system has been used with a twofold aim:

- to manage all the steps of an evaluation campaign, i.e. the ingestion of the document collections, the creation of the topics, the submission of the experiments, the relevance assessments, and the computation of performance measures and statistical analyses; and
- to make all the managed information online accessible to registered users and to preserve them over the time.

Since 2010, DIRECT provides access to an entire decade of CLEF data (2000-2009) [Agosti et al, 2010], that means:

- more than 5.6 million documents;
- more than 3.1 million relevance assessments for more than 600 topics made by over 200 assessors in 15 different countries;
- more than 3,500 experiments, which amounts to about 167 million tuples, submitted by over 170 participants in about 30 different countries;
- over 10.1 million performance measures and descriptive statistics;
- about 35,000 plots and statistical analyses graphs.

2.1 Functionalities of the DIRECT System

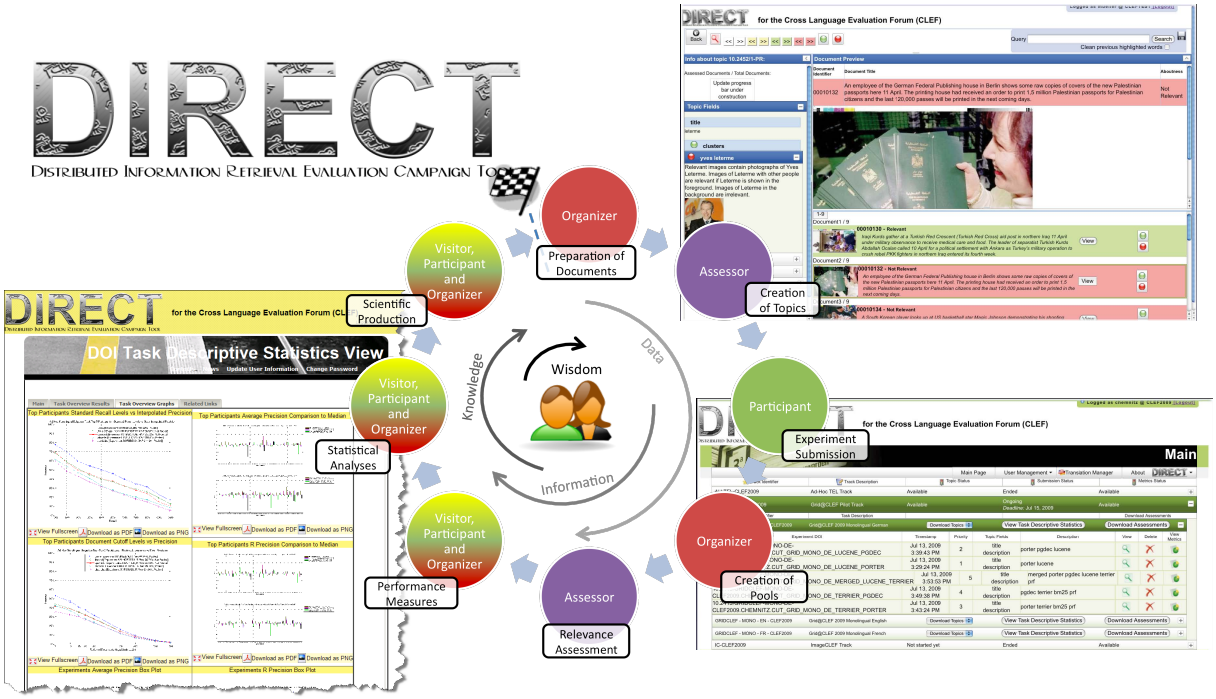


Figure 1: Functionalities supported by the DIRECT system.

Figure 1 summarizes the main functionalities of the DIRECT systems, which are reported below and related to the traditional DIKW (Data – Information – Knowledge – Wisdom) hierarchy [Ackoff, 1989; Zeleny, 1987]:

- *acquisition and preparation of documents*: the organizers are responsible for acquiring, formatting, and preparing the set of documents that are released to the participants. These documents are part of the data on which the experiments are built.
- *creation of topics*: the organizers and the assessors cooperate to create the topics for the test collection. For each topic, this step usually requires preparing a first draft of the topics and searching the set of document to verify that there are relevant documents for that topic; then the topics are refined by discussing their content and facets until a final version is reached. These topics are part of the data on which the experiments are built.
- *experiment submission*: the participants submit their experiments, which are built using the documents and the topics created in the previous steps. The result of each experiment is a list of retrieved documents in decreasing order of relevance for each topic and represents the output of the execution of the Information Retrieval System (IRS) developed by the participant. The experiments are part of the data that are produced during an evaluation campaign.
- *creation of pools*: the organizers collect all the experiments submitted by the participants and, using some appropriate sampling technique, select a subset of the retrieved documents to be manually assessed in the next step to determine their actual

relevance. The pools are midway between data and information, since they are still raw elements but represent a first form of processing of the experiments.

- *relevance assessment*: the organizers and the assessors cooperate to assess each document in the pool with respect to the topic, i.e. for determining whether the document is relevant or not for the given topic. As in the case of the pools, the relevance judgments are midway between data and information, since they are raw elements which constitute an experimental collection but represent human-added information about the relationship between the topics and documents of an experiment.
- *measures and statistics*: the organizers exploit the relevance assessments to compute the performance measures and plots about each experiment submitted by a participant; then, these measurements are used for computing descriptive statistics about the overall behaviour of both an experiment and all the experiments in a given task; furthermore, these measurements are also employed for conducting statistical analyses and tests on the submitted experiments. As discussed above, performance measures are information, since they are the results of data processing; descriptive statistics and hypothesis tests are knowledge, since they provide some more insights into the meaning of the obtained performance.
- *scientific production*: both organizers and participants prepare reports where the former describe the overall trends and provide an overview for the evaluation campaign and the latter explain their experiments, the techniques that have been adopted, and the findings. This work usually continues even after the conclusion of the campaign, since the investigation and understanding of the experimental results require deep analysis and reasoning, which usually takes the form of conference papers, journal articles, talks, and discussion among researchers. Furthermore, not only the organizers and the participants but also external visitors may exploit the information resources produced during the evaluation campaign to carry out their research activity. As explained above, the outcomes of this process are wisdom.

2.2 Architecture of the DIRECT System

DIRECT has been designed to meet the following goals:

- to be cross-platform and easily deployable to end users;
- to be as modular as possible, clearly separating the application logic from the interface logic;
- to be intuitive and capable of providing support for the various user tasks described in the previous section, such as experiment submission, consultation of metrics and plots about experiment performances, relevance assessment, and so on;
- to support different types of users, i.e. participants, assessors, organizers, and visitors, who need to have access to different kinds of features and capabilities;
- to support internationalization and localization: the application needs to be able to adapt to the language of the user and his country or culturally dependent data, such as dates and currencies.

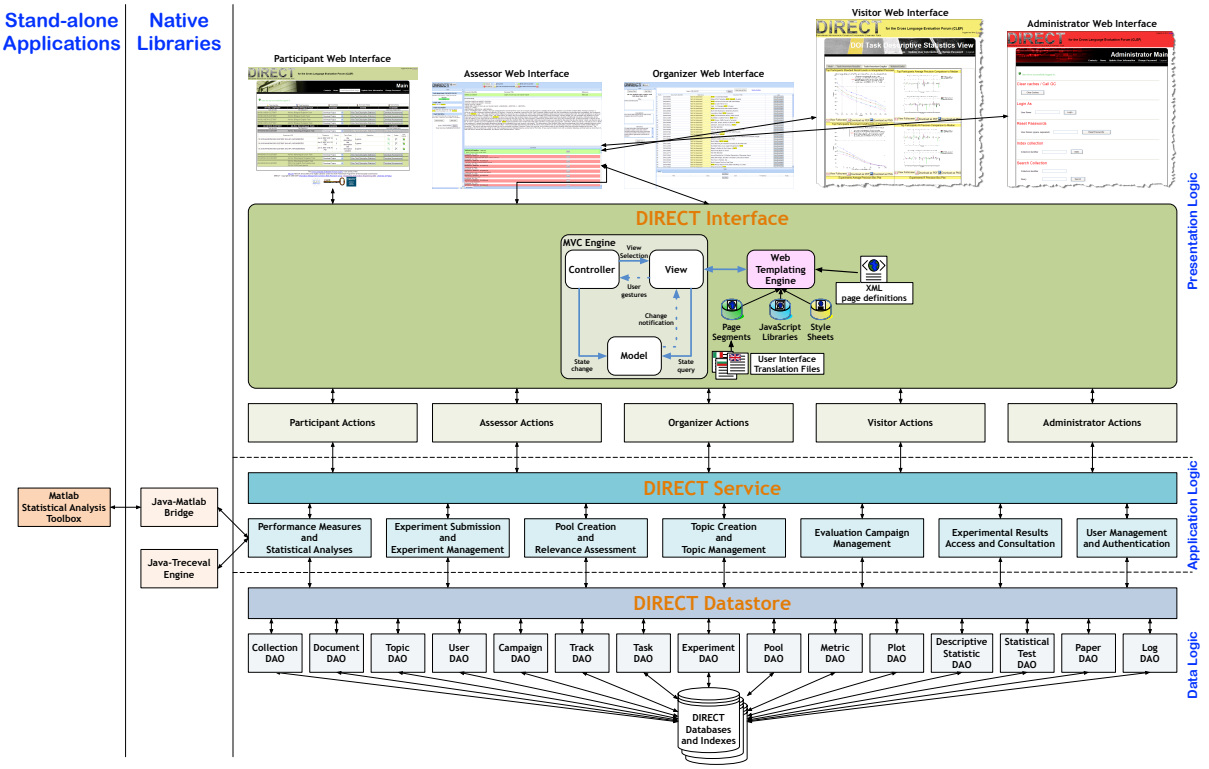


Figure 2: Architecture of the DIRECT system.

As shown in Figure 2, DIRECT adopts a three-tiers architecture – data, application and interface logic layers – in order to achieve improved modularity and to properly describe the behavior of the service by isolating specific functionalities at the proper layer:

- **data logic**: deals with the persistence of the different information objects coming from the upper layers. There is a set of "storing managers" dedicated to storing the submitted experiments, the relevance assessments and so on. The Data Access Object (DAO) and the Transfer Object (TO) design patterns have been adopted. The DAO implements the access mechanism required to work with the underlying data source, acting as an adapter between the upper layers and the data source. If the underlying data source implementation changes, this pattern allows the DAO to adapt to different storage schemes without affecting the upper layers. On top of the various DAOs there is the "DIRECT Datastore" which hides the details about the storage management to the upper layers. In this way, the addition of a new DAO is totally transparent for the upper layers.
- **application logic**: deals with the flow of operations within DIRECT. It provides a set of tools capable of managing high-level tasks, such as experiment submission, pool assessment, and statistical analysis of an experiment. The "DIRECT Service" provides the interface logic layer with uniform and integrated access to the various tools. As with the case of the "DIRECT Datastore", thanks to the "DIRECT Service" the addition of new tools is transparent for the interface logic layer too.

- **Interface logic**: the architecture of the DIRECT user interface is Web-based and designed to be cross-platform and easily deployable and accessible without the need of installing any software on the end-user machines. The system also supports the internationalization and localization of the user interface by adapting it to the language and country of the user. The correct language and country are initially loaded according to the browser settings and, in the case of not supported locales, it falls back to a default configuration. The Model-View-Controller (MVC) [Krasner and Pope, 1988] approach has been used to clearly separate the following three layers: the *model layer* contains the underlying data structures of the application and keeps the state of the application; the *view layer* concerns the way the model is presented to the user; the *controller layer* manages the interaction between the view and the input devices, such as the keyboard or the mouse, and updates the model accordingly.

3 Evaluation Infrastructure: Hardware and Software

The amount of data and the complex structure of the organization have required a migration of the DIRECT system in two parts: the migration of the database, and the migration of the application. The two migrations are described in the following.

3.1 PROMISE Infrastructure Cluster

In order to host the PROMISE infrastructure, the hardware and storage equipment already available at University of Padua has been extended.

In particular:

- 7 TByte of new storage were added to the NAS Netapp 3020, already owned by UNIPD;
- Two new blades were added to the DELL PowerEdge M1000e Modular Blade Enclosure, already owned by UNIPD;
- Each new blade is a Dell PowerEdge M610 equipped with two Intel Xeon Six-core X5680 processors at 3.33 GHz with 12MByte cache, supporting 24 threads in parallel, 48 Gbyte of DDR3 RAM at 1333MHz, and two RAID-1 mirrored 500 GByte SAS 15K rpm drives for the local storage (for the operating system, the software, and so on);
- Linux Fedora[5] 14 is used as operating system and the two blades are operated as a cluster managed by Pacemaker[6] 1.1.4 in order to ensure redundancy and uptime.

Various services critical and core to the PROMISE project run on this cluster:

- **Database**: PostgreSQL[7] version 9.0.3 for the storage of all the data managed by the PROMISE infrastructure;

[5] http://fedoraproject.org
[6] http://www.clusterlabs.org/
[7] http://www.postgresql.org/

- **Web services**: Apache Tomcat[8] 7.0.5 for hosting the Web services offered by the PROMISE infrastructure;
- **Evaluation infrastructure user interface**: Liferay[9] Portal Community Edition 6.0.5 for the user interfaces of the evaluation infrastructure;
- **Project portal**: Liferay Portal Community Edition 6.0.5 (a different instance from the previous one) for the development and operation of the PROMISE project portal[10], also described in deliverable D7.1 "Project Web Site and Project Fact Sheet";
- **Collaborative code development**: Apache Subversion[11] 1.5.9 is used as code repository for managing the development of shared code; Apache Ant[12] is used for managing the build process; and, Hudson[13] 1.3 is used for the continuous integration.

3.2 DIRECT Database Migration

The DBMS which was used in the year 2010 to store all the data of the DIRECT system was PostgreSQL version 8.3.6. In terms of tables and space occupied on disk, the size of the whole database was 1,024 tables for a total of 56 GB of data, most of the data being compressed. For the new database the newest version 9.0.3 of the PostgreSQL DMBS was installed.

The migration of the DIRECT database was split in the following steps:

- creation of a new role;
- creation of a new database and privileges;
- dump of the old DIRECT database; and
- restoration of the data in the new database.

The backup and the restoration of the entire DIRECT system was planned on the third week of February, and, as planned, it started on February 14th and ended on February 18th 2011, requiring two days for the dump of the database and two more days for the restore.

3.3 DIRECT Web Application Migration

After the migration of the data, in order to test that all the data were effectively transferred from the old to the new database, the DIRECT Web application was installed on the new server cluster as one of the Web applications of the Apache Tomcat server.

The latest version 7.0.5 of Apache Tomcat was installed and run on the new server cluster. The transfer of the Web application from the old server to the new cluster was programmed on February 18th and completed the same day without any particular problem. It was

[8] http://tomcat.apache.org/
[9] http://www.liferay.com/
[10] http://www.promise-noe.eu/
[11] http://subversion.apache.org/
[12] http://ant.apache.org/
[13] http://hudson-ci.org/

necessary to update the JDBC[14] driver to the latest version JDBC4 PostgreSQL Driver, version 9.0-801, in order to read and write byte array valued fields, such as documents of a collection, or topics of a task, in the new hexadecimal format of PosgreSQL 9.0.

The DIRECT system is available at the following address:

http:// direct.dei.unipd.it/

4 Preparation of CLEF 2011 Labs Use Cases

The initial prototype of the evaluation infrastructure supports two labs of the CLEF 2011 evaluation campaign that are also two use cases of PROMISE:

- Search for innovation (CLEF-IP)[15];
- Visual clinical decision support (ImageCLEF)[16].

The tasks of each Lab are described and the data made available for each of the tasks. Some screen dumps of the actual interface are presented in the following of this document to show the current development of the prototype.

It is important to recall that the Labs are still under development and some details are currently being defined. For this reason, some of the datasets may change; consequently, the interface will be adapted according to the final specifications of the Labs.

4.1 CLEF-IP Data

The CLEF-IP Lab was first launched in 2009 to investigate information retrieval techniques for multilingual patent retrieval. After a successful first year, the track continued in 2010 as a benchmarking activity at the CLEF 2010 conference (22 September 2010) in Padua, Italy. CLEF-IP is a benchmarking activity and Lab also at the CLEF 2011 conference (September 2011) in Amsterdam, The Netherlands[17].

CLEF-IP uses a large data collection of patent documents derived from EPO (European Patent Office) sources, covering English, French, and German patents. In 2011 the CLEF-IP data include patent images as well.

The tasks of CLEF-IP that will be managed by the prototype are still under discussion.

The dataset available for this Lab consists of:

- a collection of about 2,6 million patents in XML format;
- a set of 2,000 topics for the Prior Art Candidate Search Task;

[14] http://jdbc.postgresql.org/

[15] http://www.promise-noe.eu/search-for-innovation/

[16] http://www.promise-noe.eu/visual-clinical-decision-support/

[17] http://www.ir-facility.org/clef-ip

- a set of 2,000 topics for the Classification Task plus 300 training topics.

4.2 ImageCLEF Data

ImageCLEF is the cross-language image retrieval Lab of CLEF. The medical retrieval task of ImageCLEF 2011 will most likely use a database of PubMed Central[18] or a subset of this resource containing over 1 million images. ImageCLEF proposes five main tasks. One of them, called the "Medical retrieval" task, will be supported by the prototype.

The subtasks of the medical retrieval task are:

- Modality Classification;
- Image-based retrieval;
- Case-based retrieval.

The dataset available for this Lab consists of (as specified for ImageCLEF 2011[19]):

- an XML file, which contains a summary of the medical articles consisting of images, image captions, and a link to the original article;
- a collection of the original texts of the medical articles;
- a collection of all the images linked by the articles;
- a collection of topics for each task.

The exact number of images and articles are not yet available since the processing of all the datasets that are going to be uploaded in the DIRECT system is under way. Most likely a collection of 240,000 images will be used.

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

[19] http://www.imageclef.org/2011

5 Interfaces of the Initial Prototype

In this section we present some functions of the prototype and some screenshots of the interface currently available for the two CLEF labs. At present, the prototype supports two types of users:

- Participants: the interface allows a participant to manage the submission of the experiments for the tasks he is registered to, and to download the data available for each task (i.e. topics);
- Topic creators: the interface allows the organizers of a Lab to manage the creation of the topics for each task;
- Assessors: the interface allows the organizers of a Lab to manage the relevance assessments for each task;
- Administrators: the interface allows the administrators and organizers of a Lab to manage the translations of the labels in different languages.

The following sections describe the functionalities of the current version of the prototype.

5.1 Experiments Submission Management

Figure 3, present the main page for the management of the submission of experiments that allows the participants to access all the information about the Lab he or she registered to.

From the main page the participant can access:

- the experiments submitted for each task;
- the topics available for each task;
- the relevance assessment produced by assessors (when available during the campaign);
- the performance measures about each experiment (when available during the campaign);
- the performance measures for each task (when available during the campaign).

The interface is based on a set of folding tables, which allow participants to access their experiments, by structuring them into different levels based on a tree structure – labs, tasks, and experiments – well known to the user. Figure 3 shows the main page with one task unfolded. Therefore, the participant can manage his own data by simply selecting and expanding the right level in the tree in order to facilitate the submission, editing, or deletion of an experiment.

Besides experiments, further data are associated with each level of the tree in order to support the participant in accessing additional resources: the prototype makes available only those topics and relevance assessments that are pertinent for the task currently selected by the participant.

From this interface, the participant can access the information about the metrics of the

experiment by clicking on the “View Metrics” button. A new page will be shown to the participant with a summary of the information about the experiment, the metrics, the descriptive statistics, and the available plots. In Figure 4 and Figure 5, examples of the metrics and plots views are shown.

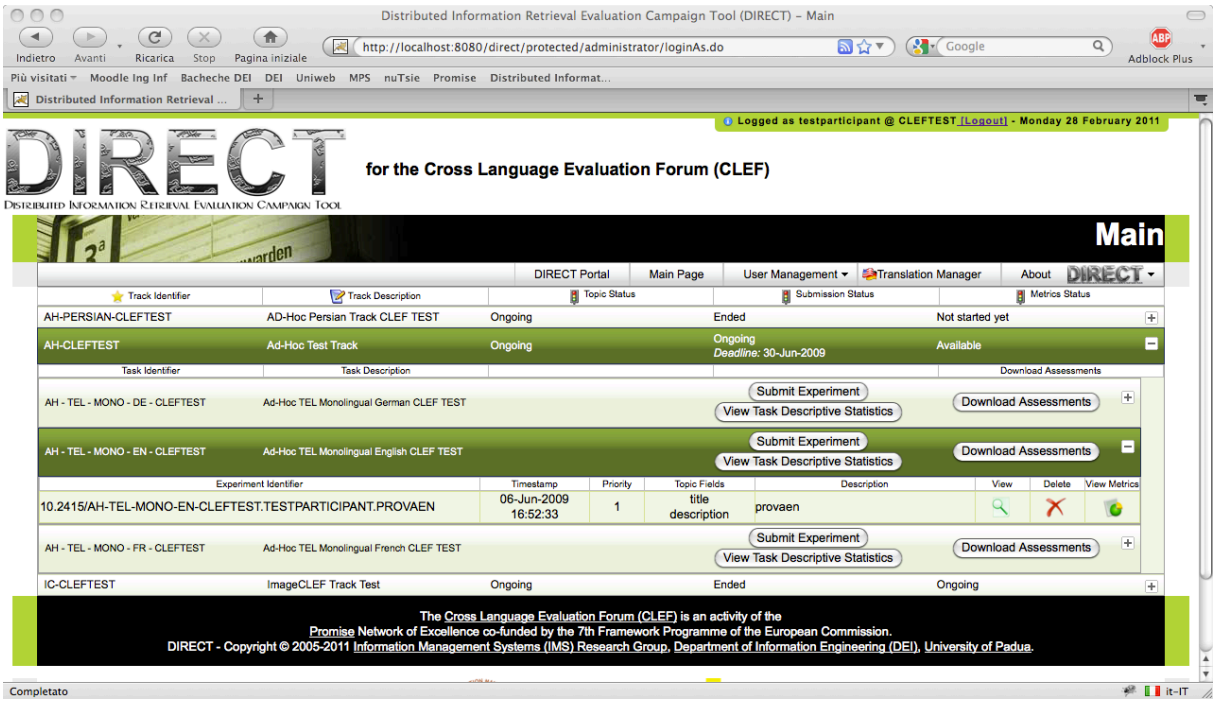


Figure 3: Experiment submission page with one task unfolded.

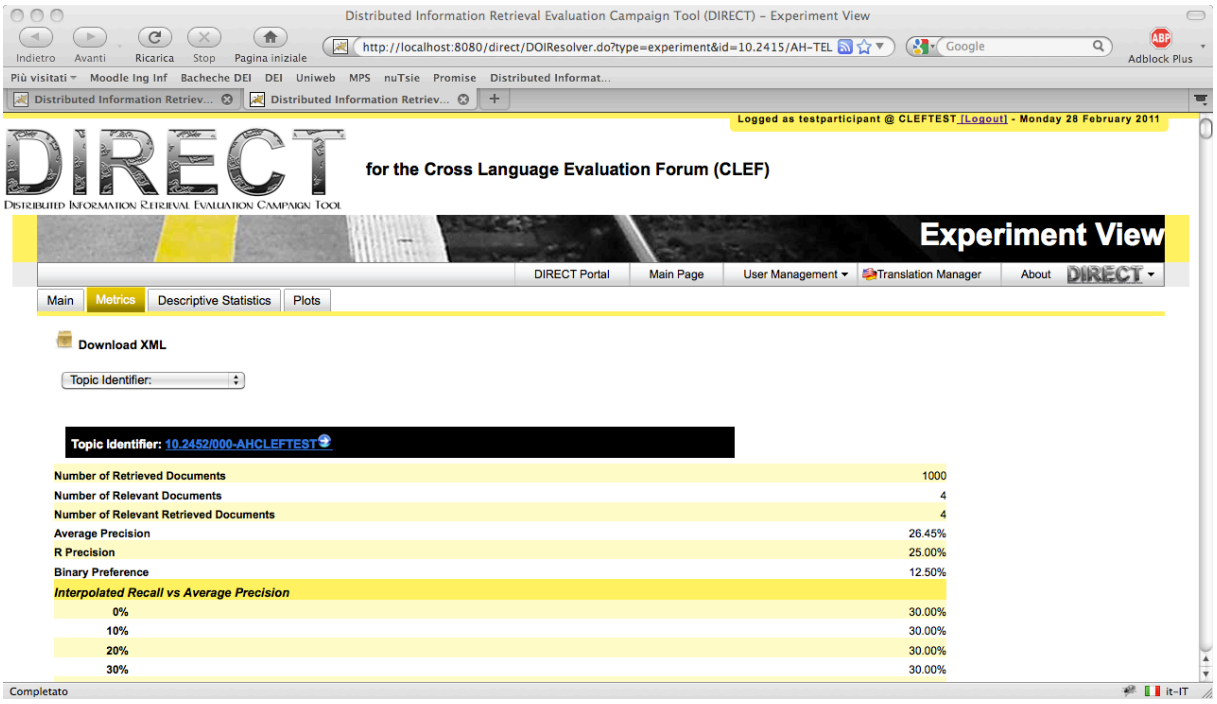


Figure 4: Metrics view for one experiment in the participant interface.

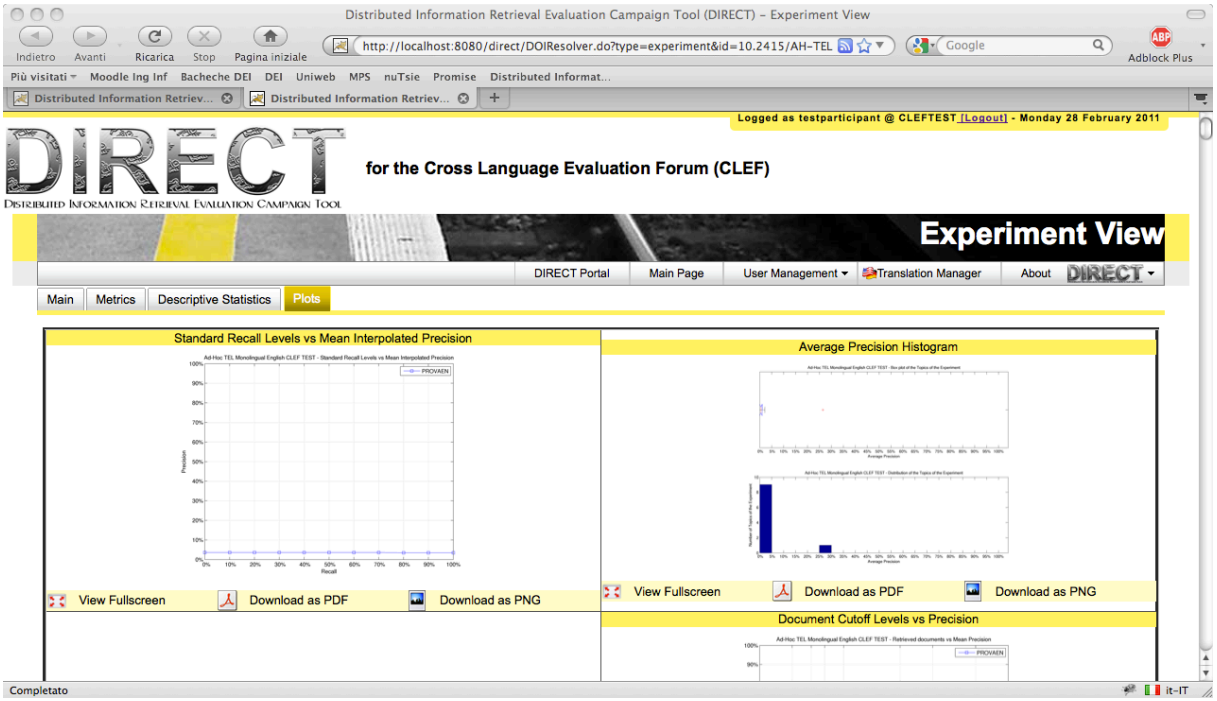


Figure 5: Plots view for one experiment in the participant interface.

From the same interface the user can submit one experiment by first uploading a file and then by filling in a submission form, shown in Figure 6. Then the system validates the

experiment by checking the correctness of the format and printing warnings in case some data are missing, see Figure 7.

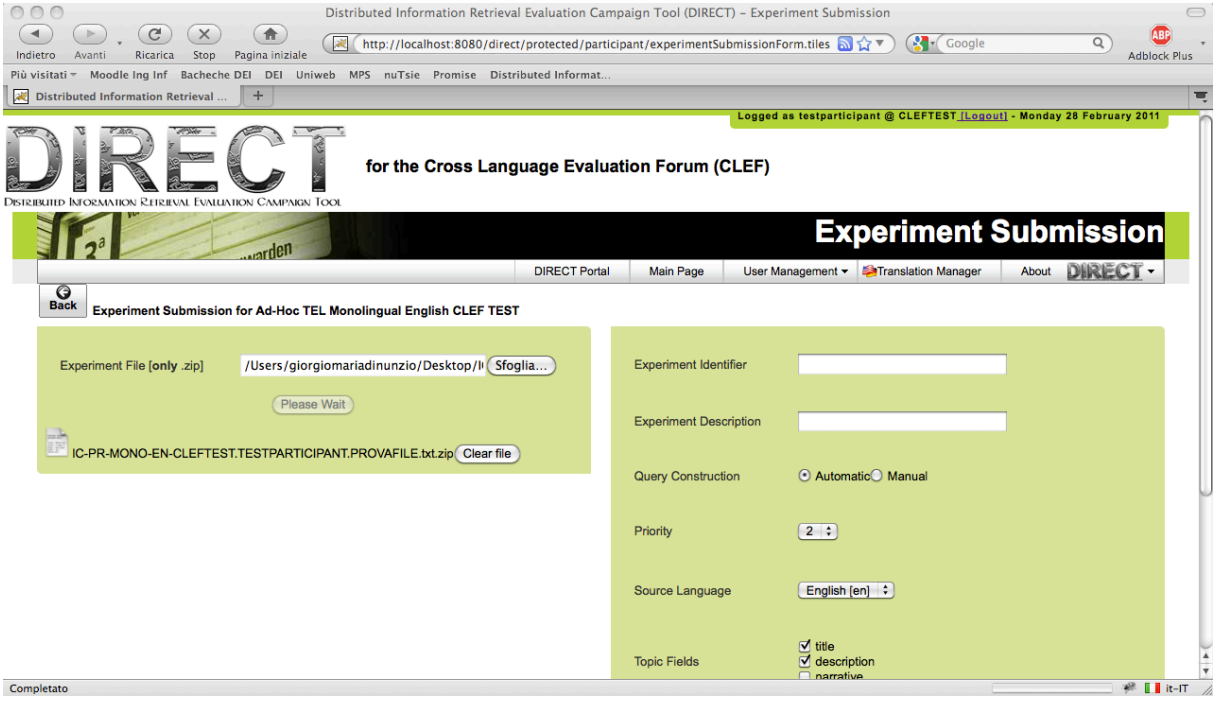


Figure 6: Submission form of one experiment.

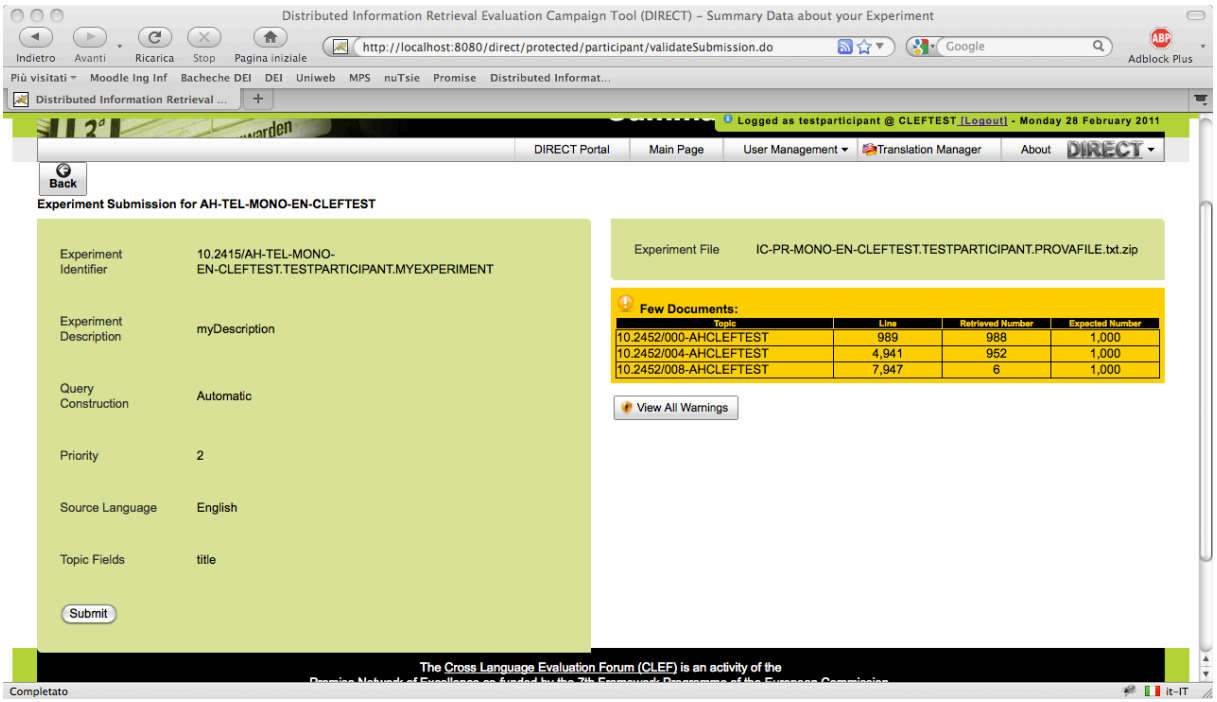


Figure 7: Validation of the submission of one experiment.

5.2 Topic Creation Management

Figure 8, Figure 9, and Figure 10 show the main pages that the topic creator uses for creating a set of topics for a task.

In Figure 8, for each topic the interface shows:

- the identifier of the topic;
- a short summary of the topic;
- the number of relevant and not relevant documents assessed so far;
- a button to edit the topic.

Figure 9, and Figure 10 present the main page for creating or modifying one topic. On the left side of the screen, the creator can access some information about the history of the query, the notes exchanged with other users, the relevant or not relevant document for the created topic. In the main area, the creator can issue a query and look in the collection for relevant or not relevant documents.

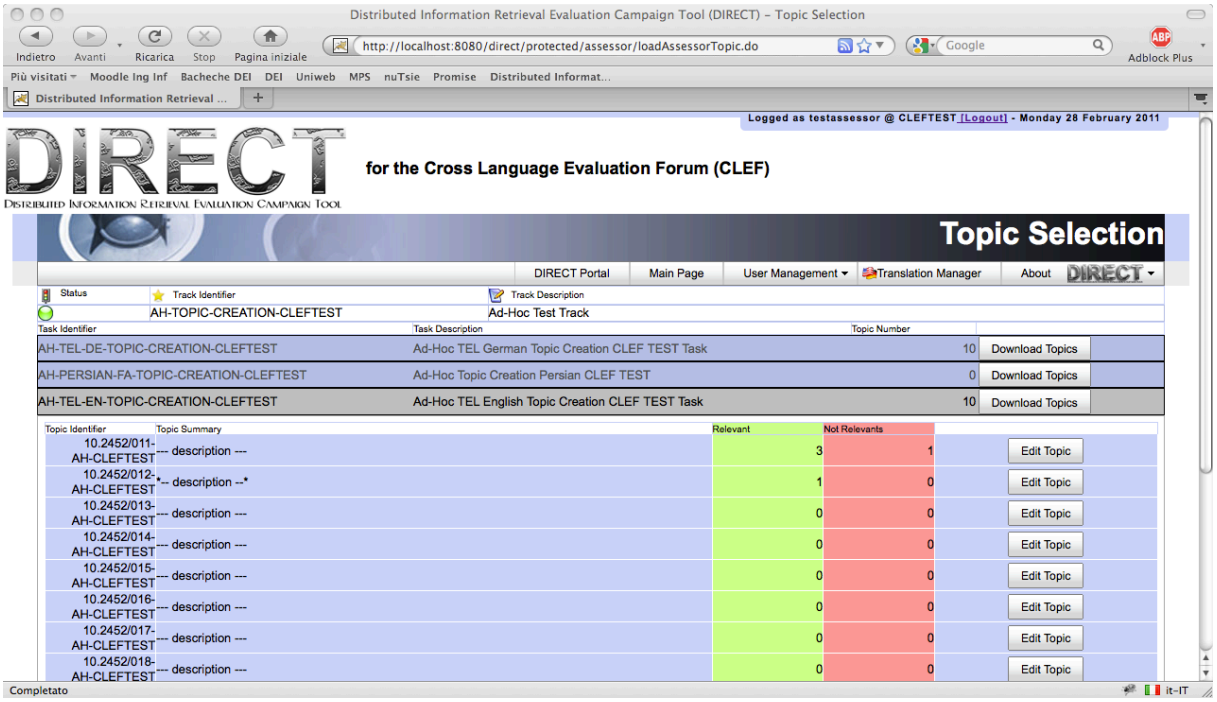


Figure 8: Main page of the creation of topics interface.

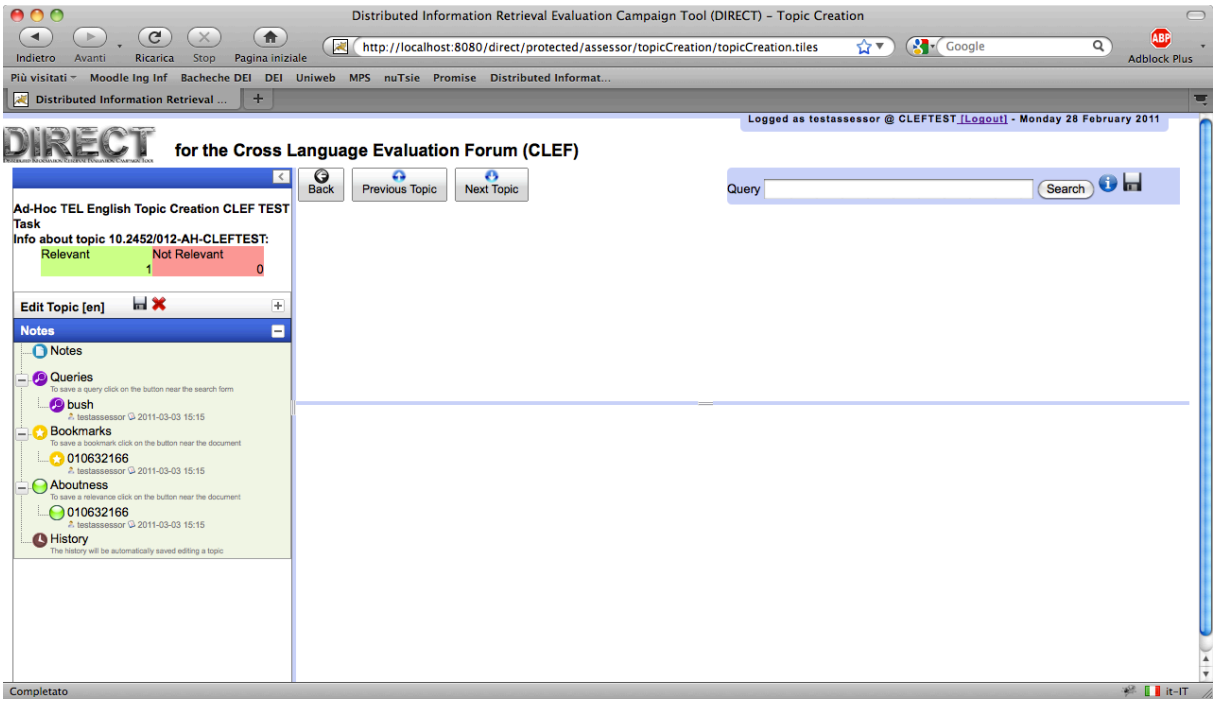


Figure 9: Topic creation page. On the left, a set of information for the creator is displayed.

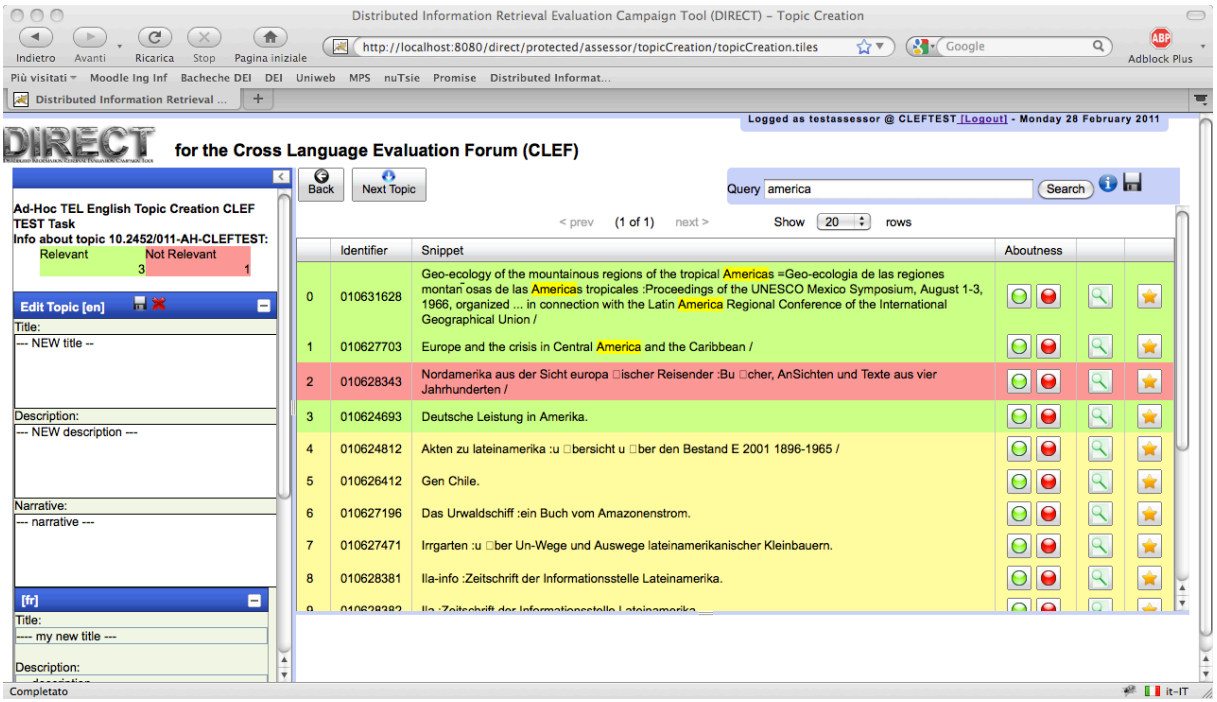


Figure 10: Topic creation phase, in the main area of the page a list of documents is presented after a query issued by the creator.

5.3 Relevance Assessment Management

Figure 11, Figure 12, and Figure 13 show the two main pages that the assessor uses for performing the relevance assessment. The list of topics available for each task of a Lab is presented in table form in order to display the information in a compact and coherent way.

For each topic, the interface shows:

- the identifier of the topic;
- a short summary of the topic;
- the total number of documents to assess for a topic;
- the number of relevant and not relevant documents assessed so far;
- the number of documents still to be judged.

The table can be folded and expanded so that the user can concentrate on the contents of interest to him without having to read all the data or scroll the whole page. Figure 11 shows one of the lists unfolded.

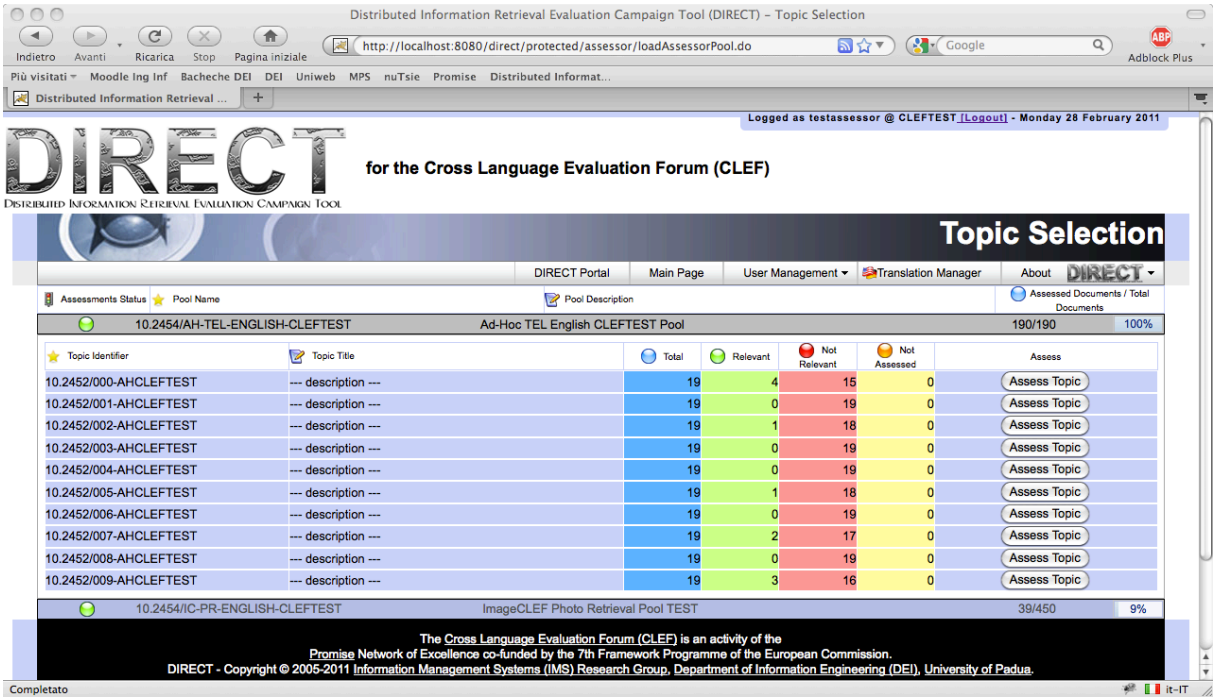


Figure 11: Assessment interface with one task unfolded.

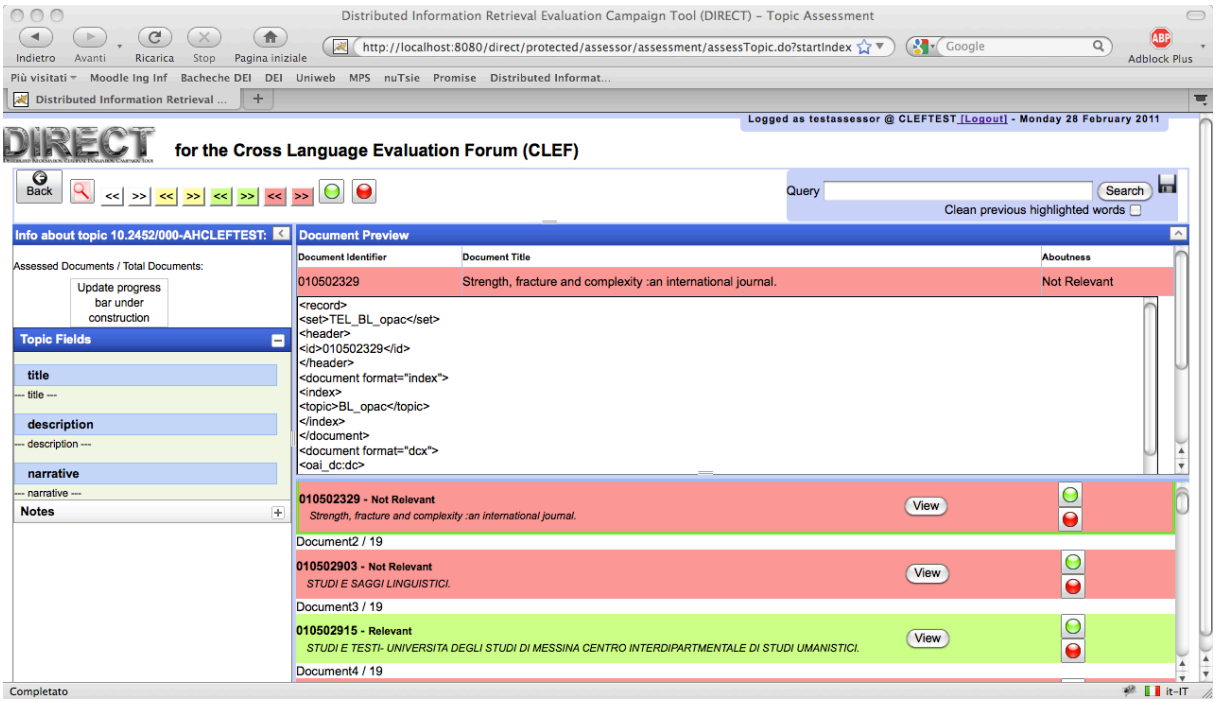


Figure 12: Example of the assessment interface for a multilingual retrieval task (similar to the one of AH TEL Task).

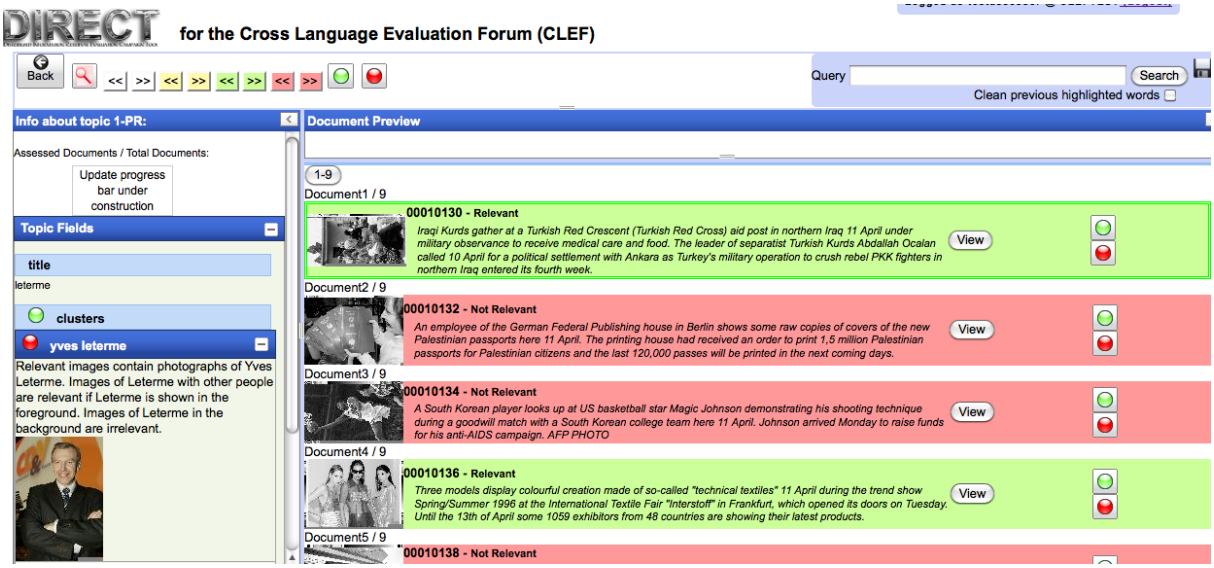


Figure 13: Example of the assessment interface for a multilingual and multimedia task (similar to the one of ImageCLEF).

Figure 12 and Figure 13 show the interface for relevance assessments, which is displayed after pressing the “Assess Topic” button. Figure 12 shows another example where topics and documents are text, Figure 13 shows an example for topics and documents that contain images. The documents that were pooled for relevance assessment are shown in

the main area of the page to the assessor. In the top left corner it is possible to read information about the topic: the title, a description, and if available, one or more images that describe the topic. A progress bar reminds to the assessor of the percentage of documents that have not yet been judged. There is a "search" function which gives the user the possibility to look for specific words that are automatically highlighted in yellow in the interface, in this way the user can have a quick help to spot the documents to be judged which contain the searched keywords. Occurrences of the query terms are highlighted in yellow both in the topic and in the document content in order to facilitate the work of the assessor.

Navigation through the documents is made easy by a set of buttons in the top bar allowing the user to quickly find the next not assessed, relevant or not relevant document.

The selected document is shown at the centre of the page, reporting its identifier, title and relevance status. In addition, a highlighting frame flowing up and down over the list of the documents at the bottom of the page shows which document the user is reading in relation to all the documents pooled for that topic.

Specific sets of buttons are also provided at the top of the page to help the user make the assessment task in an intuitive, quick, and useful way. They are characterized by the use of two colours: green to set the relevant status, and red for not relevant. When an assessment is performed, the row at the bottom of the page for the assessed document changes colour accordingly, and the highlighting frame automatically moves to the next document to be assessed.

5.4 Internalization and Localization

Administrators and organizers of a lab can manage the translation of the interface in different languages.

Figure 14, and Figure 15 show the interface for the translation of the labels of the interface: in Figure 14 an organizer of a Lab can see the list of subscribed task for the translation on the left, while on the main area of the page the whole list of translation task is displayed. The organizer can ask to join one translation task or report errors in the translation phase. In Figure 15 the interface for the translation from one language to another is shown. The first column shows the keys the system uses to substitute the labels with the actual language, for each key the second column shows default language (English in the figure) and the third column the provided translation.

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

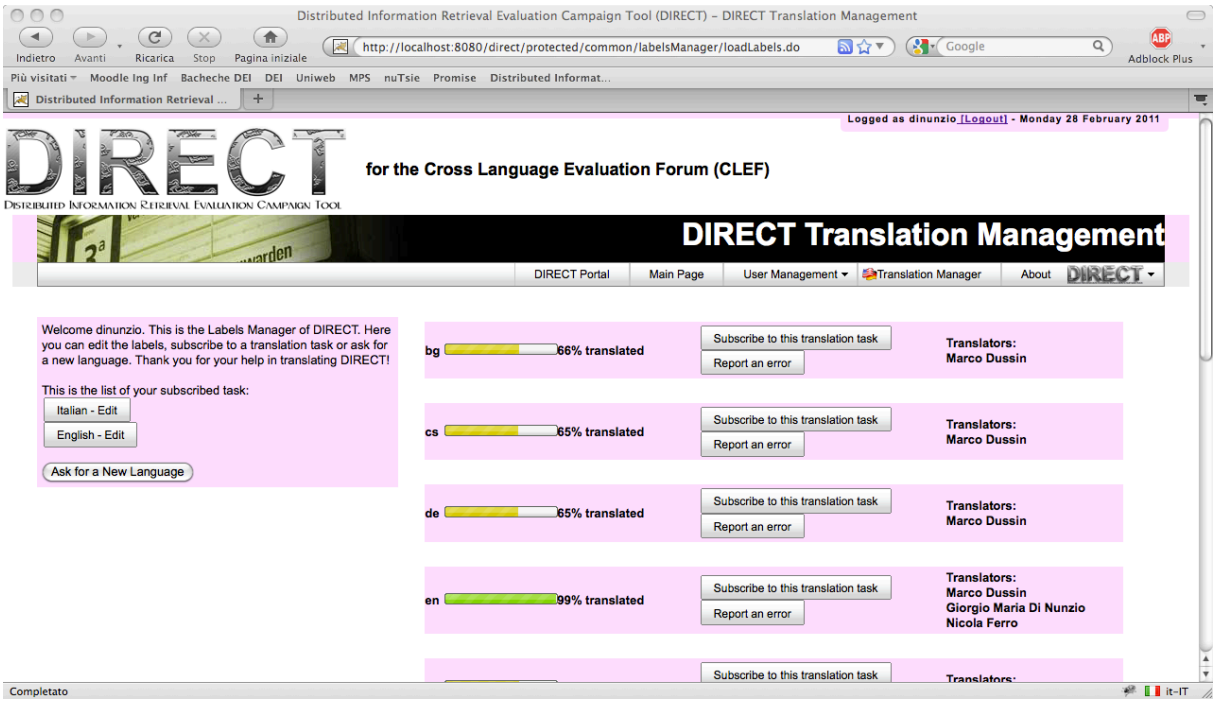


Figure 14: Management of the translation of the labels of the interface in different languages.

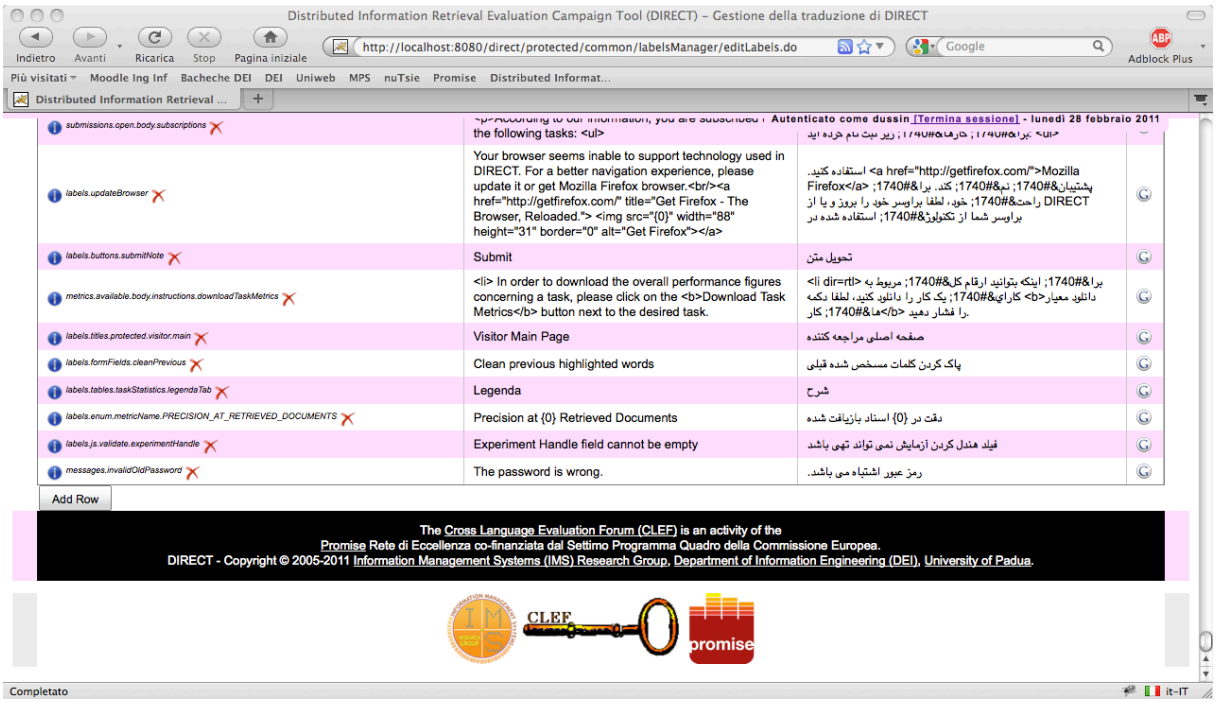


Figure 15: Interface for managing the translation of the interface from one language to another.

5.5 Access and Browsing of the Scientific Data

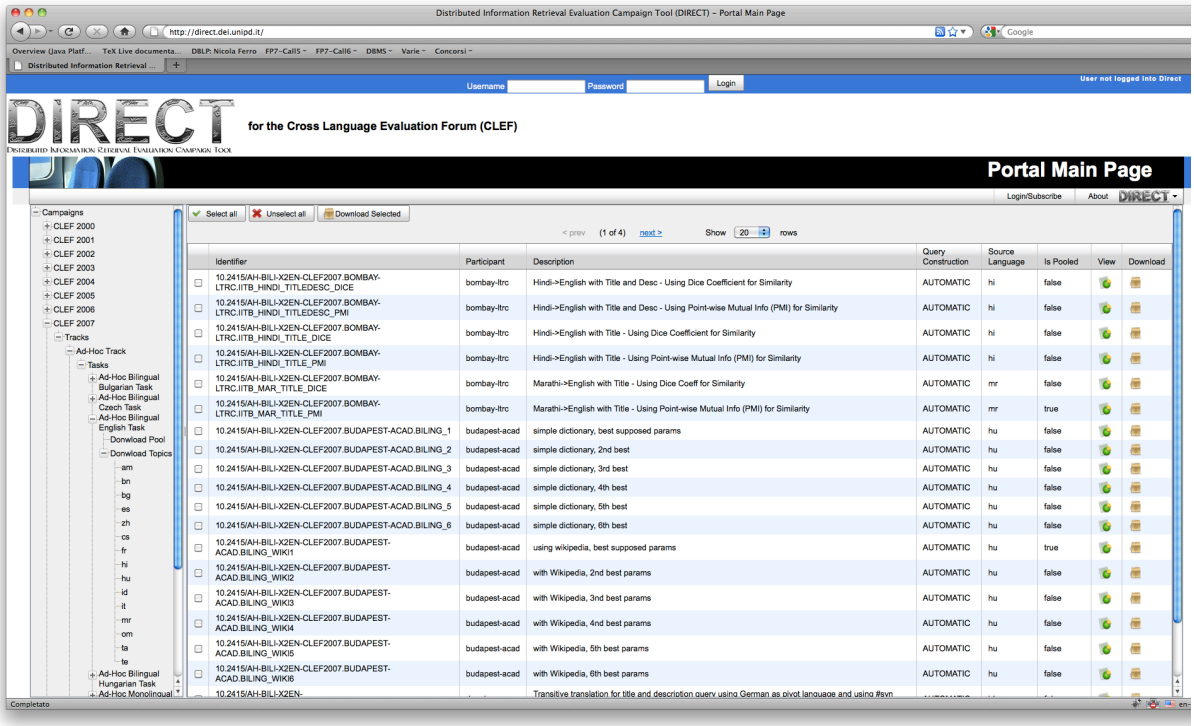


Figure 16: Interface for accessing the history of ten years of CLEF data.

Figure 16 shows a screenshot of the interface for accessing the whole set of scientific data produced during the history of CLEF. On the left, there is a tree which allows the user to browse thorough the CLEF campaigns from 2000 to 2009 and, for each campaign, it is possible to see what tracks and tasks are available.

Once the user has selected a task, he is presented, in the right pane, with the list of experiments submitted for that task plus specific information about each experiment, such as a description, the source language, which fields of the topic have been used to construct the query, whether it is an automatic or manual experiment, whether it has been pooled or not, and so on. At this point, the user can decide to view and access metrics and performance measures about the experiment, as shown in Figure 4 or to directly download

it for further re-use and analysis. It is also possible to select multiple experiments and download them at once. In the left pane, once a task has been selected, it is possible either to download the relevance assessments and pool corresponding to that task, as well as the topics used in the task, in multiple languages if available and visualize overall statistics about the task.

References

[Ackoff, 1989] R. L. Ackoff. From Data to Wisdom. *Journal of Applied Systems Analysis*, 16, pp. 3-9, 1989.

[Agosti, 2008] M. Agosti (Ed). *Access through Search Engines and Digital Libraries*. Springer-Verlag, Heidelberg, Germany, 2008.

[Agosti and Ferro, 2009] M. Agosti, N. Ferro. Towards an infrastructure for digital library performance evaluation. In: G. Tsakonas, C. Papatheodorou (Eds). *Evaluation of Digital Libraries*. Chandos Publishing, Oxford, UK, 2009, 93-120.

[Agosti et al. 2007] M. Agosti, G.M. Di Nunzio, N. Ferro. The Importance of Scientific Data Curation for Evaluation Campaigns. In: *Digital Libraries: Research and Development. First International DELOS Conference*. Revised Selected Papers, 157–166. LNCS 4877, Springer, Heidelberg, Germany, 2007.

[Agosti et al, 2010] M. Agosti, G. M. Di Nunzio, M. Dussin, N. Ferro. 10 Years of CLEF Data in DIRECT: Where We Are and Where We Can Go. In: T. Sakay, M. Sanderson, W. Webber (Eds), *Proc. 3rd International Workshop on Evaluating Information Access (EVIA 2010)*, 16-24. National Institute of Informatics, Tokyo, Japan, 2010.

[Di Nunzio and Ferro 2005] G.M. Di Nunzio, N. Ferro. DIRECT: a System for Evaluating Information Access Components of Digital Libraries. In: *9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, 483–484. Lecture Notes in Computer Science (LNCS) 3652, Springer, Heidelberg, Germany, 2005.

[Dussin and Ferro, 2008a] M. Dussin, N. Ferro. DIRECT: Applying the DIKW Hierarchy to Large-Scale Evaluation Campaigns. In: R. Larsen, A. Paepcke, J. Borbinha, M. Naaman (Eds), *Proc. 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008),* 424. ACM Press, New York, USA, 2008.

[Dussin and Ferro, 2008b] M. Dussin, N. Ferro. The Role of the DIKW Hierarchy in the Design of a Digital Library System for the Scientific Data of Large-Scale Evaluation Campaigns. In: R. Larsen, A. Paepcke, J. Borbinha, M. Naaman (Eds), *Proc. 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008),* 450. ACM Press, New York, USA, 2008.

[Dussin and Ferro, 2008c] M. Dussin, N. Ferro. Design of a Digital Library System for Large-Scale Evaluation Campaigns. In: B. Christensen-Dalsgaard, D. Castelli, J. K. Lippincott, B. Ammitzbøll Jurik (Eds), *Proc. 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008)*, 400-401. Lecture Notes in Computer Science (LNCS) 5173, Springer, Heidelberg, Germany, 2008.

[Dussin and Ferro, 2009a] M. Dussin, N. Ferro. DIRECT: Applying the DIKW Hierarchy to Large-Scale Evaluation Campaigns. *Bulletin of IEEE Technical Committee on Digital Libraries (IEEE-TCDL)*, 2009, 5(1), http://www.ieee-tcdl.org/Bulletin/v5n1/Dussin/dussin1.html.

[Dussin and Ferro, 2009b] M. Dussin, N. Ferro. The Role of the DIKW Hierarchy in the Design of a Digital Library System for the Scientific Data of Large- Scale Evaluation Campaigns. *Bulletin of IEEE Technical Committee on Digital Libraries (IEEE-TCDL)*, 2009, 5(1), http://www.ieee-tcdl.org/Bulletin/v5n1/Dussin/dussin2.html.

[Dussin and Ferro, 2009c] M. Dussin, N. Ferro. Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In: M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, G. Tsakonas, editors, *Proc. 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, 63–74. Lecture Notes in Computer Science (LNCS) 5714, Springer, Heidelberg, Germany, 2009.

[Ferro, 2011] N. Ferro. PROMISE: Advancing the Evaluation of Multilingual and Multimedia Information Systems. *ERCIM News*, 2011, 84:49.

[Krasner and Pope, 1988] G. E. Krasner and S.T. Pope. A Cookbook for Using the Model-View-Controller User Interface Paradigm in Smalltalk-80. *Journal of Object-Oriented Programming*, 1 (3), pp. 26-49, 1988.

[Zeleny, 1987] M. Zeleny. Management Support Systems: Towards Integrated Knowledge Management. *Human Systems Management*, 7 (1), 59-70, 1987.