PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Researcher Exchange Report

Managing Evaluation Tasks and Measuring their Impact

HESSO – UNIPD

May 2-4, 2011

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

Document Information

Report title:	Managing Evaluation Tasks and Measuring their Impact
Researcher Exchange date:	May 2-4, 2011
Visitor(s):	Theodora Tsikrika, HESSO
Host(s):	Nicola Ferro, UNIPD
Preparation date:	14/06/2011
Author(s):	Theodora Tsikrika

Table of Contents

1 Introduction

The purpose of the visit was to strengthen the cooperation between WP3 (*Evaluation Infrastructure*) and WP6 (*Evaluation Activities*) of the PROMISE[1] network of excellence

[1] http://www.promise- noe.eu

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

with the goals: 1) to provide support for managing the evaluation tasks for the PROMISE use cases via the PROMISE evaluation infrastructure, and 2) to investigate methodologies for measuring the impact of evaluation activities.

2 Planned Work

The first goal of this researcher exchange was the appropriate setup and adaptation of the PROMISE evaluation infrastructure to the requirements of the "Visual Medical Information Retrieval as Clinical Decision Support" PROMISE use case. In particular, the main objective was to configure the evaluation infrastructure for the Medical Image Retrieval task of the ImageCLEF[2] evaluation campaign by supporting the following steps of this evaluation task: experiments submission, creation of pools, relevance assessments, and metrics computation. Furthermore, the aim was to also investigate the possibility to enrich the acquired knowledge base of the collected experimental data with the experimental data gathered from past ImageCLEF evaluation campaigns related to the PROMISE use cases.

The second goal was to investigate the impact of each PROMISE evaluation task in the context of an evaluation campaign by developing methodologies that measure the scholarly/academic impact of such evaluation activities. The decision to focus on the scholarly impact lies in the reasonable assumption that evaluation activities are successful in their objectives if the resources they provide make possible a significant amount of research that is then published and cited. This activity is part of the Impact Metrics Group of PROMISE.

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

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

3 Conducted Work

3.1 Configuring the evaluation infrastructure to support the ImageCLEF medical image retrieval task

The ImageCLEF medical image retrieval task is currently managed through the ImageCLEF management system [1] that supports all steps of the evaluation cycle from user registration to experiment submission, while the creation of pools, relevance assessments, and metrics computation are supported by the OHSU relevance assessment system described in [2]. To facilitate the integration of the medical image retrieval task to the PROMISE evaluation infrastructure, and in particular to the part directly inherited from the DIRECT system [3], a report listing all important configuration details of the two aforementioned systems regarding the medical image retrieval task, as well as all important additional requirements was prepared by the visitor prior to the meeting. During the meeting, the hosts presented the DIRECT system, to which the medical image collection had already been uploaded at an earlier stage, and based on the report prepared by the visitor, they all worked towards its configuration for supporting the medical image retrieval task, and more specifically the steps in the evaluation cycle corresponding to the experiments submission, creation of pools, relevance assessments, and metrics computation

In particular, requirements regarding the following aspects of the evaluation task were identified, and after discussions solutions were proposed and agreed upon.

- **Topics format.**

  Currently, the ImageCLEF medical image retrieval task sets topic IDs to start from 1 each year, while DIRECT requires that topic IDs are unique across all years and all tasks of an evaluation campaign. To this end, the topic IDs for the medical retrieval

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

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

task need to be homogenised (and made unique) by transforming them appropriately. E.g., the topic from 2009 with ID equal to 5 could be transformed into med-2009-5.

Furthermore, DIRECT requires that the topics are self-contained; this means that the image samples should be provided within the topic, e.g.,

```
<images>
   <image> 34_1.jpg </image>
   <image> 34_2.jpg </image>
 </images>
```

or in a similar format, and not independently from the topics in a separate directory as currently done in the ImageCLEF medical image retrieval task.

Also, given that the topics are provided in several languages, the topic language should be specified in an appropriate XML format, such as:

```
<description xml:lang="en"> Doppler ultrasound images </description>.
```

The 2011 topics for the ImageCLEF medical retrieval task were then formatted according to the above requirements and uploaded into DIRECT. The formatting of the topics used for this task in past evaluation campaigns will be performed as soon as possible, most likely during the uploading of the collected experimental data from past runs of this task, since the submitted runs should also be formatted accordingly to reflect the changes in the topics format.

- **Experiments submission.**

  During the visit, work was performed towards setting up DIRECT's submission interface for accepting runs for the subtasks of ImageCLEF's medical image retrieval task.

  - *Ad hoc image retrieval /Case-based retrieval.* The DIRECT interface was

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

updated so that all the information requested during submission by the ImageCLEF management system [1] would also be requested by DIRECT (run type, query languages, additional info). Submissions for these substasks in DIRECT will be accepted in the trec_eval format and validated. If a run does not conform to the format, a warning is issued and the run is not stored in the system.

- *Modality classification.* The submissions for this subtask are required to adhere to an own format that consists of 3 columns, with the first column corresponding to the image ID, the second to the modality class, and the third to the classification score. DIRECT can handle and validate these submissions and adapt its submission interface accoridngly. Regarding the evaluation, there are two options: adopting either a document-pivoted classification view (for one document tell me the class codes that are relevant) or a category-pivoted classification (for one class-code tell me the documents that are relevant). Given that they were arguments supporting both views, it was decided to try out both options and see how they compare in practice.

- **Relevance assessments system.**

  During the discussions, it was decided that it would be best to not replicate the OHSU GUI [2] in DIRECT (at least for the time being), but to simply integrate DIRECT with the existing OHSU system and just store the data in DIRECT. What this means in essence is that the OHSU system is the one that will be used for the relevance assessments and what will be built is a "communication/data exchange" between the two systems. The advantages of this are manifold and allow PROMISE to fulfil its promise as an open infrastructure. Also, it allows the integration of a consolidated system (the OHSU relevance assessments interface) with which assessors are familiar and which has been widely tested for the requirements of the medical image retrieval task.

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

Overall, there was significant progress towards the integration of the ImageCLEF medical image retrieval task in the PROMISE evaluation infrastructure and the work needed to achieve full integration is expected to be carried out within the next few months.

3.2 Assessing the scholarly impact of evaluation campaigns

The visitor presented the preliminary investigation performed by HES-SO on assessing the scholarly impact of ImageCLEF. This preliminary investigation is described in detail in [4] and it has been accepted for publication in the forthcoming CLEF 2011 conference. A reference was also made to a similar study performed in the context of the TRECVID[4] evaluation campaign [3].

Following the presentation and initial discussions, it was decided to extend HES-SO's preliminary investigation towards two directions: (i) the automation of the current methodology, and (ii) the enlargement of the set of publications being considered in the analysis.

The next steps for this work are described in detail in a wiki page within the PROMISE portal at: http://www.promise-noe.eu/wiki/-/wiki/Main/ImageCLEF+scholarly+impact+analysis

The goal is to perform the extended scholarly impact analysis for ImageCLEF in the coming months and submit the findings as a publication to an appropriate journal, such as *JASIST*, by the end of 2011. This will allow us to automate and consolidate the methodology and to apply it to assess the impact of the whole of CLEF.

[4] http://trecvid.nist.gov/

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

4 References

[1] I. Eggel and H. Müller. The ImageCLEF Management System. In *Multilingual Information Access Evaluation II - Multimedia Experiments, Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum (CLEF 2009)*, pages 332-339, Springer, 2010.

[2] J. Kalpathy-Cramer. Relevance Judgements for Image Retrieval Evaluation. In *ImageCLEF - Experimental evaluation in visual information retrieval*. The Information Retrieval Series, Vol. 32, Springer, 2010.

[3] C.V. Thornley, A.C. Johnson, A.F. Smeaton, and H. Lee. The scholarly impact of TRECVid (2003–2009). JASIST, 62(4):613–627, 2011.

[4] T. Tsikrika, A. G. Seco de Herrera, and H. Müller. Assessing the Scholarly Impact of ImageCLEF. In *Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2011)*, 19-22 September, Amsterdam, The Netherlands, 2011. (to appear)