First International Workshop on Evaluating Information Access (EVIA 2007)
Tokyo, May 15, 2007

A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns

Maristella Agosti, Giorgio Maria Di Nunzio, Nicola Ferro

Information Management Systems (IMS) Research Group
Department of Information Engineering
University of Padua, Italy

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Outline

- Background on experimental evaluation
- Motivations and objectives
- Discussion on the current methodology
- Possible extensions to the current methodology
- A concrete example: the DIRECT system

Experimental Evaluation Background

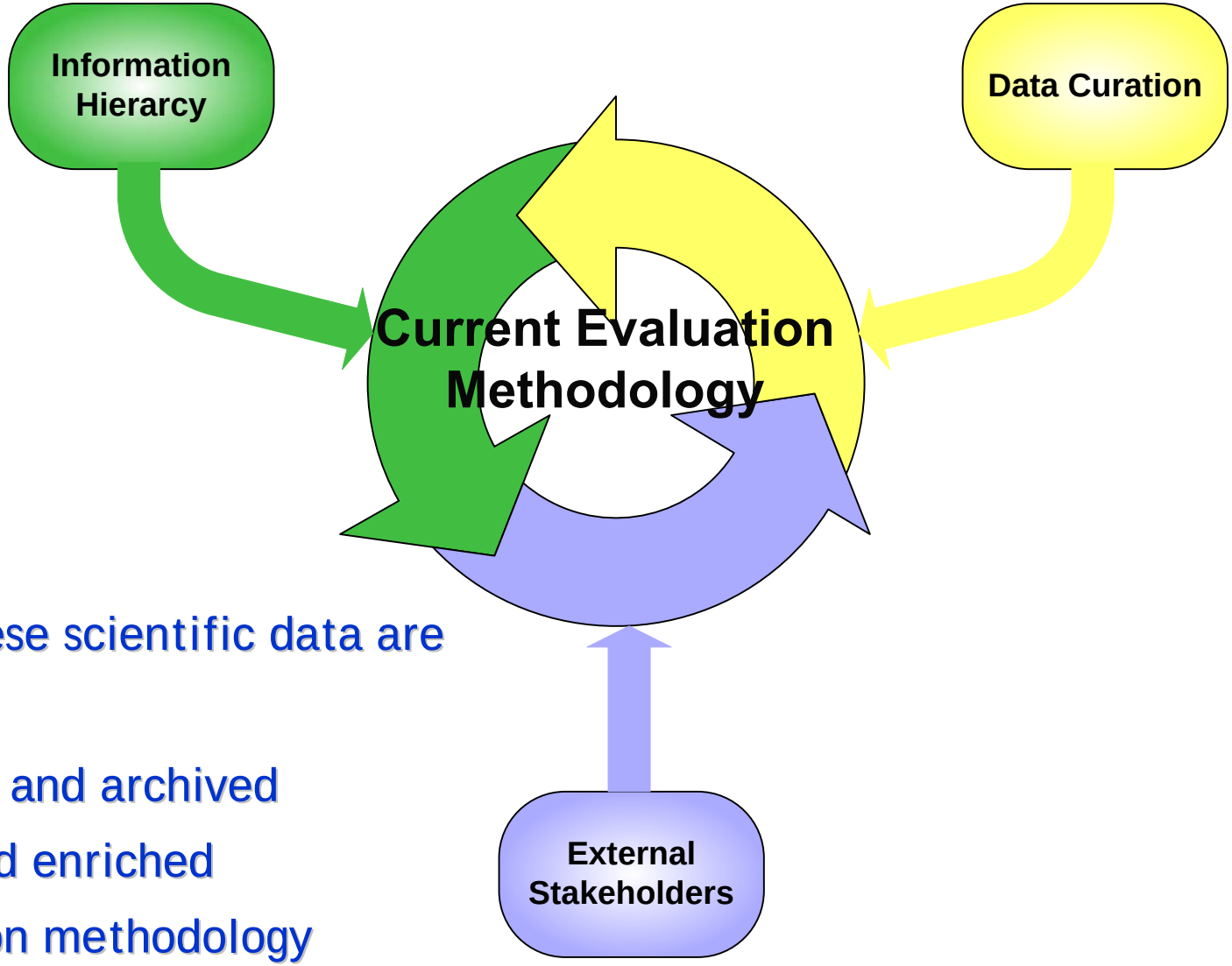
- The Cranfield evaluation methodology is a very well understood paradigm
  - its main focus is on **experiment comparability** and **performance evaluation**
- Several successful evaluation initiatives (TREC, CLEF, NTCIR, ...) have adopted this paradigm
  - they have produced a **huge amount of data**, **promoted the research** in the IR field, and favoured the creation of **cross-disciplinary communities**
- Steve Robertson, in his keynote at ECIR 2007, pointed out
  - the IR field has a **long tradition in evaluation** and "the tradition that began with Cranfield is **alive and kicking**, half a century later"
  - we need to understand "how and when to **push its boundaries**, how and when to **transcend it without throwing it away**"
  - such **evaluation initiatives** are an "**extremely valuable infrastructure** for the [IR] field".

Motivation

The experimental evaluation is a **scientific activity** and, as such, we have to realise that its outcomes are very valuable **scientific data**.

- Comparable experiments
- Performance measurements concerning the experiments
- Descriptive statistics about a collection of experiment
- Hypothesis tests for in-depth analysis of the experiments

Objectives

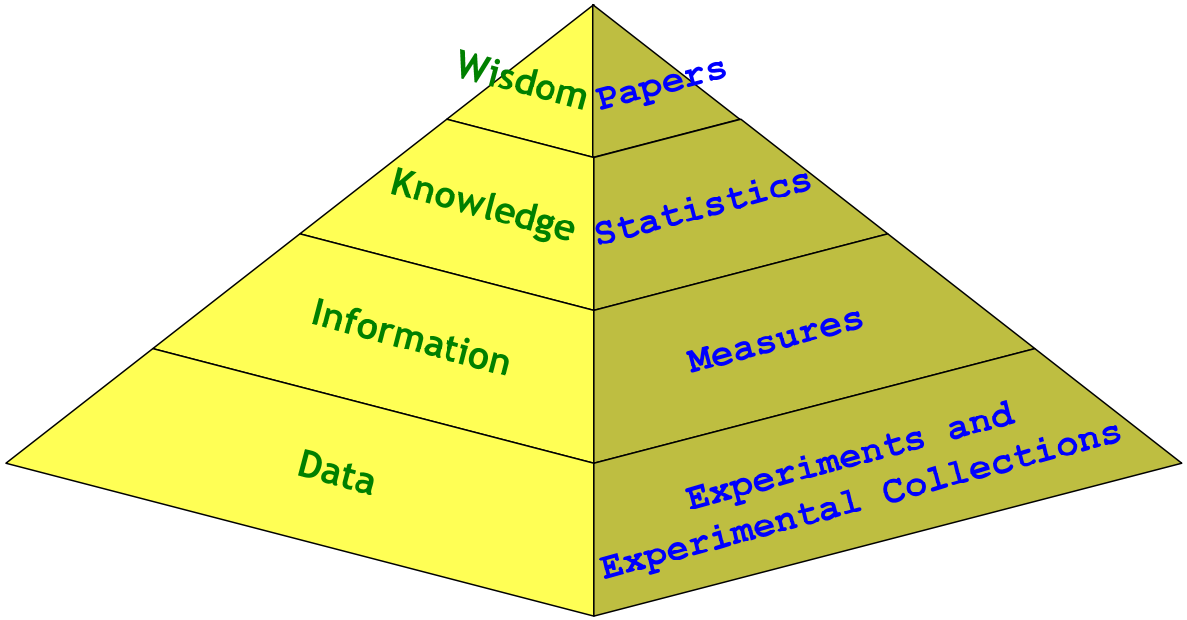


Investigate whether these scientific data are

- properly modelled
- effectively managed and archived
- carefully curated and enriched

in the current evaluation methodology

Information Hierarchy



- **experimental collections** and the **experiments** are **data**, since they are the raw, basic elements needed for any further investigation
- **performance measurements** are **information**, since they are the result of computations and processing on the data,
- **descriptive statistics** and the **hypothesis tests** are **knowledge**, since they are a further elaboration of the information carried by the performance measurements
- **theories, models, algorithms, and techniques** are **wisdom**, since they provide interpretation, explanation, and formalization of the content of the previous levels.

Data Curation

- Scientific data should be **archived, preserved**, **maintained over the time**, and **made easily accessible** to interested users;
  - Re-use of data for new research
  - Retention of expensive or difficult to generate data
- Their **lineage** should be tracked since it allows us to judge the quality and applicability of information for a given use;
  - Validation of published research results
- Scientific data should be **enriched** progressively adding further analyses and interpretations on them;
  - Enhancement of existing data available for research projects
- It should be possible to **cite** scientific data and their further elaboration
  - cross-dissemination of scientific results to research communities and industrial partners

External Stakeholders

- The EC in the 7FP i2010 Digital Library Initiative states that
  - digital repositories of scientific information are essential elements to build European eInfrastructure for knowledge sharing and transfer, feeding the cycles of scientific research and innovation up-take
- The US National Scientific Board points out that
  - organizations make choices on behalf of the current and future user community on issues such as collection access; collection structure; technical standards and processes for data curation; ontology development; annotation; and peer review
- The Australian Working Group on Data for Science suggests to
  - establish a nationally supported long-term strategic framework for scientific data management, including guiding principles, policies, best practices and infrastructure

Extending the Approach to the Evaluation (1/2)

- Introduce a **conceptual model**
  - it makes clear what are the **entities** entailed by the information space of an evaluation campaign, their **features**, and their **relationships**
  - **logical models** can be derived from it to **manage** and **preserve** the experimental data
  - commonly agreed **data formats** for **exchanging information** can be derived from it
- Develop common **metadata formats**
  - they provide meaning to the data, and thereby enable their **sharing** and **re-use**
  - they allow to keep track of the **lineage** of the managed information
- Adopt a **unique identification** mechanism
  - it allows for explicit **citation** and **easy access** to the scientific data and it supports the **enrichement** of the scientific data

Extending the Approach to the Evaluation (2/2)

- Provide **common tools for statistical analyses**
  - they allow for judging whether measured differences between retrieval methods can be considered statistically significant
  - a uniform way of performing statistical analyses on experiments make the analysis and assessment of the experiments comparable too

- Design and develop a **Digital Library System (DLS) for IR scientific data**
  - it is well suited for managing and making accessible the scientific data and the experiments produced during the course of an evaluation campaign
  - it also provides tools for analyzing, comparing, and citing the scientific data of an evaluation campaign, as well as curating, preserving, annotating, enriching, and promoting the re-use of them

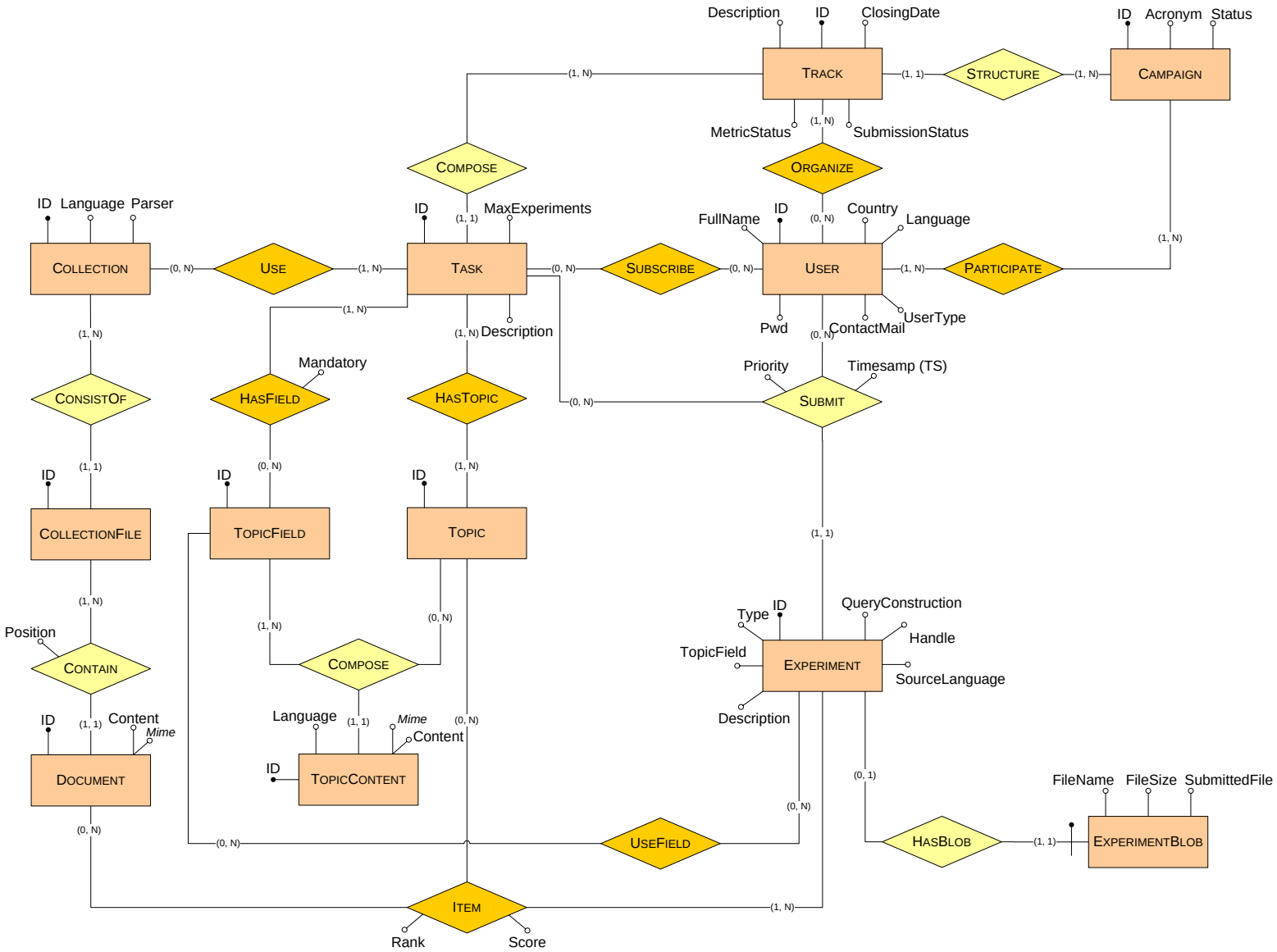
- Give to **organizations** responsible for evaluation initiatives an **active role** in this process
  - they should take a leadership role in developing a comprehensive strategy for long-lived digital data collections and drive the research community through this process in order to improve the way of doing research
  - they should take care also of defining guiding principles, policies, best practices for making use of the scientific data produced during the evaluation campaign itself

DIRECT: a DLS for IR Scientific Data

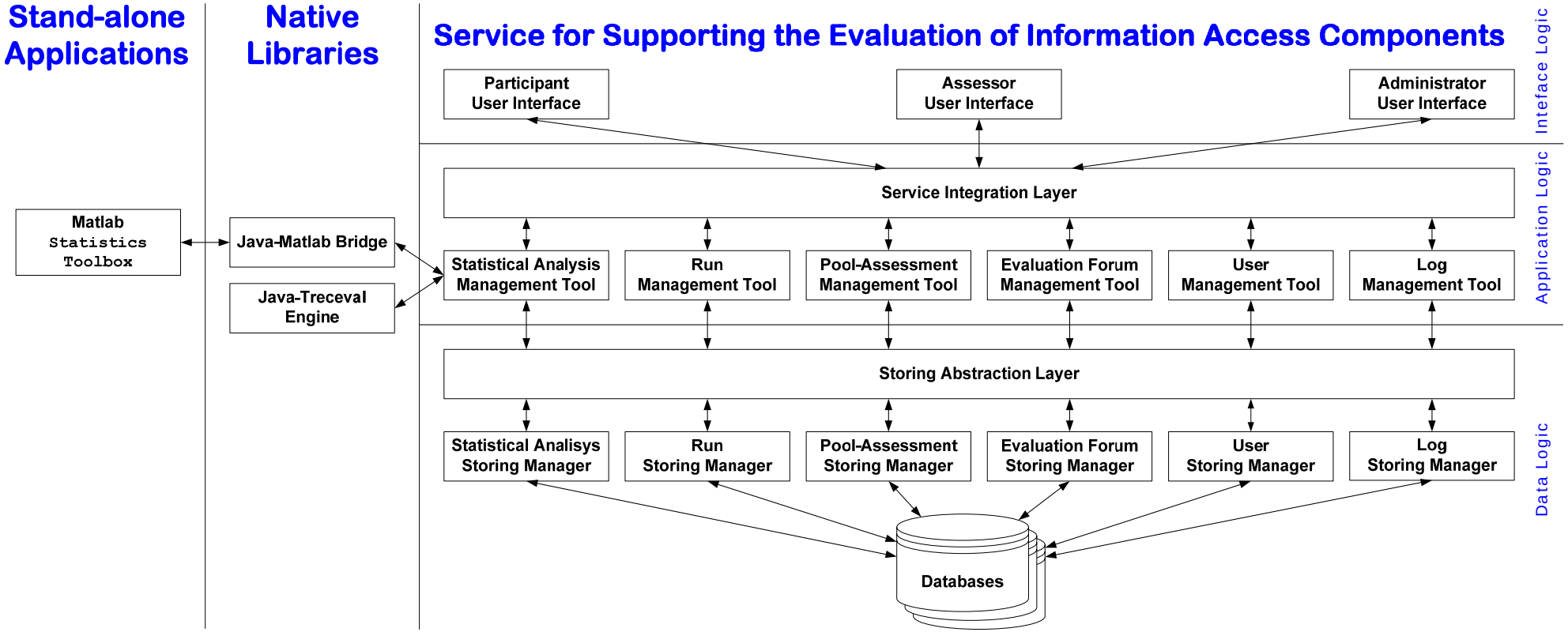
- DIRECT (Distributed Information Retrieval Evaluation Campaign Tool) is a digital library system for managing the scientific data produced during an evaluation campaign

- DIRECT has been adopted in
  - **CLEF 2005**: 30 participants spread over 15 different nations submitted more than 530 experiments; 15 assessors assessed more than 160,000 documents in 7 different languages (Latin and Cyrillic alphabets)
  - **CLEF 2006**: nearly 75 participants spread over 25 different nations submitted around 570 experiments; 40 assessors assessed more than 198,500 documents in 9 languages (Latin and Cyrillic alphabets)
  - **CLEF 2007**: ongoing

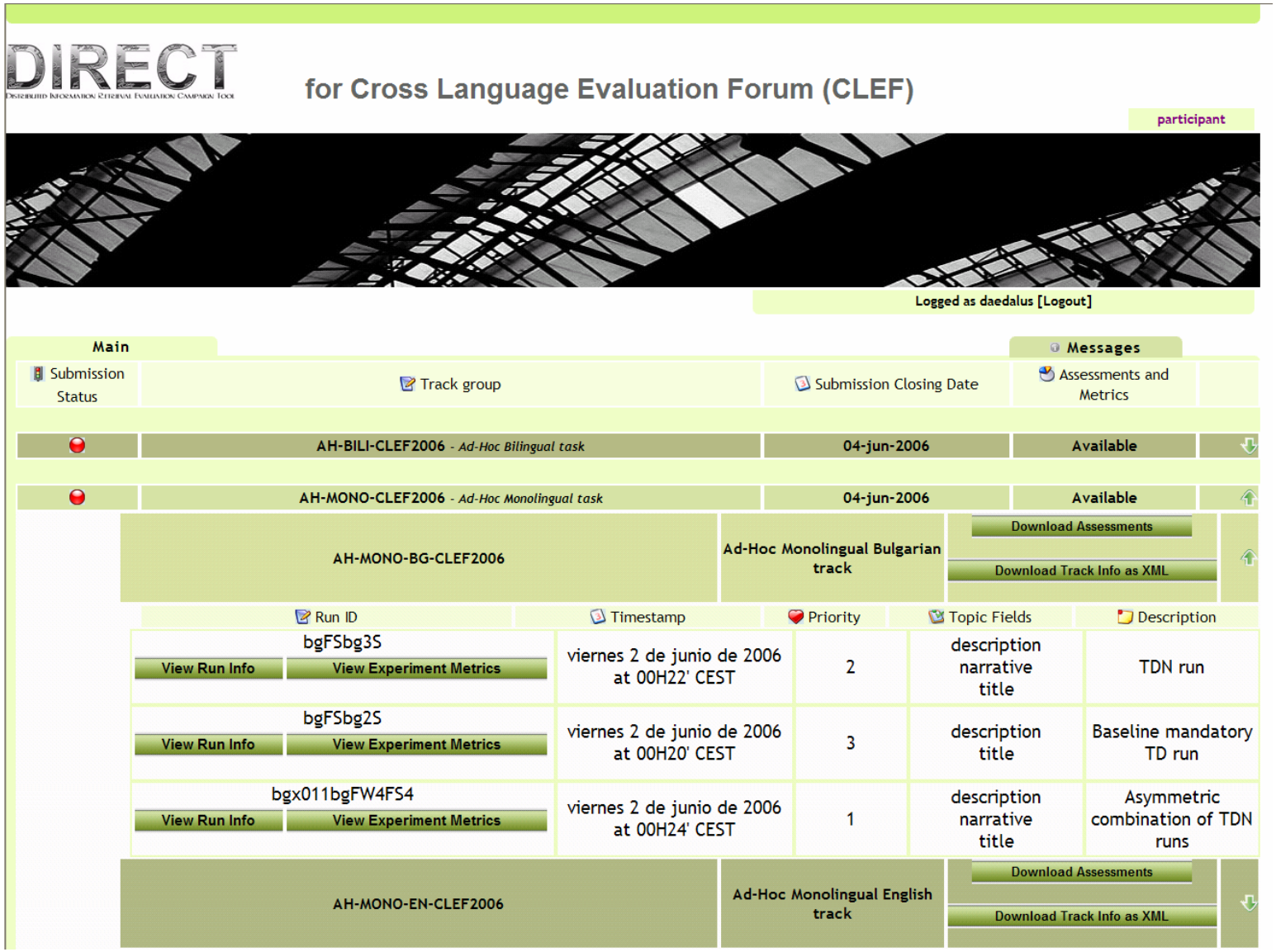
DIRECT Conceptual Model



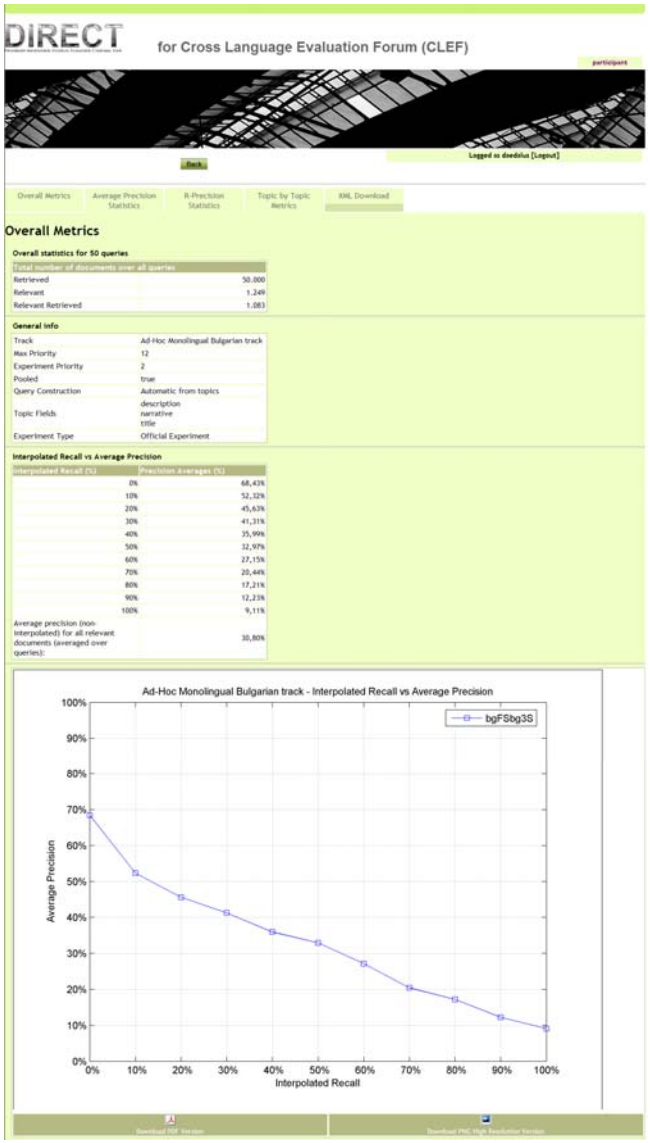
DIRECT Architecture

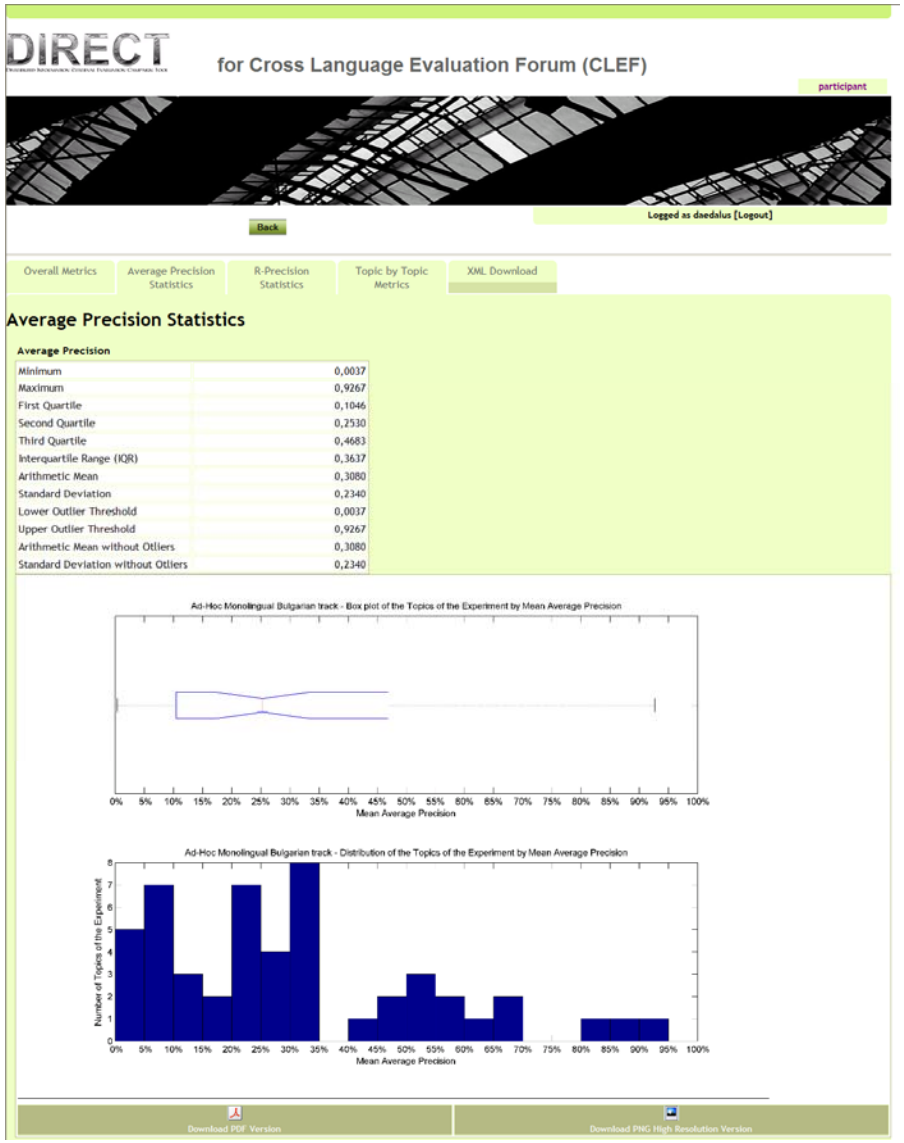


DIRECT Participant Interface (1/2)



DIRECT Participant Interface (2/2)





DIRECT Assessor Interface (1/2)

DIRECT

for Cross Language Evaluation Forum (CLEF)

assessor

Logged as petya [Logout]

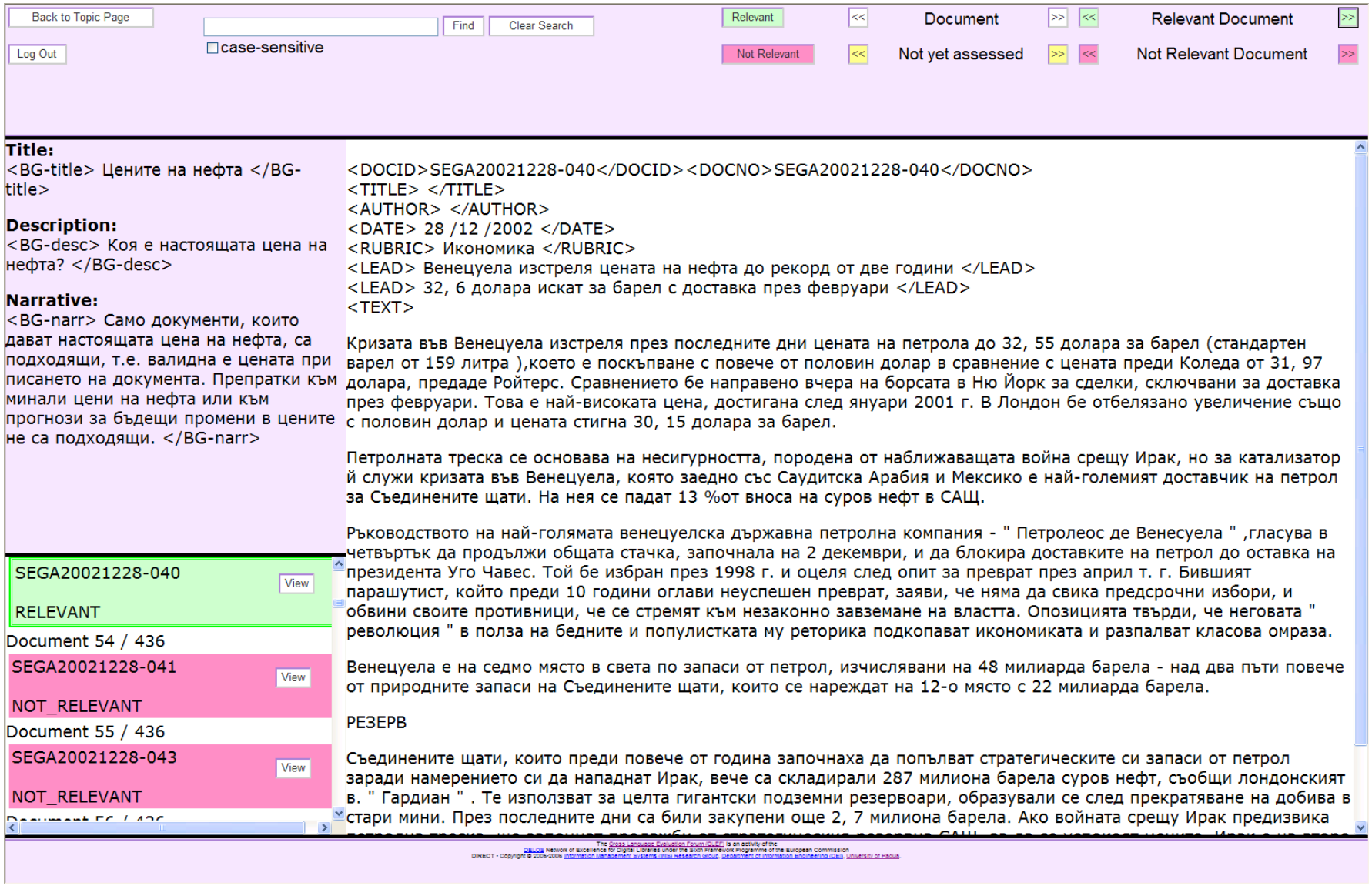
Assessments | **Messages**

Reload Pools

Assessments Status	Pool Name	Pool Description	Assessed Documents / Total Documents
	AH-BULGARIAN-CLEF2006	**Bulgarian Pool for AH-MONO-BG, and AH-BILI-X2BG**	**17 308 / 17 308**

Topic ID	Topic Summary	Total	Relevant	Not Relevant	Not Assessed	Assess Topic
301	Какви продукти на Нестле се продават по света?	487	2	485	0	
302	Намерете документи, които описват или дискутират въздействието на бойкоти от страна на потребителите.	413	6	407	0	
303	Намерете информация за места, където са изложени италиански картини.	367	9	358	0	
304	Дайте имената и/или нахождението на местата, които са били посочени от ЮНЕСКО за места на световното наследство с изключителна красота или важност.	402	14	388	0	
305	Коя е настоящата цена на нефта?	436	16	420	0	
306	Намерете документи, описващи дейността на баската терористична група ЕТА във Франция.	330	3	327	0	
307	Намерете каквато и да е информация за филми, чието действие се развива в Шотландия.	274	3	271	0	
308	Намерете документи, които споменават за пълно или частично слънчево затъмнение.	483	11	472	0	
309	Какви са опасностите за здравето от нелегалните "твърди" наркотици, като например хероин и кокаин, в сравнение с "леките" наркотици?	250	25	225	0	
310	Намерете документи, които описват методи, използвани за обработката или съхранението на индустриалните отпадъци.	321	8	313	0	
311	Намерете информация и цифри за нивото на безработица в европейските страни.	259	5	254	0	
312	Какви са последствията от нападенията на кучета над хора?	299	26	273	0	
313	Намерете документи, описващи чествания по случай 100-годишнината на някое събитие.	318	42	276	0	
314	Какви мерки се вземат в Европа, за да се защитят животните или растенията, които са изложени на риска да изчезнат?	490	30	460	0	
315	В кои спортове се използват непозволени лекарствени средства?	536	66	470	0	

DIRECT Assessor Interface (2/2)



Conclusions

- We have discussed the issues concerning the management, enrichment and curation of the scientific data produce during the evaluation
- We have proposed some possible extension to the current evaluation methodology
- We have designed and developed the DIRECT system, a DLS for IR scientific data, which has been tested in the context of CLEF 2005 and 2006