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

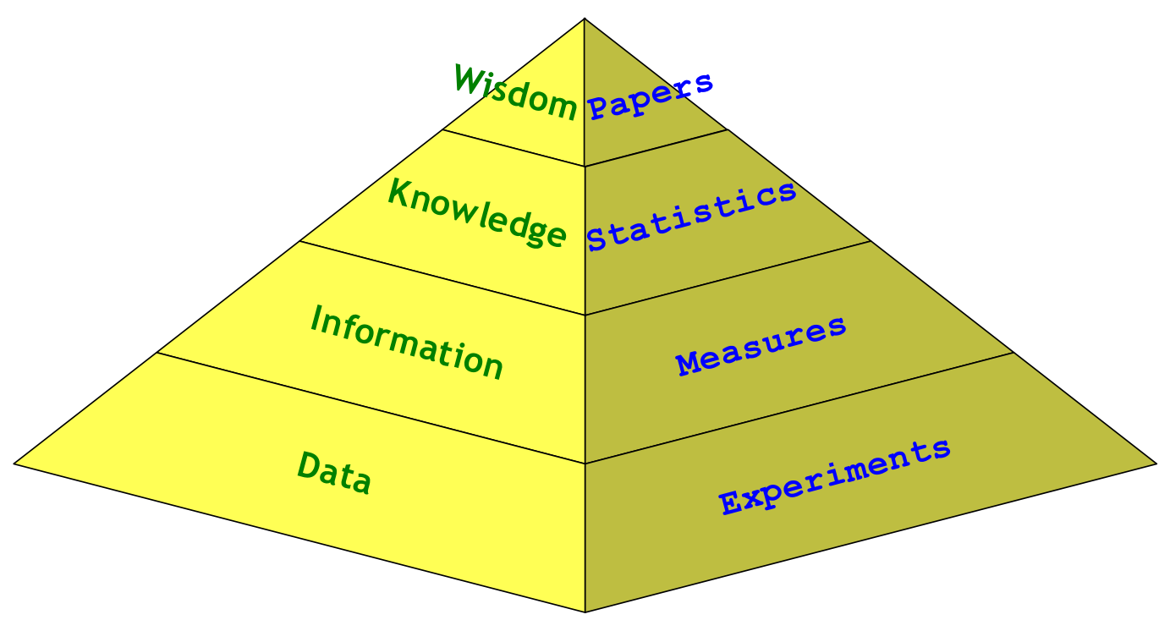
Maristella Agosti, Giorgio M. Di Nunzio, and Nicola Ferro

Information Management Systems (IMS) Research Group
Department of Information Engineering – University of Padova
{agosti, dinunzio, ferro}@dei.unipd.it

Methodology

The current approach for laboratory evaluation of information access systems relies on the Cranfield methodology, which makes use of *experimental collections*.

- An experimental collection $\mathcal{C}$ allows the comparison of information access systems according to some measurements which quantify their performances;
- If we reasoning about this evaluation paradigm, a first step is to point out that the experimental evaluation in the *Information Retrieval (IR)* field is a scientific activity and, as such, its outcomes are different kinds of valuable *scientific data*
- Using the experimental data, we produce different performance measurements, such as precision and recall, that are standard measures that are used to evaluate the performances of an *Information Retrieval System (IRS)* for a given experiment. Starting from these performance measurements, we can compute descriptive statistics, such as mean or median, used to summarize the overall performances achieved by an experiment or by a collection of experiments. Finally, we can perform hypothesis tests and other statistical analyses to conduct an in-depth analysis and comparison over a set of experiments.



- *data*: the *experimental collections* and the *experiments* correspond to the “data level” in the hierarchy, since they are the raw, basic elements needed for any further investigation and they would have little meaning by themselves. In fact, an experiment and the list of results obtained conducting it are almost useless without a relationship with the experimental collection with respect to which the experiment has been conducted and the list of results produced; those data constitute the basis for any subsequent computation;
- *information*: the *performance measurements* correspond to the “information level” in the hierarchy, since they are the result of computations and processing on the data, so that we have associated a meaning to the data by way of some kind of relational connection. For example, precision and recall measures are obtained by relating the list of results contained in an experiment with the relevance judgements J;
- *knowledge*: the *descriptive statistics* and the *hypothesis tests* correspond to the “knowledge level” in the hierarchy, since they are a further elaboration of the information carried by the performance measurements and provide us with some insights about the experiments;
- *wisdom*: *theories*, *models*, *algorithms*, *techniques*, and *observations*, which are usually communicated by means of papers, talks, and seminars, correspond to the “wisdom level” in the hierarchy, since they provide interpretation, explanation, and formalization of the content of the previous levels.

Infrustructure

- The experimental evaluation is usually carried out in important international evaluation campaigns which bring research groups together, provide them with the means for measuring the performances of their systems, discuss and compare their results.
- There are issues raised at international level which suggest that the IR experimental evaluation as a source of scientific data requests and the evaluation methodology itself need to be reconsidered to be properly supported by an organizational, hardware, and software infrastructures which allow for management, search, access, curation, enrichment, and citation of the produced scientific data.
  - ▷ The EC in the i2010 Digital Library Initiative clearly 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 Evaluation

- Scientific data, their curation, enrichment, and interpretation are essential components of scientific research. These issues are better faced and framed in the wider context of the *curation of scientific data*, which plays an important role on the systematic definition of a proper methodology to manage and promote the use of data.
- Therefore, we have to take into consideration the possibility of information enrichment of scientific data, meant as archiving and preserving scientific data so that the experiments, records, and observations will be available for future research, as well as provenance, curation, and citation of scientific data items.
- Furthermore, the importance of some of the many possible reasons for which keeping data is important are for example:
  - ▷ re-use of data for new research, including collection based research to generate new science;
  - ▷ retention of unique observational data which is impossible to re-create;
  - ▷ retention of expensively generated data which is cheaper to maintain than to re-generate;
  - ▷ enhancing existing data available for research projects;
  - ▷ validating published research results.

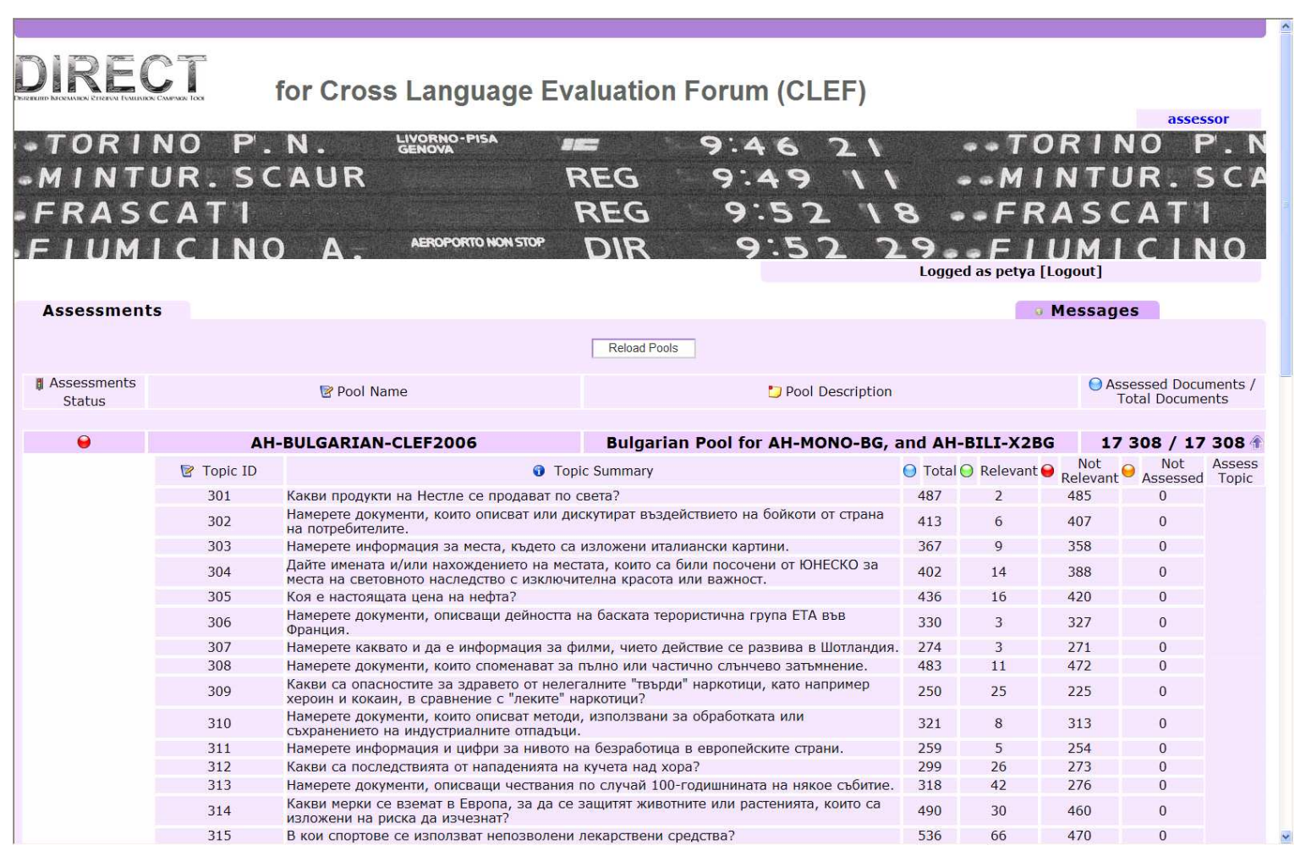
Key Points

Conceptual model and metadata the information space implied by an evaluation campaign needs an appropriate conceptual model which takes into consideration and describes all the entities involved by the evaluation campaign. From a conceptual model we can derive also appropriate data formats for exchanging information among organizers and participants.

Unique Identification Mechanism the lack of a conceptual model also implies that there is no common mechanism for uniquely identify the different digital objects involved in an evaluation campaign. Indeed, the possibility of citing scientific data and their further elaboration is an effective way for making scientists and researchers an active part of the digital curation process.

Statistical Analyses in developing an infrastructure, it is advisable to add some form of support and guide to participants for adopting a more uniform way of performing statistical analyses on their own experiments. If this support is added, participants can not only benefit from standard experimental collections which make their experiments comparable, but they can also exploit standard tools for the analysis of the experimental results, which would make the analysis and assessment of their experiments comparable too.

Running System

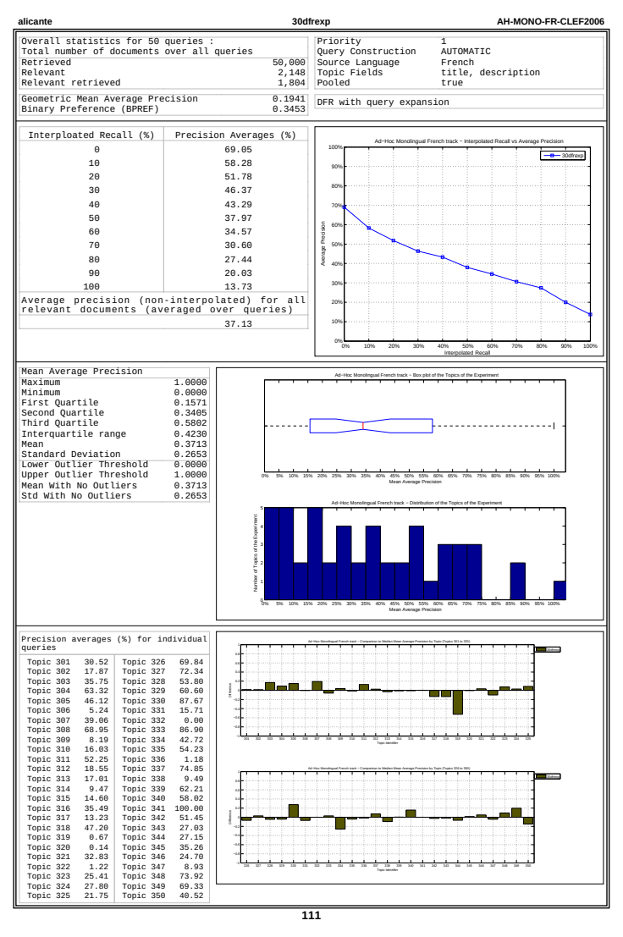

DIRECT for Cross Language Evaluation Forum (CLEF)

assessor

Logged as petya [Logout]

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 бойкота 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 съхраняването 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	


alicante 30dfrexp AH-MONO-FR-CLEF2006

```
Overall statistics for 50 queries :
Total number of documents over all queries
Retrieved                                      50,000
Relevant                                        2,148
Relevant retrieved                              1,804
Geometric Mean Average Precision               0.1941
Binary Preference (BPREF)                      0.3453
```

```
Priority                1
Query Construction      AUTOMATIC
Source Language         French
Topic Fields            title, description
Pooled                  true
DFR with query expansion
```

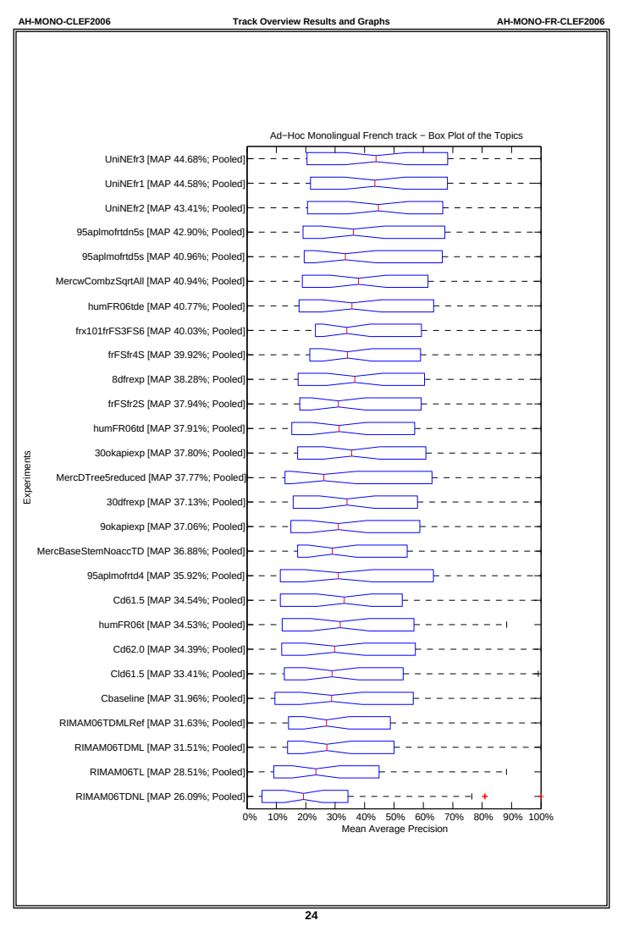
Interpolated Recall (%)	Precision Averages (%)
0	69.05
10	58.28
20	51.78
30	46.37
40	43.29
50	37.97
60	34.57
70	30.60
80	27.44
90	20.03
100	13.73
Average precision (non-interpolated) for all relevant documents (averaged over queries)	
	37.13

Mean Average Precision	
Maximum	1.0000
Minimum	0.0000
First Quartile	0.1571
Second Quartile	0.3405
Third Quartile	0.5802
Interquartile range	0.4230
Mean	0.3713
Standard Deviation	0.2653
Lower Outlier Threshold	0.0000
Upper Outlier Threshold	1.0000
Mean With No Outliers	0.3713
Std With No Outliers	0.2653

Precision averages (%) for individual queries

Topic 301	30.52	Topic 326	69.84
Topic 302	17.87	Topic 327	72.34
Topic 303	35.75	Topic 328	53.80
Topic 304	63.32	Topic 329	60.60
Topic 305	46.12	Topic 330	87.67
Topic 306	5.24	Topic 331	15.71
Topic 307	39.06	Topic 332	0.00
Topic 308	68.95	Topic 333	86.90
Topic 309	8.19	Topic 334	42.72
Topic 310	16.03	Topic 335	54.23
Topic 311	52.25	Topic 336	1.18
Topic 312	18.55	Topic 337	74.85
Topic 313	17.01	Topic 338	9.49
Topic 314	9.47	Topic 339	62.21
Topic 315	14.60	Topic 340	58.02
Topic 316	35.49	Topic 341	100.00
Topic 317	13.23	Topic 342	51.45
Topic 318	47.20	Topic 343	27.03
Topic 319	0.67	Topic 344	27.15
Topic 320	0.14	Topic 345	35.26
Topic 321	32.83	Topic 346	24.70
Topic 322	1.22	Topic 347	8.93
Topic 323	25.41	Topic 348	73.92
Topic 324	27.80	Topic 349	69.33
Topic 325	21.75	Topic 350	40.52

111


AH-MONO-CLEF2006 Track Overview Results and Graphs AH-MONO-FR-CLEF2006

24

References

[1] Agosti, M., Di Nunzio, G. M., and Ferro, N. (2006a). A Data Curation Approach to Support In-depth Evaluation Studies. In *Proc. International Workshop on New Directions in Multilingual Information Access (MLIA 2006)*, pages 65–68. http://ucdata.berkeley.edu/sigir2006-mlia.htm.

[2] Agosti, M., Di Nunzio, G. M., and Ferro, N. (2006b). Scientific Data of an Evaluation Campaign: Do We Properly Deal With Them? In *Working Notes for the CLEF 2006 Workshop*. http://www.clef-campaign.org/2006/working_notes/workingnotes2006/agostiCLEF2006.pdf.

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

[4] Di Nunzio, G. M. and Ferro, N. (2006). Scientific Evaluation of a DLMS: a service for evaluating information access components. In *Proc. 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006)*, pages 536–539. Lecture Notes in Computer Science (LNCS) 4172, Springer, Heidelberg, Germany.

The 6th NTCIR Workshop 2007, May 15-18, National Center of Science, Tokyo, Japan