PROMISE Retreat Report Prospects and Opportunities for Information Access Evaluation

Brainstorming workshop held on May 30–31, 2012, Padua, Italy

Editors
Nicola Ferro[1], Richard Berendsen[2], Allan Hanbury[8], Mihai Lupu[8],
Vivien Petras[12], Maarten de Rijke[2], and Gianmaria Silvello[1]

Authors
Maristella Agosti[1], Richard Berendsen[2], Toine Bogers[3], Martin Braschler[4], Paul Buitelaar[5], Khalid Choukri[6], Giorgio Maria Di Nunzio[1], Nicola Ferro[1], Pamela Forner[7], Allan Hanbury[8], Karin Friberg Heppin[9], Preben Hansen[10], Anni Järvelin[10], Birger Larsen[3], Mihai Lupu[8], Ivano Masiero[1], Henning Müller[11], Simone Peruzzo[1], Vivien Petras[12], Florina Piroi[8], Maarten de Rijke[2], Giuseppe Santucci[13], Gianmaria Silvello[1], and Elaine Toms[14]

[1]University of Padua, Italy
[2]University of Amsterdam, The Netherlands
[3]Royal School of Library and Information Science, Denmark
[4]Zurich University of Applied Sciences, Switzerland
[5]National University of Ireland, Galway Ireland
[6]Evaluations and Language resources Distribution Agency (ELDA), France
[7]Centre for the Evaluation of Language and Communication Technologies (CELCT), Italy
[8]Vienna University of Technology, Austria
[9]University of Gothenburg, Sweden
[10]Swedish Institute of Computer Science, Sweden
[11]University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[12]Humboldt University Berlin, Germany
[13]Sapienza, University of Rome, Italy
[14]University of Sheffield, United Kingdom

Abstract

The PROMISE network of excellence organized a two-days brainstorming workshop on 30th and 31st May 2012 in Padua, Italy, to discuss and envisage future directions and perspectives for the evaluation of information access and retrieval systems in multiple languages and multiple media. This document reports on the outcomes of this event and provides details about the six envisaged research lines: search applications; contextual evaluation; challenges in test collection design and exploitation; component-based evaluation; ongoing evaluation; and signal-aware evaluation. The ultimate goal of the PROMISE retreat is to stimulate and involve the research community along these research lines and to provide funding agencies with effective and scientifically sound ideas for coordinating and supporting information access research.

1 Introduction

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation (PROMISE)[1] is a network of excellence funded under the European Seventh Framework Programme which aims at advancing the experimental evaluation of complex multimedia and multilingual information systems in order to support individuals, commercial entities, and communities who design, develop, employ, and improve such complex systems.

PROMISE organizes a wide range of activities which span from the methodological aspects of experimental evaluation, e.g. proposing new metrics or new ground-truth creation techniques, the organization and running of large-scale evaluation exercises, i.e. the successful *Conference and Labs of the Evaluation Forum (CLEF)*[2] series [8, 20, 24, 32, 33, 49], to designing and developing an evaluation infrastructure for actually carrying out experimentation and evaluation campaigns, i.e. the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)*[3] system [2–4, 7, 11, 27, 29], and dissemination and knowledge transfer, e.g. formulation of best practices [21], organization of brainstorming workshops, tutorials, and summer schools.

Along this line of actions, the PROMISE retreat among network members has been organized as a two-days brainstorming workshop held in Padua, Italy, on 30-31 May 2012, with the aim of discussing and envisioning future research directions for the experimental evaluation of multilingual and multimedia information access and retrieval systems.

The ultimate goal of the PROMISE retreat is to sow the seeds to enlarge the discussion on these topics to the broader research community, to reach a wider consensus on what are key challenges as far as experimental evaluation is concerned, to solicit and stimulate senior and junior researchers at exploring and progressing along these research lines, and to provide funding agencies with scientifically sound ideas and information for coordinating and supporting information access research.

This is not an isolated effort but falls in the track of other similar events that characterized the last decade in the information access and retrieval field. The most recent one has been "The Second Strategic Workshop on Information Retrieval in Lorne" (SWIRL 2012)[4] [10], which has been organized in Lorne, Australia, on 15-17 February 2012, with a broader focus on the overall information access and retrieval field. What makes the PROMISE retreat different from previous events is its specific focus on all the different aspects of the experimental evaluation – both laboratory and interactive – of information systems in multiple languages and multiple media with an outlook at its boundaries and possible links with other disciplines outside computer science, like psychology or neuro-sciences.

The PROMISE retreat has been hosted in the *Academic Senate* room of the *Palazzo del Bo*, the main historical building of University of Padua. 25 researchers from 10 different European countries attended the event, covering many different research areas – information retrieval, information extraction, natural language processing, human-computer interaction, semantic technologies, information visualization and visual analytics, system architectures, and so on.

Prior to the event, all the participants have been asked to think about possible interesting

[1] `http://www.promise-noe.eu/`, contract n. 258191

[2] `http://www.clef-initiative.eu/`

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

[4] `http://www.cs.rmit.edu.au/swirl12/`

challenges, to summarize them in a couple of slides with two or three significant bibliographic references and to send this material ahead of the workshop, so that everyone can have a look and start thinking about.

About three quarters of the first day of the workshop have been devoted to plenary brainstorming where each participant introduced his/her own research statements and proposed possible relevant research directions. All the proposals have been discussed all together, topics by both senior and junior researchers have been discussed and received the same care, and many questions have been raised by the participants. This turned out to be a lengthy process which produced several positive effects: (i) it allowed the participants to go deeper and deeper in a smooth way as the discussion progressed over the day; (ii) it allowed participants to gain a better understanding of each other viewpoints and expertises, which was not so obvious when considering the really wide range of competencies that the retreat brought around the table; (iii) it ensured that no idea was left over and contributed to the formation of a consensus around the main challenges to focus on in the next steps.

The remaining part of the first day has been devoted to the selection and grouping up of the topics and to the creation of workgroups which would have worked on each topic the next day. Six main topics have been identified and a workgroup of 4-5 people has been assigned to each topic.

The first part of the morning of the second day has been devoted to separate workgroups with the task of sketching and highlighting the main items for each of the selected topics. In the second part of the morning the workgroups met up again and there was a plenary presentation and discussion on the initial outcomes of each workgroup. Then, a template to be followed for describing each topic has been discussed and agreed on.

In the afternoon of the second day, the workgroups met up again and started to deeper working on their topics according to the agreed template. In the second part of the second day afternoon, a final plenary session took place where the outcomes of each group have been gathered and put together giving origin to the first draft skeleton of this report. Afterwards, a responsible for each topic has been identified – and they act as editors of the present report – and homework has been assigned to everybody in order to come to the present version of the report.

Sections from 2 to 7 provide details for each of the six identified topics. Each of these sections follows a similar structure in order to facilitate reading and comprehension: first the motivations for the topic are presented; then, the research questions and challenges that stem from the motivations are discussed; the competencies and the cross-disciplinary aspects needed for facing the identified challenges are presented; possible steps and a roadmap for addressing the challenges is then envisaged; and, finally, an outlook of the potential impact of the topic is pointed out. Section 8 wraps up the discussion.

The present report represents a condensed version of the full report [1] available online at: `http://www.promise-noe.eu/promise-retreat-report-2012/`.

2 Search Applications

Before beginning this discussion, let us take a moment to put forward a note on terminology. For the purposes of this text, we decided to use the term ‘Application’, even though discussions also proposed ‘System’. In this case, we take the system to be the implementation of the IR method.

The application is then an instance of a system—the end-user software. Another name used for this, particularly in the commercial world is 'Solution'.

2.1 Motivation

IR research tends to go deep in one aspect, rather than take a general perspective, which may be viewed as superficial. Consequently, evaluation benchmarks are also generally designed to evaluate the core elements of search, rather than look at an application as a whole. Ultimately, evaluation needs to be done at all the different levels of detail, and this section is certainly not arguing that application-level evaluation is better or more useful than any other type of evaluation. In fact, Section 5 looks at an even lower level of detail than most current benchmarks, while Section 4 looks at how to improve current test collections and procedures.

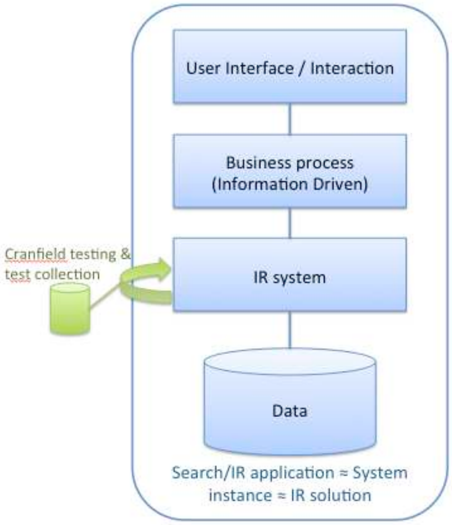

User Interface / Interaction
Business process (Information Driven)
Cranfield testing & test collection
IR system
Data
Search/IR application ≈ System instance ≈ IR solution

Figure 1: A common view of a search application.

Instead, our motivation for this section is to simply obtain a broader understanding of search applications in a scientific, reproducible and justifiable manner. In doing so, we hope to improve take-up of evaluation best practices by a larger audience, including a set of users currently not targeted by existing benchmarks (e.g. system administrators, decision makers) [50].

Our understanding of search applications departs from a common view, as show in Figure 1, but can be seen from two perspectives:

Users A search application is an application that models an information intensive process and is powered by IR technology. This is the predominant view of this section.

Components A search application is an instance of an IR system, a service layer based on an underlying model of the process, its specific data, and the user interaction.

2.2 Research Questions and Challenges

A list of challenges which we believe have to be addressed in order to obtain the broader, scientific understanding of search applications.

Who is the "Consumer"? Evaluations and benchmarks are not done for their own sake, but rather in order to assist somebody (we call this person *"a consumer"*) in taking a decision with respect to the search application.

We can argue that this "consumer" of evaluation results is a system implementer, maintainer, a decision maker, perhaps a CTO (Chief Technology Officer). The adoption of a particular search application is not a one step process, but rather a series of steps. This is a "consumer" who no longer just buys, but also enriches and participates in the design of the evaluation and the evolution of the application. For data intensive departments or companies, the customer fills a new position in the organisation: Search Application Administrator, complementing the Database Administrator—an already established position in many organisations.

The idea that the consumer/user is potentially deeply involved in the evaluation of the application and even enriches it, creating value for both herself and for the application providers, creates a link to Service Sciences [54], where research questions regarding the evolution of services through user involvement are studied [40].

Abstracting from the user Having obtained a better understanding of the consumer, the question is whether we can derive a *user* profile from user studies, or formulate one from functional requirements. Note the difference between the consumer of the evaluation, discussed above, and the user of the system. While the former is a technical decision maker, the latter may not be a technical person. Nonetheless, the consumer can only make informed decisions if the evaluation properly models its users. Furthermore, many applications have to cater to different user profiles and appropriate measures to evaluate a search application in such situations are needed.

Paradigm expansion Upon considering the problems and research questions outlined above, it becomes fairly clear that there needs to be an expansion of the test collection paradigm. In addition to core effectiveness values, we must be able to reliably monitor other aspects influencing user satisfaction and customer adoption. The objective is to maintain as much as possible the reproducibility of the test collections based evaluation, but expanded to include these new aspects.

The observation of the application as a whole provides new opportunities. We are no longer restricted to test collection, but rather agree on a set of guidelines on how to test/score search applications. Where necessary, these guidelines may consider other type of information practices such as recommendations, user generated content. It should not be unimaginable to look at specialist fora and derive measures of application quality based on the mentions there. The challenge is to make this process verifiable and repeatable.

Domain instantiation When mentioning *specialist* fora, we raise a significant research problem: How much of the user and customer modelling described above can be done generically and how much needs to be adapted from domain to domain? How does this change the evaluation scheme?

We can imagine defining a generic model as above, and for each component of the evaluation scheme defining its instantiation in a particular domain. In this derivation, the links between the components are equally important as the components themselves.

Other considerations Other aspects or components of such an evaluation may refer to the correctness of the implementation (as simple as checking that the boolean query ‘A and B’ returns less than or equal to ‘A’ and less than or equal to ‘B’), user perception (already explored in interactive systems [45]), and continuous evaluation.

2.3 Competencies and Cross-Disciplinary Aspects

Addressing the challenges described above requires a set of skills difficult to find within one group, and even less within one person.

A good understanding of **IR evaluation** practices would provide the necessary design know-how for new test collections and effectiveness measurements. **Service Sciences** to study the interaction between different entities (generally viewed as a consumer and a provider) in order to generate value for both. Experience in **User Modelling** and **Human-Computer Interaction** would allow us to design the necessary tests to go beyond effectiveness evaluation. Putting everything together requires mastering **Software Engineering** principles. Finally, a **Communication Science** professional is perhaps needed in order to make sure that communication between the three groups of people (researchers, users, consumers), is efficient and without misunderstandings.

2.4 Roadmap

The research challenges outlined above, indicate a large action space. Rather than a linear roadmap, we can think of the necessary actions as a true city map, with different roads intersecting each other. We identify four destinations:

1. **Measures** to combine standard effectiveness and efficiency with potentially new measures that a decision maker may use in considering a search application (e.g. reliability, costs)
2. A new type of **benchmark** is to be developed. The test collection is an important component, but not the only one. Identifying a generic grid of tests and applying them in a scripted fashion is a first step, followed up by improvements in scalability and reliability.
3. We need to identify a potential **infrastructure** for performing this kind of evaluation.
4. Comprehensive **analysis** from the perspective of the customer or user

2.5 Impact

In the IR community, we need to raise awareness for the needs of search application customers and users. In particular, adaptation of evaluation infrastructures, such as DIRECT, to handle [parts of] such search application evaluation are probably needed. Steps in the directions listed here involve customers and users in the definition of a best practice of search application evaluation, with the ultimate goal of increasing the adoption, for the long term, of these best practices. Ultimately, this will tighten the relationship between IR evaluation and real IR applications.

3 Contextual Evaluation

3.1 Motivation

Contextual evaluation means integrating users, tasks, search applications and underlying information retrieval systems in a holistic perspective. Encompassing the

- **user context** (i.e. location, situation, environmental or seasonal factors as well as devices),
- **user tasks and goals** (not just an individual, specific information need expressed in a query but commonly part of a larger task to be achieved),
- the **search application** (including the interfaces, tools and middle layer – see Section 2), and

- the **underlying information retrieval system**

ensures that the global impact of an IR system on a user - in terms of work place productivity or quality of life - can be assessed.

3.2 Research Questions and Challenges

Five research challenges or questions seem of particular interest and will be discussed in the following.

User vs. Task For more than a decade, IR research has moved its attention from solely analyzing the contents of documents and extracting relationships from among the components of documents to mining user-system interaction behavior patterns. Although task (or topic) has been a key element in IR systems evaluation since the Cranfield days, task is a term that is abused; the concept is sometimes considered equivalent to the experimental task in human experiments, the topic created for systems evaluation, or the query issued by the user. But which drives systems success: knowing more about the user, or knowing more about the task that the user needs to complete?

How systems are developed and the data used to predict relevant documents will be influenced by which of these (or perhaps both) influence our ability to predict system success. Part of this question also includes determining whether neither adds sufficient power/variance to make a difference to search success, and determining whether both offer equal value.

Complex Tasks and Usage Scenarios Most IR evaluation tackles fairly simple search tasks, but these often form part of complex task solutions, where search is one activity in a sequence of events where the user switches between several systems and tools over time [46]. These tasks can occur both in relation to work settings but also in a leisure or entertainment environments. We can optimize a retrieval component in this sequence, but currently we know very little about how an improvement in e.g. precision affects the whole complex work task, and even if standard performance measures are suitable for evaluating the success of a given component.

An understanding of how subtasks contribute to overall task solving and the development of success criteria and measures for each subtask type, both on its own merits and in relation to the sequence of subtasks, will be necessary.

Individual vs. Collaborative Information Retrieval It is commonly assumed that searching is an individual activity only. Recent research has shown that people also collaborate during the search process: collaborative information handling and information sharing can be found in both professional work settings and in everyday situations and contexts [36, 37].

Collaboration may occur at different stages of the information seeking process: planning a work task, defining the information need or problem, query formulation, result assessments and information use. One challenge will be to investigate and classify manifestations of collaborative information handling activities during the interactive search process and extract collaborative features. Another challenge will be to add features of collaborative information searching and integrate them into the standard individual-based models of information seeking and information retrieval.

Query-free Retrieval IR has traditionally focused on aiding users in their information seeking process by requiring them to *explicitly* (re)formulate their information need as queries and matching those against the documents in an index. However, there are many situations where users would benefit from having their more *implicit* information needs satisfied without having to specify them as explicit queries. Instead, the system could use knowledge about the user and their context to proactively satisfy those implicit information needs.

Query-free approaches to IR can be challenging to evaluate as they do not fit into the traditional IR evaluation paradigm. The relevance of suggestions is entirely dependent on the task, situation and environment, and assessing the relevance of a suggestion against a contextual snapshot is considerably more difficult than against a static query. Evaluation of query-free suggestions needs to be done in context, as they are being shown to the user. In addition, the contextual snapshot of the user against which recommendations are generated should not only contain information about the current situation, but also about the past interactions between the user and the system.

Pre- vs. Post- Retrieval Stages When evaluating an IR system with respect to their impact on the work tasks or activity of a user, the pre-retrieval context, i.e. information need, situation, skills of the users should determine the offered functionalities and interactions of the system, however, the post-retrieval context, i.e. how the information retrieval changed the outcomes, work environment and advanced the users in achieving their goals should determine the assessment of the system. Both user needs (pre retrieval) and user intentions or outcomes (post retrieval) determine the quality of the system.

Approaches for evaluating IR systems need to take their place in the work environment and task processing stage into account. The task, particular information need and usage intentions and actual usage (post-retrieval stage) need to be acquired and monitored. Measures will therefore go beyond the area of topical relevance and measures based on it (e.g. precision and recall) and move toward utility-based measures, for example user satisfaction, task achievement percentage or degree of interruptiveness for the task process.

3.3 Competencies and Cross-Disciplinary Aspects

Contextual evaluation requires an interdisciplinary research team with methodological capabilities in experimental, lab-based and longitudinal observational user studies, domain experts, HCI & usability experts, and, of course, IR system experts combining qualitative and quantitative methodologies for gathering data using ‘real-life’ approaches. Some of the innovative approaches for studying contextual factors are:

- **work task and subtask analysis** to assess the context in which the system needs to fit
- **automatic capture** of user intentions and interactions,
- **search data / process mining** for more innovative analysis of search sequences
- formal mapping of evidence and **combination of evidence** from many diverse sources

3.4 Roadmap

One of the bigger challenges for contextual evaluation is the question whether answers can still be found in a component-based way. Can we determine tasks, context, intentions and then determine individual evaluation measures before we aggregate everything? We suggest that contextual evaluation can only work when pilot studies in **robust and controlled environments** are supplemented by longitudinal experiments with **large pools of real-life user groups in their natural environments**. For this to work, prototype systems need to be deployed in real-world domains, making it necessary to develop production-type systems. Before these experiments can be implemented, preliminary theoretical approaches, methodologies and technologies for capturing context and its impact on IR systems need to be studied. Among these, we count:

- solutions to **capture user interaction**,
- **identification of real-life user groups** and methods for **observing** their systems use,
- **identification of scenarios involving different work and leisure tasks**,
- identification of **success and non-success criteria** for tasks and subtasks,
- development of **controlled environments that resemble real-life domains**, and
- development of **measures** for IR system effectiveness **incorporating contextual factors**.

3.5 Impact

A deeper understanding of the effect of context on individual systems and components and their interplay on overall task performance can aid in prioritizing which components to focus research and development on. Improvements in a particular IR component may only result in user benefit if embedded at the right stages of task solving. Extended knowledge on search behavior, both for professional and everyday searching should aid in developing innovative applications that are grounded in deeper understanding of contextual issues and should integrate more easily into users' environments.

For the field of IR evaluation, insights on how to design experiments and tools that support more realistic search scenarios and better reflect the different needs users may provide additional insights on how to model search processes and success measure. At the end, revised versions of the Cranfield model may be achieved.

4 Back on TREC: Challenges in Test Collection Design and Exploitation

4.1 Motivation

Large-scale evaluation campaigns have witnessed the rise of many of the state of the art retrieval methods that we take for granted today [53]. Still, there are recurring issues in test collection design and exploitation that deserve our ongoing attention. We can define three main themes regarding test collections:

- **Increasing efficiency** Test collection creation is a labour-intensive and costly process, which often poses an upper bound to the collections size. We need to foresee ways to overcome these issues in order to keep experimental collections aligned with the increasing number of application domains, growing sizes, and capability of quickly providing suitable and robust collections as new trends and research areas emerge.

- **Evaluating according to end user preferences** Benchmarking style evaluation usually abstracts away a great deal of variance in end users and their contexts. If outcomes of benchmarking experiments reflect end user preferences better, benchmarking will become more relevant for industrial search service providers.

- **Impact in the research community and industry** Large-scale evaluation campaigns are costly exercises which produce several benefits on industry and research in terms of both experimental collections made available and advancements in the state-of-the-art. We need appropriate means to make this impact explicit, both for research and industry.

4.2 Research Questions and Challenges

Increasing efficiency Among the several way of increasing efficiency [17], we focus on:

- **Crowdsourcing** [23, 31, 47] can be a means to create test collections more efficiently. Yet there are hidden costs associated with crowdsourcing. For *quality control*, a well thought out *experimental design* is essential. In addition, economical and ethical aspects play a role. Aspects we are interested in include assessor agreement, spammer control, payment policy, gamification, hit design, worker reputation, post-processing of judgments, communication with and among workers, hiring people based on desired demographics, and more. Are workers paid a living wage? Can they communicate with each other? Can we communicate with them? Is there a need for a new crowdsourcing platform?

- **Generating ground truth** by adopting pseudo test collections [15, 16, 41], for tuning, training [14, 18] and evaluating retrieval algorithms. This leads to several research questions: does evaluating on generated pseudo judgments rank retrieval algorithms in similar way as evaluating on editorial judgments? Do retrieval models tuned on trained on generated pseudo judgments generalize well to editorial judgments? Do they generalize better than models tuned on editorial judgments? What is the impact of bias in generated pseudo judgments toward a retrieval algorithm that we are evaluating or training?

- **Semi-automatic test collection creation** in contrast to generating ground truth fully automatically, an important area of research focuses on semi-automatic processes for test collection creation, where the human is kept in the loop. Research questions here include: What is the impact of interface design on the quality of annotations? What is the impact of the quality of the suggestions from the automatic method feeding the interface?

A use case framework for information access applications There is a need of a broad and commonly agreed use case framework for explicitly describing use cases underlying evaluation tasks [42] and enabling researchers to think about end users of information access applications in a structured way. The framework allows for describing very different use cases, broadening the scope of the traditional ad hoc search evaluation and, for test collection generation, this means that evaluation can be tailored to foreseen usage scenarios.

The envisage use cases need to be validated in the sense that they should reflect usage by real end users of real services through interviewing these end users and service providers. This leads to the following research questions: can search service providers satisfactorily describe use cases of their service in our use case framework? Can organizers of benchmarking tasks successfully translate properties of the use case underlying their evaluation task to decisions about the task setup, evaluation metrics, and so on?

Impact of CLEF in the research community and industry Several attempts have been made to assess the impact of evaluation campaigns from both a research perspective [55, 56] and an industrial/economic one [52]. On the other hand, these has been rather isolated effort, each one build on its own. Therefore, a first natural question that pops up is whether, as research field, we need to develop and adopt commonly agree and understood ways of assessing the impact of the research in the fields.

For example, as far as the scholarly impact is concerned, the accumulated research in the experimental evaluation community has certainly had a major impact on the development of information retrieval systems and beyond, but how to measure this? Citation analysis is an obvious and useful way, but may very well be enriched with other techniques such as text mining for analyzing the spread of research topics across the publications in large-scale evaluation campaigns and beyond.

Another key point is how to turn this from a sporadic "one-shot" activity into a continuos process, able to trace and monitor the impact of evaluation activities over the time. This clearly calls for the design and development of evaluation infrastructures able to provide the support for running and implementing impact analysis policies over the time [3].

4.3 Competencies and Cross-Disciplinary Aspects

The questions we have raised and the methods we intend to use to answer them span a wide range of skills and competencies. Our research into crowdsourcing will first and foremost be a literature review. Generating ground truth for evaluating and training retrieval algorithms, and estimating the success of this requires sound experimental design, use of appropriate statistical tests, implementing many retrieval algorithms, and employing machine learning. Semi-automatic creation of test collections requires interface design, experimental design, machine learning, setting up and organizing an evaluation campaign, analysing the results and so on: this work is interdisciplinary on its own. To validate and promote the use of our use case framework we need to compose questionnaires, interviews and summaries of the framework for distribution. Techniques from the social sciences will be employed in the analysis of such interviews. Impact analysis draws upon bibliometrics, expertise mining, machine learning, text mining, and so on: it is a highly challenging endeavour.

4.4 Roadmap

We extend our work on generating pseudo test collections with methods to learn from editorial judgments the characteristics of high quality content. We take our methods to different collections and domains to understand its robustness.

Tools we employ to estimate the impact the CLEF campaign has on academia and industry community include citation analysis and expertise and topic mining. The Saffron expert finder system provides insights in a research community or organization by analyzing its main topics of investigation and the experts associated with these topics. Saffron analysis is fully automatic and is based on text mining and linked data principles.

4.5 Impact

If our use case framework receives uptake in the academic and industrial community, we believe it can have a positive impact on the quality of evaluation experiments in information retrieval. The main effect should be that outcomes of such experiments would better reflect end user preferences. The impact of our research on generating ground truth for training and evaluating retrieval algorithms is encouraging researchers to tune and train evaluation algorithms on generated ground truth. Moreover, the quantity of generated ground truth can lead to an advantage over training on editorial data. Finally, a better understanding of where we are in the world wide information retrieval research community; as well as the contribution we make to innovation in industry.

5 Component-based Evaluation

5.1 Motivation

Information Retrieval has a strong tradition in empirical evaluation, as exemplified by the many evaluation campaigns [38]. The majority of IR evaluation campaigns today are based on the TREC organisation model [39], which is based on the Cranfield paradigm [25]. One of the most important parts in demonstrating the utility of evaluation campaigns is to show the improvement that they have brought to IR. As discussed in [35], the widely-adopted TREC approach has a number of disadvantages [51]. The most pertinent to component-based evaluation are: fixed timelines and cyclic nature of events; evaluation at system-level only; and difficulty in comparing systems and elucidating reasons for their performance.

Evaluation campaigns today focus on leaving a legacy in the form of the availability of the evaluation resources (datasets, topics, relevance judgements). This certainly makes an impact by removing the necessity of research groups to construct their own evaluation resources. The availability and use of these shared resources should lead to the objective comparability of techniques and focusing of researchers on promising techniques while avoiding typical mistakes of the past. However, assessment of the results obtained in papers using standardised evaluation resources [13] lead to the conclusion that it is not clear from results in published papers that IR systems have improved over the last decade — claimed improvements are often compared to weak baselines, or improvements are not statistically significant. The current emphasis on quantity of publishing also tends to lead to the publishing of minimal changes in systems, often in a non-reproducible way.

There is therefore a need for an approach to the experimental data which favors their curation and in-depth studies over them, to be able to assess the progress over long periods of time [6, 9].

5.2 Research Questions and Challenges

One of the first research questions to consider is the creation of guidelines for future publications containing empirical IR results. These guidelines should specify which information about IR systems and experimental results should be published and in what format, in order to allow rigorous comparison between the systems and results. Following the guidelines should in particular facilitate the use of *systematic reviews* on IR results. A systematic review is a methodology to rigorously and systematically locate, assess and aggregate the outcomes from all relevant empirical studies related to a particular scientific question, in order to provide an objective study of the relevant evidence [22]. The aim of the guidelines will be to encode the pertinent information explicitly so as to reduce the amount of manual intervention required in the creation of systematic reviews. The guidelines could be divided into those that can be implemented with minimal change in the current approach to empirical IR evaluation (e.g. introducing a stricter format for IR evaluation campaign papers to make important information more explicit) and those that will require a major shift in the approach to publishing empirical IR evaluation results. For the latter, the suitability of nano-publications to IR evaluation could be investigated. The concept of nano-publications has been introduced in the area of genetics, but should be applicable to all data intensive sciences [48]. The central idea is to encode the central findings of a paper as RDF triples, instead of "burying" the findings in narrative text that has to be mined to re-obtain the findings.

An initial collection of components for text and image retrieval for which it would be useful to measure the interactions and their effects on the retrieval results is shown in Figure 2. The optimal way of facilitating the use of component-based evaluation will be through the introduction of an infrastructure for component-based evaluation. A good basis for the infrastructure is a standard open source workflow tool [28], which will allow straightforward integration of components into the infrastructure, as well as a straightforward approach to combine components into workflows. A promising candidate is Taverna[5], as it is the most widely used tool on the myExperiment portal

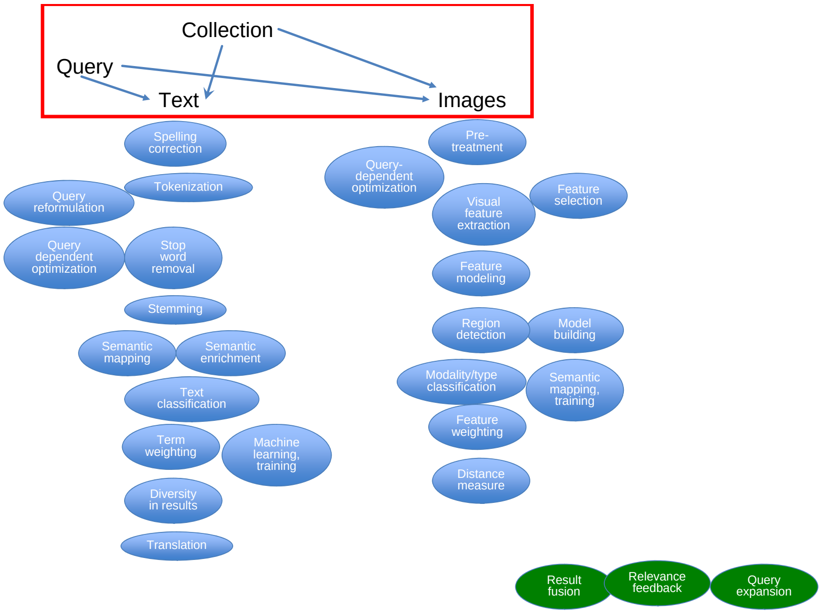


Figure 2: Text and image search components.

[5] http://taverna.org.uk/

for sharing scientific workflows[6]. Taverna also has the advantage that it has been integrated with the U-Compare UIMA-based text mining and natural language processing system[7] [44]. The infrastructure for component-based evaluation, including all protocols and a workflow system for combining the components, will have to be designed and implemented. The Cloud is a promising environment in which to host such an infrastructure — evaluation data can be stored on the cloud, and components could be programmed in computation instances of the cloud infrastructure which could then be registered with the infrastructure [34]. The infrastructure should allow IR scientists to design a workflow for an experiment specifying a specific class of component at each point. New metrics and visualisations will have to be developed to allow effective interpretation of complex experimental results. A further advantage of such an infrastructure is that experiments could be done on private data (such as medical records) in a relatively straightforward way, as it would not be necessary to distribute the data — components could be run on data on the cloud infrastructure without the data being seen by the person doing the experiments. Questions to be answered by the experiments on the component-based evaluation infrastructure include: why do some components work badly some of the time, why do combinations of sub-optimal components sometimes give better results than combinations of optimal components, what is the effect of varying parameters of the components on the results, can individual component performance predict the overall performance of the system? It will likely be necessary to take this even further and consider various parameter values for the various components.

A further challenge is providing motivations and incentives for IR scientists to make use of the component-based evaluation infrastructure and to add components to the infrastructure. These could be in the form of both “carrots” and “sticks”. Carrots include access to test data, tasks and relevance judgements, as well as to extensive evaluation results and comparison to the state-of-the-art with statistical significance tests and visualisations. Sticks include requirements to use and contribute components to the infrastructure in order for papers to be accepted at conferences and workshops, or requirements of research funders that the results of the research be contributed in this way. Papers would have to be linked to the evaluation infrastructure and include a detailed specification of the setup, leading to executable papers [30]. This would link well to current initiatives for the digital preservation of scientific data and results[8].

5.3 Competencies and Cross-Disciplinary Aspects

Political influence is needed, especially to obtain funding for the infrastructure and to implement the “stick” parts of the incentives. People with IR evaluation skills and experience are of course needed. Finally, experts on large-scale computing, cloud computing, etc. will be needed for designing, specifying and implementing the infrastructure.

5.4 Roadmap

The most immediate step is to require for the next CLEF that participants submit descriptions of the components of their systems. This will require the definition of a template, such as in XML

[6] http://www.myexperiment.org/

[7] http://u-compare.org/

[8] For example, http://www.alliancepermanentaccess.org/

format, so that it can be used for statistical analysis. This template should be short but detailed. Then we need to discuss with funding agencies, especially for obtaining the funding for putting the infrastructure in place. We also could start working out more detailed specifications for the requirements of the infrastructure.

5.5 Impact

Given the recently expressed concerns that the current IR evaluation experimental protocol leads to difficulty in comparing systems and elucidating reasons for their performance [51] and the apparent lack of improvement in ad-hoc IR systems over the previous decade [13], it is clearly time to develop improvements in the way that IR evaluation is done. At the basic level, introducing the guidelines for publishing IR experimental results will facilitate the use of a systematic review approach, which has been successfully used to obtain stronger conclusions from many experimental studies in other areas with a strong empirical tradition.

At the next level, implementation of the framework proposed has the potential to lead to a new way of working in IR by introducing concepts from e-Science into IR evaluation. Publications will contain full specifications of the system components and data used, implying that experiments will be easily reproducible and easily comparable to state-of-the-art results [5]. It will this be immediately clear how the results fit into the state-of-the-art, and where the innovation lies. This could open the possibility of doing reviewing in a more efficient and effective way.

6 Ongoing Evaluation

6.1 Motivation

The traditional IR evaluation activities have a static character, referring to fixed data and a relatively small set of questions to be answered (i.e. information needs). Once the evaluation activity has been carried-out the analysis of the results is crucial to understand the behaviour of complex systems. Unfortunately, this is an especially challenging and time-consuming activity, requiring vast amounts of human effort to inspect query-by-query the output of a system in order to understand what went well or bad [12].

Although time is a key factor, IR evaluation and the consequent analysis of the results are demanding also from the resources point-of-view. Indeed, evaluation procedures are required to handle huge amount of input data and in return they output other data that need to be efficiently managed, preserved, accessed and re-used. Furthermore, there are plenty of retrieval algorithms, paradigms, services and applications to be tested and compared under the lens of an increasingly higher number of evaluation metrics.

In this context a flexible and effective environment is needed to manage big amount of data, compose and test different services and applications, deal with heterogeneous sources of data and systems and support time-consuming and resource-demanding evaluation activities. To this end, experimental evaluation could take advantage by moving the evaluation work-flow into the *cloud* [34] where it can exploit loosely-coupled parallel applications and leverage on abstraction for work-flow description to obtain "ease of use, scalability, and portability" [43].

However, the time factor, even by moving the evaluation work-flow into the cloud is not eliminated because researchers still have to wait for experiments to finish until they can evaluate their results, without having any intermediate hint on the "direction the results are taking" which is useful to analyze and, if necessary, to revise things on the way.

As a consequence, it is mandatory to address at least three basic issues: (i) to provide the user with a partial result (i.e., visualizations and metrics) as soon as a minimal amount of data has been processed; (ii) to provide the user with an estimation of the approximation (e.g., the confidence interval) of the current results and a suitable visualization of that; and (iii) to define formal Visual Analytics solutions able to detect significant changes in the visualizations, in order to raise the user attention only when the visualization is significantly changed.

Dealing with data stream visualizations is a complex activity, since it requires the knowledge of aspects closely related to data streaming details, like bandwidth, transfer rate, sampling, memory management and so on, and the knowledge of aspect related to visualization of big amount of data (e.g., pixel oriented techniques). The complexity of data stream analysis leads to deal with the change detection from different point of view. An interested result is reported in [57], in which authors propose a framework to visualize data streams with the goal to show significant pattern changes to users.

6.2 Research Questions and Challenges

The main goal of on-going evaluation is to increase the efficiency of an evaluation activity in terms of user effort when setting up and performing IR experiments. Our main concern is to speed-up the execution of the experiments by cutting-down the waiting time, and the analysis of the results by providing new analysis methodologies and software tools.

We envision a methodology which allows for continuous evaluation [19] and assessment of retrieval results also during the retrieval process, in a *trial—take decision—trial* cyclic setting. This cyclic setting presents three main challenges.

The first challenge concerns **"Output inspection".** There are several types of output that can be generated in an IR experiment. The basic one is a list of answers to a question. Another type of output could be a set of numerical values that expresses attributes of the (components of the) retrieval system. In an on-going evaluation experiment the researcher should be able to to take two main actions: (i) to indicate the output that should be monitored during the execution of an IR experiment; and, (ii) to specify which parts or components of the IR experiment should be included in the monitoring phase and how they could be combined.

To this end, it is necessary to provide a flexible and customizable evaluation system which allows also for taking account of the semantic descriptions of the components and their relationships.

The second challenge is **"Process interruption".** We need to know how and when to stop the *trial* process and give the researcher the chance to *take a decision*. Furthermore, it is necessary to stop (or to pause) the process when the output data is accurate enough for the experiment purposes. If the process is stopped too early the data could be insufficient to drive any (even partial) conclusion or it could lead to a wrong interpretation of the results.

The third challenge is **"Error estimation".** Upon process interruption, the monitored outputs are only a part of the whole result set that should be obtained at the end of the experiment.

It is necessary to estimate the error range of the considered outputs at the interruption point when compared to the complete output. The definition of error thresholds is required to let researchers estimate the accuracy of their conclusion at the various levels of the on-going process.

6.3 Competencies and Cross-Disciplinary Aspects

Addressing the challenges described above requires a set of skills difficult to find within one group, and even less within one person. The basic requirement is to have a good understanding of IR evaluation practices provides the necessary design know-how for envisioning an evolution of the traditional Cranfield paradigm to be adopted in a highly dynamic environment. A deep knowledge of statistics is also highly recommendable in order to establish the minimal amount of data required to get some valid conclusions from the experiments. Creating good visualizations is important when conveying the essence of information. The inner dynamic of on-going evaluation need to be captured by the proposed visualizations that need to advance the current state-of-the-art in the visual analytics field. Lastly, the cloud and *Service-oriented Architectures (SoA)* environments require for a deep understanding of software design. The definition of new services and applications out of re-used components requires for advanced knowledge in software engineering.

6.4 Roadmap

A possible roadmap to address the above mentioned research challenges is to define or select an evaluation task well-suited for on-going evaluation purposes. This task should involve a big experimental collection and require a wide-spectrum of experimental analyses. A second step is to analyze the available cloud and SoA environments and verify how they allow for service and application composition and how they manage the produced data and provenance information. A consequence of this is to design and develop a software system allowing for stopping or pausing the evaluation process at intermediate steps. The last challenging step is to perform interactive and visual analysis in which visual analytics provides innovative solutions to infer meaning from the data produced in the evaluation process.

6.5 Impact

On-going evaluation is destined to change and evolve the current experimental evaluation practice. The dynamics considered by this new paradigm could influence not only the way in which results are produced and analyzed, but also how the experimental collection is built. Indeed, we can envision dynamic changes in relevance judgments set, thus producing a compound result set. As a consequence metrics and statistical analyses on the results will need to be revised and adapted to handle these changes.

On-going evaluation could allow for new, highly complex and resource demanding result analyses. Furthermore, it will change the way in which the researcher approaches experimental evaluation by transforming it in a continuous and highly dynamic process.

7 Signal Aware Evaluation

7.1 Motivation

Users and systems do not live in a vacuum. They interact in a complex environment. They are exposed to many kinds of signal and continuously emit signals themselves. When interacting with digital content, users generate huge amounts of *signals* that represent the concrete traces or their activity and actions on digital assets within information access systems. These user signals can be either implicit or explicit: examples of the former type are page views, clicks, purchases, dwell time, bookmarks; examples of the latter types are searches, annotations, "likes," tags, and different kinds of user generated content. All these user signals surround and enhance the digital contents and they come in volumes and growth rates that are often much bigger than the volumes and growth rates of the digital contents themselves.

State-of-the-art information access systems are designed to record and exploit many of these user signals, for example by means of Web access logs, query logs, clickstream logs, update streams, and they often foster and support the creation of explicit ones, such as allowing users to annotate or tag their assets. A lot of research is being carried out to understand how to exploit these user signals for many different purposes such as log analysis, creation of user profiles, implicit or explicit relevance feedback to improve search, or recommendation. However, these efforts are far from being effectively combined and jointly integrated in real systems with near real-time reaction capabilities to interpret incoming streams of user signals in order to improve the interaction and the experience of the users with the digital material.

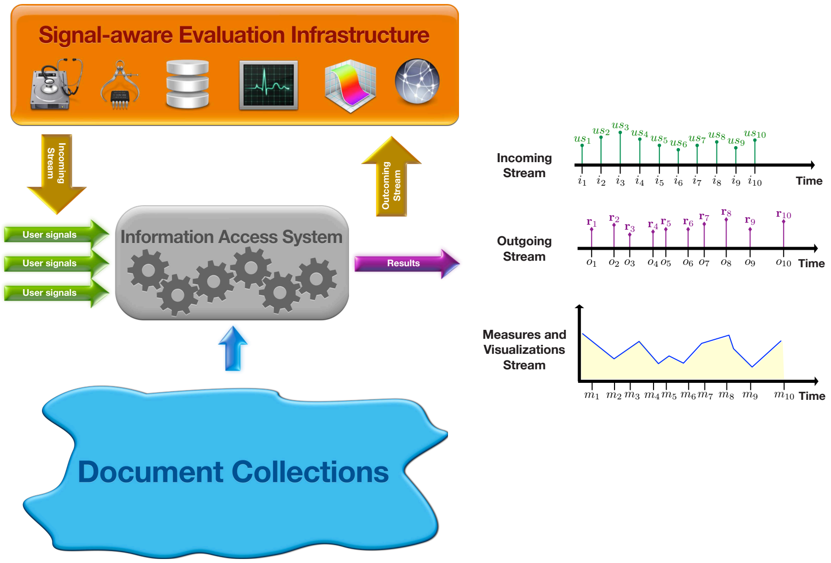
Tackling this challenge raises many issues. What happens when it comes to reacting in real-time to these user signals and to actively exploiting them in relation to the tasks the systems support? Can we use streams of clicks for entity linking or for enriching the collections by means of automatically generated annotations? Can we reliably interpret noisy page views and dwell time together with streams of clicks to understand user interests, visualize trends in their behaviour and adapt the system response to them? What happens when one user tags digital material? Should these tags be automatically linked to other cultural material providing alternative browsing paths for the other users? Should these tags be translated into several target languages and properly linked to the other information resources? And, what is the latency of the system for reacting and adapting to these user signals? How long should it take before users can experience variations in the responses of the information access systems due to the effect of user signals?

7.2 Research Questions and Challenges

There is a need to build innovative benchmarking frameworks where a target information access system is evaluated, with respect to target collections of its reference domain, against a stream of incoming user signals and produces a stream of outgoing responses which will then be measured and analysed. The benchmarking framework will manage incoming streams fed to the information access system which will have to react in near real time to them and produce responses accordingly, i.e., outgoing streams. These system responses will then trigger further streams of performance measures, analyses, and visualizations.

In order to realize such a benchmarking framework, we need to transform the current evaluation paradigm at a methodological level, extending it to model, represent, manipulate, combine, process, analyze, and interpret different noisy streams of signals, in a theoretically transparent manner.

Consider Figure 3: on the left, one can see the conceptual architecture of the framework while, on the right, one can see an example of an incoming stream where each incoming event represents a user signal of the kind discussed above. The stream of incoming events produces a stream of outgoing events that corresponds to a vector of results and responses produced by the system. The stream of outgoing events, in turn, originates a stream of measure and visualization events to which several effectiveness and efficiency measures are associated.



Signal-aware Evaluation Infrastructure
Incoming Stream
Outcoming Stream
User signals
User signals
User signals
Information Access System
Results
Document Collections
Incoming Stream
us_1 us_2 us_3 us_4 us_5 us_6 us_7 us_8 us_9 us_{10}
i_1 i_2 i_3 i_4 i_5 i_6 i_7 i_8 i_9 i_{10}
Time
Outgoing Stream
$\mathbf{r}_1$ $\mathbf{r}_2$ $\mathbf{r}_3$ $\mathbf{r}_4$ $\mathbf{r}_5$ $\mathbf{r}_6$ $\mathbf{r}_7$ $\mathbf{r}_8$ $\mathbf{r}_9$ $\mathbf{r}_{10}$
o_1 o_2 o_3 o_4 o_5 o_6 o_7 o_8 o_9 o_{10}
Time
Measures and Visualizations Stream
m_1 m_2 m_3 m_4 m_5 m_6 m_7 m_8 m_9 m_{10}
Time

Figure 3: Signal-aware evaluation of real-time information access systems.

7.3 Competencies and Cross-Disciplinary Aspects

Several competencies are needed to envision this shift in the evaluation paradigms: information retrieval; stochastic processes and signal analysis; online learning; measurement theory; system architectures and data management; human behavior and communication; real-time systems.

7.4 Roadmap

A possible roadmap to address the research challenges mentioned above is:

- to extend the current evaluation methodology by providing the concepts and the formal tools to be able to represent, deal with, and reliably interpret noisy signals;
- to build a scalable streaming evaluation framework: a ground-breaking methodology for carrying out streaming evaluation partnered with a streaming evaluation infrastructure able to automatically manage incoming and outgoing streams, store, preserve and make accessible the produced experimental data, compute performance measures, and conduct analyses;
- to perform interactive streaming analysis: a visual analytics environment where innovative and intuitive visual techniques, specifically targeted for addressing the continuous stream of generate data, will allow researchers and developers to interactively analyse the experimental results;

- to rely on the long-standing IR tradition and test and experiment the newly proposed ideas in the context of open, public, and large-scale evaluation initiatives where participants from academia and industry will have the possibility of performing experimentation with their systems and solutions in order to compare them and to improve them over the time.

7.5 Impact

Signal-aware evaluation represents a ground-breaking departure from traditional benchmarking and evaluation methodologies, which are common to all the evaluation campaigns and are based on the Cranfield paradigm [26] dating back to early 60s of last century. This evaluation paradigm has been originally designed to be carried out by hand and by using still snapshots of collections, topics, and systems so that the evaluation tasks are operated in batches [38].

Nevertheless, if we wish to really promote and push for the development of next-generation information access systems able to react in real-time to incoming streams of user signals, also the evaluation methodologies needed to support this advancement need to go real-time. They need to be able to cope with incoming streams of user signals that produce outgoing streams of system responses which have to measured and assessed originating streams of performance measures, indicators, analyses, and visualizations. Importantly, in this streaming setting everything becomes subject to change: collections, information needs, relevance judgments, user satisfaction, . . .

This represent a radical innovation with respect to the traditional evaluation paradigm because we will move from a still snapshots-based approach to a real-time evaluation framework.

8 Conclusions

Measuring is a key to scientific progress. This is particularly true for research concerning complex systems, whether natural or human-built. Multilingual and multimedia information systems are increasingly complex: they need to satisfy diverse user needs and support challenging tasks. Their development calls for proper evaluation methodologies to ensure that they meet the expected user requirements and provide the desired effectiveness.

Information access and retrieval is a discipline strongly rooted in experimentation, dating back the fundamental Cranfield paradigm in the mid of the previous century. Since then, large-scale worldwide experimental evaluations provided fundamental contributions to the advancement of state-of-the-art techniques through common evaluation procedures, regular and systematic evaluation cycles, comparison and benchmarking of the adopted approaches, and spreading of knowledge. In the process, vast amounts of experimental data are generated that beg for analysis tools to enable interpretation and thereby facilitate scientific and technological progress.

Nevertheless the discussions, the enthusiasm, and the ideas that emerged during the PROMISE retreat show that there is still a long way ahead of us for improving and advancing the experimental evaluation of information system in multiple languages and multiple media under several aspects.

PROMISE will do its best to disseminate and transfer this ideas to the research community, trying to stimulate take-up and investigation by junior and senior researchers, as well as to raise awareness about the need for an appropriate funding strategy to support the research community in this endeavor.

References

[1] M. Agosti, R. Berendsen, T. Bogers, M. Braschler, P. Buitelaar, K. Choukri, G. M. Di Nunzio, N. Ferro, P. Forner, A. Hanbury, K. Friberg Heppin, P. Hansen, A. Järvelin, B. Larsen, M. Lupu, I. Masiero, H. Müller, S. Peruzzo, V. Petras, F. Piroi, M. de Rijke, G. Santucci, G. Silvello, and E. Toms. *PROMISE Retreat Report – Prospects and Opportunities for Information Access Evaluation.* PROMISE network of excellence, ISBN 978-88-6321-039-2, `http://www.promise-noe.eu/promise-retreat-report-2012/`, September 2012.

[2] M. Agosti, M. Braschler, E. Di Buccio, M. Dussin, N. Ferro, G. L. Granato, I. Masiero, E. Pianta, G. Santucci, G. Silvello, and G. Tino. Deliverable D3.2 – Specification of the evaluation infrastructure based on user requirements. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. `http://www.promise-noe.eu/documents/10156/fdf43394-0997-4638-9f99-38b2e9c63802`, August 2011.

[3] M. Agosti, E. Di Buccio, N. Ferro, I. Masiero, M. Nicchio, S. Peruzzo, and G. Silvello. Deliverable D3.3 – Prototype of the Evaluation Infrastructure. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. `http://www.promise-noe.eu/documents/10156/3783730a-bce3-481b-83df-48e209c6286a`, September 2012.

[4] M. Agosti, E. Di Buccio, N. Ferro, I. Masiero, S. Peruzzo, and G. Silvello. DIRECTions: Design and Specication of an IR Evaluation Infrastructure. In Catarci et al. [24].

[5] M. Agosti, G. M. Di Nunzio, and N. Ferro. A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns. In T. Sakay, M. Sanderson, and D. K. Evans, editors, *Proc. 1st International Workshop on Evaluating Information Access (EVIA 2007)*, pages 62–73. National Institute of Informatics, Tokyo, Japan, 2007.

[6] M. Agosti, G. M. Di Nunzio, and N. Ferro. The Importance of Scientific Data Curation for Evaluation Campaigns. In C. Thanos, F. Borri, and L. Candela, editors, *Digital Libraries: Research and Development. First International DELOS Conference. Revised Selected Papers*, pages 157–166. Lecture Notes in Computer Science (LNCS) 4877, Springer, Heidelberg, Germany, 2007.

[7] M. Agosti and N. Ferro. Towards an Evaluation Infrastructure for DL Performance Evaluation. In G. Tsakonas and C. Papatheodorou, editors, *Evaluation of Digital Libraries: An insight into useful applications and methods*, pages 93–120. Chandos Publishing, Oxford, UK, 2009.

[8] M. Agosti, N. Ferro, C. Peters, M. de Rijke, and A. Smeaton, editors. *Multilingual and Multimodal Information Access Evaluation. Proceedings of the International Conference of the Cross-Language Evaluation Forum (CLEF 2010).* Lecture Notes in Computer Science (LNCS) 6360, Springer, Heidelberg, Germany, 2010.

[9] M. Agosti, N. Ferro, and C. Thanos. DESIRE 2011 Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation. *SIGIR Forum*, 46(1):51–55, June 2012.

[10] J. Allan, J. Aslam, L. Azzopardi, N. Belkin, P. Borlund, P. Bruza, J. Callan, C. Carman, M. Clarke, N. Craswell, W. B. Croft, J. S. Culpepper, F. Diaz, S. Dumais, N. Ferro, S. Geva, J. Gonzalo, D. Hawking, K. Järvelin, G. Jones, R. Jones, J. Kamps, N. Kando, E. Kanoulos, J. Karlgren, D. Kelly, M. Lease, J. Lin, S. Mizzaro, A. Moffat, V. Murdock, D. W. Oard, M. de Rijke, T. Sakai, M. Sanderson, F. Scholer, L. Si, J. Thom, P. Thomas, A. Trotman, A. Turpin, A. P. de Vries,

W. Webber, X. Zhang, and Y. Zhang. Frontiers, Challenges, and Opportunities for Information Retrieval – Report from SWIRL 2012, The Second Strategic Workshop on Information Retrieval in Lorne, February 2012. *SIGIR Forum*, 46(1):2–32, June 2012.

[11] M. Angelini, N. Ferro, G. L. Granato, and G. Santucci. Deliverable D5.3 – Collaborative User Interface Prototype with Annotation Functionalities. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. `http://www.promise-noe.eu/documents/10156/8c475e6c-36b5-4822-9fbc-d7d116b3a897`, September 2012.

[12] M. Angelini, N. Ferro, G. Santucci, and G. Silvello. Visual Interactive Failure Analysis: Supporting Users in Information Retrieval Evaluation. In J. Kamps, W. Kraaij, and N. Fuhr, editors, *Proc. 4th Symposium on Information Interaction in Context (IIiX 2012)*, pages 195 – 203. ACM Press, New York, USA, 2012.

[13] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin, editors, *Proc. 18th International Conference on Information and Knowledge Management (CIKM 2009)*, pages 601–610. ACM Press, New York, USA, 2009.

[14] N. Asadi, D. Metzler, T. Elsayed, and J. Lin. Pseudo test collections for learning web search ranking functions. In W.-Y. Ma, J.-Y. Nie, R. Baeza-Yaetes, T.-S. Chua, and W. B. Croft, editors, *Proc. 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*, pages 1073–1082. ACM, ACM Press, New York, USA, 2011.

[15] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors, *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, pages 455–462. ACM Press, New York, USA, 2007.

[16] S.M. Beitzel, E.C. Jensen, A. Chowdhury, and D. Grossman. Using titles and category names from editor-driven taxonomies for automatic evaluation. In D. Kraft, O. Frieder, J. Hammer, S. Qureshi, and L. Seligman, editors, *Proc. 12th International Conference on Information and Knowledge Management (CIKM 2003)*, pages 17–23. ACM Press, New York, USA, 2003.

[17] R. Berendsen, M. Braschler, M. Gäde, M. Kleineberg, M. Lupu, V. Petras, and S. Reitberger. Deliverable D4.3 – Final Report on Alternative Evaluation Methodology. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. `http://www.promise-noe.eu/documents/10156/0092298d-892b-45c0-a534-b9a3d0c717b1`, September 2012.

[18] R. Berendsen, E. Tsagkias, M. de Rijke, and E. Meij. Generating pseudo test collections for learning to rank scientific articles. In Catarci et al. [24].

[19] M. Braschler, K. Choukri, N. Ferro, A. Hanbury, J. Karlgren, H. Müller, V. Petras, E. Pianta, M. de Rijke, and G. Santucci. A PROMISE for Experimental Evaluation. In Agosti et al. [8], pages 140–144.

[20] M. Braschler, D. K. Harman, and E. Pianta, editors. *CLEF 2010 Labs and Workshops, Notebook Papers*. MINT srl, Trento, Italy. ISBN 978-88-904810-0-0., 2010.

[21] M. Braschler, S. Reitberger, M. Imhof, A. Järvelin, P. Hansen, M. Lupu, M. Gäde, R. Berendsen, and A. Garcia Seco de Herrera. Deliverable D2.3 – Best Practices Report. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. http://www.promise-noe.eu/documents/10156/086010bb-0d3f-46ef-946f-f0bbeef305e8, August 2012.

[22] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, 80:571–583, 2007.

[23] V. R. Carvalho, M. Lease, and E. Yilmaz. Crowdsourcing for search evaluation. *SIGIR Forum*, 44(2):17–22, 2011.

[24] T. Catarci, P. Forner, D. Hiemstra, A. Peñas, and G. Santucci, editors. *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. Proceedings of the Third International Conference of the CLEF Initiative (CLEF 2012).* Lecture Notes in Computer Science (LNCS) 7488, Springer, Heidelberg, Germany, 2012.

[25] C. W. Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, Aslib Cranfield Research Project, 1962.

[26] C. W. Cleverdon. The Cranfield Tests on Index Languages Devices. In K. Spärck Jones and P. Willett, editors, *Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA, 1997.

[27] M. Croce, E. Di Buccio, E. Di Reto, M. Dussin, N. Ferro, G. L. Granato, P. Hansen, M. Lupu, M. Perlorca, A. Pronesti, A. Sabetta, G. Santucci, G. Silvello, G. Tino, and T. Tsikrika. Deliverable D5.2 – User interface and Visual analytics environment requirements. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. http://www.promise-noe.eu/documents/10156/21f1512a-5b47-48ae-834a-89d6441d079e, August 2011.

[28] E. Deelman, D. Gannon, M. Shields, and I. Taylor. Workflows and e-science: An overview of workflow system features and capabilities. *Future Generation Computer Systems*, 25(5):528–540, 2009.

[29] N. Ferro. DIRECT: the First Prototype of the PROMISE Evaluation Infrastructure for Information Retrieval Experimental Evaluation. *ERCIM News*, 86:54–55, July 2011.

[30] N. Ferro, A. Hanbury, H. Müller, and G. Santucci. Harnessing the Scientific Data Produced by the Experimental Evaluation of Search Engines and Information Access Systems. *Procedia Computer Science*, 4:740–749, 2011.

[31] A. Foncubierta Rodríguez and H. Müller. Ground truth generation in medical imaging,a crowdsourcing-based iterative approach. In W. T. Chu, M. Larson, W. T. Ooi, and K.-T. Chen, editors, *Proc. International ACM Workshop on Crowdsourcing for Multimedia (CrowdMM 2012)*, 2012.

[32] P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, and M. de Rijke, editors. *Multilingual and Multimodal Information Access Evaluation. Proceedings of the Second International Conference of the Cross-Language Evaluation Forum (CLEF 2011).* Lecture Notes in Computer Science (LNCS) 6941, Springer, Heidelberg, Germany, 2011.

[33] P. Forner, J. Karlgren, and C. Womser-Hacker, editors. *CLEF 2012 Labs and Workshops, Notebook Papers.* MINT srl, Trento, Italy. ISBN 978-88-904810-1-7., 2012.

[34] A. Hanbury, H. Müller, G. Langs, M. A. Weber, B. H. Menze, and T. S. Fernandez. Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis. In Catarci et al. [24].

[35] Allan Hanbury and Henning Müller. Automated component-level evaluation: Present and future. In Agosti et al. [8], pages 124–135.

[36] P. Hansen, G. L. Granato, and G. Santucci. Collecting and Assessing Collaborative Requirements. In C. Shah, P. Hansen, and R. Capra, editors, *Proc. Workshop on Collaborative Information Seeking: Briding the Gap between Theory and Practice (CIS 2011)*, 2011.

[37] P. Hansen and A. Järvelin. Collaborative Information Retrieval in an Information-intensive Domain. *Information Processing & Management*, 41(5):1101–1119, September 2005.

[38] D. K. Harman. *Information Retrieval Evaluation.* Morgan & Claypool Publishers, USA, 2011.

[39] D. K. Harman and E. M. Voorhees, editors. *TREC. Experiment and Evaluation in Information Retrieval.* MIT Press, Cambridge (MA), USA, 2005.

[40] B. Hefley and W. Murphy, editors. *Service Science, Management, and Engineering: Education for the 21st Century.* Springer, Heidelberg, Germany, 2008.

[41] B. Huurnink, K. Hofmann, M. de Rijke, and M. Bron. Validating query simulators: An experiment using commercial searches and purchases. In Agosti et al. [8], pages 40–51.

[42] A. Järvelin, G. Eriksson, P. Hansen, T. Tsikrika, A. Garcia Seco de Herrera, M. Lupu, M. Gäde, V. Petras, S. Rietberger, M. Braschler, and R. Berendsen. Deliverable D2.2 – Revised Specification of Evaluation Tasks. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. `http://www.promise-noe.eu/documents/10156/a0d664fe-16e4-4df6-bcf9-1dc3e5e8c18e`, February 2012.

[43] G. Juve and E. Deelman. Scientific Workflows and Clouds. *ACM Crossroads*, 16(3):14–18, 2010.

[44] Y. Kano, P. Dobson, M. Nakanishi, J. Tsujii, and S. Ananiadou. Text mining meets workflow: linking u-compare with taverna. *Bioinformatics*, 26(19):2486–2487, 2010.

[45] D. Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval (FnTIR)*, 3(1-2), 2009.

[46] S. Kumpulainen and K. Järvelin. Information Interaction in Molecular Medicine: Integrated Use of Multiple Channels. In N. J. Belkin and D. a Kelly, editors, *Proc. 3rd Symposium on Information Interaction in Context (IIiX 2010)*, pages 95–104. ACM Press, New York, USA, 2010.

[47] M. Lease and E. Yilmaz. Crowdsourcing for information retrieval. *SIGIR Forum*, 45(2):66–75, 2012.

[48] B. Mons, H. van Haagen, C. Chichester, P.-B. 't Hoen, J. T. den Dunnen, G. van Ommen, E. van Mulligen, B. Singh, R. Hooft, M. Roos, J. Hammond, B. Kiesel, B. Giardine, J. Velterop, P. Groth, and E. Schultes. The value of data. *Nature Genetics*, 43:281–283, 2011.

[49] V. Petras, P. Forner, and P. Clough, editors. *CLEF 2011 Labs and Workshops, Notebook Papers.* MINT srl, Trento, Italy. ISBN 978-88-904810-1-7., 2011.

[50] S. Reitberger, M. Imhof, M. Braschler, R. Berendsen, A. Järvelin, P. Hansen, A. Garcia Seco de Herrera, T. Tsikrika, M. Lupu, V. Petras, M. Gäde, M. Kleineberg, and K. Choukri. Deliverable D4.2 – Tutorial on Evaluation in the Wild. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. http://www.promise-noe.eu/documents/10156/3f546a0b-be7c-48df-b228-924cc5e185cb, August 2012.

[51] S. E. Robertson. On the history of evaluation in IR. *Journal of Information Science*, 34(4):439–456, 2008.

[52] B. R. Rowe, D. W. Wood, A. L. Link, and D. A. Simoni. *Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program.* RTI Project Number 0211875, RTI International, USA. http://trec.nist.gov/pubs/2010.economic.impact.pdf, July 2010.

[53] M. Sanderson. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR)*, 4(4):247–375, 2010.

[54] J. Spohrer. Editorial Column—Welcome to Our Declaration of Interdependence. *Service Science*, 1(1):i–ii, 2009.

[55] C. V. Thornley, A. C. Johnson, A. F. Smeaton, and H. Lee. The Scholarly Impact of TRECVid (2003–2009). *Journal of the American Society for Information Science and Technology (JASIST)*, 62(4):613–627, April 2011.

[56] T. Tsikrika, A. Garcia Seco de Herrera, and H. Müller. Assessing the Scholarly Impact of ImageCLEF. In Forner et al. [32], pages 95–106.

[57] Z. Xie, M. O. Ward, and E. A. Rundensteiner. Visual exploration of stream pattern changes using a data-driven framework. In *Proceedings of the 6th international conference on Advances in visual computing - Volume Part II*, ISVC'10, pages 522–532, Berlin, Heidelberg, 2010. Springer-Verlag.