PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

PROMISE Retreat Report
Prospects and Opportunities for Information Access Evaluation

Editors

Nicola Ferro, University of Padua, Italy
Richard Berendsen, University of Amsterdam, The Netherlands
Allan Hanbury, Vienna University of Technology, Austria
Mihai Lupu, Vienna University of Technology, Austria
Vivien Petras, Humboldt University Berlin, Germany
Maarten de Rijke, University of Amsterdam, The Netherlands
Gianmaria Silvello, University of Padua, Italy

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

PROMISE
Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

SEVENTH FRAMEWORK PROGRAMME

PROMISE Retreat Report
Prospects and Opportunities for Information Access Evaluation

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

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. 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. The event has been organized as a "retreat" allowing researchers to work back to back and propose hot topics where to focus research in the field in the coming years. 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.

Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement n. 258191

PROMISE
Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

Volume Editors

Nicola Ferro, University of Padua, Italy
ferro@dei.unipd.it

Richard Berendsen, University of Amsterdam, The Netherlands
r.w.berendsen@uva.nl

Allan Hanbury, Vienna University of Technology, Austria
hanbury@ifs.tuwien.ac.at

Mihai Lupu, Vienna University of Technology, Austria
lupu@ifs.tuwien.ac.at

Vivien Petras, Humboldt University Berlin, Germany
vivien.petras@ibi.hu-berlin.de

Maarten de Rijke, University of Amsterdam, The Netherlands
derijke@uva.nl

Gianmaria Silvello, University of Padua, Italy
silvello@dei.unipd.it

ISBN 978-88-6321-039-2

Printed on September 2012
http://www.promise-noe.eu/

Contents

1 Introduction **7**
1.1 Format of the Retreat . . . 8
1.2 Summary of the Retreat Outcomes . . . 9
1.3 Organization of the Report . . . 11

2 Search Applications **13**
2.1 Motivation . . . 13
2.2 Research Questions and Challenges . . . 14
2.2.1 Who is the “Consumer”? . . . 14
2.2.2 Abstracting from the user . . . 14
2.2.3 Paradigm expansion . . . 15
2.2.4 Domain instantiation . . . 15
2.2.5 Other considerations . . . 15
2.3 Competencies and Cross-Disciplinary Aspects . . . 16
2.4 Roadmap . . . 16
2.5 Impact . . . 17

3 Contextual Evaluation **19**
3.1 Motivation . . . 19
3.2 Research Questions and Challenges . . . 19
3.2.1 User vs. Task . . . 19
3.2.2 Complex Tasks and Usage Scenarios . . . 20
3.2.3 Individual vs. Collaborative Information Retrieval . . . 20
3.2.4 Query-free Retrieval . . . 21
3.2.5 Pre- vs. Post- Retrieval Stages . . . 21
3.3 Competencies and Cross-Disciplinary Aspects . . . 22
3.4 Roadmap . . . 22
3.5 Impact . . . 23

4 Back on TREC: Challenges in Test Collection Design and Exploitation **25**
4.1 Motivation . . . 25
4.2 Research Questions and Challenges . . . 26
4.2.1 Crowdsourcing . . . 26
4.2.2 Generating pseudo test collections . . . 26
4.2.3 Semi-automatic creation of test collections . . . 27
4.2.4 A use case framework for information access applications . . . 27
4.2.5 Impact of CLEF in the research community and industry . . . 28
4.3 Competencies and Cross-Disciplinary Aspects . . . 28
4.4 Roadmap . . . 28
4.5 Impact . . . 29

5 Component-based Evaluation **31**
5.1 Motivation . . . 31
5.2 Research Questions and Challenges . . . 33
5.3 Competencies and Cross-Disciplinary Aspects . . . 35
5.4 Roadmap . . . 35
5.5 Impact . . . 35

6 Ongoing Evaluation **37**
6.1 Motivation . . . 37
6.2 Research Questions and Challenges . . . 39
6.3 Competencies and Cross-Disciplinary Aspects . . . 40
6.4 Roadmap . . . 41
6.5 Impact . . . 41

7 Signal Aware Evaluation **43**
7.1 Motivation . . . 43
7.2 Research Questions and Challenges . . . 44
7.3 Competencies and Cross-Disciplinary Aspects . . . 44
7.4 Roadmap . . . 45
7.5 Impact . . . 45

8 Conclusions **47**

References **49**

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.

We consider experimental evaluation – both laboratory and interactive – a key means for supporting and fostering the development of multilingual and multimedia information systems which are more adherent to the new user needs in order to ensure that they meet the expected user requirements, provide the desired effectiveness and efficiency, guarantee the required robustness and reliability, and operate with the necessary scalability.

In order to achieve its goals, 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 [Agosti et al., 2010, Forner et al., 2011, Catarci et al., 2012, Braschler et al., 2010b, Petras et al., 2011, Forner et al., 2012], 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 [Agosti and Ferro, 2009, Agosti et al., 2012a, Ferro, 2011, Agosti et al., 2011, Agosti et al., 2012b, Angelini et al., 2012a], and dissemination and knowledge transfer, e.g. formulation of best practices, 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] [Allan et al., 2012], 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

[1] http://www.promise-noe.eu/
[2] http://www.clef-initiative.eu/
[3] http://direct.dei.unipd.it/
[4] http://www.cs.rmit.edu.au/swirl12/

Figure 1: Group photo of the participants at the PROMISE retreat.

its boundaries and possible links with other disciplines outside computer science, like psychology or neuro-sciences.

1.1 Format of the Retreat

The PROMISE retreat has been hosted in the *Academic Senate* room of the *Palazzo del Bo*, the main historical building of University of Padua, which was originally occupied by an inn, at the “sign of the Ox" (which means *Bo*), given to a butcher by Francesco da Carrara, Lord of Padua, in repayment for the meat supplied during the 1405 siege of the city. This venue provided a really pleasant setting which favored discussion and concentration, abstracting people away from their day-to-day business, making the brainstorming workshop a real “retreat" and not a project meeting. 25 researchers from 10 different European countries attended the event.

As far as the format of the retreat is concerned, we took an approach very similar to the one adopted in the recent SWIRL 2012 workshop, which turned out to be quite effective in raising ideas and coming to concrete outcomes.

Prior to the event, all the participants have been asked to think about possible interesting challenges, to summarize them in a couple of slides with two or three significant bibliographic references

Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement n. 258191

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 – information retrieval, information extraction, natural language processing, human-computer inter- action, semantic technologies, information visualization and visual analytics, system architectures, and so on; (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 homeworks have been assigned to everybody in order to come to the present version of the report.

1.2 Summary of the Retreat Outcomes

Figure 2 shows the six main challenges that have been identified and deeply discussed during the PROMISE retreat.

Search Applications deals with moving a step forward from laboratory settings and being able to evaluate search functionalities of an information system in the wild and in real settings, both from the user/consumer point of view and from the organization/institution providing the system. Further details are provided in Section 2.

Contextual Evaluation concerns the integration of users, tasks, search applications and underlying information retrieval systems in a holistic perspective to ensure that the global impact of an

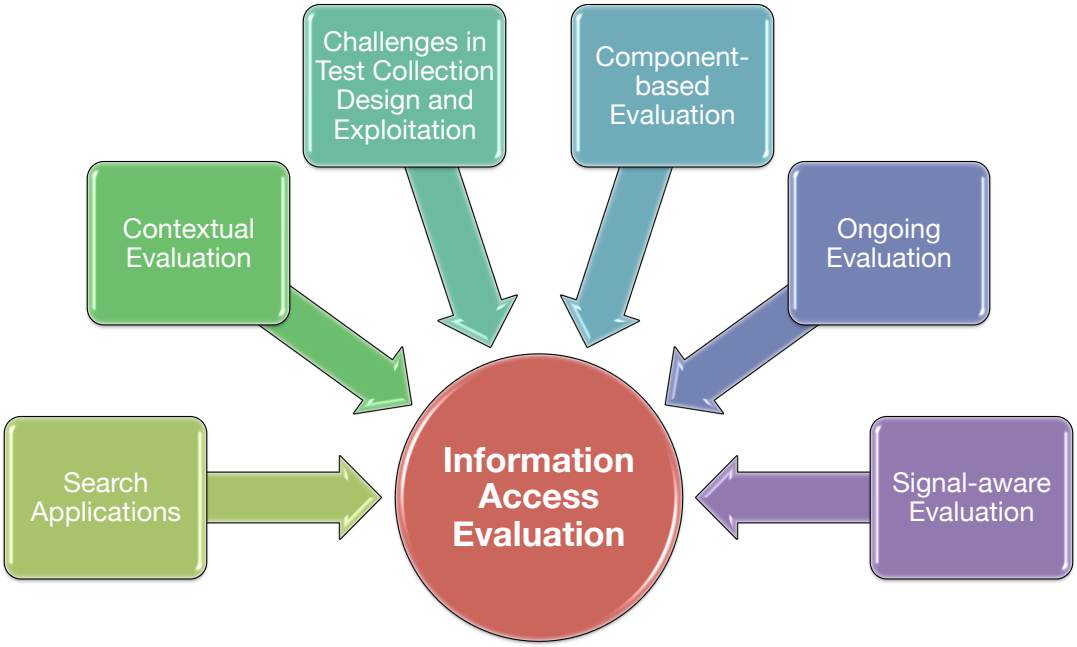

Figure 2: Six main challenges emerging from the PROMISE retreat.

information retrieval system on a user – in terms of work place productivity or quality of life – can be assessed. Further details are provided in Section 3.

Challenges in Test Collection Design and Exploitation regards how to make evaluation based on test collections step forward in terms of increased efficiency, capability of evaluating according to end user preferences, and impact in the research community and industry. Further details are provided in Section 4.

Component-based Evaluation examines how to best design and develop an infrastructure for conducting experimentation on a component basis rather than a whole system basis and how to exploit the experimental results in the information retrieval scientific production. Further details are provided in Section 5.

Ongoing Evaluation studies how to allow researchers to decide earlier on the course of their experiments and to develop an experimental design which allows the continuous evaluation and assessment of retrieval results during the retrieval process. Further details are provided in Section 6.

Signal-aware Evaluation stems from the consideration that users and systems are not in a vacuum but they interact in a complex environment. They are immersed in a kind of field where they are exposed to many kinds of signals and emit signals themselves as well – biological signals, changes

Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement n. 258191

in the datasets e.g. due to user-generated contents, system interaction signals like clicks. Evaluation has to be aware of these different types of signals, their interaction, and nature in order to correlate them, assess their effect, quantify it, and correctly interpret them, also with respect to traditional evaluation concepts, such as relevance. Further details are provided in Section 7.

1.3 Organization 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.

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

Research in Information Retrieval, as in any other field, 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.

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) [Reitberger et al., 2012].

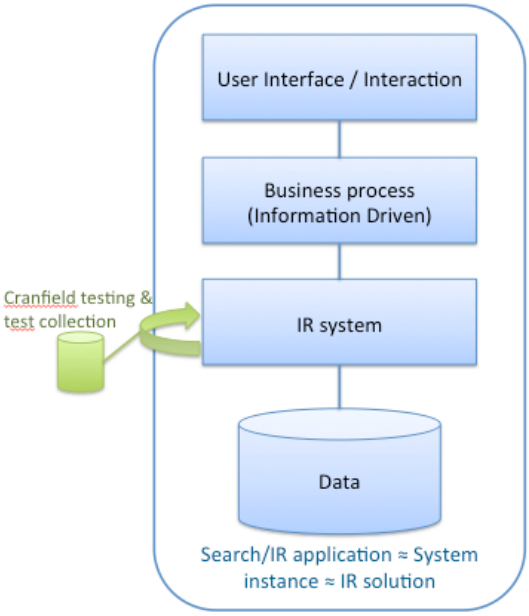

Figure 3: A common view of a search application.

Our understanding of search applications departs from a common view, as show in Figure 3, but can be seen from two perspectives:

From the users perspective A search application is an application that models a process that is information intensive or driven and is powered by an information retrieval technology.

From the components' perspective 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.

For this section, the predominant view is that of the user.

2.2 Research Questions and Challenges

In this section we propose a list of challenges which we believe have to be addressed in order to obtain the broader, scientific understanding of search applications.

2.2.1 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). In this sense, 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. The search application may indeed be enriched by the stakeholder and by the testing. 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 search application and even enriches it, creating value for both herself and for the application providers creates a link to Service Sciences [Spohrer, 2009]. There, research questions regarding the evolution of services through user involvement are studied [Hefley and Murphy, 2008] and serve as a potential source of inspiration for our current study.

2.2.2 Abstracting from the user

After having obtained a better understanding of the consumer and its role in the evaluation of the search application, the question is whether we can derive a *user* profile from user studies, or formulate one from functional requirements. To note here is the difference between the consumer of the evaluation, about whom we have discussed above, and the user of the system. While the consumer is, as mentioned above, a technical decision maker, the user may not necessarily be a technical person at all. Nonetheless, the consumer of the evaluation can only make informed decision if the evaluation properly models its users. Furthermore, many applications have to cater to different

user profiles and therefore another question here is whether we can find appropriate measures to evaluate a search application in such situations?

> *Example.* Two user profiles with different requirements must be balanced: is it better to have a great search application for one, which is really bad for the other, or one which is mediocre for both?

2.2.3 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, and challenge here 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.

2.2.4 Domain instantiation

In the previous paragraph we mentioned specialist fora in reference to groups of users of a search application in a particular domain. In fact, this is 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.

2.2.5 Other considerations

We could think of additional aspects to consider in the evaluation of a search application:

Correctness of the implementation It may sound obvious, but an important aspect is that the application conforms to explicit or implicit service contracts (e.g. boolean ‘A and B’ should return less than or equal to ‘A’ and less than or equal to ‘B’)

User perception has received a fair amount of attention for interactive systems [Kelly, 2009], and the question is how to expand it to this more general view of applications.

Continuous evaluation or constantly monitoring may be needed in some cases. We might argue that in fact any evaluation we can think of, can be used in a continuous way, but there are specific issues relating to the impact this evaluation has on the production system.

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:

IR Evaluation A good understanding of IR evaluation practices provides the necessary design know-how for new test collections and effectiveness measurements

Service Sciences As described above, Service Sciences study the interaction between different entities (generally viewed as a consumer and a provider) in order to generate value for both.

Human-Computer Interaction Experience in User Modelling would allow us to design the necessary tests to go beyond effectiveness evaluation

Software Engineering The interplay between different components of the search application is often quite complex, and incorporating customer feedback into it requires a deep understanding of this interplay and how changes affect the system as a whole.

Communication Science A communication professional is perhaps needed in order to make sure that when the three groups of people communicate (researchers, users, consumers), this is done in an efficient manner and without misunderstandings.

2.4 Roadmap

Based on the research challenges outlined above, we can say that there is a significantly large action space in this domain. 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.** Define what and how to measure through interaction with customers and users. Such measures 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. **Benchmarks.** A new type of benchmark is to be developed. The test collection is still an important component, but not necessarily the only tool for evaluation. Identifying a generic grid of tests and applying them in a scripted fashion is a first step. Subsequently, we need to identify ways to make these tests even more scalable and reliable.

3. **Infrastructure.** We need to identify a potential infrastructure for performing this kind of evaluation. Most existing ones are designed around the Cranfield paradigm.

4. **Analysis.** We always analyse the results, but are we always doing int from the perspective of the customer or user?

Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement n. 258191

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 in the previous section will 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.

Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement n. 258191

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. Even if the underlying system stays the same, changes in user intentions or offered tools have a profound influence on the perceived helpfulness or utility of the system and therefore need to be taken into account. Only if the users and their context are understood, then the system can be improved.

3.2 Research Questions and Challenges

Assessing the global impact of an IR system requires leaving the well-understood ad-hoc information retrieval scenario (*single query input* $\rightarrow$ *black-box information retrieval system* $\rightarrow$ *ranked document list*) behind. Instead, a high number of factors (context, task etc.) need to be added to the evaluation model. It is unclear how these varied factors can be acquired and merged with a Cranfield-related paradigm. What measures can be defined to aggregate these different factors? When can we know that an IR system is successful - with respect to a user, a task, a context or compared with another system? Five research challenges or questions seem of particular interest and will be discussed in the following.

3.2.1 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. With the exception of the contribution of relevance feedback, results have been mixed, with little gain in system performance. At the same time, the concept of task has emerged as a contender. 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. Yet IR systems are developed to support particular task environments from supporting the work of patent analysts, to helping people find consumer health information, and students finding evidence to support term papers. But which drives systems

success: knowing more about the user, or knowing more about the task that the user needs to complete? Do users searching for the same task search in approximately the same way and make similar decisions about relevant documents?

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. Without a full understanding of both, we may be needlessly modifying systems, or missing valuable variables that may influence search outcome. Which measures or metrics will enable us to differentiate between placing a priority on analyzing user behavior patterns, or customizing the design of search systems to support particular task environments needs to be determined. 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.

3.2.2 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 [Kumpulainen and Järvelin, 2010]. 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. However, in electronic settings we can now more easily collect behavioral data across systems - data that might aid in understanding complex tasks and their success criteria.

A large number of performance measures exist already that may be appropriate for search subtasks, but collecting sufficient amounts of data from complex tasks for understanding complex task solving and finding patterns across different persons and types of tasks is a different challenge. The actual task and its subtasks as well as which parts of the surrounding context to analyze need to be clearly defined and delimited. 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.

3.2.3 Individual vs. Collaborative Information Retrieval

It is commonly assumed that searching is an individual activity only. The majority of IR systems today do not facilitate collaborative information searching. Recent research has provided insight that people also collaborate during the search process, however, the concept is not yet fully understood. Collaborative information handling and information sharing can be found in both professional work settings and in everyday situations and contexts. Organizations are paying more and more attention to their own information searching and management practices in order to be more effective as well as to disseminate, share and collaborate using common resources. Most of the research on collaborative information searching has been focused on developing tools and systems for collaboration (computer science and HCI research). Efforts in developing theories, models and frameworks, including some empirical studies in order to investigate certain domains show the research interest.

One important issue would be to identify dimensions and motivations for collaborative searching. The main focus for IR evaluation would be to take the models and insights from collaborative information searching and to extract features for evaluation purposes.

For example, 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. For evaluation purposes, measures that may be used and investigated can be traditional IR measures such as precision, recall, coverage and usability measures such as ease of use, cognitive load. In what way can traditional measures be appropriate and when are they insufficient?

3.2.4 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. A successful example from the field of recommender systems are the location-aware recommendations for mobile phones that recommend interesting places to visit based on a profile of the user's interests, some rudimentary contextual information (location and time), and information about other users' preferences. However, more traditional task-centered information seeking behavior could also benefit from such query-free suggestions.

Imagine a digital writing assistant that aids the user in writing documents by recommending relevant literature based on the active working document as well as previously read and authored documents, and the behavior or other similar users. Such a marriage of the fields of IR and recommender systems could serve to more accurately and pro-actively support the different stages of information seeking.

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.

3.2.5 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.

In order to provide a holistic approach to evaluation, both pre-stage and post-stage context need to be taken into account in order to assess the true success factors of a system. Both user needs (pre retrieval) and user intentions or outcomes (post retrieval) determine the quality of the system. If IR evaluation takes all stages into account, the assessment will be more realistic and more closely correlated with true user satisfaction of the overall user-system interaction.

Different users have different contexts (or tasks or information needs on a more concrete level) and approach the system with different intentions. Different intentions require different evaluation requirements. For example, if a user is looking for a quick overview of the ten touristic highlights for a town, this poses different requirements in terms of desired document types, content and requires focusing the evaluation measures on precision-based characteristics. If the user has been offered ten touristic highlights, but finds that he cannot use them because they are all closed on the day he visits, then the pre-stage requirements need to be adjusted as well as the evaluation measure (both precision- and time-based).

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. For instance, improvements in a particular IR component may only result in user benefit if embedded at the right stages of task solving. Depending on the outcome of the research, it could lead to a new breed of IR system, particularly if the task factor becomes the critical element. The broader societal impact is more efficient information access in workplace settings. 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

In this chapter, we raise research questions relating to test collection based evaluation of information retrieval algorithms. These research questions shape our ongoing work in this area. The main forum for IR evaluation for the past decades has been *Text REtrieval Conference (TREC)*. The Cranfield style benchmarking it features has been an experimental platform where large and small, improvements could be accumulated. TREC has witnessed the rise of many of the state of the art retrieval methods that we take for granted today [Sanderson, 2010]. Still, there are recurring issues in test collection design and exploitation that deserve our ongoing attention.

Several sets of work in the PROMISE project touch on test collections, and they can be grouped under these three themes:

Increasing efficiency Test collection creation is a labour-intensive and therefore costly process. We study several ways of alleviating this [Berendsen et al., 2012a]:

- **Crowdsourcing** Crowdsourcing [Lease and Yilmaz, 2012] can be a means to create test collections more efficiently. We review issues in experimental design as well as economical and ethical issues.
- **Generating ground truth** Pseudo test collections for tuning, training and evaluating retrieval algorithms. We investigate ways to generate ground truth for different retrieval tasks.
- **Semi-automatic test collection creation** In addition to generating pseudo test collections, an important area of research focuses on semi-automatic processes for test collection creation, where the human is kept in the loop. Key issues here are the impact interface design has on the quality of annotations, and the quality of the suggestions of the automatic method feeding the interface.

Evaluating according to end user preferences In benchmarking style evaluation of retrieval algorithms, often a great deal of variance in end users and their contexts is ignored or abstracted away. We study ways of improving this situation. If outcomes of benchmarking experiments reflect end user preferences better, benchmarking will become more relevant for industrial search service providers.

- **Use case framework** We have developed a use case framework [Järvelin et al., 2012] which enables researchers to think about end users of information access applications in a structured way. For test collection generation this means that evaluation can be tailored to foreseen usage scenarios.
- **Crowdsourcing** In crowdsourcing, if workers could be selected such that their demographics match foreseen end users of a search engine, these workers would make good assessors. In our review of crowdsourcing communication with workers is one of the aspects we will pay attention to.

Impact in the research community and industry We analyse the impact the CLEF-campaign with its rich variety of multilingual test collections has on the research community and industry, as for example in [Thornley et al., 2011, Tsikrika et al., 2011]. In addition, we work to increase collaboration between academia and industry.

4.2 Research Questions and Challenges

In this section we highlight the main research questions, challenges and goals in our current efforts. The reader should take note how these questions relate to the three themes discussed in the previous section.

4.2.1 Crowdsourcing

Recent years have seen an increased interest in crowdsourcing as a way of obtaining annotated data fast, at a reduced cost [Carvalho et al., 2011, Lease and Yilmaz, 2012, Foncubierta Rodríguez and Müller, 2012]. 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. This translates in the following research questions:

What is the current state of the art in best practices for experimental design of crowdsourced evaluation experiments?
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. Note that hiring people based on desired demographics is a way of improving evaluation such that it correlates better with end user preferences. For a scientfic article search service, we would like to hire scientists, students, researchers and the like as assessors. This is because these people will be representative of the population that will use the search systems we evaluate.

How do currently available crowdsourcing platforms compare on a variety of economical and ethical aspects?
For example: 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?

4.2.2 Generating pseudo test collections

Another way of obtaining annotated data on the cheap is to generate pseudo test collections. These can be used for training and evaluation purposes. Examples of work on evaluation are [Beitzel et al., 2003, Azzopardi et al., 2007, Huurnink et al., 2010]. Training is the goal in [Asadi et al., 2011, Berendsen et al., 2012b]. Research questions in this area include:

Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement n. 258191

Does evaluating on generated pseudo judgments rank retrieval algorithms in similar way as evaluating on editorial judgments? This is a central question in work that aims to benchmark using generated ground truth.

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?
Tuning and training retrieval algorithms is can be done on a training set. Oftentimes, these are editorial topics with relevance judgments from the same collection as the editorial test topics. Because such training material is expensive to acquire, training sets are small. However, there is huge variation in topics, certainly for general purpose ad hoc search. If we can generate pseudo judgments of sufficient quality, we can potentially generate thousands of pseudo topics with pseudo judgments. The size of the training set is then expected to lead to better performance [Berendsen et al., 2012b].

What is the impact of bias in generated pseudo judgments toward a retrieval algorithm that we are evaluating or training?
To generate ground truth from a collection automatically, we often use methods that might also be used by retrieval algorithms. This leads to the following questions: Are such retrieval algorithms unfairly favoured in evaluation? Are they given too much weight in models that were learned on biased generated ground truth?

4.2.3 Semi-automatic creation of test collections

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?

4.2.4 A use case framework for information access applications

We haved developed a use case framework for explicitly describing use cases underlying evaluation tasks [Järvelin et al., 2012]. The framework allows for describing very different use cases, broadening the scope of the traditional ad hoc search evaluation. We plan to validate use cases 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?
Interviews with search service providers and end users of these search services should help us gain

insights here.

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?
Evaluation tasks specificy a task, a *search service*, which is to be fullfilled by the systems of task participants. We believe that a careful study of how this search service will be used—and by whom it will be used, and in what context—should lead to insights about what aspecst to take into account when benchmarking systems. This is why we approach organizers of evaluation tasks and ask them to share their thoughts on this.

4.2.5 Impact of CLEF in the research community and industry

The accumulated research in the CLEF 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 CLEF publications and beyond. A collaboration between NUIG and ZHAW is currently investigating this by use of the Saffron system.

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

CLEF 2012 will be a milestone event in our work with the use case framework. We intend to interview organizers of evaluation tasks, meet industry representatives, and analyse the output of these encounters. Independent from CLEF we are continuously approaching industry stakeholders to share our use case framework.

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. The current version of Saffron[5] analyzes the following research communities in the analytics field: Information Retrieval, Natural Language Processing, Semantic Web. The Information Retrieval instance of Saffron[6] analyzes this research community based on the proceedings of the conferences organized by CLEF, TREC, and *NII-NACSIS Test Collection for IR Systems (NTCIR)*. Other options we investigate are mining papers for the use of test collections.

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. This in turn would predict economic succes of a search service better, even though it must be recognized that retrieval quality is just one aspect contributing to the success of a search service.

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: the lack of editorial training data should not be an excuse to optimize algorithms on test data. Moreover, the quantity of generated ground truth can lead to an advantage over training on editorial data. Finally, editorial and generated ground truth may be combined into an even better training set.

What will be the impact of our impact analysis into the CLEF campaign? 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] http://saffron.deri.ie/
[6] http://saffron.deri.ie/ir

5 Component-based Evaluation

5.1 Motivation

Information Retrieval has a strong tradition in empirical evaluation, as exemplified by the many evaluation campaigns [Harman, 2011]. The majority of IR evaluation campaigns today are based on the TREC organisation model [Harman and Voorhees, 2005], which is based on the Cranfield paradigm [Cleverdon, 1962]. 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 [Hanbury and Müller, 2010], the widely-adopted TREC approach has a number of *disadvantages* [Robertson, 2008]. The most pertinent to component-based evaluation are:

1. fixed timelines and cyclic nature of events;
2. evaluation at system-level only;
3. difficulty in comparing systems and elucidating reasons for their performance.

The first disadvantage is the cyclic nature of events, with a fixed deadline for submitting runs and a period of time during which the runs are evaluated before the evaluation results are released. At the end of each cycle, the data, topics and relevance judgements are usually made available to permit further "offline" evaluation. However, evaluating a system on a large number of test datasets still involves much effort on the part of the experimenter. A solution that has been proposed is online evaluation of systems, as implemented in the EvaluatIR system[7] [Armstrong et al., 2009]. This system makes available testsets for download, and allows runs in the standard TREC run format to be uploaded (they are private when uploaded, but can be shared with other users). It maintains a database of past runs submitted to TREC, benchmarks of IR systems and uploaded runs that have been shared, and supports a number of methods for comparing runs. The EvaluatIR system has however received little use from IR researchers, demonstrating that an effort must be made to convince IR researchers of the advantages of proposed solutions. Therefore, a more comprehensive effort is needed to develop an evaluation infrastructure able to cover different aspects and uses of information access system evaluation. A first effort in this direction is represented by the DIRECT system [Agosti and Ferro, 2009, Agosti et al., 2012a], but a deeper discussion involving experts from information retrieval, databases, and knowledge management is needed, as witnessed by a recent workshop on data infrastructures for IR [Agosti et al., 2012c].

The second disadvantage is the evaluation at system-level only. An IR system contains many components (e.g. stemmer, tokeniser, feature extractor, indexer, etc.), but it is difficult to judge the effect of each component on the final result returned for a query. In evaluation campaigns, usually only evaluation metrics for the outcome of a complete IR system are provided, with information on the components used in a system usually not provided. However, even if evaluations on each component individually were available, extrapolating the effect on a complete IR system from an evaluation of a single component is impossible. As pointed out by Robertson [Robertson, 1981], to choose the optimal component for a task in an IR system, alternatives for this component should be

[7] http://www.evaluatir.org/

evaluated while keeping all other components constant. However, this does not take into account that interactions between components can also affect the retrieval results. For research groups that are experts on one particular component of an IR system, the requirement to evaluate a full system could mean that their component is never fully appreciated, as they do not have the expertise to get a full IR system including their component to perform well. Some attempts at introducing component-level evaluation have been made. TRECVid has separate evaluation of selected components, e.g. visual feature detectors, the output of which can be made available to participants for use in their own systems [Smeaton et al., 2006]. Further examples are the MediaMill challenge [Snoek et al., 2006] and Grid@CLEF 2009[8] [Ferro and Harman, 2010]. However, there is currently a very low acceptance of component-level evaluation. For MediaMill, browsing the papers citing [Snoek et al., 2006] gives the idea that while many researchers make use of data and ground truth, few use the system framework. There were only two participants in Grid@CLEF 2009, and the task has not been repeated in subsequent years. A recent paper [Kürsten and Eibl, 2011] was the first paper to run extensive (over 140,000) experiments comparing combinations of 5 stemming approaches, 13 ranking algorithms and 2 feedback models with different sets of parameters. While this study allowed very useful conclusions about optimal combinations of components to be drawn, the experimental setup is not accessible for further use and modification by other IR researchers, who could find it useful to test their own components in the system used, or change the workflow by for example adding another type of component or modality.

The third disadvantage due to the system-level approach is that it leads to difficult reproducibility of the results. Furthermore, when reviewing a number of years of an evaluation task, it is often difficult to go beyond superficial conclusions based on complete system performance and textual descriptions of the systems. Little information on where to concentrate effort so as to best improve results can be obtained.

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 [Armstrong et al., 2009] 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 [Agosti et al., 2007b].

[8] http://ims.dei.unipd.it/websites/gridclef

Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement n. 258191

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 [Brereton et al., 2007]. 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 and discussed in the area of genetics, but should be applicable to all data intensive sciences [Mons et al., 2011]. The central idea is to encode the central findings of a paper as a set of 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 4. 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 [Deelman et al., 2009], 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[9], as it is the most widely used tool on the myExperiment portal for sharing scientific workflows[10]. Taverna also has the advantage that it has been integrated with the U-Compare UIMA-based text mining and natural language processing system[11] [Kano et al., 2010]. 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 [Hanbury et al., 2012a]. The infrastructure should allow IR scientists to design a workflow for an experiment specifying a specific class of component at each point. It will have to be investigated how to make the granularity of the components at which the evaluation is done adjustable — higher granularity will imply less scalability and vice versa. For example, if the researcher plans to fine-tune components of a search engine, then the granularity can be at the level of stemmers and indexers, while for a more high-level task such as adapting a search system to a domain, a selection of well-performing complete search engines could be used

[9] http://taverna.org.uk/
[10] http://www.myexperiment.org/
[11] http://u-compare.org/

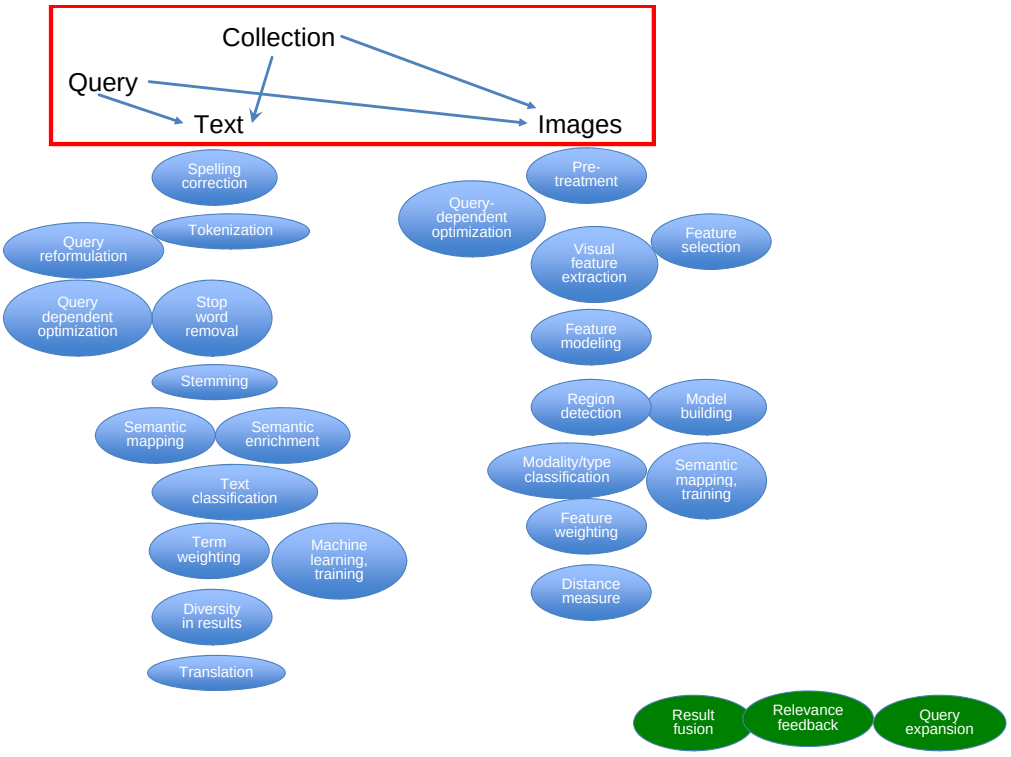

Figure 4: Text and image search components.

along with some components for domain adaption. New metrics and visualisations will have to be developed to allow effective interpretation of complex experimental results. One could imagine that once components have been contributed to the infrastructure, setting up a workflow will start the infrastructure continuously running tasks using different combinations of components on various data sources — the job of the IR scientist would then become the interpretation of the output of the large number of component combinations applied by the infrastructure (making this aspect perhaps more like particle physics to a certain extent). A further advantage of such an infrastructure is that it would allow experiments to be done on private data (such as medical records) in a relatively straightforward way, as it would not be necessary to distribute the data to the people doing the experiments — 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. Components would not have to made open source (although this would be encouraged), but could also be provided as web services. A practical problem is what to do about multiple different implementations of the same algorithms. Papers would have to be linked to the evaluation infrastructure and include a detailed specification of the setup, leading to executable papers, meaning that a format for doing this will have to be defined [Ferro et al., 2011]. This would link well to current initiatives for the digital preservation of scientific data and results[12].

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 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 [Robertson, 2008] and the apparent lack of improvement in ad-hoc IR systems over the previous decade [Armstrong et al., 2009], 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 [Agosti et al., 2007a]. Through

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

the use of this approach, it will 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. In the long term, this could even be through community reviewing, where comments on papers are placed on an open publishing site. Papers that are useful will be cited and the developed components will be used; papers that are less useful will sink into oblivion; errors in papers will be found by the community of reviewers and will likely be retracted by authors. This could potentially serve as a template for experimental computer science in general.

6 Ongoing Evaluation

6.1 Motivation

Current rates of producing digital information have incited research in developing new IR methods able to cope with large data collections. Testing and evaluating new IR methods is non-trivial: both IR engines and the data they work with are changing at high rates and require for some revision and evolution of Cranfield-based [Cleverdon, 1997] IR evaluation techniques.

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). An IR system being assessed will give a set of answers to the questions which, in turn, will be valuated by some measures and eventually compared with other sets of answers [Sanderson, 2010, Agosti et al., 2010, Forner et al., 2011]. Often, the results of the evaluation results trigger modifications in the IR system with the aim of improving the quality of the system and consequently of the retrieval process (see Figure 5).

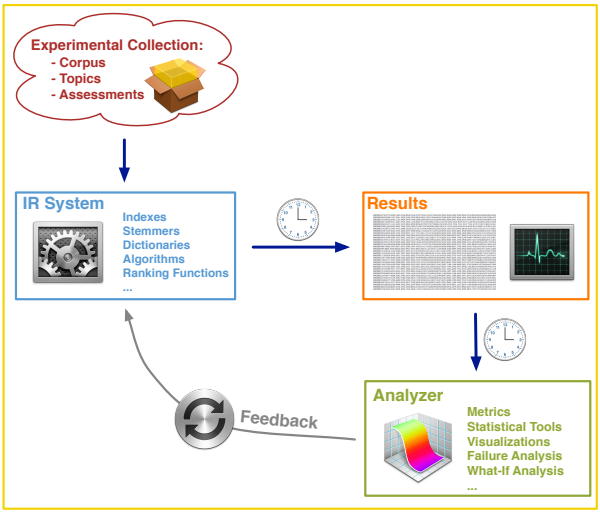

Figure 5: A simplified view of an IR experiment

Experimental evaluation is the focus of a high-valued but time-consuming process which aim is improving IR systems; setting-up a traditional experimental environment is a time-consuming process which involves the creation of the corpus, the specification of the information needs then transformed into topics, the assessment procedure, the management and coordination of evaluation campaigns along with many other demanding and challenging intermediate steps. 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 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 [Angelini et al., 2012b]. Moreover, once you understand the reason behind a failure, you still need to conduct an analysis to understand what among the different possible solutions is most promising and effective before actually starting to modify your system.

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 in order 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* [Hanbury et al., 2012b] where it can exploit loosely-coupled parallel applications and leverage on abstraction for work-flow description to obtain "ease of use, scalability, and portability" [Juve and Deelman, 2010].

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. In cases when the execution time has to be payed for (e.g. server and program instantiations in a non-public cloud), it is desirable to reduce costs by reducing execution time and to be able to understand how experimentations are going while they are on-going in order to take actions preventing the waste of resources.

Moreover, in this scenario, problems experimented in the context of data streams may arise. In particular, due to the computational time and network latency the user may experiment a very long delay in getting the result. As a consequence, it is mandatory to address at least three basic issues:

- provide the user with a partial result (i.e., visualizations and metrics) as soon as a minimal amount of data has been processed;
- provide the user with an estimation of the approximation (e.g., the confidence interval) of the actual result and a suitable visualization of that;
- define a formal Visual Analytics solution able to detect significative 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 very 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 [Xie et al., 2010], in which authors propose a framework to visualize data streams with the goal to show significant pattern changes to users. A similar approach is presented in [Hao et al., 2009] with the goal of detecting anomalies.

In the context of ongoing-evaluation we do not focus on pattern changes, but on detecting changes in cumulated data and on evaluating them in order to show to user only relevant ones.

Therefore, we are interested in representing all the stream cumulated data and in proposing measures that allow to understand when cumulated data is significant with respect to the past elaborations.

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.

A key factor decreasing the efficiency of the evaluation process is the time required to execute an experiment, to output and present the results in a readable way. While a process is on-going researchers cannot analyze the experiment and cannot have a partial view of what is happening. For example, before measuring the effectiveness of a retrieval process, the researcher has to wait until all retrieval results are returned which can be problematic for long computations using large retrieval results such as effectiveness measurements or statistical significance tests. Therefore, let researchers decide and take action on the course of their experiments is a key aspect of an evaluation activity. We envision, in this sense, a methodology which allows for continuous evaluation [Braschler et al., 2010a] and assessment of retrieval results also during the retrieval process. In a *trial—take decision—trial* cyclic setting, where the *trial* phase could be extremely long, it is desirable to *take decisions* while the *trial* is going on. In the following we present some challenges in achieving this latter goal.

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) answer(s) 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 put in the conditions to have the capacity to take the following actions:

- indicate the output that should be monitored during the execution of an IR experiment;
- 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.

In order to achieve these challenges a promising solution could be envisaged by exploiting "cloud" and *Service-oriented Architectures (SoA)* environments [Vouk, 2008]. Indeed, these environments provide us with potential benefits which can be exploited in on-going evaluation: (i) *Dynamic provisioning* is a useful feature because it can handle the resource requirements changing over time providing application-specific solutions and auto-scaling for web-servers which leads to a more rationale and customized use of the resources. (ii) The user has control over the software environment enabling the re-use of components. (iii) Experimental environments and experiments are reproducible. (iv) *Virtualization* abstraction and isolation of lower level functionalities and underlying hardware enabling portability of higher level functions.

On the other hand, there are also potential drawback telling us that the cloud could be part of the solution, but it is not the solution itself: (i) The user has more control on the evaluation process, on the software and on the applications and this lead to more complexity concerning the use and the setting-up of experimental environment. (ii) How to collect provenance information in a standardized way and with minimal overhead design and integrated provenance recording. (iii) How to store this information in a permanent way so that one can come back to it at anytime and re-use the data and experiment performed in this environment. (iv) How to present data and information to the user in a fruitful manner.

Thus, the design and development of a *software* system suitable for on-going evaluation is another big research challenge. This system should be equipped with advanced visual analytics techniques which could help to infer meaning from the produced data. The proposed visualizations should be able to change dynamically on the basis of the new data that will be available during the on-going evaluation 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:

- **IR**: 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. Furthermore, the knowledge of evaluation metrics will be requested in order to select or define new metrics well-suited for the requirements of on-going evaluation. Output inspection needs for a deep understanding of the components of an IR system.

- **Statistics**: The error estimation requires a deep knowledge of statistics, required to establish the minimal amount of data required to get some valid conclusions from the experiments and to analyze the stability and the reliability of the performed analyses. A deep understanding of the evaluation process and the related metrics and their statistical properties is also necessary.

Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement n. 258191

- **Visual Analytics**: 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 field. Moreover, formal methods and algorithms are needed, in order to del with partial elaborations.
- **Software Engineering**: The cloud and 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 this field. Furthermore, this competency is required to understand where the cloud could be used and where we need to find out other solutions.
- **Database Systems**: Knowledge in this field is essential to manage all the data produced by the on-going evaluation, to allow for repeatability of the experiments, and re-use of the produced data.

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;
- analyze the available cloud and SoA available environment verifying how they allow for service and application composition, test and how they manage the produced data and they handle provenance;
- design and develop a software system allowing for stopping or pausing the evaluation process at intermediate steps;
- perform interactive and visual analysis in which visual analytics provides innovative solutions to infer meaning from the data produced in the evaluation process. Some sort of dynamic visualization environment will be needed.

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

Nowadays multilingual and multimedia information access systems are not only called on to manage and provide access to huge amounts of heterogeneous data and content stored in large silos. Increasingly, they also need to address ever more complex interactions of users with such contents. To a large degree this is due that we increasingly live our lifes online. There is a meteoric rise of outlets for new forms of instant, multifaceted "user generated content" or "social media," as exemplified by services such as Facebook, Twitter, and Flickr, through which we share many aspects of day-to-day existence. Blogs, discussion forums, comments left behind on news sites, instant messages—they all carry valuable and timely information, which companies and individual users wish to interact with. User expectations about the capability of information access systems to incorporate such information and to adapt to continuous interactions are increasing.

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?

Even when today's state-of-the-art information access systems try to tackle one or more of the issues discussed above, they mainly operate in batches or have limited incremental update capabilities. Instead, what is needed to fully support a fruitful interaction between users, digital material, and user signals are systems that are able to react in near real-time, and in a uniform and coherent manner, to streams of information that is very diverse in nature, volumes, and growth rates. Such information is potentially heterogeneous multilingual and multimedia digital material, implicit user signals, explicit user annotations and tags. We need to address the challenge of interpreting and linking such information to already managed information resources, in order to make them readily available for subsequent search, access, and navigation by other users.

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 6: 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.

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.

Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement n. 258191

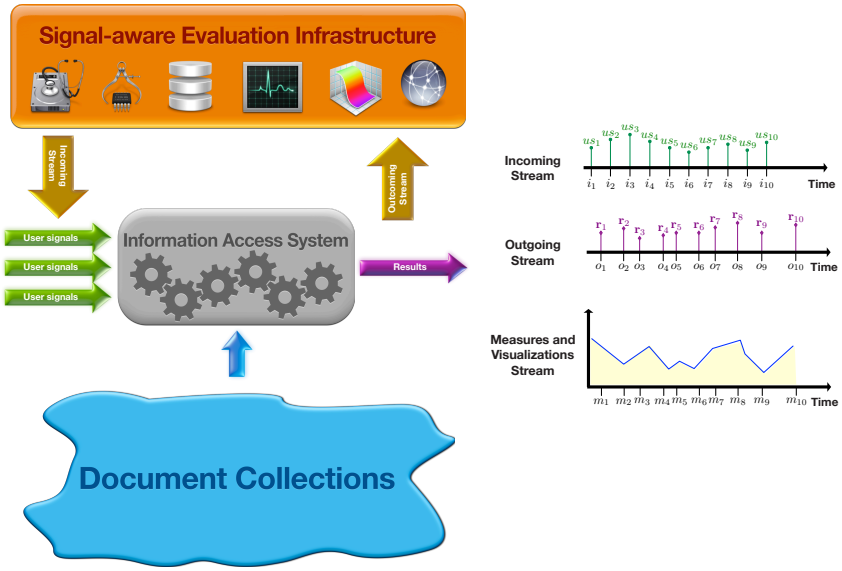

Figure 6: Signal-aware evaluation of real-time information access systems.

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 [Cleverdon, 1997] dating back to early 60s of last century. The Cranfield paradigm is based on the notion of experimental collection, i.e., a triple (D, T, J) where D is a collection of documents representative of a given domain, T is a set of topics, i.e., surrogates of the user information needs in that domain, and J is a set of relevance judgements, i.e. the ground-truth which determines which documents in D are relevant for each topic in T. 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 [Harman, 2011].

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.

Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement n. 258191

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. The topics and challenges proposed during this two-days brainstorming workshop hit the boundaries of the discipline and go beyond them, often calling for competencies from other fields.

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

[Agosti et al., 2011] Agosti, M., Braschler, M., Di Buccio, E., Dussin, M., Ferro, N., Granato, G. L., Masiero, I., Pianta, E., Santucci, G., Silvello, G., and Tino, G. (2011). 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.

[Agosti et al., 2012a] Agosti, M., Di Buccio, E., Ferro, N., Masiero, I., Peruzzo, S., and Silvello, G. (2012a). DIRECTions: Design and Specication of an IR Evaluation Infrastructure. In [Catarci et al., 2012].

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

[Agosti et al., 2007b] Agosti, M., Di Nunzio, G. M., and Ferro, N. (2007b). The Importance of Scientific Data Curation for Evaluation Campaigns. In Thanos, C., Borri, F., and Candela, L., 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.

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

[Agosti et al., 2012b] Agosti, M., Ferro, N., Masiero, I., Nicchio, M., Peruzzo, S., and Silvello, G. (2012b). Deliverable D3.3 – Prototype of the Evaluation Infrastructure. PROMISE Network of Excellence, EU 7FP, Contract N. 258191.

[Agosti et al., 2010] Agosti, M., Ferro, N., Peters, C., de Rijke, M., and Smeaton, A., editors (2010). *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.

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

[Allan et al., 2012] Allan, J., Aslam, J., Azzopardi, L., Belkin, N., Borlund, P., Bruza, P., Callan, J., Carman, M. Clarke, C., Craswell, N., Croft, W. B., Culpepper, J. S., Diaz, F., Dumais, S., Ferro, N., Geva, S., Gonzalo, J., Hawking, D., Järvelin, K., Jones, G., Jones, R., Kamps, J., Kando, N., Kanoulos, E., Karlgren, J., Kelly, D., Lease, M., Lin, J., Mizzaro, S., Moffat, A., Murdock, V., Oard, D. W., de Rijke, M., Sakai, T., Sanderson, M., Scholer, F., Si, L., Thom, J., Thomas, P.,

Trotman, A., Turpin, A., de Vries, A. P., Webber, W., Zhang, X., and Zhang, Y. (2012). 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.

[Angelini et al., 2012a] Angelini, M., Ferro, N., Granato, G. L., and Santucci, G. (2012a). Deliverable D5.3 – Collaborative User Interface Prototype with Annotation Functionalities. PROMISE Network of Excellence, EU 7FP, Contract N. 258191.

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

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

[Asadi et al., 2011] Asadi, N., Metzler, D., Elsayed, T., and Lin, J. (2011). Pseudo test collections for learning web search ranking functions. In Ma, W.-Y., Nie, J.-Y., Baeza-Yaetes, R., Chua, T.-S., and Croft, W. B., 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.

[Azzopardi et al., 2007] Azzopardi, L., de Rijke, M., and Balog, K. (2007). Building simulated queries for known-item topics: an analysis using six european languages. In Kraaij, W., de Vries, A. P., Clarke, C. L. A., Fuhr, N., and Kando, N., 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.

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

[Berendsen et al., 2012a] Berendsen, R., Braschler, M., Gäde, M., Kleineberg, M., Lupu, M., Petras, V., and Reitberger, S. (2012a). 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.

[Berendsen et al., 2012b] Berendsen, R., Tsagkias, E., de Rijke, M., and Meij, E. (2012b). Generating pseudo test collections for learning to rank scientific articles. In [Catarci et al., 2012].

[Braschler et al., 2010a] Braschler, M., Choukri, K., Ferro, N., Hanbury, A., Karlgren, J., Müller, H., Petras, V., Pianta, E., de Rijke, M., and Santucci, G. (2010a). A PROMISE for Experimental Evaluation. In [Agosti et al., 2010], pages 140–144.

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

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

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

[Catarci et al., 2012] Catarci, T., Forner, P., Hiemstra, D., Peñas, A., and Santucci, G., editors (2012). *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.

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

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

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

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

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

[Ferro and Harman, 2010] Ferro, N. and Harman, D. (2010). CLEF 2009: Grid@CLEF Pilot Track Overview. In Peters, C., Di Nunzio, G. M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., and Roda, G., editors, *Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF 2009). Revised Selected Papers*, pages 552–565. Lecture Notes in Computer Science (LNCS) 6241, Springer, Heidelberg, Germany.

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

[Forner et al., 2011] Forner, P., Gonzalo, J., Kekäläinen, J., Lalmas, M., and de Rijke, M., editors (2011). *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.

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

[Hanbury and Müller, 2010] Hanbury, A. and Müller, H. (2010). Automated component-level evaluation: Present and future. In [Agosti et al., 2010], pages 124–135.

[Hanbury et al., 2012a] Hanbury, A., Müller, H., Langs, G., Weber, M., Menze, B. H., and Salas Fernandez, T. (2012a). Bringing the algorithms to the data: Cloud-based benchmarking for medical image analysis. In [Catarci et al., 2012].

[Hanbury et al., 2012b] Hanbury, A., Müller, H., Langs, G., Weber, M. A., Menze, B. H., and Fernandez, T. S. (2012b). Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis. In [Catarci et al., 2012].

[Hao et al., 2009] Hao, M. C., Dayal, U., Keim, D. A., Sharma, R. K., and Mehta, A. (2009). Visual analytics of anomaly detection in large data streams. In Börner, K. and Park, J., editors, *Proc. Visualization and Data Analysis (VDA 2009)*. SPIE Proceedings 7243.

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

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

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

[Huurnink et al., 2010] Huurnink, B., Hofmann, K., de Rijke, M., and Bron, M. (2010). Validating query simulators: An experiment using commercial searches and purchases. In [Agosti et al., 2010], pages 40–51.

[Järvelin et al., 2012] Järvelin, A., Eriksson, G., Hansen, P., Tsikrika, T., Garcia Seco de Herrera, A., Lupu, M., Gäde, M., Petras, V., Rietberger, S., Braschler, M., and Berendsen, R. (2012). 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.

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

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

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

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

[Kürsten and Eibl, 2011] Kürsten, J. and Eibl, M. (2011). A large-scale system evaluation on component-level. In Clough, P., Foley, C., Gurrin, C., Jones, G. J. F., Kraaij, W., Lee, H., and Mudoch, V., editors, *Advances in Information Retrieval. Proc. 33rd European Conference on IR Research (ECIR 2011)*, pages 679–682. Lecture Notes in Computer Science (LNCS) 6611, Springer, Heidelberg, Germany.

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

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

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

[Reitberger et al., 2012] Reitberger, S., Imhof, M., Braschler, M., Berendsen, R., Järvelin, A., Hansen, P., Garcia Seco de Herrera, A., Tsikrika, T., Lupu, M., Petras, V., Gäde, M., Kleineberg, M., and Choukri, K. (2012). 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.

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

[Robertson, 1981] Robertson, S. E. (1981). The methodology of information retrieval experiment. In Spärck Jones, K., editor, *Information Retrieval Experiment*, pages 9–31. Butterworths, London, United Kingdom.

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

[Smeaton et al., 2006] Smeaton, A. F., Over, P., and Kraaij, W. (2006). Evaluation campaigns and trecvid. In Wang, J., Boujemaa, N., and Chen, Y., editors, *Proc. 8th ACM International Workshop on Multimedia Information Retrieval (MIR 2006)*. ACM Press, New York, USA.

[Snoek et al., 2006] Snoek, C. G. M., Worring, M., van Gemert, J. C., Geusebroek, J.-M., and Smeulders, A. W. M. (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia. In Nahrstedt, K., Turk, M., Rui, Y., Klas, W., and Mayer-Patel, K., editors, *Proceedings of the 14th annual ACM international conference on Multimedia (MM 2006*, pages 421–430. ACM Press, New York, USA.

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

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

[Tsikrika et al., 2011] Tsikrika, T., Garcia Seco de Herrera, A., and Müller, H. (2011). Assessing the Scholarly Impact of ImageCLEF. In [Forner et al., 2011], pages 95–106.

[Vouk, 2008] Vouk, M. A. (2008). Cloud computing - issues, research and implementations. In *30th International Conference on Information Technology Interfaces*, pages 31–40.

[Xie et al., 2010] Xie, Z., Ward, M. O., and Rundensteiner, E. A. (2010). 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. Springer-Verlag.

ISBN 978-88-6321-039-2

Printed on September 2012

Available online at:
http://www.promise-noe.eu/promise-retreat-report-2012/