Proposal for an ESF Research Networking Programme – Call 2009

Section I

Programme title: Evaluating Information Access Systems

Programme acronym: ELIAS

Name and full coordinates of principal applicant(s):

Kalervo Järvelin Department of Information Studies and Interactive Media, University of Tampere, FIN-33014 University of Tampere, Finland. Email: kalervo.jarvelin@uta.fi. URL: http://www.uta.fi/~likaja/. Phone: +358 3 355 111/+358 50 547.

Maarten de Rijke Intelligent Systems Lab Amsterdam, University of Amsterdam, Science Park 107, 1098 XG Amsterdam, The Netherlands. Email: derijke@uva.nl. URL: http://www.science.uva.nl/~mdr. Phone: +31 20 525 5358/+31 6 51 938 523

Mark Sanderson Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello St, Sheffield, S1 4DP, UK. Email: m.sanderson@shef.ac.uk. URL: http://dis.shef.ac.uk/mark/. Phone: +44 114 22 22648.

Indication of which of the principal applicants is the contact person: Maarten de Rijke

Keywords relating to the topic of the proposal: Evaluation methodology, Living laboratories, Information access, Interactive information retrieval, Information seeking

Abstract: Although the speed of computers and volume of information has grown exponentially since the first search engines were created in the 1950s, nearly 6 decades on, the methods used to test state of the art information access systems have changed little. However, there is growing scientific evidence of a need for a complete testing overhaul: producing novel evaluation methods that take a user-oriented perspective on assessing the effectiveness of information access systems. The proposed network will establish a research and training programme for the evaluation of information access systems which will define a new measurement paradigm based on so-called living laboratories. This paradigm involves (i) exploitation of novel market places and forums where large numbers of users are recruited into early stage evaluation experiments to test a particular aspect of an information access system; and (ii) using operational systems as experimental platforms on which to conduct user-based experiments at scale. The network will achieve an evaluation methodology directed at user-oriented interactive evaluation, a new and tested set of interactive evaluation metrics, infrastructure and test suites based on these and an ongoing collaborative forum with evaluation cycles with a focus on evaluating information access systems, as well as producing a European network of new researchers trained in the improved methodologies.

Previous or concurrent applications to the ESF for any of the ESF instruments: There have not been any previous or concurrent applications to the ESF of this programme.

Section II: *(not more than 8 A4 single pages / max. length 3000 words)*

II.1 Status of the relevant research field; scientific context, objectives and envisaged achievements of the proposed Programme:

Performance evaluation in Information Retrieval (IR) has a long tradition that has greatly boosted the development of information access systems (IASs). Over the past few years, however, these well established IR evaluation methodologies were criticized over a number of important failings: failures in addressing users, search interfaces, the scale of testing environments, the diversity of IASs, and the diversity of search tasks and search processes. Below we discuss these shortcomings and point out alternatives in evaluation, before discussing the goals and approach of the proposed Network.

II.1.1 Evaluation in information access

The goals of a research area may be classified as (a) theoretical understanding, (b) empirical description, prediction and explanation, and (c) technology development. Evaluation is a key ingredient of the scientific method: researchers formulate a hypothesis, design an experiment that permits it to be tested and then measure the degree to which the hypothesis is confirmed. In research related to IASs, such as search engines, evaluation is often related to the question of whether the eventual day-to-day users of a system will be more successful at (or simply prefer) using one IAS over another. It is impractical for scientists to continually ask such users to help in evaluation; therefore, means of simulating these users were devised.

The primary approach to such simulation was to create a benchmark, known as a *test collection* to assess the *effectiveness* of an IAS. The collections are most commonly used to compare the number of relevant documents retrieved by competing systems. The origins of this approach date back to library research in the 1950s, which was synthesized into the creation of the first test collection for computer-based IASs in the early 1960s: the so-called Cranfield experiments. Evaluating competing systems on test collections proved to be a powerful way to improve the state of the art. However, this decades old design remains an extremely crude simulation of user behavior.

II.1.2. Gaps in the evaluation of information access

As a simulation, test collections have a number of important failings: failure to address differences in users; failure to consider search interfaces; failure to scale as the amount of information and IAS users grow; failure to address the diversity of information access systems; and failure to address different types of search tasks and the longitudinal process of searching.

1. Individual differences between users such as location and personalization are strong, but not currently catered for in test collections; the arrival of new types of content (especially user generated content) makes this painfully clear. The potential for users having different interpretations of a query are barely addressed in current testing. The same applies to users' learning in the search process.
2. We need to do more than count relevance to explain and predict searcher behavior. Given that graphical user interfaces have been ubiquitous for the past 25 years, it is surprising that result presentation is not part of test collection evaluation. Aspects such as document snippets, speed, query suggestions, and the layout of the result page as a whole, hardly ever play a role in today's evaluation.
3. While document collections used for experimentation have grown in size, the number of test queries has remained small, most often in the double digits, despite strong scientific evidence that this is insufficient.
4. Testing has focused on very traditional information seeking activities. However, information access addresses a much broader range of tasks, ranging from simple navigation to find known resources to complex explorative searching with initially vague ideas of the possible targets to access. To remain relevant, benchmarking

activities for information access need to consider tasks that naturally correspond to these needs, and the evaluation metrics need to enable researchers to assess their theories, models and system's successes and failures along relevant dimensions.

5. Searchers using IR systems often use queries that individually have poor performance but when used in sessions, may be surprisingly effective. Test collection-based evaluation ignores the search session. To be able to test models and theories of searchers' underlying strategies we need more data—more queries and queries organized in sessions. In addition, we need to recognize different types of search topics (e.g., known-item vs. simple factual vs. complex relational; focused vs. open) and retrieval tasks (low effort/high precision vs. high effort/relaxed precision; dynamic explorative vs. stable and focused)—and have sufficient numbers of tasks in each category.

The current test collection approach fosters research that focuses on the narrow goal of optimizing the output of individual queries in terms of relevance-based metrics. Such research is unhelpful in understanding or supporting human behavior or performance in information access. The shortcomings of test collections cannot be overcome by fine-tuning current evaluation practices. Evaluation methodologies need to be completely overhauled to provide a better means of validating, refuting, or deciding between, the theories and models of information access.

Alternatives have been explored, but they too have been found wanting. Many researchers conducted so-called *interactive evaluation* of information access systems. Such evaluations tended to take place in a lab setting, with controlled tasks, relatively small sets of users, and queries, testing known tasks and apriori known target outcomes. This approach did not allow evaluation of search processes where the user learns on the way, i.e., the desired outcome is a moving target. Although the methodology produces valuable results, it is slow to set up, and expensive to run, which prevents it from often being used.

Static (transaction) log files recording user query and click behavior have been analyzed by researchers since the late 1960s. Log-based methodologies that attempt to evaluate new information access systems have been devised (e.g., *click prediction*). However, the utility of static log files is limited primarily because access to the logs is difficult owing to data protection and privacy concerns. In real life, searchers often interleave the use of several information access systems. Server side logs at each access system cannot therefore reveal more than a distorted and perhaps uninterpretable picture of what takes place at the searcher side.

II.1.3 Living laboratories

Very recently, new approaches in evaluation have emerged, that we class under the common name *living laboratories* or *living labs*. A living lab is a new research paradigm integrating both a user-centric multidisciplinary research approach and user community driven innovation based on real life experiments. Among others, it is intended to increase the understanding of occurring phenomena, explore and evaluate new ideas, concepts and implementations, confront new ideas, concepts and implementations with users' value model, and enable re-usable experiments (i.e., dataset, research protocols and methods). Moreover, living labs may contribute to bringing science and innovation closer to the general public. A living lab is more than an experimental facility as its philosophy is to turn users, from being traditionally considered as a problem, into value creation. In a living lab, users are exposed to new innovative solutions in (semi)realistic contexts, as part of medium- or long-term studies.

The ELIAS Network Programme envisages two types of living labs. The first type of laboratory involves exploitation of market places and forums where large numbers of users can be recruited to be involved in early stage evaluation experiments to test a particular aspect of a information access system.

The second type of laboratory involves using operational systems as experimental platforms on which to conduct user-based experiments at scale. Building a deployed

system provides an interesting interactive IR setting as well as a useful complement to static click logs. In particular, the resulting click logs can be analyzed but dynamic studies can also be performed where the system is changed on the fly—thus enabling researchers to conduct controlled experiments *in situ*, either in the form of A/B testing (establishing preference between two alternatives—see below) or by interleaving results from different methods.

Both approaches have been explored, but there is as yet no consensus on how living labs can best be exploited, which types of information access will best be evaluated using such laboratories and, most importantly, what the limitations of the living labs methodology will be.

In many ways, today's information access research community is at the same stage as the test collection researchers of the 1950s. A basic methodology has been established, but it needs to be synthesized into a refined approach that other researchers can use. We contend that an ESF Research Network Programme will provide an ideal forum for like-minded researchers to collaborate and interact so as to share experiences in their use of living labs such that at the conclusion of the network a new approach to evaluation of IASs will be defined. This is necessary for the development of IASs that are effective in their actual contexts of use—a goal that cannot be achieved by continued use of methods that evaluate IASs in isolation under incomplete and/or artificial conditions.

II.1.4 Research objectives

We aim to *study living laboratories as a platform for the evaluation of information access in the large*. We organize our research objectives along two dimensions. In the *horizontal dimension* we examine a range of domains and application areas that drive our research; in each situation we explain the underlying information access scenario, set out the challenges and explain how the living labs approach would fit. In the *vertical dimension* of fundamental questions, we consider the methodological and user simulation issues to be addressed. This two-dimensional, cross-cutting methodology is designed to ensure that the research of the network will develop into a coherent field that is methodologically sound and grounded in practice.

II.1.4.1 Horizontal dimension: domains and applications

Numerous tasks in information access can be assessed in a living lab. We have selected to present five examples of access scenarios. Each of these motivates the study of a different living lab—in each case the living lab perspective facilitates a type of analysis that is simply beyond the scope of evaluation methodologies currently being practiced in academia.

User generated content such as blogs, discussion forums, comments, is often produced in response to world events or reports on personal experiences. Searchers of this content are not just looking for relevant documents, but are in search of "the big picture": the event or product being commented on, long-term developments, outliers, stakeholders, emerging phenomena etc. By running an operational system for searching user generated content as an experimental platform we will be able to determine what mix of activities users of user generated content engage in (time-based ranking, tracking information re-use, assessing the online impact of an event, etc) and how to best present the multi-faceted results and materials pertaining to their activities.

Political data (parliamentary proceedings, *written answers and statements*, *bills*, etc) are particularly interesting from an information access point of view. With a lot of implicit structure (rhetoric, topical, social, etc.), a focus on entities (politicians and their parties) as well as on topics, its users—whether professional or layman—are characterized by strong individual differences (covering the entire political spectrum and a broad range of interest groups) and numerous tasks (fact finding, topic tracking, identifying stakeholders, etc.). The traditional Cranfield set-up, with its limited set of assessors and a limited set of queries, simply cannot do justice to the richness of the

domain—but with a living lab as an experimental platform one can overcome this problem and assess search facilities in this context.

Europe's rich cultural heritage is increasingly being made available online. Much of the material is shared across country and language boundaries and many searchers for cultural heritage materials are poly-lingual, switching from a query in one language to digesting results in *multiple languages*, and navigating across languages, across sources, from authoritative ones to fan sites and other user contributed experience reports, often supported by visual materials and online translation tools, thus lowering any language barriers that might exist, a far cry from the abstract and static scenarios in which multilingual information access is being evaluated today.

Wikipedia is increasingly being used as a general knowledge source, to which organizations link their data and in which individuals find background information to answer their informational queries. It attracts a broad spectrum of users, from (consuming) novices new to a knowledge area to (producing) experts sharing their expertise. Many Wikipedia users engage in long sessions, relying on extensive semantic information made available in the form of, e.g., categories or info-boxes. Given the range of users and, therefore, preference criteria, Cranfield-based evaluations of models that attempt to capture those criteria are severely underpowered. In a setting where large numbers of users (on forums and market places) are called upon to aid in producing assessments, we can begin to reliably test our models of access to semantically rich knowledge sources such as Wikipedia.

In knowledge-intensive domains, experts have a complex information environment that provides multiple IASs. In carrying out their work tasks, they utilize the systems in a concerted, interleaved manner, where one system feeds into the access of another. *Scientific research in molecular medicine and bioinformatics* provides a prominent example of such an information environment. On-going research based on logging, SenseCam photography, shadowing and interviews has clearly shown, that if real work task performance is examined from the point of view of any single system or collection (such as web search engine logs), understanding of searcher access behavior remains very narrow and distorted. In real life, IASs are rarely used in isolation; the challenge is to devise an evaluation methodology appropriate for such demanding contexts.

II.1.4.2 Vertical dimension: methods and models

A suitably developed methodology for evaluating information access in the domains described above can play an enabling role, in much the same way as the Cranfield-tradition has done for system-oriented evaluation of IASs. We aim to develop evaluation methodologies and the underlying foundations for measuring success and failures of information access theories and models in a range of domains. Here, we explain a number of fundamental questions to be addressed by the proposed Network.

Traditional test collection-based evaluation of IR systems is based on absolute judgments of individual documents (either binary or graded), independent of other documents. Real searchers use *relative judgments* or *preferences* ("Is document A better than document B?") and prefer diverse results that are novel ("Is this document novel when added to a result set?"). Asking people which of two results they prefer is faster to digest and more reliable than asking them to make absolute judgments about the relevance of each result. A challenging goal is to develop standard methodologies for reliably collecting such judgments and metrics for using them in evaluation.

Also, the traditional Cranfield-type studies have proven very useful in core relevance but are lacking extensions to *set-level properties* like diversity or novelty. Diversification can not only be achieved by diversifying a ranked result list but also by altering the result presentation. E.g., site collapsing, background information, related searches can all be seen as ways of diversifying—ways that can be explored using living labs, but not in a Cranfield-type setting. Extending our evaluation methodology to result set-level properties (or even result page-level properties) involves formalizing the utility that matters as well as the cost of getting judgments (e.g., if costs must be made per set or page of results, different queries may have vastly different costs).

Furthermore, relevance, especially *topical relevance* does not fully capture whether the user's information need has been satisfied or measure the user's performance on a task. In particular, it is important to determine what components other than relevance play a role in measuring satisfaction and task performance and how one can reliably measure these quantities.

Unlike the "abstract" user being modeled in test collection-based evaluation, real searchers often have *various interaction strategies and expectations* and may learn on the way and have a moving target, usually attempting to avoid examining sequences of non-relevant documents and sometimes going for just a few, but good, documents. We aim to facilitate the study of this behavior in living labs and to use the outcomes for developing an evaluation methodology that is compatible with these features.

In the simplest controlled experiment, often referred to as an *A/B test*, users are exposed to one of two variants of a system: control (A) or treatment (B). In traditional evaluations based on test-collections, considerable attention has been devoted to methods for reliably ascertaining the significance of experimental outcomes. An important challenge in A/B testing in a living lab environment is the sampling of users (for either the A or B group) in such a way that we can reliably ascertain the significance of our outcomes and guarantee replicability of the results.

Searchers' search strategies in sessions may be analyzed from client-side search logs obtained from real task processes in context. When both the test searchers and their search tasks are carefully controlled, the variation in initial queries and in subsequent moves can be identified. Such findings can be distilled into a range of session strategy types. The latter can be used to generate *simulated searcher sessions* for the sake of novel retrieval tasks, which can then be tested using a traditional test collection approach. This method has recently been introduced and has several desirable features: great numbers of sessions can be simulated for various kinds of retrieval tasks, similar sessions can be rerun cost-effectively and without learning effects/biases. A thorough analysis addressing the question how systems' rankings resulting from this approach to evaluation compares to systems' rankings obtained by traditional evaluation means will be very important for our methodology.

II.1.5 Envisaged achievements

The main envisaged achievement of the proposed Network is the establishment of a new evaluation paradigm for information access. Specifically, we take this to mean the following:

(A) A new test collection/living lab *methodology*, comparable to the Cranfield methodology, but now directed at user-oriented interactive evaluation of IR systems in the large.

(B) A new and tested set of interactive evaluation *metrics* that measure the costs (efforts) and benefits of information access system users.

(C) *Infrastructure* and *test suites* based on the above A+B

(D) An ongoing community-based *forum* with tracks and annual evaluation tasks/rounds with a focus on evaluating information access systems

---

II.2 Facilities and expertise which would be accessible by the Programme:

The ELIAS Network Programme will be closely allied with the annual CLEF evaluation exercise and the wide range of European and worldwide research groups associated with it. CLEF (the Cross-Language Evaluation Forum, http://www.clef-campaign.org) has run for 10 years, creating and sharing test collections to be used in shared evaluation tasks. The exercise has involved hundreds of research groups and tackled a wide range of search tasks.

The living labs evaluation approach taken by the ELIAS Network Programme will complement the test collection approach of CLEF. ELIAS Network members will be able to exploit the *facilities* and *expertise* within CLEF, which has a large set of test

collections and extensive expertise in designing IR experiments. The forum also has recently made publicly available a large repository of *run data*, which records the submissions of research groups who participated in CLEF over the years.

Such facilities will provide a valuable source of gold standard material against which a range of living lab experiments can be calibrated. The CLEF organizers also have access to licensed corpora and to storage, analysis and evaluation systems that aid researchers in determining the number of relevant documents there are in such corpora. The extensive range of participants in CLEF provides an enthusiastic pool of researchers who will participate in ELIAS activities. The long experience of CLEF organizers will serve as a springboard to help ensure an effective and efficient design and running of experiments in ELIAS.

II.3 Expected benefit from European collaboration in this area:

The ELIAS Network Programme includes European groups that play a leading role in IR evaluation research, having proposed innovative tasks, metrics and evaluation scenarios. Our main goal is the coordinated development of the field, while consolidating the already strong European presence and tradition in the evaluation of information access systems. The programme will create a European network of excellence and significantly enhance the visibility of European research in the field. We are confident that it will make Europe an even more attractive place for researchers and prospective PhD students in the areas covered by this Network Programme.

Europe has a long history in the evaluation of information retrieval systems: the Cranfield tradition was born in the UK and since the early 1990s, two (complementary) IR evaluation initiatives have been running in Europe, with 100+ participating research groups each (in 2009): the Cross-Language Evaluation Campaign (CLEF) and the Initiative for the Evaluation of XML Retrieval (INEX). The ELIAS Networking Programme aims to exploit the informal networks and dynamics created by CLEF and INEX; it will complement the traditional system-oriented evaluation methodology now dominant at CLEF and INEX with its new user-oriented evaluation paradigm.

To realize its ambitions, the ELIAS Network Programme needs to bring together expertise in traditional evaluation methodology, in user studies and user-oriented evaluation, in evaluating access to social media, in mining user data as well as in deploying live systems as experimental platforms. To this end, crossing European borders is inevitable, and, conversely, by bringing together experts in these areas from around Europe, each will benefit from the complementary insights brought in by others. The evaluation activities envisaged by ELIAS need critical mass with (typically) at least a dozen participating groups per experiment to be able to arrive at well-founded conclusions, thus requiring collaboration at a European scale.

The ELIAS Networking Programme is the ideal instrument to introduce, experiment with and analyze a new evaluation paradigm. The applicants and committees that support this Network Programme are strongly rooted in today's main evaluation efforts, in Europe (CLEF, INEX), the US (TREC, TRECVid, TAC), India (FIRE) and Japan (NTCIR). Rather than proposing an incremental change, this Network proposes a departure with far-reaching methodological innovations. The envisaged *Spring planning workshops* will serve as planning events for setting up collaborative experiments based on living labs and for team integration. At the envisaged *Fall reporting workshops*, participants will report on their experimental outcomes; the Fall workshops will also serve for dissemination of results of the Network to the wider research community. These will be further supported by high-level networking and training activities in the form of *research visits* and *short-term fellowships* for junior researchers and *summer schools*. We expect the high visibility and innovative character of the Network to attract leading scientists in neighboring disciplines to collaborate with members of the Network. In summary, the proposed Network can play a major role in coordinating research efforts in this field all over Europe.

II.4 European context

The ELIAS Network Programme will be allied with CLEF. CLEF—both the annual meeting and the cycle of evaluation efforts that lead up to it—has been supported under the EU IST program. CLEF 2000–2001 and CLEF 2004–2007 have been run as an activity of the DELOS Network of Excellence under FP5 and FP6; CLEF 2002–2003 was run independently as an accompanying measure (IST-2000-31002); CLEF 2008–2009 is being run as an activity of the TrebleCLEF Coordination Action (215231). The ELIAS Network Programme will also coordinate with INEX, the Initiative for the Evaluation of XML Retrieval; in the past, INEX has been supported as an activity of the DELOS Network of Excellence under FP5 and FP6. The recently launched SEALS FP7 project aims to provide services for evaluating semantic technologies, without special focus on users or information access. At present, the ESF or FP7 has no activity that is centered on the topics proposed by the ELIAS Network Programme.

II.5 Proposed activities, key targets and milestones

The proposed activities will be organized around five consecutive annual evaluation cycles with planning workshops in the early Spring, and reporting and dissemination workshops in early Fall. High-level networking and training activities in the form of research visits and short-term fellowships for junior researchers and summer schools will complement these annual evaluation cycles. While none of the activities is primarily aimed at providing a publication venue, we will explore opportunities for the production of publications (special journal editions, edited volumes). To reduce travel costs, the regular meetings of the steering committee will be co-located with the annual Fall workshops; additional meetings by the executive group will be co-located with the Spring workshops.

II.5.1 Annual Spring planning workshops

During the lifetime of the project, each year three specific collaborative evaluation tasks will be run by members of the ELIAS Network. These three tasks will be selected from the “horizontal dimension” discussed above (II.1.4.1) with a view to answering questions from the “vertical dimension” (II.1.4.2). Early on in the annual cycle, the Spring workshops—one for each task—will determine the real world work task(s) being modeled; the corpora to be used; (if appropriate) the test queries to be used and/or the traffic to be generated; the metrics to be used; the way(s) of acquiring ground truth; etc. We envisage that 10 (senior) people will take part per planning workshop: this number includes 1 of the 3 project applicants. For each task/workshop, two members of the community will be asked to lead the benchmarking activities.

II.5.2 Annual Fall reporting workshops (co-located with CLEF)

Co-located with the annual CLEF conferences, the annual ELIAS Fall reporting workshops will report on outcomes of the ELIAS experiments that were run. The focus will be on lessons learned and on refining the experimental set-up and methodology. We envisage that 20 participants will take part per reporting workshop.

II.5.3 Summer schools

We plan to organize five summer schools for PhD students and young researchers from within and outside the ELIAS Network Programme. The teaching event will be modeled on the successful formula of the TrebleCLEF Coordination Action. Each school will have a target of about 30 participants and 5–8 main lecturers. The schools will be designed to support the integration of different research traditions and methodologies in evaluating IASs on the one hand and the variety of application domains and approaches to addressing the needs of those domains on the other.

II.5.4 Individual short visits

In order to foster scientific collaboration between partners of the Network, exchange of ideas and results, and facilitate preparation of joint *living laboratories*, other evaluation efforts and joint papers, we plan short visits (of about one week) of both junior and senior researchers to other teams in the Network.

II.5.5 Short fellowships for young researchers

To help junior researchers acquire knowledge and experience, we plan to award short fellowships of about one month to PhD students and postdocs within the Network. These fellowships will give the recipients the opportunity to interact with leading research teams and serve as the basis for future collaboration.

II.6 Duration (48 or 60 months)**:** 60 months

II.7 Budget estimate (in €) by type of activities and per year of the Programme

							Total
		Year 1	**Year 2**	**Year 3**	**Year 4**	**Year 5**	
Steering committee meetings	Steering committee meetings	8000	8000	8000	8000	8000	**40000**
	Executive group meetings	5000	5000	5000	5000	5000	**25000**
	Chair travel	2000	2000	2000	2000	2000	**10000**
Science meetings	Spring ws's	20500	20500	20500	20500	20500	**102500**
	Fall ws's	22000	22000	22000	22000	22000	**110000**
	Summer schools	24000	24000	24000	24000	24000	**120000**
Grants	Short visits	10000	10000	10000	10000	10000	**50000**
	Exchange grants	20000	20000	20000	20000	20000	**100000**
Publicity	Programme brochure	500					**500**
	Web site	500	500	500	500	500	**2500**
	Publications	1000	1000	1000	1000	1000	**5000**
Database costs	Creation of ground truth	5000	5000	5000	5000	5000	**25000**
	Acquisition of session data	5000	5000	5000	5000	5000	**25000**
Programme coordinator	Salary	10000	10000	10000	10000	10000	**50000**
Total		**133500**	**133000**	**133000**	**133000**	**133000**	**665500**

We assume that approximately half of the steering committee will be self-funded for their trip to the annual steering committee meeting. The executive group will meet at the Spring and Fall workshops. Costs for the yearly Spring workshops comprise full expenses of two organizers plus up to two external experts per workshop as well as (limited) support for the other attendees. For the yearly Fall workshops ELIAS will fully fund two organizers per workshop and provide (limited) support for attendees. Creation of ground truth (under "Database costs") will be done using market places and forums (such as "Mechanical Turk"). Session data (again under "Database costs") will be acquired both from external partners and from general public-facing and/or highly specialized experimental platforms; for the latter case, we seek to recruit student volunteers from the participating institutes; both types of data (ground truth and session data) will be used to populate a database for measurement purposes.

Section III

III.1 List of names and full coordinates of the **envisaged Steering Committee members:**

Austria – John Tait. The Information Retrieval Facility, Operngasse 20b, 1040 Vienna, Austria. Phone: +43-1-236 94 74/6053. Email: john.tait@ir-facility.org

Belgium – Francine Moens. Katholieke Universiteit Leuven, Department of Computer Science, Informatics section, Celestijnenlaan 200A, B-3001 Heverlee. Phone: +32 (0)16 32 53 83. Email: sien.moens@cs.kuleuven.be

Denmark – Pia Borlund. Danmarks Biblioteksskole, Fredrik Bajers Vej 7K, 9220 Aalborg Øst, Phone: +45 98 15 79 22. Email: pb@db.dk

Finland – Kal Jarvelin (applicant).

France – Gregory Grefenstette. Exalead S.A.. 10, place de la Madeleine, 75008 Paris. Phone: +33 (0)1 55 35 26 26. Email: ggrefens@exalead.com

Germany – Norbert Fuhr. Faculty of Engineering Sciences, Department of Computational and Cognitive Sciences, University of Duisburg-Essen, Duisburg, 47048. Phone: +49 (0) 203 / 379 – 2524. Email: norbert.fuhr@uni-due.de

Ireland – Alan Smeaton. CLARITY: Centre for Sensor Web Technologies, Centre for Digital Video Processing and School of Computing Dublin City University, Glasnevin, Dublin 9. Phone: +353 - 1 – 7005262. Email: alan.smeaton@dcu.ie.

Italy – Nicola Ferro. Department of Information Engineering, University of Padova. Via Gradenigo, 6/a, 35131 Padova. Phone +39 049 827-7939. Email: ferro@dei.unipd.it.

Netherlands – Maarten de Rijke (applicant).

Norway – Nils Pharo. Oslo University College, Faculty of Journalism, Library and Information Science, Postboks 4 St. Olavs plass, 0130 Oslo. Phone: +47 22 45 26 84. Email: nils.pharo@jbi.hio.no.

Portugal – Mário J. Gaspar da Silva. Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa. Phone: +351-21-750-0128. Email: mjs@di.fc.ul.pt.

Spain – Julio Gonzalo. E.T.S.I. Informática de la UNED, c/ Juan del Rosal, 16 (Ciudad Universitaria), 28040 Madrid. Phone: +34 913987922. Email: Julio@ls.uned.es.

Sweden – Jussi Karlgren. Swedish Institute for Computer Science (SICS). Box 1263, SE-164 29 Kista. Phone: +46 (0)8 633 15 52 Email: jussi@sics.se.

Switzerland – Henning Müller. Business Information Systems, University of Applied Sciences Western Switzerland, Sierre (HES SO), TechnoArk 3, 3960 Sierre. Phone: +41 27 606 9036. Email: Henning.Mueller@hevs.ch.

United Kingdom – Mark Sanderson (applicant).

An **Executive Group** (consisting of five people, the three applicants plus two to be selected by and from the Steering Committee) will act between the annual meetings of the Steering Committee.

III.2 Programme Collaborations

The following list of research teams represent active European teams at CLEF 2009.

Austria – ICG, TU Graz; Knowledge Relationship Discovery Dept., Know-Center

Belarus – Informatics, National Academy of Sciences Belarus

Belgium – GIGA – Bionformatics and Modeling, Université Liége

Bosnia & Herzegovina – Electrotechnics, University of Banja Luka

Finland – The FIRE Research Group, Dept. of Information Studies and Interactive Media, University of Tampere

France – Synapse Developpment; Information Science, Université du Sud Toulon; Centre CEA de Saclay; Ecole Centrale de Lyon; Department of Computer Science, UPMC-LIP6; Lear team, INRI; LIMSI; Xerox Research; Laboratoire Informatique Grenoble; Computer Science and Image Processing, Université Jean Monnet

Germany – Department of Computer Science, Universität München; Information Systems and Semantic Web, Universität Koblenz; Information Science, Universität Hildesheim; Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt; Medical Informaitcs (IRMA), RWTH Aachen; AI Research (AGKI), Universität Koblenz-Landau; Inst. AIFB, Universität Karlsruhe; Dept. Interactive Media, Fraunhofer Inst. Telecommunications; Chemnitz TU; Computer Science and Engineering, Fernuniversität Hagen; Bauhaus Universität Weimar

Greece – Information Processing, Athens U. Economics and Business; Informatics, Alexander TEI Thessaloniki

Hungary – MTA SZTAKI, Computer and Automation Research Institute, Hungarian Academy of Sciences

Ireland – Computer Science, Trinity College Dublin; School of Computing, Dublin City University; Creative Language Systems, University College Dublin;

Italy – Computer Science, University Bari; Department of Mathematics and Computer Science, University of Udine; Imaging and Vision Lab, University Milano

Macedonia – Faculty Electrical Engineering & IT, UKIM

The Netherlands – ISLA, University of Amsterdam; Centre for Language and Speech Technology, Radboud University; Information and Communication, Delft University of Technology; Information Science, University of Groningen; Interactive Information Access, Center for Mathematics and Computer Science (CWI)

Poland – Informatics, Wroclaw University of Technology

Portugal – Informatics, University of Lisbon; Informatica, U, Evora; Electronics and Telematics Engineering, U. Aveiro; INESC;

Romania – Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi; Institutul de Cercetari pentru Inteligenta Artificiala, Academia Romana

Spain – Intelligent Systems, U. Jaen; Natural Language Engineering, U. Politecnica Valencia; TALP, UP Catalonia; Computer Science, U. de Catilla-L Mancha; REINA, U. Salamanca; Dept. Lienguatges, U. Alicante; Dept. Electronica y Computacion, Santiage de Compostela U.; NLP Group, UNED; Yahoo! Research Barcelona

Sweden **–** Swedish Institute for Computer Science

Switzerland – University Hospitals, Geneve; Computer Science, U. Neuchatel; IDIAP Research Institute; Information Systems, University Basel; University Geneva;

United Kingdom – Knowledge Media Institute, Open University; Information and communications, Manchester Metropolitan University; Electronics and Computer Science, University of Southampton; Computing, University of Surrey; Computing, University of Glasgow; Engineering and Information Science, University of Middlesex; Information Science, University of Sheffield; Information and Software Systems, University of Westminster; Computational Linguistics, University of Wolverhampton

---

III.3 Global dimension

The ELIAS Network Programme has a clear global dimension. The ELIAS activities and experiences will be shared and coordinated with organizers of benchmarking efforts for IASs outside Europe (TREC: Ellen Voorhees; NTCIR: Noriko Kando; FIRE: Prasenjit Majumder), with whom longstanding collaborations exist. However, no corresponding non-ESF proposal for funding is currently being prepared.

Section IV

CVs

IV.1 Curriculum vitae of Kalervo Järvelin

Address: Department of Information Studies and Interactive Media, University of Tampere, FIN-33014 University of Tampere, Finland.

Email: kalervo.jarvelin@uta.fi

Phone: +358 3 355 111/+358 50 547 9062

Web: http://www.uta.fi/~likaja

Date and place of birth: 1953, Sulkava, Finland

Areas of interest: Information Seeking, Information retrieval, Evaluation methodology.

Education: PhD in Information Studies, University of Tampere (1987)

Working experience: Associate Professor in Information Studies 1987 -; Professor, 1997-; Academy Professor, 2004-2009.

Academic activities: KJ's research covers information seeking and retrieval, and database management; and linguistic and conceptual methods in IR. He has authored some 250 scholarly publications and supervised fifteen doctoral dissertations. His H-index is 23 in Google Scholar (October 12, 2009).

He is principal investigator of numerous research projects funded by EU, industrial organizations and the Academy of Finland.

KJ has served the ACM SIGIR Conferences as a program committee member (1992-2005), Conference Chair (2002) and Program Co-Chair (2004, 2006). He is an Associate Editor of *Information Processing and Management* (USA).

Five selected publications

1. H. Keskustalo, K. Järvelin, A. Pirkola, T. Sharma and M. Lykke Nielsen. Test Collection-Based IR Evaluation Needs Extension Toward Sessions – A Case of Extremely Short Queries. *5th Asia Information Retrieval Symposium* (AIRS 2009), October 2009 (to appear)
2. H. Keskustalo, K. Järvelin, A. Pirkola and J. Kekäläinen. Intuition-Supporting Visualization of User's Performance Based on Explicit Negative Higher-Order Relevance. *31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR 2008), pages 675–682, July 2008.
3. K. Järvelin, S. Price, L. Delcambre and M. Nielsen. Discounted Cumulated Gain based Evaluation of Multiple-Query IR Sessions. *30th European Conference on Information Retrieval* (ECIR 2008), pages 4–15, April 2008. (Recipient of the ECIR 2008 Best Paper Award).
4. P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Dordrecht, The Netherlands: Springer, 2005.
5. K. Järvelin and J. Kekäläinen. Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems* (ACM TOIS) 20(4): 422–446, 2002

IV.2 Curriculum vitae of Maarten de Rijke (contact person)

Address: Intelligent Systems Lab Amsterdam, University of Amsterdam, Science Park 107, 1098 XG Amsterdam, The Netherlands.

Email: mdr@science.uva.nl

Phone: +31 20 525 5358/+31 6 51 938 523

Web: http://www.science.uva.nl/~mdr

Date and place of birth: 1961, Vlissingen, The Netherlands

Areas of interest: Information retrieval, Social media, Evaluation methodology, System deployment.

Education: PhD in Computer Science, University of Amsterdam (1993)

Working experience: Post-doctoral researcher (Center for Mathematics and Computer Science, 1994–1995); Warwick Research Fellow (University of Warwick, 1996–1997), Assistant professor (University of Amsterdam, 1998–2001), Associate professor (University of Amsterdam, 2001–2004), Full professor (University of Amsterdam, 2004–present).

Academic activities: Editor of *Foundations and Trends in Information Retrieval*, *Research on Language and Computation*, *The Information Retrieval Series*, and *Cambridge Studies in Natural Language Processing*.

Published over 400 papers, books and edited volumes. H-index in Google Scholar (October 7, 2009): 35. Has supervised close to 20 doctoral dissertations.

Principal investigator of numerous research projects funded by NWO, EU, industrial and governmental organizations.

Director of the Intelligent Systems Lab Amsterdam (ISLA) and of the Center for Creation, Content and Technology. In April 2009, a spin-off based on his research was launched by the University of Amsterdam.

Invited talks at 20+ conferences and workshops. Over a dozen invited tutorials at international conferences and summer schools

Five selected publications

1. K. Hofmann, K. Balog, T. Bogers, and M. de Rijke, Contextual Factors for Finding Similar Experts. *Journal of the American Society for Information Science and Technology*, 2010 (to appear).
2. L. Azzopardi, M. de Rijke, and K. Balog, Building Simulated Queries for Known-Item Topics: An Analysis using Six European Languages. In: *30th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR 2007)*, July 2007.
3. V. Jijkoun, M. Marx, M. de Rijke, and F. van Waveren, Electoral Search Using the VerkiezingsKijker: An Experience Report. In: *16th International World Wide Web Conference (WWW 2007)*, May 2007.
4. I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff, Overview of the TREC-2006 Blog Track. In: *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, NIST, February 2007.
5. G. Mishne and M. de Rijke. A Study of Blog Search. In: *Proceedings 28th European Conference on Information Retrieval (ECIR 2006)*, LNCS 3936, pages 289-301, April 2006.

IV.3 Curriculum vitae of Mark Sanderson

Address: Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello St, Sheffield, S1 4DP, UK.

Email: m.sanderson@shef.ac.uk

Phone: +44 114 22 22648

Web: http://dis.shef.ac.uk/mark/

Date and place of birth: 1966, St. Andrews, Scotland

Areas of interest: Evaluation of Information retrieval, Multimedia search, Cross language information retrieval.

Education: PhD in Computer Science, University of Glasgow (1996)

Working experience: Research assistant (University of Glasgow, 1988-1990); Post doctoral researcher (University of Glasgow, 1995-1997), Post doctoral researcher (Center for Intelligent Information Retrieval, University of Massachusetts, 1998–1999), Lecturer (University of Sheffield, 1999–2003), Senior Lecturer (University of Sheffield, 2004–2006), Reader (University of Sheffield, 2007–present).

Academic activities: Associate editor of 2 major journals, *ACM Transactions of Information Systems*, *ACM Transactions on the Web*; on the editorial board of *Information Retrieval* and *Information Processing & Management*. General Chair of ACM SIGIR 2004 and PC-Chair of ACM SIGIR 2009.

Published around 100 papers, books and edited volumes. H-index in Google Scholar (October 19, 2009): 23. Has or is in the process of supervising 8 doctoral dissertations.

Principal investigator of numerous research projects funded by EU, industrial and governmental organizations.

11 keynote or invited talks at major conferences or symposia. Six invited tutorials at international conferences and summer schools

Five selected publications

1. M. Sanderson, J. Tang, T. Arni and P. Clough. What else is there? Search Diversity Examined. 31st *European Conference on IR Research on Advances in Information Retrieval* (ECIR 2009), LNCS 5478, pages 562–569, 2009
2. S. Sedghi, M. Sanderson and P. Clough. A study on the relevance criteria for medical images. *Pattern Recognition Letters* 29(15): 2046–2057, 2008
3. M. Sanderson. Ambiguous Queries: Test Collections Need More Sense., *31st ACM SIGIR Conference* on Research *& Development on Information Retrieval* (SIGIR 2008), pages 499–506, 2008
4. A. Al-Maskari, M. Sanderson and P. Clough. The Good and the Bad System: Does the Test Collection Predict Users' Effectiveness? *31st ACM SIGIR Conference* on Research *& Development on Information Retrieval* (SIGIR 2008) pages 59–66, 2008
5. M. Sanderson and J. Zobel. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. *28th ACM SIGIR Conference* on Research *& Development on Information Retrieval* (SIGIR 2005), pages 162–169, 2005