INEX iTrack Revisited: Exploring the Potential for Re-use

Nils Pharo
Oslo Metropolitan University
Oslo, Norway
nilsp@oslomet.no

ABSTRACT

This paper presents the experiences from the INEX iTrack experiments conducted over a period of seven years. The purpose is to present the infrastructure of the experiments with the aim to identify its potential for re-use in new experiments. The paper discusses the terminology, research design, methodology, resources and reporting from the Inex iTrack in light of this.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; • **Human-centered computing → User studies**; **Empirical studies in HCI**.

KEYWORDS

Interactive information retrieval, methodology, open science

1 INTRODUCTION

The Initiative for Evaluation of XML retrieval (INEX) started in 2002 as a set of experiments following the Cranfield model. The purpose of INEX was initially to test the potential of XML elements as items for retrieval, as an alternative to full text documents, document parts and document passages. The INEX interactive track (iTrack) was run as a subtrack from 2004 to 2010 [3, 5, 8, 10–12], with the goal to study how end-users query, interact with, and evaluate documents and document parts. The iTrack was organized in a distributed way. Participating groups from universities and other research institutions across the world collected data following a standardised procedure for data collection in an experimental setting. In this way, it was possible to collect rather large data sets of user-system interaction.

In this paper we shall investigate the methodological approach used in INEX iTrack. The intention is to explore its potential for re-use and the experience that can be of value for establishing a common methodology for interactive information retrieval (IR) experiments. The paper is structured in the following way; the first part contains the method, we present iTrack infrastructure, i.e. the terminology, research design, methodology, resources and reporting used. Thereafter follows a discussion of challenges, before the final part with summary and conclusions.

2 METHOD

In order to identify the infrastructure of the INEX iTrack we investigate the reports published in the proceedings from 2004 to 2010. The structure of the iTrack reports was kept fairly consistent across the years. The experimental set-up included presentation of the tasks, the search system, the document corpus, and the procedure for data collection. In varying degree, results were presented in the proceedings report, some years the experiments had not ended at the time of proceedings report deadlines.

We do not report any of the findings, these can be found in the proceedings reports and a summary of the seven years of iTrack experiments [6].

Table 1: Consistent INEX iTrack terminology over time

Year/period	Common terms
2005-2010	Document corpus, relevance assessments, experimental procedure
2006-2010	Search system, logging
2008-2010	Tasks, participating groups

3 THE INEX ITRACK INFRASTRUCTURE

3.1 Terminology

During the iTrack years, the terminology used went through some changes. In particular, the first year (2004) stands out with an idiosyncratic terminology. Table 1 shows the distribution of central terms used over the period, compared according to their intended use, i.e. the concept (infrastructure element) they represent. This means, e.g., that from 2005 to 2010 the term "document corpus" was used consistently to refer to the collection of documents used in the experiments, whereas the term "Tasks" was used consistently from 2008 to 2010.

Table 2 provides an overview of central concepts, definitions, and the terminology where term use have changed over time. This does not represent an exhaustive overview, only concepts used over several years of experiments are included.

Although term use has changed over time, it is easy to identify the common infrastructure elements from the proceedings report. Most confusing is the different uses of the term "Task", which was used to refer to different experimental tasks in 2005 and 2006. In 2006, e.g., three different tasks were described as "Task A - Common Baseline System with IEEE Collection", "Task B - Participation with Own Element Retrieval System" and "Task C - Searching the Lonely Planet Collection", respectively.

3.2 Research design

The research design used in the iTrack experiments has been stable. A generic representation of the experimental procedure can be described in the following way:

Workshop on Barriers to Interactive IR Resources Re-use at the ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2019), 14 March 2019, Glasgow, UK

Table 2: INEX iTrack terminologial changes over time

Concept	Definition	Distribution
Task	The task(s) assigned to participants, what they are asked to find information about and its context	Topics (2004), tasks/topics (2005), search tasks (2006), tasks (2008-2010)
Search system	The system(s) designed to be used in the experiments	System (2004, 2005), search system (2006-2010)
Document corpus	The documents searchable in the search system	Document corpus (2005-2010)
Experimental procedure	The procedure used for performing the experiment	Experimental protocol (2004), experimental procedure (2005-2010)

(1) General questionnaire. The participant fills out a questionnaire on background knowledge, demographic data etc. Questionnaires were on paper (2004-2006) or online (2008-2010)
(2) Training task. The participant is given a training task to introduce them to the system's design and functionalities.
(3) Task 1
   (a) Task specific questionnaire. The participant fills out a questionnaire on task specific knowledge
   (b) Search session. The participant interacts with the system in order to perform the task.
   (c) Post task questionnaire. The participant fills out questionnaires related to the experience with the system, difficulty in solving the task etc.
(4) Additional tasks performed as described in step 3.
(5) Post experiment questionnaire. The participant fills out a questionnaire to provide feedback about the search system.

In addition to a common experimental procedure, the participating groups had the opportunity to perform their own experiments. In 2005 and 2006 it was explicitly organized so that research groups could use their own systems and compare their results to the system developed for the experiments as a baseline.

Very little analysis was performed as part of the iTrack work. Studies performed on iTrack data and reported in journal articles and conference proceedings papers have used transaction log analysis, statistical analysis of questionnaire data, screen capturing and eye-tracking. The studies have, e.g., investigated users preference with respect to element granularity [2, 4, 7] and the effect of task types on preferred elements [9].

3.3 Methodology

The initial purpose of the iTrack was twofold: "to investigate the behaviour of users when interacting with components of XML documents, and secondly to investigate and develop approaches for XML retrieval which are effective in user-based environments". In the first two years, the iTrack was closely connected with the INEX ad hoc-track, using the ad hoc-track's document corpus and topics/tasks. The tasks have been formulated as simulated work task situations [1] during the whole period. During the years, changes in methodology include changes in: document corpus, search systems, task types, relevance scales and analysis. Also the overall research questions have changed. Some examples of iTrack research questions are:

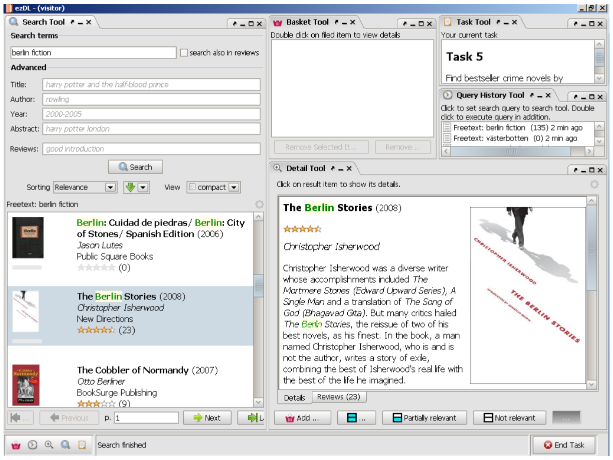
- What element types / level of granularity do searchers chose to see? In what sequence?
- How do users make use of document structure
  - in making relevance judgements?
  - in choosing level of granulaity to view?
- What level of element granularity constitutes the basis of a relevance decision? With what degree of certainty?
- How do factors such as topic knowledge influence
  - choice of element granularity?
  - number of elements viewed / amount read?
  - relevance judgements?

3.3.1 *Document corpus.* In 2004 and 2005 the corpus was a collection of journal articles published by IEEE (also used in other INEX tracks), in addition, a collection of Lonely Planet travel guides was used in 2005. In 2006 and 2008 the Wikipedia collection, consisting of more than 650 000 XML-formatted encyclopaedic articles, was used in the iTrack as well as other INEX tracks. In 2009 and 2010 a collection of Amazon and Librarything book reviews, was specifically collected for the iTrack. This collection has later been adopted by CLEF's Social Book Search Lab.

3.3.2 *Search system.* Several search systems were developed by the iTrack organizers. In 2004 and 2005 the HyREX retrieval engine [1] was used as backend in the baseline system. In 2006 two different backends were used to test the difference between passage and element retrieval, CSIRO's Panoptic/Funnelback platform as passage retrieval backend and TopX [2] from Max Planck Institute for Informatics for the element retrieval backend. In 2008 and 2009 a retrieval system built within the Daffodil framework developed at the University of Duisburg-Essen [3]) was used. In 2010 Daffodil was replaced with a system based on the ezDL framework [4]. The system interface design was quite consistent throughout the whole period. It was built within the Daffodil framework. In 2009-2010 the design consisted of three main components (see Figure 1): a query panel, a result list, and a window showing the details of the item retrieved from the result list. Previous years the document was shown in a separate interface.

3.3.3 *Task types.* Table 3 contains an overview of iTrack task categories. The iTrack experiments' task categories typically have changed from year to year with categories differing in complexity. In particular the 2006 tasks should be noted, where tasks were two-dimensional combining type and structure. This is an example of a 2006 fact-finding hierarchical task:
"A friend has just sent an email from an Internet café in the southern USA where she is on a hiking trip. She tells you that she has

[1] The system can be downloaded from http://www.is.informatik.uni-duisburg.de/projects/hyrex/.
[2] Only the TopX backend is available for download: http://topx.sourceforge.net/.
[3] more details are available on http://www.is.informatik.uni-duisburg.de/projects/daffodil/index.html
[4] More information on ezDL can be found on http://www.is.informatik.uni-duisburg.de/projects/ezdl/.



Search terms: berlin fiction
□ search also in reviews

Title: harry potter and the half-blood prince
Author: rowling
Year: 2000-2005
Abstract: harry potter london
Reviews: good introduction

Sorting: Relevance
View: compact

Freetext: berlin fiction

Berlin: Cuidad de piedras/ Berlin: City of Stones/ Spanish Edition (2006)
Jason Lutes
Public Square Books
(0)

The Berlin Stories (2008)
Christopher Isherwood
New Directions
(23)

The Cobbler of Normandy (2007)
Otto Berliner
BookSurge Publishing
(9)

p. 1

Search finished

Double click on filed item to view details

Your current task

Task 5

Find bestseller crime novels by

Click to set search query to search tool. Double click to execute query in addition.

Freetext: berlin fiction (135) 2 min ago
Freetext: västerbotten (0) 2 min ago

Click on result item to show its details.

The Berlin Stories (2008)

Christopher Isherwood

Christopher Isherwood was a diverse writer whose accomplishments included *The Mortmere Stories (Edward Upward Series)*, *A Single Man* and a translation of *The Song of God (Bhagavad Gita)*. But many critics hailed *The Berlin Stories*, the reissue of two of his best novels, as his finest. In the book, a man named Christopher Isherwood, who is and is not the author, writes a story of exile, combining the best of Isherwood's real life with the best of the life he imagined.

Figure 1: Inex iTrack 2009-2010 interface

just stepped into an anthill of small red ants and has a large number of painful bites on her leg. She wants to know what species of ants they are likely to be, how dangerous they are and what she can do about the bites. What will you tell her?"

The task types used in the 2010 iTrack was designed to simulate searchers at different stages of the search process, as defined by Kuhlthau. Below is an example of a 2010 explorative task:

"You are at an early stage of working on an assignment, and have decided to start exploring the literature of your topic. Your initial idea has led to one of the following three research needs:

(1) Find trustworthy books discussing the conspiracy theories which developed after the 9/11 terrorist attacks in New York.
(2) Find controversial books discussing the climate change and whether it is man-made or not.
(3) Find highly acclaimed novels that treat issues related to racial discrimination."

Semi self-selected tasks were used in 2009 and 2010. The participants were asked to "[t]ry to find books about a specific topic or of a certain type, but do not look for a specific title you already know."

3.3.4 *Relevance scales.*

A variety of relevance scales have been used in the iTrack. The complexity of the scales have varied a lot. In 2005, 2009 and 2010 a simple trinary relevance scale was used, the searchers were asked to assess elements as "relevant", "partially relevant" or "not relevant". In 2004 a ten point relevance scale was used:

A Very useful and Very specific
B Very useful and Fairly specific
C Very useful and Marginally specific
D Fairly useful and Very specific
E Fairly useful and Fairly specific
F Fairly useful and Marginally specific
G Marginally useful and Very specific
H Marginally useful and Fairly specific
I Marginally useful and Marginally specific
J Contains no relevant information

2004	Background Comparison	"Find background information about..." "Find differences between..."
2005	General/Challenging	The "general" challenges were designed as simpler than the "more complex" challenging tasks
2006	Types: Decision making; Fact finding; Information gathering Structure: Hierarchical; Parallel	The tasks were combined on two dimensions: type and structure.
2008	Fact finding/Research	The tasks were designed to represent information needs typical for Wikipedia users, finding facts, such as the "biggest airport" or perform research to write a paper.
2009	Broad/Narrow/Semi self-selected	Broad tasks represented needs that lead to thematic exploration. Narrow tasks represented relatively narrow topical information needs.
2010	Explorative/Data gathering/Semi self-selected	The tasks were designed to represent different stages in information seeking processes.

Table 3: iTrack task categories

In 2005 the author noted concerns that the 2004 scale "was far too complex for the test persons to comprehend", thus choosing the simple scale in 2005. In 2006 and 2008 a two-dimensional scale with five possible scores was used, with the following definitions: **Relevant, but too broad**, contains relevant information, but also a substantial amount of other information. **Relevant**, contains highly relevant information, and is just the right size to be understandable. **Relevant, but too narrow**, contains relevant information, but needs more context to be understood. **Partially relevant**, has enough context to be understandable, but contains only partially relevant information. **Not relevant**, does not contain any relevant information that is useful for solving the task.

3.3.5 *Analysis methods.*

iTrack data analysis has been performed using a combination of transaction logs and questionnaire data. Studies have been performed investigating the types of transactions taking place, typical transaction patterns, and factors influencing transaction patterns.

3.4 Resources

The INEX iTrack evolved from 2004 to 2010. In the first years, it complemented the research goals of the ad hoc-track, re-using topics, with some modifications, from the ad hoc-track, with the intention to identify how end-user react to element-based IR systems. The software used for the search system, which was developed at the University of Duisburg-Essen, gradually developed and interface design was kept consistent. Questionnaires were also kept fairly consistent, addressing the same background factors from year to year.

3.5 Reporting

The iTracks proceeding reports document the study design. The software is documented at the web sites. The questionnaires are not well documented. The biggest issue is the availability of transaction logs and questionnaire data. These are not openly available at the time of writing. The intention of the iTrack was that the data should be available only to the research groups for a limited period and then become available for others upon request. Unfortunately, the iTrack web sites are no longer available, which leaves us with the track reports as the main official documentation.

4 DISCUSSION

The experiences from the INEX iTrack have been manifold. With the collaborative effort of several research groups collecting data in a standardized manner, the iTrack resulted in large interactive IR datasets. The maximum number of participating research groups were 11 (in 2004 and 2005), with 119 searchers taking part in the 2005 experiment. The data can be compared across countries and, to a certain degree, across different user groups (although the majority of participants have, however, been students in computer science and library and information science). In addition, rich background data on many searchers have been collected.

The major challenges of the experiments are the design of tasks. These should be relevant for the participants and tailored following Borlund's simulated work task situation method [1]. This can be done either by agreeing upon a very specific user group to collect participants from or by making very generic tasks. To design realistic experiments we should also take into account that today's information searchers search all the time, in a fragmented way and on various platforms.

Other challenges include the identification of factors that influence interaction. We need to be able to identify the degree in which we can make valid analysis based on the data.

Specific challenges related to re-use and data sharing in interactive IR include establishing standardized ways of documenting experiments, which is what the BIIRRR workshop addresses. It is also necessary to establish a forum for discussions and coordination of IIR experiment efforts

5 SUMMARY AND FUTURE WORK

The INEX interactive track organized collaborative interactive information retrieval experiments from 2004 to 2010. In all, the iTrack initiated six rounds of experiments with changes in tasks, collections and search systems. The experiments resulted in data in the form of transaction logs and questionnaires. All experiments were documented in the INEX proceedings.

Although experiments evolved throughout the period, with significant impact on elements such as task types and relevance scales, the documentation is fairly consistent. The data are, however, at present not publicly available and the systems that were used are only partially available. This raises the following questions and challenges for securing re-use of Inex iTrack experiments, which will also be of value for resuse of interactive IR experiments in general:

- the need for a data repository for preservation of research designs, including transaction logs and questionnaires along with code books and necessary documentation for re-use
- a common repository for document corpuses and search systems
- a discussion on the need for standardized questions in questionnaires in order to compare across experiments

6 ACKNOWLEDGEMENTS

I would like to thank Norbert Fuhr and Thomas Beckers for valuable information about the current status of the iTrack systems and data.

REFERENCES

[1] Pia Borlund. 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. 8, 3 (2003). http://informationr.net/ir/8-3/paper152.html

[2] Barbara Hammer-Aebi, Kirstine Wilfred Christensen, Haakon Lund, and Birger Larsen. 2006. Users, structured documents and overlap: interactive searching of elements and the influence of context on search behaviour. In *Proceedings of the 1st international conference on Information interaction in context (IIiX)*. ACM, New York, NY, USA, 46–55. https://doi.org/10.1145/1164820.1164833

[3] Birger Larsen, Saadia Malik, and Anastasios Tombros. 2006. The interactive track at INEX 2005. In *Advances in XML Information Retrieval and Evaluation*, Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Gabriella Kazai (Eds.). Springer, Berlin, 398–410. http://dx.doi.org/10.1007/978-3-540-34963-1_30

[4] Birger Larsen, Anastasios Tombros, and Saadia Malik. 2006. Is XML retrieval meaningful to users?: searcher preferences for full documents vs. elements. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06)*. ACM, New York, NY, USA, 663–664. https://doi.org/10.1145/1148170.1148306

[5] Saadia Malik, Anastasios Tombros, and Birger Larsen. 2007. The Interactive Track at INEX 2006. In *Comparative Evaluation of XML Information Retrieval Systems*, Norbert Fuhr, Mounia Lalmas, and Andrew Trotman (Eds.). Vol. 4518. Springer, Berlin, 387–399. http://www.springerlink.com/content/d4rv145135659g38/

[6] Ragnar Nordlie and Nils Pharo. 2012. Seven Years of INEX Interactive Retrieval Experiments – Lessons and Challenges. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics (Lecture Notes in Computer Science)*, Tiziana Catarci, Pamela Forner, Djoerd Hiemstra, Anselmo Peñas, and Giuseppe Santucci (Eds.). Springer Berlin Heidelberg, 13–23.

[7] Nils Pharo. 2008. The effect of granularity and order in XML element retrieval. *Information Processing and Management* 44, 5 (Sept. 2008), 1732–1740. https://doi.org/10.1016/j.ipm.2008.05.004

[8] Nils Pharo, Thomas Beckers, Ragnar Nordlie, and Norbert Fuhr. 2011. Overview of the INEX 2010 Interactive Track. In *Comparative Evaluation of Focused Retrieval*, Shlomo Geva, Jaap Kamps, Ralf Schenkel, and Andrew Trotman (Eds.). Vol. 6932. Springer, Berlin, 227–235.

[9] Nils Pharo and Astrid Krahn. 2011. The effect of task type on preferred element types in an XML-based retrieval system. *Journal of the American Society for Information Science and Technology* 62, 9 (Sept. 2011), 1717–1726. https://doi.org/10.1002/asi.21587

[10] Nils Pharo, Ragnar Nordlie, and Khairun Nisa Fachry. 2009. Overview of the INEX 2008 Interactive Track. In *Advances in Focused Retrieval*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Shlomo Geva, Jaap Kamps, and Andrew Trotman (Eds.). Vol. 5631. Springer, Berlin, 300–313.

[11] Nils Pharo, Ragnar Nordlie, Norbert Fuhr, Thomas Beckers, and Khairun Nisa Fachry. 2010. Overview of the INEX 2009 Interactive Track. In *Focused Retrieval*

and Evaluation, Shlomo Geva, Jaap Kamps, and Andrew Trotman (Eds.). Vol. 6203. Springer, Berlin, 303–311.

[12] Anastasios Tombros, Birger Larsen, and Saadia Malik. 2005. The interactive track at INEX 2004. In *Advances in XML Information Retrieval*, Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Zoltán Szlávik (Eds.). Springer, Berlin, 410–423.