Overview of INEX 2006

Saadia Malik[1], Andrew Trotman[2], Mounia Lalmas[3], and Norbert Fuhr[1]

[1] University of Duisburg-Essen, Duisburg, Germany
{malik,fuhr}@is.informatik.uni-duisburg.de
[2] University of Otago, Dunedin, New Zealand
andrew@cs.otago.ac.nz
[3] Queen Mary, University of London, London, UK
mounia@dcs.qmul.ac.uk

Abstract. Since 2002, INEX has been working towards the goal of establishing an infrastructure, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of content-oriented XML retrieval systems. This paper provides an overview of the work carried out as part of INEX 2006.

1 Introduction

The continuous growth in XML[1] information repositories has been matched by increasing efforts in the development of XML retrieval systems (e.g. [1,2]), in large part aiming at supporting content-oriented XML retrieval. These systems exploit the available structural information, as marked up in XML, in documents, in order to return document components – the so-called XML elements – instead of complete documents in response to a user query. This is of particular benefit for information repositories containing long documents or documents covering a wide variety of topics (e.g. books, user manuals, legal documents), where users' effort to locate relevant content can be reduced by directing them to the most relevant parts of these documents. For example, in response to a user's query on a collection of scientific articles marked-up in XML, an XML retrieval system may return a mixture of paragraph, section, article, or other elements that have been estimated as best answers to the user's query. As the number of XML retrieval systems increases, so does the need to evaluate their effectiveness.

The INitiative for the Evaluation of XML retrieval (INEX)[2], which was set up in 2002, established an infrastructure and provided means, in the form of large test collections and appropriate scoring methods, for evaluating how effective content-oriented XML search systems are. As a result of a collaborative effort during the course of 2006, the INEX test collection has been further extended with the addition of the Wikipedia collection, new topics, and new assessments. Using the constructed test collection and the developed set of measures, the retrieval effectiveness of the participants' XML search engines were evaluated and their results compared.

This paper presents an overview of INEX 2006. Section 2 gives a brief summary of this year's participants. Section 3 provides an overview of the test collection. Section 4

[1] http://www.w3.org/XML/

[2] http://inex.is.informatik.uni-duisburg.de/

N. Fuhr, M. Lalmas, and A. Trotman (Eds.): INEX 2006, LNCS 4518, pp. 1–11, 2007.

outlines the retrieval tasks in the main ad hoc track. Section 5 reports some statistics of the submitted runs. Section 6 describes the relevance assessment phase. The different measures used to evaluate retrieval performance are described in a separate paper [6]. Section 7 provides a short description of the tracks at INEX 2006.

2 Participating Organizations

In reponse to the call for participation, issued in March 2006, 68 organizations registered. Throughout the year a number of groups dropped out due to resource requirements, while 23 groups joined later during the year. The final 50 active groups along with details of their participation is summarized in Table 1.

3 The Test Collection

Test collections consist of three parts: a set of documents, a set of information needs called topics and a set of relevance assessments listing the relevant documents for each topic. Although a test collection for XML retrieval consists of the same three parts, each component is rather different from its traditional information retrieval counterpart.

In traditional information retrieval test collections, documents are considered as units of unstructured text, queries are generally treated as bags of terms or phrases, and relevance assessments provide judgments whether a document as a whole is relevant to a query or not. XML documents organize their content into smaller, nested structural elements. Each of these elements in the document's hierarchy, along with the document itself (the root of the hierarchy), represent a retrievable unit. In addition, with the use of XML query languages, users of an XML retrieval system can express their information need as a combination of content and structural conditions, e.g. users can restrict their search to specific structural elements within the collection. Consequently, the relevance assessments for an XML collection must also consider the structural nature of the documents and provide assessments at different levels of the document hierarchy.

3.1 Documents

INEX 2006 uses a different document collection than in previous years [9], made from English documents from the Wikipedia[3]. The collection is made up of the XML full-texts of 659,388 articles of the Wikipedia project, covering a hierarchy of 113,483 categories, and totaling more than 60 Gigabytes (4.6 Gigabytes without images) and 30 million elements. The collection has a structure containing text, more than 300,000 images and some structured parts corresponding to the Wikipedia templates (about 5000 different tags). The collection has a structure similar to the IEEE collection, which was used up to 2005 in INEX. On average an article contains 161.35 XML nodes, where the average depth of an element is 6.72. For a detailed description of the document collection used for the ad hoc and other tracks at INEX 2006 see [3].

[3] http://en.wikipedia.org

Table 1. List of active INEX 2006 participants

Organizations	Submitted topics	Submitted runs	Assessed topics
Utrecht University	6	11	3
University of California, Berkeley	1	2	3
University of Otago	6	0	3
Queensland University of Technology	6	24	3
Queen Mary University of London	4	12	3
Ecoles des Mines de Saint-Etienne	6	9	3
University of Granada	6	2	3
Indian Statistical Institute	0	2	3
University of Tampere	6	0	3
La Trobe University	6	0	3
University of Kaiserslautern,	6	24	3
City University London	6	13	3
RMIT University	6	12	3
IRIT	9	23	3
Max-Planck-Institut fuer Informatik	6	19	3
University of Cambridge	6	0	3
CSIRO	4	8	3
University of Wollongong in Dubai	5	7	3
University of Amsterdam	8	13	3
Fondazione Ugo Bordoni	6	6	3
The Hebrew University of Jerusalem	6	24	3
Royal School of LIS	6	0	3
University of Toronto	6	1	3
Universität Duisburg-Essen	2	0	1
Oslo University College	3	7	3
University of Waterloo	0	0	3
University of Massachusetts Amherst	6	0	3
Kyungpook National University	0	0	3
University of Rostock	6	3	3
LIP6	5	12	3
CWI and University of Twente	6	23	3
University of Helsinki	4	3	3
The Robert Gordon University	6	6	3
IBM Haifa Research Lab	0	18	3
LIPN	1	0	3
CLIPS-IMAG	6	0	3
Université de Saint-Etienne	6	3	3
Justsystem Corporation	0	12	3
University of South-Brittany	0	20	3
Joint Research Centre	0	0	3
University of Minnesota Duluth	6	14	3
Huazhong University of Science & Technology	0	0	3
Dalhousie University	0	0	3
University College of Boras	0	0	3
Université Libre de Bruxelles	0	0	3
Universidad de Chile			
Organizations participated only in XML document mining track			
INRIA			
Western Kentucky University			
University of Wolongong			
Organization participated only in interactive track			
Rutgers University			

3.2 Topics

Querying XML documents with respect to content can be with or without respect to structure. Taking this into account, INEX identifies two types of topics:

Table 2. Statistics on CO+S topics on the INEX 2006 test collection

	CO+S
No. of topics	125
Average length of title (in words)	4.2
Use of boolean operators (and/or) in title	14
Use of (+/-) in title	61
Use of phrases in title	120
Use of boolean operators (and/or) in castitle	65
Use of (+/-) in castitle	49
Use of phrases in castitle	120
Average length of narrative (in words)	94
Average length of topic description (in words)	14
Average length of topic ontopic_keywords (in words)	6

- Content-only (CO) topics are requests that do not include reference to the document structure. They are, in a sense, the traditional topics used in information retrieval test collections. In XML retrieval, the results to such topics can be elements of various complexity, e.g. at different levels of the XML documents' structure.
- Content-and-structure (CAS) topics are requests that contain conditions referring both to content and structure of a document. These conditions may refer to the content of specific elements (e.g. the elements to be returned must contain a section about a particular topic), or may specify the type of the requested answer elements (e.g. sections should be retrieved).

In previous years a distinction was made between CO and CAS topics. Topic were also designed for use in multiple tracks (such as the natural language track and interactive track) and so contained multiple variant queries for each purpose. Since 2006, these have all been combined into a single topic type: the Content Only + Structure (CO+S) topic. All the information needed by the different tasks and tracks are expressed in each topic, but in different parts of that topic.

Topic Format. Topics are made up of several parts; these parts explain the same information need, but for different purposes.

<narrative>:A detailed explanation of the information need and the description of what makes an element relevant or not. The <narrative> explains not only what information is being sought, but also the context and motivation of the information need, i.e., why the information is being sought and what work-task it might help to solve. Assessments are made on compliance to the <narrative> alone.

<title>: A short explanation of the information need. It serves as a summary of the content of the user's information need. A word in the <title> can have a + or − prefix, where + is used to emphasize an important concept, and − is used to denote an unwanted concept.

<castitle>: A short explanation of the information need, specifying any structural requirements. As with a topic <title>, a word in the <castitle> can have a + or − prefix, where + is used to emphasize an important concept, and − is used to denote an unwanted concept. The <castitle> is expressed in the NEXI query language [14].

<description>: A brief description of the information need written in natural language – used in the natural language track. The description is as precise, concise, and as informative as the <title> and <castitle> combined.

<ontopic_keywords>: Terms and phrases that are likely to appear in most relevant documents. For example, if the user is searching for information about element retrieval and the query has the <title> "INEX" then <ontopic_keywords> might be: "element, XML".

The DTD of the topics is shown in Figure 1. The attributes of a topic are: topic_id (which in INEX 2006 ranges from 289 to 413) and ct_no, which refers to the candidate topic number (ranging from 1 to 218[4]). An example topic can be seen in Figure 2.

```
<!ELEMENT inex_topic  (title,castitle?,description,
    narrative,ontopic_keywords)>
<!ATTLIST inex_topic
  topic_id     CDATA     #REQUIRED
  ct_no        CDATA     #REQUIRED
>
<!ELEMENT title             (#PCDATA)>
<!ELEMENT castitle          (#PCDATA)>
<!ELEMENT description       (#PCDATA)>
<!ELEMENT narrative         (#PCDATA)>
<!ELEMENT ontopic_keywords  (#PCDATA)>
```

Fig. 1. Topic DTD

Topics were created by participating groups. Each participant was asked to submit up to 6 candidate topics. A detailed guideline was provided to the participants for the topic creation [10]. Several steps were identified for this process: 1) initial topic statement creation, 2) exploration phase, 3) topic refinement, and 4) topic selection. The first three steps were performed by the participants themselves while the selection of topics was performed by the INEX organizers.

During the first step, participants created their initial topic statement. These were treated as a user's description of their information need and were formed without regard to system capabilities or collection peculiarities to avoid artificial or collection biased queries. During the collection exploration phase, participants estimated the number of relevant results to their candidate topics. The TopX XML retrieval system [12] was provided to participants to help with this task. Participants were asked to judge the top 25 retrieved results and record for each found relevant result its file name and its XPath. Those topics having at least 2 relevant results but less than 20 results were to be submitted as candidate topics. In the topic refinement stage, the topics were finalised ensuring coherency and that each part of the topic could be used in stand-alone fashion.

[4] This number is exceeding the total candidate topic number (203) due to the deletion of some candidate topics by topic authors.

```
<inex_topic topic_id="408"  ct_no="202">

<title>
"electroconvulsive therapy" depression
</title>

<castitle>
//*[about(.,"electroconvulsive therapy" depression)]
</castitle>

<description>
Find me information about the treatment of depression with
electroconvulsive therapy
</description>

<narrative>
An old friend of mine suffers from depressions. Usually
medication keeps him well, but  occasionally he is admitted
to hospital with heavy depressions. The treatments often
involve shock therapy (electroconvulsive therapy or ECT).
I am worried about the long term effect of ECT and would
like to learn why passing an electrical current through
the brain can help cure depressions, how it  works, and
if there are any alternatives. Relevant elements will
discuss one of these issues. Elements  that deal with ECT
for other mental illnesses than depresseion are not relevant.
The purpose of the  search is to find information that will
make me better capable of understanding my friends illness.
</narrative>

<ontopic_keywords>
ECT, electroshock, "induced convulsion",  seizure
</ontopic_keywords>

</inex_topic>
```

Fig. 2. A CO+S topic from the INEX 2006 test collection

After the completion of the first three stages, topics were submitted to INEX. A total of 203 candidate topics were received, of which 125 topics were selected. The topic selection was based on a combination of criteria such as 1) balancing the number of topics across all participants, 2) eliminating topics that were considered too ambiguous or too difficult to judge, 3) uniqueness of topics, 4) considering their suitability to the different tracks, and 5) syntactic correctness.

4 Retrieval Tasks

The retrieval task to be performed by the participating groups of INEX 2006 is defined as the ad hoc retrieval of XML elements. In information retrieval literature [15], ad hoc retrieval is described as a simulation of how a library might be used and involves the searching of a static set of documents using a new set of topics. Here the collection consists of XML documents composed of different granularities of nested XML elements, each of which represents a possible unit of retrieval. The user's query may also contain structural constraints or hints in addition to the content conditions. In addition, the output of an XML retrieval system may follow the traditional ranked list presentation, or may extend to non-linear forms, such as grouping of elements per document.

Within the ad hoc XML retrieval task, four sub-tasks were defined based on the different assumptions regarding a search system's output and learning aims.

4.1 Thorough Task

The core system task underlying most XML retrieval strategies is the ability to estimate the relevance of retrievable elements in the collection. Hence, the thorough task asks systems to return elements ranked by their relevance to the topic of request. Since the retrieved elements are meant for further processing (either by a dedicated interface, or by other tools) there are no display-related assumptions nor user-related assumptions underlying the task.

The aims for this task included establishing: How good systems are at estimating the relevance of XML elements, how well systems can locate all the relevant elements in the collection, and how much structural constraints improve retrieval.

4.2 Focused Task

The scenario underlying this task is the return, to the user, of a ranked list of elements for the topic of request. The task requires systems to find the most focused elements that satisfy an information need, without returning "overlapping" elements (e.g. a paragraph and its container section). That is, for a given topic, elements in the result list may not contain text already contained in previous element. The task is similar to the thorough task in that it requires a ranking of XML elements, but here systems are required not only to estimate the relevance of elements, but also to decide which elements are the most focused non-overlapping.

The learning aims for this task include establishing: how the focused task differs from the thorough task, if the focused task can be reduced to a straightforward filter on the thorough task, which techniques are effective at early ranks, and how structural constraints help retrieval.

4.3 Relevant in Context Task

The scenario underlying this task is the return of relevant information (captured by a set of elements) within the context of the full article. A result, an article devoted to the topic of request, will contain a lot of relevant information across many elements. The task requires systems to find a set of elements that corresponds to (all) relevant information in each article. The set of result elements should not contain overlaps.

The learning aims for this task include establishing: how the relevant in context task differs from the thorough and focused tasks, which techniques are effective at locating relevance within an article, and how structural constraints help retrieval.

4.4 Best in Context Task

The scenario underlying this task is finding the best entry point from which to start reading a relevant document. Even a document completely devoted to the topic of request will only have one best starting point to read, even if this is the start of the document. This task requires systems to find the XML elements that correspond to these best entry points.

The learning aims for this task include establishing: how the best in context task differs from the relevant in context task, how best entry points relate to the relevance of elements, and how structural constraints help retrieval.

5 Submissions

Participating organizations evaluated the 125 INEX 2006 topics against the Wikipedia document collection and produced an ordered list of XML elements as the retrieval results for each topic. Participants could use either the <title> or <castitle> of the CO+S topics. The top 1500 elements of each topic's retrieval results were then submitted to INEX. For each topic, around 500 articles along with their elements were pooled from all the submissions in a round robin fashion for assessment. Table 3 shows the pooling effect on the CO+S topics.

Table 3. Pooling effect for CO+S topics

	CO+S topics
number of documents submitted	126111
number of documents in pools	63684
number of elements submitted	281761
number of elements in pools	137559

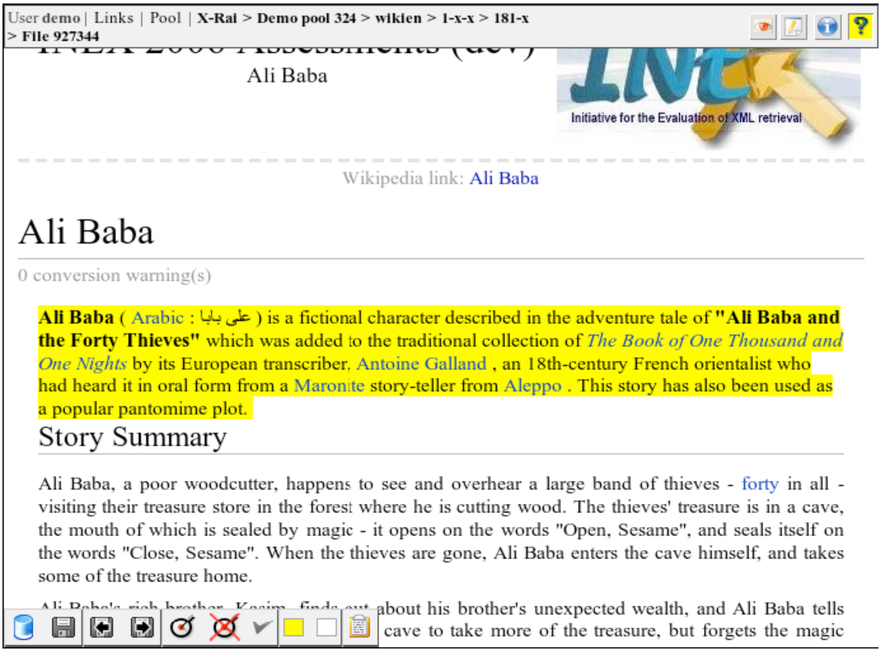
Table 4. Number of runs submitted to the four ad hoc tasks

Tasks	runs
Thorough	106
Focused	85
Relevant in context	65
Best in context	77

6 Assessments

Relevance in INEX is defined according to the notion of specificity, which is the extent to which an element focuses on the topic. Previously, INEX used a more complex definition of relevance but a number of studies showed that specificity alone was sufficient to determine an unambigious rank order of search systems with respect to their effectiveness (see, for example, [11]).

The specificity of an element is determined by an assessor using the highlighting method introduced at INEX 2005. In this approach the specificity of any (partially highlighted) elements can be calculated automatically as some function of the contained relevant and irrelevant content (for example the ratio of one to the other). Specificity is, thus, measured on a continuous scale in the range $[0, 1]$, where 1 represents a fully specific (relevant) element and 0 a non relevant element.


User demo | Links | Pool | X-Rai > Demo pool 324 > wikien > 1-x-x > 181-x > File 927344
Ali Baba
Initiative for the Evaluation of XML retrieval

Wikipedia link: Ali Baba

Ali Baba

0 conversion warning(s)

Ali Baba (Arabic : على بابا) is a fictional character described in the adventure tale of **"Ali Baba and the Forty Thieves"** which was added to the traditional collection of *The Book of One Thousand and One Nights* by its European transcriber, Antoine Galland , an 18th-century French orientalist who had heard it in oral form from a Maronite story-teller from Aleppo . This story has also been used as a popular pantomime plot.

Story Summary

Ali Baba, a poor woodcutter, happens to see and overhear a large band of thieves - forty in all - visiting their treasure store in the forest where he is cutting wood. The thieves' treasure is in a cave, the mouth of which is sealed by magic - it opens on the words "Open, Sesame", and seals itself on the words "Close, Sesame". When the thieves are gone, Ali Baba enters the cave himself, and takes some of the treasure home.

Ali Baba's rich brother, Kasim, finds out about his brother's unexpected wealth, and Ali Baba tells
cave to take more of the treasure, but forgets the magic

Fig. 3. X-RAI Article view

Assessment was done using an online assessment tool developed specifically for this purpose (see Figure 3). A relevance assessment guide [8] explaining how to assess relevance was distributed to the participants. In short, the assessors had only to highlight relevant text fragments from articles identified as candidates using the pooling strategy. These highlighted passages were then automatically converted into element specificity scores. Assessors were also asked to identify best entry points, one per relevant article.

7 INEX 2006 Tracks

In addition to the main ad hoc track, six research tracks were included, each studying a different aspect of XML information retrieval: Interactive, Relevance Feedback, Heterogeneous, Natural Language Processing, Multimedia, and Document Mining. Two new tracks were added in 2006: Use Case and Entity ranking.

In its fourth year, the **Interactive Track** (iTrack) put emphasis on comparing XML element retrieval with passage retrieval, and on investigating differences between multiple dimensions of the search tasks. This year the track required substantial work and data collection, so was not completed before the INEX 2006 workshop. The track continues into 2007 [7].

The **Relevance Feedback track** investigated approaches to relevanc feedback that considered the structural hints. To limit the number of submissions a subset of ad hoc track tasks were chosen for participants to test their algorithms. These include the thorough task with CO and CAS topics. The reported evaluation score for each relevance feedback submission measures the relative and absolute improvement of the relevance feedback run over the original base run and the significance level under the t-test and the Wilxocon signed-rank test.

The **Heterogeneous Collection track** was setup to cope with the challenges posed by heterogenous collections, which are syntactic (collections based on different DTDs), semantic (collections covering diverse topics) and genre (different document types) in heterogeneity. This year the track focused on finalising the heterogeneous collection and on topic definition. Based on this, the track will continue in 2007 with the evaluation of submitted runs. This year's track details can be found in [5].

The **Natural Language Processing (NLP) track** focused on whether it is possible to express topics in natural language, to be then used as a basis for retrieval. Two tasks were defined NLQ2NEXI and NLQ. NLQ2NEXI requires the translation of a natural language query, provided in the <description> element of a topic, into a formal INEX <castitle> element. The NLQ task has no restrictions on the use of any NLP technique to interpret the queries as they appear in the <description> element of a topic.The objective is not only to compare between different NLP based systems, but to also compare the results obtained with natural language queries with the results obtained with NEXI queries by any other system in the ad hoc track. During the topic creation stage, it was ensured that the description component of the topics were equivalent in meanings to their corresponding NEXI title, so it was possible to re-use the same topics, relevance assessments and evaluation procedures as in the ad hoc track. The descriptions were used as input to natural language processing tools, which would process them into representations suitable for XML search engines.

The main objective of the **Multimedia track** was to provide an evaluation platform for structured document access systems that do not only include text in the retrieval process, but also other types of media, such as images, speech, and video. Full details of the track can be found in [16].

The aim of the **Document Mining track**, run in collaboration with the PASCAL network of Excellence[5], was to develop machine learning methods for structured data mining and to evaluate these methods for XML document mining tasks. Full details of the track can be found in [4].

The aim of the **Use Case track** was to identify the potential users, scenarios and use-cases of XML retrieval systems. As a result, commercial XML search engines have

[5] http://www.pascal-network.org/

now been identified. XML (and other semi-structured formats) are being used behind the scenes in some on-line search engines without user knowledge. Book search engines that follow a thorough retrieval strategy have also been identified [13].

The aim of the **Entity Ranking track** was to examine list completion and associative ranking. In the former, entities of the same kind as those in a given list were to be extracted from the document collection. In the latter, a similar list to a given list, but on a different topic, was to be extracted from the document collection. Guidelines were drafted, a wide range of potential participants actively approached us, but this did not materialize into sufficient support to run a full track. Given the interest in the track at the INEX workshop, the track will continue in 2007.

References

1. Baeza-Yates, R., Fuhr, N., Maarek, Y.: XML and information retrieval, SIGIR workshop SIGIR Forum, vol. 36(2) (2002)
2. Blanken, H., Grabs, T., Schek, H.-J., Schenkel, R., Weikum, G. (eds.): Intelligent Search on XML Data, Applications, Languages, Models, Implementations, and Benchmarks (2003)
3. Denoyer, L., Gallinari, P.: Wikipedia XML corpus at INEX 2006. In: INEX 2006 Proceedings (2007)
4. Denoyer, L., Gallinari, P.: Report on the XML Mining Track at INEX 2005 and INEX 2006 - Categorization and Clustering of XML Documents. In: INEX 2006 Proceedings (2007)
5. Frommholz, I., Larson, R.: The heterogeneous collection track at INEX 2006. In: INEX 2006 Proceedings (2007)
6. Lalmas, M., Kazai, G., Kamps, J., Pehcevski, J., Piwowarski, B., Robertson, S.: INEX 2006 Evaluation Measures. In: INEX 2006 Proceedings (2007)
7. Larsen, B., Malik, S., Tombros, A.: The interactive track at INEX 2006. In: INEX 2006 Proceedings (2007)
8. Lalmas, M., Piwowarski, B.: INEX 2006 relevance assessment guide. In: INEX 2006 Preroceedings (2006)
9. Lalmas, M., Tombros, A.: INEX 2002 - 2006: Understanding XML Retrieval Evaluation. In: DELOS Conference on Digital Libraries, Tirrenia, Pisa, Italy (2007)
10. Larsen, B., Trotman, A.: INEX 2006 guidelines for topic development. In: INEX 2006 Preproceedings (2006)
11. Ogilvie, P., Lalmas, M.: Investigating the exhaustivity dimension in content-oriented XML element retrieval evaluation. In: CIKM (2006)
12. Theobald, M., Schenkel, R., Weikum, G.: An efficient and versatile query engine for topx search. In: VLDB, pp. 625–636. ACM, New York (2005)
13. Trotman, A., Pharo, N., Lehtonen, M.: XML-IR users and use cases. In: INEX 2006 Proceedings (2007)
14. Trotman, A., Sigurbjornsson, B.: Narrowed extended XPATH I (NEXI). In: INEX 2004 Proceedings (2005)
15. Voorhees, E., Harman, D. (eds.): The Tenth Text REtrieval Conference (TREC 2001) (2001)
16. Westerveld, T., van Zwol, R.: The INEX 2006 multimedia track. In: INEX 2006 Proceedings (2007)