Overview of the INEX 2007 Ad Hoc Track

Norbert Fuhr[1], Jaap Kamps[2], Mounia Lalmas[3], Saadia Malik[1], and Andrew Trotman[4]

[1] University of Duisburg-Essen, Duisburg, Germany
{norbert.fuhr,saadia.malik}@uni-due.de
[2] University of Amsterdam, Amsterdam, The Netherlands
kamps@science.uva.nl
[3] Queen Mary, University of London, London, UK
lalmas@dcs.qmul.uk.ac
[4] University of Otago, Dunedin, New Zealand
andrew@cs.otago.ac.nz

Abstract. This paper gives an overview of the INEX 2007 Ad Hoc Track. The main purpose of the Ad Hoc Track was to investigate the value of the internal document structure (as provided by the XML mark-up) for retrieving relevant information. For this reason, the retrieval results were liberalized to arbitrary passages and measures were chosen to fairly compare systems retrieving elements, ranges of elements, and arbitrary passages. The INEX 2007 Ad Hoc Track featured three tasks: For the *Focused Task* a ranked-list of non-overlapping results (elements or passages) was needed. For the *Relevant in Context Task* non-overlapping results (elements or passages) were returned grouped by the article from which they came. For the *Best in Context Task* a single starting point (element start tag or passage start) for each article was needed. We discuss the results for the three tasks, examine the relative effectiveness of element and passage retrieval. This is examined in the context of content only (CO, or Keyword) search as well as content and structure (CAS, or structured) search.

1 Introduction

This paper gives an overview of the INEX 2007 Ad Hoc Track. The main research question underlying the Ad Hoc Track is that of the value of the internal document structure (mark-up) for retrieving relevant information. That is, does the document structure help in identify where the relevant information is within a document? This question has recently attracted a lot of attention. Trotman and Geva [13] argued that, since INEX relevance assessments are not bound to XML element boundaries, retrieval systems should also not be bound to XML element boundaries. Their implicit assumption is that a system returning passages is at least as effective as a system returning XML elements. This assumption is based on the observation that elements are of a lower granularity than passages and so all elements can be described as passages. The reverse, however is not

N. Fuhr et al. (Eds.): INEX 2007, LNCS 4862, pp. 1–23, 2008.

true and only some passages can be described as elements. Huang et al. [5] implement a fixed window passage retrieval system and show that a comparable element retrieval ranking can be derived. In a similar study, Itakura and Clarke [6] show that although ranking elements based on passage-evidence is comparable, a direct estimation of the relevance of elements is superior. Finally, Kamps and Koolen [7] study the relation between the passages highlighted by the assessors and the XML structure of the collection directly, showing reasonable correspondence between the document structure and the relevant information.

Up to now, element and passage retrieval approaches could only be compared when mapping passages to elements. This may significantly affect the comparison, since the mapping is non-trivial and, of course, turns the passage retrieval approaches effectively into element retrieval approaches. To study the value of the document structure through direct comparison of element and passage retrieval approaches, the retrieval results for INEX 2007 were liberalized to arbitrary passages. Every XML element is, of course, also a passage of text.

The evaluation measures are now based directly on the highlighted passages, or arbitrary best-entry points, as identified by the assessors. As a result it is now possible to fairly compare systems retrieving elements, ranges of elements, or arbitrary passages. These changes address earlier requests to liberalize the retrieval format to ranges of elements [1] and later requests to liberalize to arbitrary passages of text [13].

The INEX 2007 Ad Hoc Track featured three tasks:

1. For the *Focused Task* a ranked-list of non-overlapping results (elements or passages) must be returned. It is evaluated at early precision relative to the highlighted (or believed relevant) text retrieved.
2. For the *Relevant in Context Task* non-overlapping results (elements or passages) must be returned, these are grouped by document. It is evaluated by mean average generalized precision where the generalized score per article is based on the retrieved highlighted text.
3. For the *Best in Context Task* a single starting point (element's starting tag or passage offset) per article must be returned. It is also evaluated by mean average generalized precision but with the generalized score (per article) based on the distance to the assessor's best-entry point.

The *Thorough Task* as defined in earlier INEX rounds is discontinued. We discuss the results for the three tasks, giving results for the top 10 participating groups and discussing the best scoring approaches in detail. We also examine the relative effectiveness of element and passage runs, and with content only (CO) queries and content and structure (CAS) queries.

The rest of the paper is organized as follows. First, Section 2 describes the INEX 2007 Ad Hoc retrieval tasks and measures. Section 3 details the collection, topics, and assessments of the INEX 2007 Ad Hoc Track. In Section 4, we report the results for the Focused Task (Section 4.2); the Relevant in Context Task (Section 4.3); and the Best in Context Task (Section 4.4). Section 5 details particular types of runs (such as CO versus CAS, and element versus passage),

and on particular subsets of the topics (such as topics with a non-trivial CAS query). Finally, in Section 6, we discuss our findings and draw some conclusions.

2 Ad Hoc Retrieval Track

In this section, we briefly summarize the ad hoc retrieval tasks and the submission format (especially how elements and passages are identified). We also summarize the metrics used for evaluation. For more detail the reader is referred to the formal specification documents [2] and [9].

2.1 Tasks

Focused Task. The scenario underlying the Focused Task is the return, to the user, of a ranked list of elements or passages for their topic of request. The Focused Task requires systems to find the most focused results that satisfy an information need, without returning "overlapping" elements (shorter is preferred in the case of equally relevant elements). Since ancestors elements and longer passages are always relevant (to a greater or lesser extent) it is a challenge to chose the correct granularity.

The task has a number of assumptions:

Display the results are presented to the user as a ranked-list of results.
Users view the results top-down, one-by-one.

Relevant in Context Task. The scenario underlying the Relevant in Context Task is the return of a ranked list of articles and within those articles the relevant information (captured by a set of non-overlapping elements or passages). A relevant article will likely contain relevant information that could be spread across different elements. The task requires systems to find a set of results that corresponds well to all relevant information in each relevant article. The task has a number of assumptions:

Display results will be grouped per article, in their original document order, access will be provided through further navigational means, such as a document heat-map or table of contents.
Users consider the article to be the most natural retrieval unit, and prefer an overview of relevance within this context.

Best in Context Task. The scenario underlying the Best in Context Task is the return of a ranked list of articles and the identification of a best-entry-point from which a user should start reading each article in order to satisfy the information need. Even an article completely devoted to the topic of request will only have one best starting point from which to read (even if that is the beginning of the article). The task has a number of assumptions:

Display a single result per article.
Users consider articles to be natural unit of retrieval, but prefer to be guided to the best point from which to start reading the most relevant content.

2.2 Submission Format

Since XML retrieval approaches may return arbitrary results from within documents, a way to identify these nodes is needed.

XML element results are identified by means of a file name and an element (node) path specification. File names in the Wikipedia collection are unique so that (with the .xml extension removed), for example:

```
<file>9996</file>
```

identifies 9996.xml as the target document from the Wikipedia collection. Element paths are given in XPath, but only fully specified paths are allowed. For example:

```
<path>/article[1]/body[1]/section[1]/p[1]</path>
```

identifies the first "article" element, then within that, the first "body" element, then the first "section" element, and finally within that the first "p" element. Importantly, XPath counts elements from 1 and counts element types. For example if a section had a title and two paragraphs then their paths would be: `title[1]`, `p[1]` and `p[2]`.

A result element, then, is identified unambiguously using the combination of file name and element path, for example:

```
<result>
  <file>9996</file>
  <path>/article[1]/body[1]/section[1]/p[1]</path>
  <rsv>0.9999</rsv>
</result>
```

Passages are given in the same format, but extended for optional character-offsets. As a passage need not start and end in the same element, each is given separately. The following example is equivalent to the element result example above since it starts and ends on an element boundary.

```
<result>
  <file>9996</file>
  <passage start="/article[1]/body[1]/section[1]/p[1]"
     end="/article[1]/body[1]/section[1]/p[1]"/>
  <rsv>0.9999</rsv>
</result>
```

In the next passage example the result starts 85 characters after the start of the paragraph and continues until 106 characters after a list item in list. The end location is, of course, after the start location.

```
<result>
  <file>9996</file>
  <passage start="/article[1]/body[1]/section[1]/p[1]/text()[1].85"
     end="/article[1]/body[1]/section[1]/normallist[1]/item[2]/text()[2].106"/>
  <rsv>0.6666</rsv>
</result>
```

The result can start anywhere in any text node. Character positions count from 0 (before the first character) to the *node-length* (after the last character). A detailed example is provided in [2].

2.3 Measures

We briefly summarize the main measures used for the Ad Hoc Track (see Kamps et al. [9] for details). The main change at INEX 2007 is the inclusion of arbitrary passages of text. Unfortunately this simple change has necessitated the deprecation of element-based metrics used in prior INEX campaigns because the "natural" retrieval unit is no longer an element, so elements cannot be used as the basis of measure. We note that properly evaluating the effectiveness in XML-IR remains an ongoing research question at INEX.

The INEX 2007 measures are solely based on the retrieval of highlighted text. We simplify all INEX tasks to highlighted text retrieval and assume that systems return all, and only, highlighted text. We then compare the characters of text retrieved by a search engine to the number and location of characters of text identified as relevant by the assessor. For best in context we use the distance between the best entry point in the run to that identified by an assessor.

Focused Task. Recall is measured as the fraction of all highlighted text that has been retrieved. Precision is measured as the fraction of retrieved text that was highlighted. The notion of rank is relatively fluid for passages so we use an interpolated precision measure which calculates interpolated precision scores at selected recall levels. Since we are most interested in what happens in the first retrieved results, the INEX 2007 official measure is interpolated precision at 1% recall (iP[0.01]). We also present interpolated precision at other early recall points, and (mean average) interpolated precision over 101 standard recall points (0.00, 0.01, 0.02, ..., 1.00) as an overall measure.

Relevant in Context Task. The evaluation of the Relevant in Context Task is based on the measures of generalized precision and recall [10], where the per document score reflects how well the retrieved text matches the relevant text in the document. Specifically, the per document score is the harmonic mean of precision and recall in terms of the fractions of retrieved and highlighted text in the document. We are most interested in overall performances so the main measure is mean average generalized precision (MAgP). We also present the generalized precision scores at early ranks (5, 10, 25, 50).

Best in Context Task. The evaluation of the Best in Context Task is based on the measures of generalized precision and recall where the per document score reflects how well the retrieved entry point matches the best entry point in the

document. Specifically, the per document score is a linear discounting function of the distance d (measured in characters)

$$\frac{n - d(x, b)}{n}$$

for $d < n$ and 0 otherwise. We use $n = 1,000$ which is roughly the number of characters corresponding to the visible part of the document on a screen. We are most interested in overall performance, and the main measure is mean average generalized precision (MAgP). We also show the generalized precision scores at early ranks (5, 10, 25, 50).

3 Ad Hoc Test Collection

In this section, we discuss the corpus, topics, and relevance assessments used in the Ad Hoc Track.

3.1 Corpus

The document collection was the Wikipedia XML Corpus based on the English Wikipedia in early 2006 [3]. The Wikipedia collection contains 659,338 Wikipedia articles. On average an article contains 161 XML nodes, where the average depth of a node in the XML tree of the document is 6.72.

The original Wiki syntax has been converted into XML, using both general tags of the layout structure (like *article*, *section*, *paragraph*, *title*, *list* and *item*), typographical tags (like *bold*, *emphatic*), and frequently occurring link-tags. For details see Denoyer and Gallinari [3].

3.2 Topics

The ad hoc topics were created by participants following precise instructions given elsewhere [14]. Candidate topics contained a short CO (keyword) query, an optional structured CAS query, a one line description of the search request, and narrative with a details of the topic of request and the task context in which the information need arose. Figure 1 presents an example of an Ad Hoc topic. Based on the submitted candidate topics, 130 topics were selected for use in the INEX 2007 Ad Hoc track as topic numbers 414–543.

The INEX 2007 Multimedia track also had an ad hoc search task and 19 topics were used for both the Ad Hoc track and the Multimedia track. They were designated topics 525–543. Table 1 presents the topics shared between the Ad Hoc and Multimedia tracks. Six of these topics (527, 528, 530, 532, 535, 540) have an additional ⟨`mmtitle`⟩ field, a multimedia query.

The 12 INEX 2006 iTrack topics were also inserted into the topic set (as topics 512-514, and 516-524) as these topics were not assessed in 2006. Table 2 presents the 12 INEX 2006 iTrack topics, and their corresponding Ad Hoc track topic numbers.

```
<inex_topic topic_id="414" ct_no="3">
  <title>hip hop beat</title>
  <castitle>//*[about(., hip hop beat)]</castitle>
  <description>what is a hip hop beat?</description>
  <narrative>
    To solve an argument with a friend about hip hop music and beats, I
    want to learn all there is to know about hip hop beats. I want to know
    what is meant by hip hop beats, what is considered a hip hop beat,
    what distinguishes a hip hop beat from other beats, when it was
    introduced and by whom. I consider elements relevant if they
    specifically mention beats or rythm. Any element mentioning hip hop
    music or style but doesn't discuss abything about beats or rythm is
    considered not relevant. Also, elements discussing beats and rythm,
    but not hip hop music in particular, are considered not relevant.
  </narrative>
</inex_topic>
```

Fig. 1. INEX Ad Hoc Track topic 414

Table 1. Topics shared with the INEX 2007 Multimedia track

Topic	Title-field
525	potatoes in paintings
526	pyramids of egypt
527	walt disney land world
528	skyscraper building tall towers
529	paint works museum picasso
530	Hurricane satellite image
531	oil refinery or platform photographs
532	motor car
533	Images of phones
534	Van Gogh paintings
535	japanese garden old building -chapel
536	Ecuador volcano climbing quito
537	pictures of Mont Blanc
538	photographer photo
539	self-portrait
540	war map place
541	classic furniture design chairs
542	Images of tsunami
543	Tux

3.3 Judgments

Topics were assessed by participants following precise instructions [11]. The assessors used Piwowarski's X-RAI assessment system that assists assessors in highlight relevant text. Topic assessors were asked to mark all, and only,

Table 2. iTrack 2006 topics

iTrack	Ad hoc	Title-field	Type	Structure
1	519	types of bridges vehicles water ice	Decision making	Hierarchical
2	512	french impressionism degas monet renoir impressionist movement	Decision making	Hierarchical
3	520	Chartres Versailles history architecture travelling	Decision making	Parallel
4	516	environmental effects mining logging	Decision making	Parallel
5	521	red ants USA bites treatment	Fact finding	Hierarchical
6	513	chanterelle mushroom poisonous deadly species	Fact finding	Hierarchical
7	522	April 19th revolution peaceful revolution velvet revolution quiet revolution	Fact finding	Parallel
8	517	difference fortress castle	Fact finding	Parallel
9	523	fuel efficient cars	Info gathering	Hierarchical
10	514	food additives physical health risk grocery store labels	Info gathering	Hierarchical
11	524	home heating solar panels	Info gathering	Parallel
12	518	tidal power wind power	Info gathering	Parallel

relevant text in a pool of documents. The granularity of assessment was roughly a sentence. After assessing each article a separate best entry point decision was made by the assessor. The Focused and Relevant in Context Tasks were evaluated against the text highlighted by the assessors, whereas the Best in Context Task was evaluated against the best-entry-points.

The relevance judgments were frozen in January 2008. At this time 107 topics had been fully assessed. Moreover, 13 topics were judged by two separate assessors, each without the knowledge of the other. All results in this paper refer to the 107 topics with the judgments of the first assigned assessor.

- The 107 assessed topics were: 414-431, 433-436, 439-441, 444-450, 453, 454, 458, 459, 461-463, 465, 467, 468-475, 476-491, 495-500, 502, 503, 505-509, 511, 515-523, and 525-543.
- All 19 Multimedia topics, 525-543, were assessed.
- Only 8 of the 12 iTrack 2006 topics, 516-523, were assessed.

Table 3. Statistics over judged and relevant articles per topic

	total		# per topic				
	topics	number	min	max	median	mean	st.dev
judged articles	107	65,503	600	703	610	612	13.55
articles with relevance	107	6,491	2	479	36	61	70.91
highlighted passages	107	11,482	2	832	62	107	150.20

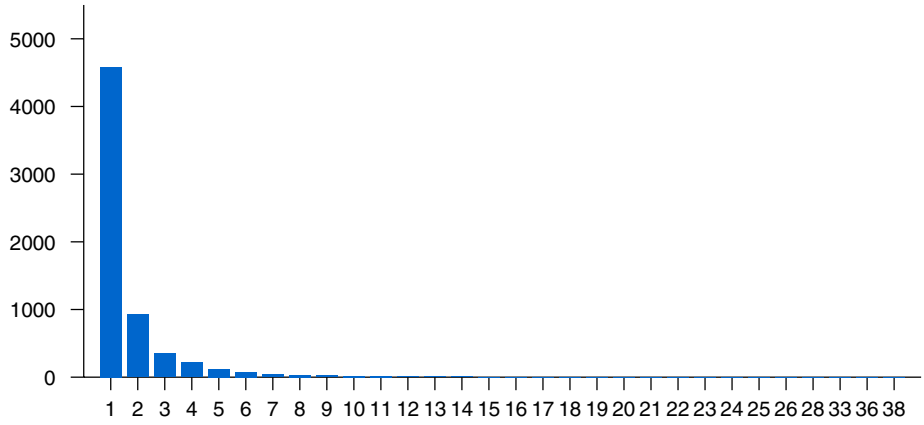

Fig. 2. Distribution of passages over articles

Table 3 presents statistics of the number of judged and relevant articles, and passages. In total 65,503 articles were judged. Relevant passages were found in 6,491 articles. The mean number of relevant articles per topic is 61, but the distribution is skewed with a median of 36. There were 11,482 highlighted passages. The mean was 107 passages and the median was 62 passages per topic.[1]

Figure 2 presents the number of articles with the given number of passages. The vast majority of relevant articles (4,581 out of 6,491) had only a single highlighted passage, and the number of passages quickly tapers off.

3.4 Questionnaires

At INEX 2007, all candidate topic authors and assessors were asked to complete a questionnaire designed to capture the context of the topic author and the topic of request. The candidate topic questionnaire (shown in Table 4) featured 20 questions capturing contextual data on the search request. The post-assessment questionnaire (shown in Table 5) featured 14 questions capturing further contextual data on the search request, and the way the topic has been judged.

The responses to the questionnaires show a considerable variation over topics and topic authors in terms of topic familiarity; the type of information requested; the expected results; the interpretation of structural information in the search request; the meaning of a highlighted passage; and the meaning of best entry points. There is a need for further analysis of the contextual data of the topics in relation to the results of the INEX 2007 Ad Hoc Track.

[1] Recall from above that for the Focused Task the main effectiveness measures is precision at 1% recall. Given that the average topic has 107 relevant passages in 61 articles, the 1% recall rougly corresponds to a relevant passage retrieved—for many systems this will be accomplished by the first or first few results.

Table 4. Candidate Topic Questionnaire

B1	How familiar are you with the subject matter of the topic?
B2	Would you search for this topic in real-life?
B3	Does your query differ from what you would type in a web search engine?
B4	Are you looking for very specific information?
B5	Are you interested in reading a lot of relevant information on the topic?
B6	Could the topic be satisfied by combining the information in different (parts of) documents?
B7	Is the topic based on a seen relevant (part of a) document?
B8	Can information of equal relevance to the topic be found in several documents?
B9	Approximately how many articles in the whole collection do you expect to contain relevant information?
B10	Approximately how many relevant document parts do you expect in the whole collection?
B11	Could a relevant result be (check all that apply): a single sentence; a single paragraph; a single (sub)section; a whole article
B12	Can the topic be completely satisfied by a single relevant result?
B13	Is there additional value in reading several relevant results?
B14	Is there additional value in knowing all relevant results?
B15	Would you prefer seeing: only the best results; all relevant results; don't know
B16	Would you prefer seeing: isolated document parts; the article's context; don't know
B17	Do you assume perfect knowledge of the DTD?
B18	Do you assume that the structure of at least one relevant result is known?
B19	Do you assume that references to the document structure are vague and imprecise?
B20	Comments or suggestions on any of the above (optional)

Table 5. Post Assessment Questionnaire

C1	Did you submit this topic to INEX?
C2	How familiar were you with the subject matter of the topic?
C3	How hard was it to decide whether information was relevant?
C4	Is Wikipedia an obvious source to look for information on the topic?
C5	Can a highlighted passage be (check all that apply): a single sentence; a single paragraph; a single (sub)section; a whole article
C6	Is a single highlighted passage enough to answer the topic?
C7	Are highlighted passages still informative when presented out of context?
C8	How often does relevant information occur in an article about something else?
C9	How well does the total length of highlighted text correspond to the usefulness of an article?
C10	Which of the following two strategies is closer to your actual highlighting: (I) I located useful articles and highlighted the best passages and nothing more, (II) I highlighted all text relevant according to narrative, even if this meant highlighting an entire article.
C11	Can a best entry point be (check all that apply): the start of a highlighted passage; the sectioning structure containing the highlighted text; the start of the article
C12	Does the best entry point correspond to the best passage?
C13	Does the best entry point correspond to the first passage?
C14	Comments or suggestions on any of the above (optional)

4 Ad Hoc Retrieval Results

In this section, we discuss, for the three ad hoc tasks, the participants and their results.

4.1 Participation

216 runs were submitted by 27 participating groups. Table 6 lists the participants and the number of runs they submitted, also broken down over the tasks (Focused, Relevant in Context, or Best in Context); the used query (Content-Only or Content-And-Structure); and the used result type (Element or Passage). Participants were allowed to submit up to three CO-runs per task and three CAS-runs per task (for all three tasks). This totaled to 18 runs per participant.[2] The submissions are spread well over the ad hoc retrieval tasks with 79 submissions for Focused, 66 submissions for Relevant in Context, and 71 submissions for Best in Context.

4.2 Focused Task

We now discuss the results of the Focused Task in which a ranked-list of non-overlapping results (elements or passages) was required. The official measure for the task was (mean) interpolated precision at 1% recall (iP[0.01]). Table 7 shows the best run of the top 10 participating groups. The first column gives the participant, see Table 6 for the full name of group, and see Appendix 6 for the precise run label. The second to fifth column give the interpolated precision at 0%, 1%, 5%, and 10% recall. The sixth column gives mean average interpolated precision over 101 standard recall levels (0%, 1%, ..., 100%).

Here we briefly summarize what is currently known about the experiments conducted by the top five groups (based on official measure for the task, iP[0.01]).

Dalian University of Technology. Using the CAS query. Only index the content contained by the tags often occur or retrieved by users. Use the BM25 retrieval model and pseudo-relevance feedback. Both document retrieval and document parts retrieval, and then combine the document score and document parts score. Further special handlings on the category of topics finding images, by removing the returned elements whose structural paths contained "image" or "figure" tags to the top one by one. Overlap was removed in the order of the resulting run.

Ecoles des Mines de Saint-Etienne. Using the CO query. Runs are based on the use of the proximity of the query terms in the documents. The proximity of an XML element to a query is based on the summation of the

[2] As it turns out, three groups submitted more runs than allowed: *mines* submitted 1 extra CO-run, and both *lip6* and *qutau* submitted 6 extra CO-runs each. At this moment, we have not decided on any repercussions other than mentioning them in this footnote.

Table 6. Participants in the Ad Hoc Track

Participant	Full name	Foc	RiC	BiC	CO	CAS	Ele	Pas	Total
cmu	Language Technologies Institute, School of Computer Science, Carnegie Mellon University	1	0	0	1	0	1	0	1
eurise	Laboratoire Hubert Curien - Université de Saint-Etienne	2	0	0	2	0	2	0	2
indstainst	Indian Statistical Institute	2	0	0	2	0	2	0	2
inria	INRIA-Rocquencourt- Axis	3	3	3	9	0	9	0	9
irit	IRIT	0	0	2	1	1	2	0	2
justsystem	JustSystems Corporation	6	6	6	9	9	18	0	18
labcsiro	Information Engineering lab, ICT Centre, CSIRO	1	0	0	1	0	1	0	1
lip6	LIP6	5	5	5	15	0	15	0	15
maxplanck	Max-Planck-Institut fuer Informatik	4	4	4	6	6	12	0	12
mines	Ecoles des Mines de Saint-Etienne, France	3	4	3	10	0	10	0	10
qutau	Queensland University of Technology	7	7	7	15	6	21	0	21
rmit	RMIT University	1	1	1	3	0	3	0	3
uamsterdam	University of Amsterdam	6	6	6	9	9	18	0	18
udalian	Dalian University of Technology	6	6	6	9	9	18	0	18
udoshisha	Doshisha University	2	0	0	1	1	2	0	2
ugrenoble	CLIPS-IMAG	3	3	3	9	0	9	0	9
uhelsinki	University of Helsinki	2	0	0	2	0	2	0	2
uminnesota	University of Minnesota Duluth	1	2	2	5	0	5	0	5
uniKaislau	University of Kaiserslautern, AG DBIS	3	3	0	6	0	6	0	6
unigordon	Information Retrieval and Interaction Group, The Robert Gordon University	3	3	3	9	0	9	0	9
unigranada	University of Granada	3	3	5	8	3	11	0	11
unitoronto	University of Toronto	2	0	0	0	2	2	0	2
uotago	University of Otago	3	3	3	9	0	0	9	9
utampere	University of Tampere	3	3	3	9	0	9	0	9
utwente	Cirquid Project (CWI and University of Twente)	3	2	1	6	0	6	0	6
uwaterloo	University of Waterloo	2	0	4	6	0	6	0	6
uwuhan	Center for Studies of Information Resources, School of Information Management, Wuhan University, China	2	2	4	8	0	8	0	8
Total	runs	79	66	71	170	46	207	9	216

normalized proximity score of each term position in the XML element. The proximity model is extended to take into account the document structure. The most simple and most used structure in document is the hierarchical one with sections, subsections, etc. where each instance at each level has got

Table 7. Top 10 Participants in the Ad Hoc Track Focused Task

Participant	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]	MAiP
udalian-15	0.5633	0.5271	0.4697	0.4041	0.1689
mines-2	0.6056	0.5164	0.3677	0.2984	0.1221
uwaterloo-0	0.5335	0.5108	0.4284	0.3916	0.1765
cmu-0	0.5924	0.5083	0.4000	0.3435	0.1351
maxplanck-3	0.5780	0.5066	0.4006	0.3430	0.1307
utampere-5	0.5460	0.4998	0.3915	0.3007	0.0981
udoshisha-0	0.5262	0.4975	0.3970	0.3360	0.1460
qutau-8	0.5120	0.4924	0.4493	0.4234	0.2025
inria-2	0.4986	0.4835	0.4540	0.4118	0.2132
rmit-0	0.4995	0.4834	0.4545	0.4172	0.2238

a title. With this kind of structure, we define the proximity to a position in a title as 1 (maximum value) over all the positions in the corresponding section.

University of Waterloo. Using the CO query. Query terms were formed by transforming each topic title into a disjunctive form, less negative query terms. Wumpus [15] was used to obtain positions of query terms and XML elements. The most frequently occurring XML elements in the corpus were listed and ranked using Okapi BM25. Nested results were removed for the Focused task.

Carnegie Mellon University. Using the CO query. XML elements are ranked using a hierarchical language model that estimates the probability of generating the query from an element. The hierarchical language models incorporate evidence from the document, its parent, and its children, using a linear combination of several language models [12].

Max-Planck-Institut für Informatik. Using the CAS query: the basis for this run is an ad hoc CAS run were the target tag was evaluated strictly, i.e., a result was required to have the tag specified as target in the query and match at least one of the content conditions, whereas support conditions were optional; phrases and negations in the query were ignored. To produce the focused run, elements were removed in case they overlap with a higher scoring element for the same topic.

Based on the information from these and other participants:

- Both the best scoring team and the fifth rank team used the CAS query. Hence using the structural hints, even strict adherence to the target tag, seemed to promote early precision
- More generally, limiting the retrieved types of elements, either at indexing time (by selecting elements based on tag type or length) or at retrieval time (by enforcing CAS target elements, or using length-priors), seems to promote early precision.
- The systems at rank nine, *inria-2*, and at rank ten, *rmit-0*, are retrieving only full articles.

Table 8. Top 10 Participants in the Ad Hoc Track Relevant in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
udalian-16	0.2566	0.2318	0.1888	0.1511	0.1552
qutau-18	0.2618	0.2223	0.1802	0.1454	0.1489
rmit-1	0.2483	0.2335	0.1792	0.1379	0.1358
uamsterdam-4	0.2403	0.2121	0.1647	0.1275	0.1323
unigordon-7	0.2531	0.2205	0.1680	0.1283	0.1302
utwente-5	0.2067	0.1838	0.1512	0.1187	0.1233
inria-5	0.2483	0.2335	0.1861	0.1358	0.1147
justsystem-14	0.2072	0.1732	0.1342	0.1023	0.1107
mines-7	0.2120	0.1913	0.1527	0.1185	0.1081
maxplanck-8	0.2168	0.1879	0.1356	0.1050	0.1077

4.3 Relevant in Context Task

We now discuss the results of the Relevant in Context Task in which non-overlapping results (elements or passages) need to be returned grouped by the article they came from. The task was evaluated using generalized precision where the generalized score per article was based on the retrieved highlighted text. The official measure for the task was mean average generalized precision (MAgP).

Table 8 shows the top 10 participating groups (only the best run per group is shown) in the Relevant in Context Task. The first column lists the participant, see Table 6 for the full name of group, and see Appendix 6 for the precise run label. The second to fifth column list generalized precision at 5, 10, 25, 50 retrieved articles. The sixth column lists mean average generalized precision.

Here we briefly summarize the information available about the experiments conducted by the top five groups (based on MAgP).

Dalian University of Technology. Using the CO query. See the description for the Focused Task above. Although submitted as CO run, image finding topics received special handling promoting elements with paths containing image of figure to the top of the ranking. Cluster the returned elements per document, and remove overlap top-down.

Queensland University of Technology. Using the CO query: plural/singular expansion was used on the query, as well as removal of words preceded by a minus sign. GPX [4] was used to rank elements, based on a leaf-node index and $tf \cdot icf$ (term frequency times inverted collection frequency) weighting modified by i) the number of unique terms, ii) the proximity of query-term matches, and iii) boosting of query-term occurrences in the name field. All leaf-node-scores were normalized by their length, and the overall article's similarity score was added. The score of elements was calculated directly from the content of the nodes, obviating the need for score propagation with decaying factors.

RMIT University. Using the CO query. This is a baseline article run using Zettair [16] with the Okapi similarity measure with default settings. The title from each topic was automatically translated as an input query to Zettair. The similarity of an article to a query determines its final rank.

University of Amsterdam. Using the CO query. Having an index with only the "container" elements – elements that frequently contain an entire highlighted passage at INEX 2006 – basically corresponding to the main layout structure. A language model was used with a standard length prior and an incoming links prior, after list-based removal of overlapping elements the final results are clustered per article on a first-come, first-served basis.

Robert Gordon University. Using the CO query. An element's score was computed by a mixture language model combining estimates based on element full-text and a "summary" of it (i.e., extracted titles, section titles, and figure captions nested inside the element). A prior was used according to an element's location in the original text, and the length of its path. For the post-processing, they filter out redundant elements by selecting the highest scored element from each of the paths. Elements are reordered so that results from the same article are grouped together.

Based on the information from these and other participants:

- Solid article ranking seems a prerequisite for good overall performance, with third best run, *rmit-1*, and the seventh best run, *inria-5*, retrieving only full articles.
- The use of the structured query does not appear to promote overall performance: all five groups submitting a CAS query run had a superior CO query run.

4.4 Best in Context Task

We now discuss the results of the Best in Context Task in which documents were ranked on topical relevance and a single best entry point into the document was identified. The Best in Context Task was evaluated using generalized precision but here the generalized score per article was based on the distance to the assessor's best-entry point. The official measure for the task was mean average generalized precision (MAgP).

Table 9. Top 10 Participants in the Ad Hoc Track Best in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
rmit-2	0.3551	0.3280	0.2554	0.1931	0.1919
qutau-19	0.3256	0.2736	0.2138	0.1734	0.1831
uwaterloo-3	0.2600	0.2467	0.2181	0.1716	0.1817
udalian-7	0.2512	0.2416	0.2024	0.1601	0.1759
unigordon-2	0.3405	0.2906	0.2278	0.1761	0.1742
uamsterdam-16	0.3325	0.2917	0.2292	0.1788	0.1731
justsystem-7	0.2904	0.2714	0.2054	0.1611	0.1661
inria-8	0.3551	0.3280	0.2610	0.1952	0.1633
maxplanck-6	0.2005	0.2053	0.1735	0.1348	0.1350
utwente-2	0.2562	0.2246	0.1821	0.1430	0.1339

Table 9 shows the top 10 participating groups (only the best run per group is shown) in the Best in Context Task. The first column lists the participant, see Table 6 for the full name of group, and see Appendix 6 for the precise run label. The second to fifth column list generalized precision at 5, 10, 25, 50 retrieved articles. The sixth column lists mean average generalized precision.

Here we briefly summarize the information available about the experiments conducted by the top five groups (based on MAgP).

RMIT University. Using the CO query. This is the exact same run as the article run for the Relevant in Context Task. See the description for the Relevant in Context Task above.

Queensland University of Technology. Using the CO query. See the description for the Relevant in Context Task above. The best scoring element was selected.

University of Waterloo. Using the CO query. See the description for the Focused Task above. Based on the Focused run, duplicated articles were removed in a post-processing step.

Dalian University of Technology. Using the CO query. See the description for the Focused Task and Relevant in Context above. Return the element which has the largest score per document.

Robert Gordon University. Using the CO query. See the description for the Relevant in Context Task above. For the best-in-context task, the element with the highest score for each of the documents is chosen.

Based on the information from these and other participants:

- As for the Relevant in Context Task, we see again that solid article ranking is very important. In fact, the full article run *rmit-2* is the most effective system. Also the eighth best participant, *inria-8*, is retrieving only full articles.
- Using the start of the whole article as a best-entry-point, as done by the top scoring article run, appears to be a reasonable strategy.
- With the exception of *uamsterdam-16*, which used a filter based on all CAS target elements in the topic set, all best runs per group use the CO query.

4.5 Significance Tests

We tested whether higher ranked systems were significantly better than lower ranked system, using a t-test (one-tailed) at 95%. Table 10 shows, for each task, whether it is significantly better (indicated by "⋆") than lower ranked runs. For example, For the Focused Task, we see that the early precision (at 1% recall) is a rather unstable measure and none of the runs are significantly different. Hence we should be careful when drawing conclusions based on the Focused Task results. For the Relevant in Context Task, we see that the top run is significantly better than ranks 3 through 10, the second best run better than ranks 4 through 10, the third ranked system better than ranks 6 through 10, and the fourth and

Table 10. Statistical significance (t-test, one-tailed, 95%)

(a) Focused Task

	1 2 3 4 5 6 7 8 9 10
udalian-15	- - - - - - - - - -
mines-2	- - - - - - - - -
uwaterloo-0	- - - - - - -
cmu-0	- - - - - -
maxplanck-3	- - - - -
utampere-5	- - - -
udoshisha-0	- - -
qutau-8	- -
inria-2	-
rmit-0	

(b) Relevant in Context Task

	1 2 3 4 5 6 7 8 9 10
udalian-16	- ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆
qutau-18	- ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆
rmit-1	- - ⋆ ⋆ ⋆ ⋆ ⋆
uamsterdam-4	- - ⋆ ⋆ ⋆ ⋆
unigordon-7	- ⋆ ⋆ ⋆ ⋆
utwente-5	- - - -
inria-5	- - -
justsystem-14	- -
mines-7	-
maxplanck-8	

(c) Best in Context Task

	1 2 3 4 5 6 7 8 9 10
rmit-2	- - - ⋆ ⋆ ⋆ ⋆ ⋆ ⋆
qutau-19	- - - - - - ⋆ ⋆
uwaterloo-3	- - - - - ⋆ ⋆
udalian-7	- - - - ⋆ ⋆
unigordon-2	- - - ⋆ ⋆
uamsterdam-16	- - ⋆ ⋆
justsystem-7	- ⋆ ⋆
inria-8	⋆ ⋆
maxplanck-6	-
utwente-2	

fifth ranked systems better than ranks 7 through 10. For the Best in Context Task, we see that the top run is significantly better than ranks 5 through 10, the second to eighth ranked systems are significantly better than those at rank 9 and 10.

5 Analysis of Run and Topic Types

In this section, we will discuss relative effectiveness of element and passage retrieval approaches, and on the relative effectiveness of systems using the keyword and structured queries.

5.1 Elements Versus Passages

We received some, but few, submissions using passage results. We will look at the relative effectiveness of element and passage runs.

As we saw above, in Section 4, for all three tasks the best scoring runs used elements as the unit of retrieval. All nine official passage submissions were from the same participant. Table 11 shows their best passage runs for the three ad

Table 11. Ad Hoc Track: Passage runs

(a) Focused Task

Participant	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]	MAiP
uotago-3	0.4850	0.4716	0.3423	0.2639	0.0902

(b) Relevant in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
uotago-1	0.1625	0.1529	0.1213	0.0955	0.1033

(c) Best in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
uotago-6	0.1377	0.1415	0.1194	0.0994	0.1064

Table 12. CAS query target elements over all 130 topics

Target Element	Frequency
*	51
article	29
section	28
figure	9
p	5
image	5
title	1
(section|p)	1
body	1

hoc tasks. As it turns out, the passage run *otago-3* would have been the 12th ranked participant (out of 26) for the Focused Task; *otago-1* would have been the 11th ranked group (out of 18) for the Relevant in Context Task; and *otago-6* would have been the 13th ranked group (out of 19) for the Best in Context Task.

This outcome is consistent with earlier results using passage-based element retrieval, where passage retrieval approaches showed comparable but not superior behavior to element retrieval approaches [5, 6].

It is hard to draw any conclusions for several reasons. First, the passage runs took no account of document structure with passages frequently starting and ending mid-sentence. Second, with only a single participant it is not clear whether the approach is comparable or the participant's runs are only comparable. Third, this is the first year passage retrieval has run at INEX and so the technology is less mature than element retrieval.

We hope and expect that the test collection and the passage runs will be used for further research into the relative effectiveness of element and passage retrieval approaches.

5.2 CO Versus CAS

We now zoom in on the relative effectiveness of the keyword (CO) and structured (CAS) queries. As we saw above, in Section 4, the best two runs for the Focused task used the CAS query, and one of the top 10 runs for the Best in Context Task used the CAS query.

All topics have a CAS query since artificial CAS queries of the form

```
//*[about(., keyword title)]
```

were added to topics without CAS title. Table 12 show the distribution of target elements. In total 111 topics had a CAS query formulated by the authors. Some authors already used the generic CAS query above. There are only 86 topics with a non-trivial CAS query.[3]

[3] Note that some of the wild-card topics (using the "*" target) in Table 12 had non-trivial about-predicates and hence have not been regarded as trivial CAS queries.

Table 13. Ad Hoc Track CAS Topics: CO runs (left-hand side) versus CAS runs (right-hand side)

(a) Focused Task

Participant	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]	MAiP	Participant	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]	MAiP
mines-2	0.6207	0.5426	0.3848	0.3016	0.1285	udalian-15	0.5503	0.5159	0.4481	0.4050	0.1795
udoshisha-0	0.5472	0.5190	0.3995	0.3454	0.1588	maxplanck-3	0.5780	0.4919	0.3834	0.3402	0.1397
cmu-0	0.6047	0.5184	0.4213	0.3679	0.1475	justsystem-3	0.5238	0.4798	0.3736	0.3087	0.1175
uwaterloo-0	0.5397	0.5140	0.4384	0.4079	0.1938	udoshisha-1	0.5337	0.4519	0.3466	0.2969	0.1319
qutau-8	0.5225	0.5124	0.4808	0.4594	0.2120	uamsterdam-10	0.4840	0.4413	0.3835	0.3443	0.1671
udalian-2	0.5343	0.5045	0.4429	0.4077	0.1903	unitoronto-0	0.4921	0.4079	0.3148	0.2680	0.1059
rmit-0	0.5115	0.5024	0.4734	0.4340	0.2351	qutau-9	0.4072	0.4033	0.3895	0.3590	0.1614
inria-2	0.5096	0.5007	0.4724	0.4315	0.2258	unigranada-0	0.3981	0.2644	0.1006	0.0637	0.0229
justsystem-0	0.5292	0.4998	0.4207	0.3599	0.1331						
unigordon-1	0.5189	0.4922	0.4297	0.3918	0.1977						

(b) Relevant in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP	Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
qutau-18	0.2798	0.2286	0.1846	0.1482	0.1654	udalian-14	0.2525	0.2217	0.1800	0.1419	0.1578
udalian-4	0.2570	0.2345	0.1871	0.1442	0.1622	uamsterdam-13	0.2473	0.2180	0.1626	0.1237	0.1351
rmit-1	0.2505	0.2356	0.1719	0.1299	0.1455	qutau-10	0.2218	0.1892	0.1507	0.1178	0.1150
uamsterdam-4	0.2428	0.2137	0.1637	0.1242	0.1416	justsystem-15	0.2005	0.1687	0.1224	0.0952	0.1136
unigordon-7	0.2708	0.2296	0.1640	0.1214	0.1407	maxplanck-5	0.2293	0.1926	0.1448	0.1022	0.1048
utwente-5	0.2068	0.1821	0.1455	0.1094	0.1313						
inria-5	0.2505	0.2356	0.1810	0.1312	0.1259						
justsystem-14	0.2145	0.1765	0.1334	0.0981	0.1219						
maxplanck-8	0.2294	0.1921	0.1353	0.1042	0.1185						
mines-7	0.2119	0.1783	0.1181	0.0864	0.1137						

(c) Best in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP	Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
rmit-2	0.3552	0.3262	0.2436	0.1842	0.2013	udalian-17	0.2523	0.2442	0.2095	0.1705	0.1800
uwaterloo-3	0.2869	0.2658	0.2259	0.1744	0.1986	uamsterdam-16	0.3233	0.2879	0.2233	0.1728	0.1768
qutau-19	0.3415	0.2777	0.2221	0.1818	0.1964	justsystem-9	0.3065	0.2712	0.2077	0.1710	0.1652
udalian-7	0.2609	0.2491	0.2105	0.1665	0.1899	qutau-3	0.2805	0.2366	0.1679	0.1299	0.1529
unigordon-2	0.3616	0.2950	0.2220	0.1683	0.1854	maxplanck-1	0.2726	0.2466	0.1965	0.1374	0.1281
justsystem-7	0.3109	0.2931	0.2183	0.1689	0.1792	unigranada-6	0.1930	0.1821	0.1548	0.1277	0.1139
uamsterdam-7	0.2706	0.2634	0.2123	0.1676	0.1760	irit-4	0.0337	0.0329	0.0316	0.0219	0.0170
inria-8	0.3552	0.3262	0.2521	0.1877	0.1735						
maxplanck-6	0.2088	0.2188	0.1790	0.1417	0.1451						
utwente-2	0.2532	0.2134	0.1592	0.1216	0.1366						

The CAS topics numbered 415, 416, 418-424, 426-432, 434-440, 442-448, 454, 459, 461, 463, 464, 466, 470, 472, 474, 476-491, 493-498, 500, 501, 507, 508, 511, 515, and 525-543. As it turned out, 77 of these CAS topics were assessed. The results presented here are restricted to only these 77 CAS topics.

Table 13 lists the top 10 participants measured using just the 77 CAS topics and for the Focused Task (a), the Relevant in Context Task (b), and the Best in Context Task (c). For the Focused Task the best two CAS runs outperform the CO runs, as they did over the full topic set. For the Relevant in Context Task, the best CAS run would have ranked fourth among CO runs. For the Best in Context Task, the best two CAS runs would rank sixth and seventh among the CO runs.

We look in detail at the Focused Task runs, where CAS submissions were competitive. Overall, the CAS submissions appear to perform similarly on the subset of 77 CAS topics to the whole set of topics. This was unexpected as these

topics do contain real structural hints. The 77 CAS topics constitute three-quarters of the full topic set, making it reasonable to get such a result. However, there are some notable performance characteristics among the CO submissions:

- Some runs (like *maxplanck-3*) perform equally well as over all topics.
- Some runs (like *rmit-0* and *udoshisha-0*) perform much better than over all topics. A possible explanation is the larger number of article-targets among the CAS queries.
- Some runs (like *utampere-5*) perform less well than over all topics.

We should be careful to draw conclusions based on these observations, since the early precision differences between the runs tend not to be significant.

6 Discussion and Conclusions

In this paper we provided an overview of the INEX 2007 Ad Hoc Track that contained three tasks: For the *Focused Task* a ranked-list of non-overlapping results (elements or passages) was required. For the *Relevant in Context Task* non-overlapping results (elements or passages) grouped by the article that they belong to were required. For the *Best in Context Task* a single starting point (element's starting tag or passage offset) per article was required. We discussed the results for the three tasks, and analysed the relative effectiveness of element and passage runs, and of keyword (CO) queries and structured queries (CAS).

When examining the relative effectiveness of CO and CAS we found that the best Focused Task submissions use the CAS query, showing that structural hints can help promote initial precision. This provides further evidence that structured queries can be a useful early precision enhancing device [8]. Although, when restricting to non-trivial CAS queries, we see no real gain for the CAS submissions relative to the CO submissions.

An unexpected finding is that article retrieval is a reasonably effective at XML-IR: an article-only run scored the eighth best group for the Focused Task; the third best for the Relevant in Context Task; and the top ranking group for the Best in Context Task. This demonstrates the importance of the article ranking in the "in context" tasks. The chosen measures were also not unfavorable towards article-submissions:

- For the Relevant in Context Task, the F-score per document equally rewards precision and recall. Article runs have excellent recall, and in the case of Wikipedia, where articles tend to be focused on a single topic, acceptable precision.
- For the Best in Context Task, the window receiving scores was 1,000 characters which, although more strict than the measures at INEX 2006, remains too lenient.

Given the efforts put into the fair comparison of element and passage retrieval approaches, the number of passage submissions was disappointing. The passage

runs that were submitted ignored document structure—perhaps the identification based on the XML structure turned out to be difficult, or perhaps the technology is just not yet mature. Although we received only passage results from a single participant, and should be careful to avoid hasty conclusions, we saw that the passage based approach was better than average, but not superior to element based approaches. This outcome is consistent with earlier results using passage-based element retrieval [5, 6]. The comparative analysis of element and passage retrieval approaches was the aim of the track, hoping to shed light on the value of the document structure as provided by the XML mark-up. Although few official submissions used passage retrieval approaches, we hope and expect that the resulting test collection will prove its value in future use. After all, the main aim of the INEX initiative is to create bench-mark test-collections for the evaluation of structured retrieval approaches.

Acknowledgments

Eternal thanks to Benjamin Piwowarski for completely updating the X-RAI tools to ensure that all passage offsets can be mapped exactly.

Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO, grants # 612.066.513, 639.072.601, and 640.001.501), and by the E.U.'s 6th FP for RTD (project MultiMATCH contract IST-033104).

References

[1] Clarke, C.L.A.: Range results in XML retrieval. In: Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, pp. 4–5, Glasgow, UK (2005)
[2] Clarke, C.L.A., Kamps, J., Lalmas, M.: INEX 2007 retrieval task and result submission specification. In: Pre-Proceedings of INEX 2007, pp. 445–453 (2007)
[3] Denoyer, L., Gallinari, P.: The Wikipedia XML Corpus. SIGIR Forum 40, 64–69 (2006)
[4] Geva, S.: GPX – gardens point XML IR at INEX 2005. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 204–253. Springer, Heidelberg (2006)
[5] Huang, W., Trotman, A., O'Keefe, R.A.: Element retrieval using a passage retrieval approach. In: Proceedings of the 11th Australasian Document Computing Symposium (ADCS 2006), pp. 80–83 (2006)
[6] Itakura, K.Y., Clarke, C.L.A.: From passages into elements in XML retrieval. In: Proceedings of the SIGIR 2007 Workshop on Focused Retrieval, University of Otago, Dunedin New Zealand, pp. 17–22 (2007)
[7] Kamps, J., Koolen, M.: On the relation between relevant passages and XML document structure. In: Proceedings of the SIGIR 2007 Workshop on Focused Retrieval, University of Otago, Dunedin New Zealand, pp. 28–32 (2007)
[8] Kamps, J., Marx, M., de Rijke, M., Sigurbjörnsson, B.: Articulating information needs in XML query languages. Transactions on Information Systems 24, 407–436 (2006)
[9] Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., Robertson, S.: INEX 2007 evaluation measures. In: Fuhr, N., Lalmas, M., Trotman, A. (eds.) INEX 2006. LNCS, vol. 4518. Springer, Heidelberg (2007)

[10] Kekäläinen, J., Järvelin, K.: Using graded relevance assessments in IR evaluation. Journal of the American Society for Information Science and Technology 53, 1120–1129 (2002)
[11] Lalmas, M., Piwowarski, B.: INEX 2007 relevance assessment guide. In: Pre-Proceedings of INEX 2007, pp. 454–463 (2007)
[12] Ogilvie, P., Callan, J.: Parameter estimation for a simple hierarchical generative model for xml retrieval. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 211–224. Springer, Heidelberg (2006)
[13] Trotman, A., Geva, S.: Passage retrieval and other XML-retrieval tasks. In: Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology, University of Otago, Dunedin New Zealand, pp. 43–50 (2006)
[14] Trotman, A., Larsen, B.: INEX 2007 guidelines for topic development. In: Pre-Proceedings of INEX 2007, pp. 436–444 (2007)
[15] Wumpus. The Wumpus search engine (2007), `http://www.wumpus-search.org`
[16] Zettair. The Zettair search engine (2007), `http://www.seg.rmit.edu.au/zettair/`

Appendix: Full Run Names

Run	Label
cmu-0	p40_nophrasebase
inria-2	p11_ent-ZM-Focused
inria-5	p11_ent-ZM-RiC
inria-8	p11_ent-ZM-BiC
irit-4	p49_xfirm.cos.01_BIC
justsystem-0	p41_VSM_CO_01
justsystem-14	p41_VSM_CO_09
justsystem-15	p41_VSM_CAS_10
justsystem-3	p41_VSM_CAS_04
justsystem-7	p41_VSM_CO_14
justsystem-9	p41_VSM_CAS_16
maxplanck-1	p25_TOPX-CAS-exp-BIC
maxplanck-3	p25_TOPX-CAS-Focused-all
maxplanck-5	p25_TOPX-CAS-RIC
maxplanck-6	p25_TOPX-CO-all-BIC
maxplanck-8	p25_TOPX-CO-all-exp-RIC
mines-2	p53_EMSE.boolean.Prox200NF.0012
mines-7	p53_EMSE.boolean.Prox200NRm.0010
qutau-10	p9_RIC_05
qutau-18	p9_RIC_07
qutau-19	p9_BIC_07
qutau-3	p9_BIC_04
qutau-8	p9_FOC_03
qutau-9	p9_FOC_04
rmit-0	p32_zet-okapi-Focused
rmit-1	p32_zet-okapi-RiC
rmit-2	p32_zet-okapi-BiC
uamsterdam-10	p36_inex07_contain_beta1_focused_clp_10000_cl_cas_pool_filter
uamsterdam-13	p36_inex07_contain_beta1_focused_clp_10000_cl_cas_pool_filter_ric_hse
uamsterdam-16	p36_inex07_contain_beta1_focused_clp_10000_cl_cas_pool_filter_bic_hse
uamsterdam-4	p36_inex07_contain_beta1_focused_clp_10000_cl_ric_hse
uamsterdam-7	p36_inex07_contain_beta1_focused_clp_10000_cl_bic_hse
udalian-14	p26_DUT_03_Relevant
udalian-15	p26_DUT_03_Focused
udalian-16	p26_DUT_01_Relevant
udalian-17	p26_DUT_03_Best
udalian-2	p26_DUT_01_Focused_3
udalian-4	p26_DUT_02_Relevant
udalian-7	p26_DUT_02_Best
udoshisha-0	p22_Kikori-CO-Focused
udoshisha-1	p22_Kikori-CAS-Focused
unigordon-1	p35_Focused-LM
unigordon-2	p35_BestInContext-LM
unigordon-7	p35_RelevantInContext-LM
unigranada-0	p4_CID_pesos_15_util_2
unigranada-6	p4_CID_pesos_15_bic
unitoronto-0	p60_4-sr
uotago-1	p10_DocsNostem-PassagesNoStem-StdDevNo
uotago-3	p10_DocsNostem-PassagesStem-StdDevNo-Focused
uotago-6	p10_DocsNostem-PassagesStem-StdDevNo-BEP
utampere-5	p55_Foc k=0.3, v=4.5, cont=2.3
utwente-2	p45_articleBic
utwente-5	p45_star_logLP_RinC
uwaterloo-0	p37_FOER
uwaterloo-3	p37_BICERGood