Overview of the INEX 2008 Ad Hoc Track

Jaap Kamps[1], Shlomo Geva[2], Andrew Trotman[3],
Alan Woodley[2], and Marijn Koolen[1]

[1] University of Amsterdam, Amsterdam, The Netherlands
{kamps,m.h.a.koolen}@uva.nl
[2] Queensland University of Technology, Brisbane, Australia
{s.geva,a.woodley}@qut.edu.au
[3] University of Otago, Dunedin, New Zealand
andrew@cs.otago.ac.nz

Abstract. This paper gives an overview of the INEX 2008 Ad Hoc Track. The main goals of the Ad Hoc Track were two-fold. The first goal was to investigate the value of the internal document structure (as provided by the XML mark-up) for retrieving relevant information. This is a continuation of INEX 2007 and, for this reason, the retrieval results are liberalized to arbitrary passages and measures were chosen to fairly compare systems retrieving elements, ranges of elements, and arbitrary passages. The second goal was to compare focused retrieval to article retrieval more directly than in earlier years. For this reason, standard document retrieval rankings have been derived from all runs, and evaluated with standard measures. In addition, a set of queries targeting Wikipedia have been derived from a proxy log, and the runs are also evaluated against the clicked Wikipedia pages. The INEX 2008 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, and 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. Finally, we look at the ability of focused retrieval techniques to rank articles, using standard document retrieval techniques, both against the judged topics as well as against queries and clicks from a proxy log.

1 Introduction

This paper gives an overview of the INEX 2008 Ad Hoc Track. There are two main research question underlying the Ad Hoc Track. The first main research question 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, first studied at INEX 2007, has attracted a lot of attention in recent years.

Trotman and Geva [11] 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 true and only some passages can be described as elements. Huang et al. [4] 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 [5] 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 [6] 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 were liberalized to arbitrary passages. Every XML element is, of course, also a passage of text. At INEX 2008, a simple passage retrieval format was introduced using file-offset-length (FOL) triplets, that allow for standard passage retrieval systems to work on content-only versions of the collection. That is, the offset and length are calculated over the text of the article, ignoring all mark-up. The evaluation measures are 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 [2] and later requests to liberalize to arbitrary passages of text [11].

The second main question is to compare focused retrieval directly to traditional article retrieval. Throughout the history of INEX, participating groups have found that article retrieval—a system retrieving the whole article by default—resulted in fairly competitive performance [e.g., 7, 10]. Note that every focused retrieval system also generates an underlying article ranking, simply by the order is which results from different articles are ranked. This is most clear in the Relevant in Context and Best in Context tasks, where the article ranking is an explicit part of the task description. To study the importance of the underlying article ranking quality, we derived article level judgments by treating every article with some highlighted text as relevant, derived article rankings from every submission on a first-come, first-served basis, and evaluated with standard measures. This will also shed light on the value of element or passage level evidence for document retrieval [1]. In addition to this, we also include queries derived from a proxy log in the topic set, and can derive judgments from the later clicks in the same proxy log, treating all clicked articles as relevant for the query at

hand. All submissions are also evaluated against these clicked Wikipedia pages, giving some insight in the differences between an IR test collection and real-world searching of Wikipedia.

The INEX 2008 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.

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 2008 ad hoc retrieval tasks and measures. Section 3 details the collection, topics, and assessments of the INEX 2008 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). Section 6 looks at the article retrieval aspects of the submissions, both in terms of the judged topics treating any article with highlighted text as relevant, and in terms of clicked Wikipedia pages for queries derived from a proxy log. Finally, in Section 7, 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 measures used for evaluation.

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 ancestor elements and longer

passages are always relevant (to a greater or lesser extent) it is a challenge to choose 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. At INEX 2008, we allowed the submission of three types of results: XML elements; ranges of XML elements; and file-offset-length (FOL) text passages.

Element Results 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 the next example identifies 9996.xml as the target document from the Wikipedia collection (with the .xml extension removed).

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

Element paths are given in XPath, but only fully specified paths are allowed. The next example 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.

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

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, as shown in the next example.

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

Ranges of Elements To support ranges of elements, elemental passages are given in the same format.[1] 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>
```

Note that this format is very convenient for specifying ranges of elements, e.g., the following example retrieves the first three sections.

```
<result>
  <file>9996</file>
  <passage start="/article[1]/body[1]/section[1]"
     end="/article[1]/body[1]/section[3]"/>
  <rsv>0.9999</rsv>
</result>
```

FOL passages Passage results can be given in file-offset-length (FOL) format, where offset and length are calculated in characters with respect to the textual content (ignoring all tags) of the XML file. A special text-only version of the

[1] At INEX 2007, and in earlier qrels, an extended format allowing for optional character-offsets was used that allowed these passages to start or end in the middle of element or text-nodes. This format is superseded with the clean file-offset-length (FOL) passage format.

collection is provided to facilitate the use of passage retrieval systems. File offsets start counting a 0 (zero).

The following example is effectively equivalent to the example element result above.

```
<result>
  <file>9996</file>
  <fol offset="461" length="202"/>
  <rsv>0.9999</rsv>
</result>
```

The paragraph starts at the 462th character (so 461 characters beyond the first character), and has a length of 202 characters.

2.3 Evaluation Measures

We briefly summarize the main measures used for the Ad Hoc Track. Since INEX 2007, we allow the retrieval of arbitrary passages of text matching the judges ability to regard any passage of text as relevant. 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 2008 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 2008 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 [9], 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 use an F_β score with $\beta = 1/4$ making precision four times as important as recall (at INEX 2007, F_1 was used). 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 $n = 500$ which is roughly the number of characters corresponding to the visible part of the document on a screen (at INEX 2007, $n = 1,000$ was used). 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. 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, 135 topics were selected for use in the INEX 2008 Ad Hoc Track as topic numbers 544–678.

```
<topic id="544" ct_no="6">
  <title>meaning of life</title>
  <castitle>
    //article[about(., philosophy)]//section[about(., meaning of life)]
  </castitle>
  <description>What is the meaning of life?</description>
  <narrative>
    I got bored of my life and started wondering what the meaning of
    life is. An element is relevant if it discusses the meaning of life
    from different perspectives, as long as it is serious. For example,
    Socrates discussing meaning of life is relevant, but something like
    "42" from H2G2 or "the meaning of life is cheese" from a comedy is
    irrelevant. An element must be self contained. An  element that is a
    list of links is considered irrelevant because it is not
    self-contained in the sense that I don't know in which context the
    links are given.
  </narrative>
</topic>
```

Fig. 1. INEX 2008 Ad Hoc Track topic 544.

In addition, 150 queries were derived from a proxy-log for use in the INEX 2008 Ad Hoc Track as topic numbers 679–828. For these topics, as well as the candidate topics without a ⟨`castitle`⟩ field, a default CAS-query was added based on the CO-query: `//*[about(., "`*CO-query*`")]`.

3.3 Judgments

Topics were assessed by participants following precise instructions. The assessors used the new GPXrai assessment system that assists assessors in highlight relevant text. Topic assessors were asked to mark all, and only, relevant text in a pool of documents. After assessing an article with relevance, 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 on October 22, 2008. At this time 70 topics had been fully assessed. Moreover, 11 topics were judged by two separate assessors, each without the knowledge of the other. All results in this paper refer to the 70 topics with the judgments of the first assigned assessor, which is typically the topic author.

- The 70 assessed topics were: 544–547, 550–553, 555–557, 559, 561, 562–563, 565, 570, 574, 576–582, 585–587, 592, 595–598, 600–603, 607, 609–611, 613, 616–617, 624, 626, 628, 629, 634–637, 641–644, 646–647, 649–650, 656–657, 659, 666–669, 673, 675, and 677.

In addition, there are clicked Wikipedia pages available in the proxy log for 125 topics:

Table 1. Statistics over judged and relevant articles per topic.

	total		# per topic				
	topics	number	min	max	median	mean	st.dev
judged articles	70	42,272	588	618	603	603.9	5.6
articles with relevance	70	4,887	2	376	49	69.8	68.9
highlighted passages	70	6,908	3	897	56	98.7	124.6
highlighted characters	70	11,471,649	1,419	1,113,578	99,569	163,880.7	202,757.2
Unique articles with clicks	125	225	1	10	1	1.8	1.5
Total clicked articles	125	532	1	24	3	4.3	3.8

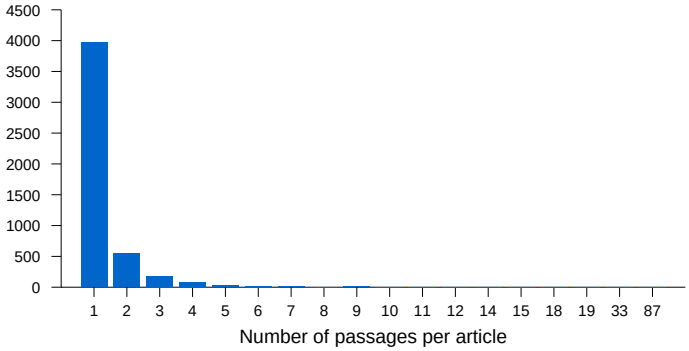

Fig. 2. Distribution of passages over articles.

– The 125 topics with clicked articles are numbered: 679–682, 684–685, 687–693, 695–704, 706–708, 711–727, 729–732, 734–751, 753–776, 778, 780–782, 784, 786–787, 789–790, 792–793, 795–796, 799–804, 806–807, 809–810, 812–813, 816–819, 821–824, and 826–828.

Table 1 presents statistics of the number of judged and relevant articles, and passages. In total 42,272 articles were judged. Relevant passages were found in 4,887 articles. The mean number of relevant articles per topic is 70, but the distribution is skewed with a median of 49. There were 6,908 highlighted passages. The mean was 99 passages and the median was 56 passages per topic.[2]

Table 1 also includes some statistics of the number of clicked articles in the proxy log. There are in total 225 clicked articles (unique per topic) over in total 125 topics, with a mean of 1.8 and a median of 1 clicked article per topic. We filtered the log for queries issued by multiple persons, and can also count the total number of clicks. Here, we see a total of 532 clicks (on the same 225 articles before), with a mean of 4.3 and a median of 3 clicks per topic. It is clear that the topics and clicked articles from the log are very different in character from the ad hoc topics.

[2] Recall from above that for the Focused Task the main effectiveness measures is precision at 1% recall. Given that the average topic has 99 relevant passages in 70 articles, the 1% recall roughly corresponds to a relevant passage retrieved—for many systems this will be accomplished by the first or first few results.

Table 2. Statistics over best entry point judgement.

	# topics	number	min	max	median	mean	st.dev
best entry point offset	70	4,887	1	87,982	14	1,738.1	4,814.3
first relevant character offset	70	4,887	1	87,982	20	1,816.1	4,854.2
fraction highlighted text	70	4,850	0.0005	1.000	0.583	0.550	0.425

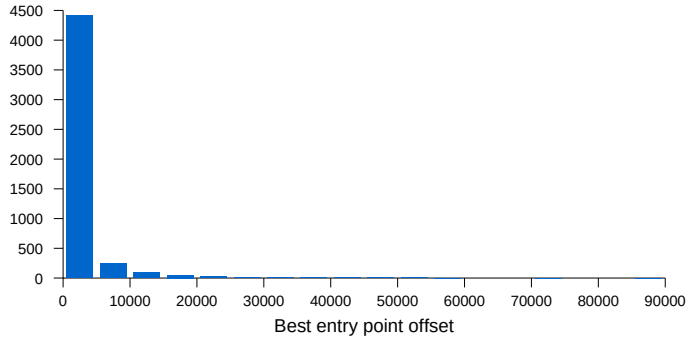

Fig. 3. Distribution of best entry point offsets.

Figure 2 presents the number of articles with the given number of passages. The vast majority of relevant articles (3,967 out of 4,887) had only a single highlighted passage, and the number of passages quickly tapers off.

Assessors where requested to provide a separate best entry point (BEP) judgment, for every article where they highlighted relevant text. Table 2 presents statistics on the best entry point offset, on the first highlighted or relevant character, and on the fraction of highlighted text in relevant articles. We first look at the BEPs. The mean BEP is well within the article with offset 1,738 but the distribution is very skewed with a median BEP offset of only 14. Figure 3 shows the distribution of the character offsets of the 4,887 best entry points. It is clear that the overwhelming majority of BEPs is at the beginning of the article.

The statistics of the first highlighted or relevant character (FRC) in Table 2 give very similar numbers as the BEP offsets: the mean offset of the first relevant character is 1,816 but the median offset is only 20. This suggests a relation between the BEP offset and the FRC offset. Figure 4 shows a scatter plot the BEP and FRC offsets. Two observations present themselves. First, there is a clear diagonal where the BEP is positioned exactly at the first highlighted character in the article. Second, there is also a vertical line at BEP offset zero, indicating a tendency to put the BEP at the start of the article even when the relevant text appears later on.

Finally, the statistics on the fraction of highlighted text in Table 2 show that amount of relevant text varies from almost nothing to almost everything. The mean fraction is 0.55, and the median is 0.58, indicating that typically over half the article is relevant. Given that the majority of relevant articles contain such a large fraction of relevant text plausibly explains that BEPs being frequently positioned on or near the start of the article.

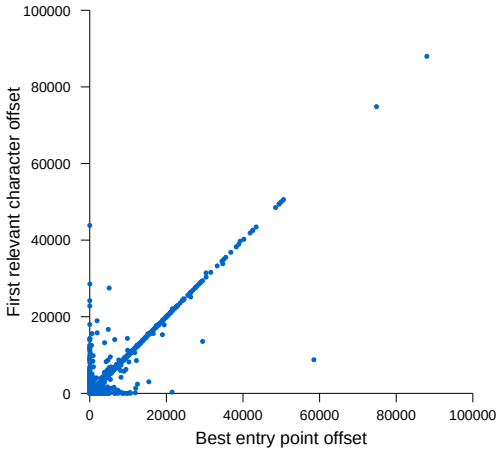

Fig. 4. Scatter plot of best entry point offsets versus the first relevant character.

Table 3. 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)

3.4 Questionnaires

At INEX 2008, 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 3) featured 20 questions capturing contextual data on the search request. The post-assessment

Table 4. 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)

questionnaire (shown in Table 4) featured 14 questions capturing further contextual data on the search request, and the way the topic has been judged (a few questions on GPXrai were added to the end).

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 2008 Ad Hoc Track.

4 Ad Hoc Retrieval Results

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

4.1 Participation

A total of 163 runs were submitted by 23 participating groups. Table 5 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, Passage or FOL). Unfortunately, no less than 27 runs turned out to be invalid and will only be evaluated with respect to their "article retrieval" value in Section 6.

Participants were allowed to submit up to three element result-type runs per task and three passage result-type runs per task (for all three tasks). This

Table 5. Participants in the Ad Hoc Track.

Id	Participant	Focused	Relevant in Context	Best in Context	CO query	CAS query	Element results	Passage results	FOL results	# valid runs	# submitted runs
4	University of Otago	0	6	0	6	0	3	3	0	6	6
5	Queensland University of Technology	6	6	6	15	3	9	9	0	18	18
6	University of Amsterdam	6	6	3	9	6	13	0	2	15	15
9	University of Helsinki	3	0	0	3	0	3	0	0	3	3
10	Max-Planck-Institut Informatik	3	1	1	5	0	5	0	0	5	5
12	University of Granada	3	3	3	9	0	9	0	0	9	9
14	University of California, Berkeley	2	0	1	3	0	3	0	0	3	3
16	University of Frankfurt	1	3	3	0	7	7	0	0	7	9
22	ENSM-SE	2	0	0	2	0	0	0	2	2	9
25	Renmin University of China	3	0	1	2	2	4	0	0	4	4
29	INDIAN STATISTICAL INSTITUTE	3	0	0	3	0	3	0	0	3	3
37	Katholieke Universiteit Leuven	6	0	0	3	3	6	0	0	6	6
40	IRIT	0	0	2	1	1	2	0	0	2	6
42	University of Toronto	2	0	0	0	2	2	0	0	2	3
48	LIG	3	2	0	5	0	5	0	0	5	5
55	Doshisha University	0	0	1	0	1	1	0	0	1	3
56	JustSystems Corporation	3	3	3	6	3	9	0	0	9	9
60	Saint Etienne University	3	0	0	3	0	3	0	0	3	9
61	Universit Libre de Bruxelles	0	0	0	0	0	0	0	0	0	2
68	University Pierre et Marie Curie - LIP6	2	0	0	2	0	2	0	0	2	2
72	University of Minnesota Duluth	2	2	2	6	0	6	0	0	6	6
78	University of Waterloo	3	3	4	10	0	8	2	0	10	13
92	University of Lyon3	5	5	5	15	0	15	0	0	15	15
	Total runs	61	40	35	108	28	118	14	4	136	163

totaled to 18 runs per participant.[3] The submissions are spread well over the ad hoc retrieval tasks with 61 submissions for Focused, 40 submissions for Relevant in Context, and 35 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 6 shows the best run of the top 10 participating groups. The first column gives the

[3] As it turns out, two groups submitted more runs than allowed: *University of Lyon3* submitted 6 extra element runs, and *University of Amsterdam* submitted 4 extra element runs. At this moment, we have not decided on any repercussions other than mentioning them in this footnote.

Table 6. Top 10 Participants in the Ad Hoc Track Focused Task.

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p78-FOERStep	0.7660	0.6897	0.5714	0.4908	0.2076
p10-TOPXCOarti	0.6808	0.6799	0.5812	0.5372	0.2981
p48-LIGMLFOCRI	0.7127	0.6678	0.5223	0.4229	0.1446
p92-manualQEin*	0.6664	0.6664	0.6139	0.5583	0.3077
p9-UHelRun394	0.7109	0.6648	0.5558	0.5044	0.2268
p60-JMUexpe142	0.6918	0.6640	0.5800	0.5071	0.2347
p14-T2FBCOPARA	0.7319	0.6427	0.4908	0.4036	0.1399
p29-LMnofb020	0.6855	0.6365	0.5566	0.5152	0.2868
p25-weightedfi	0.6553	0.6346	0.5495	0.5263	0.2661
p5-GPX1COFOCe	0.6818	0.6344	0.5693	0.5180	0.2592

participant, see Table 5 for the full name of group. 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]).

University of Waterloo Element retrieval run using the CO query. Description: the run uses the Okapi BM25 model in Wumpus to score all content-bearing elements such as sections and paragraphs using Okapi BM25. In addition, scores were boosted by doubling the tf values of the first 10 words of an element.

Max-Planck-Institut für Informatik Element retrieval run using the CO query. Description: The TopX system retrieving only article elements, using a linear combination of a BM25 content score with a BM25 proximity score that also takes document structure into accout.

LIG Grenoble An element retrieval run using the CO query. Description: Based on a language Model using a Dirichlet smoothing, and equally weighting element score and its context score, where the context score are based on the collection-links in Wikipedia.

University of Lyon3 A *manual* element retrieval run using the CO query. Description: Using indri search engine in Lemur with manually expanded queries from CO, description and narrative fields. The run is retrieving only articles.

University of Helsinki An element retrieval run using the CO query. Description: A special phrase index was created based on the detection of phrases in the collection, where the phrases are replication three times—effectively boosting query word occurrences in phrases. In addition, a standard keyword index was used. The run using BM25 is a combination of the retrieval status value on the word-index (94% of weight) and the phrase-index (6% of weight).

Saint Etienne University An element retrieval run using the CO query. Description: A probabilistic model used to evaluate a weight for each tag: "the probability that tags distinguishes terms which are the most relevant", i.e.

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

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p78-RICBest	0.4100	0.3454	0.2767	0.2202	0.2278
p92-manualQEin*	0.4175	0.3589	0.2692	0.2095	0.2106
p5-GPX1CORICe	0.3759	0.3441	0.2677	0.2151	0.2106
p10-TOPXCOallA	0.3681	0.3108	0.2386	0.1928	0.1947
p4-WHOLEDOC	0.3742	0.3276	0.2492	0.1962	0.1929
p6-inex08artB	0.3510	0.3008	0.2216	0.1741	0.1758
p72-UMDRic2	0.3853	0.3361	0.2357	0.1894	0.1724
p12-p8u3exp511	0.2966	0.2726	0.2169	0.1621	0.1582
p56-VSMRIP05	0.3281	0.2647	0.2113	0.1616	0.1500
p48-LIGMLRIC4O	0.3634	0.3115	0.2327	0.1721	0.1497

based on the fact that the tag contains relevant or non relevant passages. The resulting tag weights are incorporated into an element-level run with BM25 weighting.

Based on the information from these and other participants:

- All ten runs use the CO query. The fourth run, *p92-manualQEin*, uses a manually expanded query using words from the description and narrative fields. The eighth run, *p29-LMnofb020*, is an automatic run using the title and description fields. All other runs use only the CO query in the title field.
- All runs retrieve elements as results.
- The systems at rank second (*p10-TOPXCOarti*), fourth (*p92-manualQEin*), and eighth (*p29-LMnofb020*), are retrieving only full articles.

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 7 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 5 for the full name of group. 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).

University of Waterloo Element retrieval run using the CO query. Description: the run uses the Okapi BM25 model in Wumpus to score all content-bearing elements such as sections and paragraphs using Okapi BM25, and grouped the results by articles and ranked the articles by their best scoring element.

University of Lyon3 A manual element retrieval run using the CO query. Description: the same as the Focused run above. In fact it is literally the same article ranking as the Focused run. Recall that the run is retrieving only whole articles.

Queensland University of Technology Element retrieval run using the CO query. Description: GPX run using a `//*[about(.,keywords)]` query, serving non-overlapping elements grouped per article, with the articles ordered by their best scoring element.

Max-Planck-Institut für Informatik Element retrieval run using the CO query. Description: An element retrieval run using the new BM25 scoring function (i.e., considering each element as "document" and then computing a standard BM25 model), selecting non-overlapping elements based on score, and grouping them per article with the articles ranked by their highest scoring element.

University of Otago Element retrieval run using the CO query. Description: BM25 is used to select and rank the top 1,500 documents and whole documents are selected as the passage. The run is retrieving only whole articles.

Based on the information from these and other participants:

- The runs ranked sixth (*p6-inex08artB*) and ninth (*p56-VSMRIP05*) are using the CAS query. The run ranked second, *p92-manualQEin*, is using a manually expanded query based on keywords in the description and narrative. All other runs use only the CO query in the topic's title field.
- All runs retrieve elements as results.
- Solid article ranking seems a prerequisite for good overall performance, with second best run, *p92-manualQEin*, the fifth best run, *p4-WHOLEDOC*, and the ninth best run, *p56-VSMRIP05*, retrieving only full articles.

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 8 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 5 for the full name of group. 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).

University of Waterloo Element retrieval run using the CO query. Description: the run uses the Okapi BM25 model in Wumpus to score all content-bearing elements such as sections and paragraphs using Okapi BM25, and kept only the best scoring element per article.

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

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p78-BICER	0.3896	0.3306	0.2555	0.2019	0.2238
p92-manualQEin*	0.4144	0.3688	0.2834	0.2244	0.2197
p25-weightedfi	0.3510	0.3058	0.2531	0.2042	0.2037
p5-GPX1COBICp	0.3711	0.3395	0.2605	0.2046	0.1989
p6-submitinex	0.3475	0.2898	0.2236	0.1706	0.1709
p10-TOPXCOallB	0.2417	0.2374	0.1913	0.1550	0.1708
p12-p8u3exp501	0.2546	0.2331	0.1952	0.1503	0.1468
p72-UMDBIC1	0.3192	0.2752	0.1891	0.1474	0.1455
p56-VSMRIP08	0.2269	0.2038	0.1748	0.1403	0.1317
p55-KikoriBest	0.2041	0.1958	0.1552	0.1210	0.0960

University of Lyon3 A manual element retrieval run using the CO query. Description: the same as the Focused and Relevant in Context runs above. In fact all three runs have literally the same article ranking. This run is retrieving the start of the whole article as best entry point, in other words an article retrieval run.

Renmin University of China Element retrieval run using the CO query. Description: using language model to compute RSV at leaf level combined with aggregation at retrieval time, assuming independence.

Queensland University of Technology Run retrieving ranges of elements using the CO query. The run is always returning a whole article, setting the BEP at the very start of the article. Description: GPX run using a `//*[about(.,keywords)]` query, ranking articles by their best scoring element, but transformed to return the complete article as a passages. This is effectively an article level GPX run.

University of Amsterdam Run retrieving FOL passages using the CO query. Description: language model with local indegree prior, setting the BEP always at the start of the article. Since the offset is always zero, this is similar to an article retrieval run.

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, we see runs putting the BEP at the start of all the retrieved articles at rank two (*p92-manualQEin*), rank four (*p5-GPX1COBICp*), and rank five (*p6-submitinex*).
- The fourth ranked run, *p5-GPX1COBICp*, uses ranges of elements, albeit a degenerate case where always the full article is selected. The fifth run, *p6-submitinex*, uses fol passages, albeit again a degenerate case where the BEP is always the zero offset.
- With the exception of the runs at rank nine (*p56-VSMRIP08*) and ten (*p55-KikoriBest*), which used the CAS query, all the other best runs per group use the CO query.

Table 9. Statistical significance (t-test, one-tailed, 95%).

(a) Focused Task

	1	2	3	4	5	6	7	8	9	10
p78	-	-	-	-	-	-	-	-	-	-
p10		-	-	-	-	-	-	-	-	-
p48			-	-	-	-	-	-	-	-
p92				-	-	-	-	-	-	-
p9					-	-	-	-	-	-
p60						-	-	-	-	-
p14							-	-	-	-
p29								-	-	-
p25									-	-
p5										

(b) Relevant in Context Task

	1	2	3	4	5	6	7	8	9	10
p78	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
p92		-	-	-	⋆	⋆	⋆	⋆	⋆	⋆
p5			⋆	-	⋆	⋆	⋆	⋆	⋆	⋆
p10				-	-	-	⋆	⋆	⋆	⋆
p4					-	-	⋆	⋆	⋆	⋆
p6						-	-	-	⋆	⋆
p72							-	-	⋆	⋆
p12								-	-	⋆
p56									-	⋆
p48										

(c) Best in Context Task

	1	2	3	4	5	6	7	8	9	10
p78	-	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
p92		-	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆
p25			-	⋆	⋆	⋆	⋆	⋆	⋆	⋆
p5				⋆	-	⋆	⋆	⋆	⋆	⋆
p6					-	-	-	⋆	⋆	⋆
p10						-	-	⋆	⋆	⋆
p12							-	-	⋆	⋆
p72								-	⋆	⋆
p56									⋆	⋆
p55										

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 9 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 6 through 10, the third ranked system better than ranks 4 and 6 through 10, and the fourth and fifth ranked systems better than ranks 8 through 10. For the Best in Context Task, we see that the top run is significantly better than ranks 4 through 10, the second and third runs significantly better than than ranks 5 to 10. The fourth ranked system is better than the systems ranked 5 and 7 to 10, and the fifth ranked system better than ranks 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 18 submissions using ranges of elements of FOL-passage results, from in total 5 participating groups. 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. Table 10 shows the best runs using ranges of elements or FOL passages for the three ad hoc tasks. All these runs use the CO query. As it turns out, the best focused run using passages ranks outside the

Table 10. Ad Hoc Track: Runs with ranges of elements or FOL passages.

(a) Focused Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p5-GPX2COFOCp	0.6311	0.6305	0.5365	0.4719	0.2507
p22-EMSEFocuse*	0.6757	0.5724	0.4487	0.3847	0.1555

(b) Relevant in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p4-WHOLEDOCPA	0.3742	0.3276	0.2492	0.1962	0.1929
p5-GPX1CORICp	0.3566	0.3220	0.2430	0.1875	0.1900

(c) Best in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p5-GPX1COBICp	0.3711	0.3395	0.2605	0.2046	0.1989
p6-submitinex	0.3475	0.2898	0.2236	0.1706	0.1709
p78-BICPRplus	0.2651	0.2252	0.1666	0.1268	0.1254

top scoring runs in Table 6; the best relevant in context run using passages is ranked fifth among the top scoring runs in Table 7; and the best best in context run using passages is ranked fourth among the top scoring runs in Table 8. 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 [4, 5].

However, looking at the runs in more detail, their character is often unlike what one would expect from a "passage" retrieval run. For Focused, *p5-GPX2COFOCp* is an article run using ranges of elements; and *p22-EMSEFocuse* is a manual query run using FOL passages. For Relevant in Context, both *p4-WHOLEDOCPA* and *p5-GPX1CORICp* are article runs using ranges of elements. For Best in Context, *p5-GPX1COBICp* is an article runs using ranges of elements; *p6-submitinex* is an article run using FOL passages; and *p78-BICPRplus* is an element retrieving run using ranges of elements. So, all but two of the runs retrieve only articles. Hence, this is not sufficient evidence to warrant any conclusion on the effectiveness of passage level results. 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 look at the relative effectiveness of the keyword (CO) and structured (CAS) queries. As we saw above, in Section 4, one of the best runs per group for the Relevant in Context Task, and two 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 11 show the distribution of target

Table 11. CAS query target elements over all 135 topics.

Target Element	Frequency
*	51
section	39
article	30
p	11
figure	3
body	1

Table 12. Ad Hoc Track CAS Topics: CO runs (left-hand side) versus CAS runs (right-hand side).

(a) Focused Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP	Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p60-JMUexpe136	0.7321	0.7245	0.6416	0.5936	0.2934	p6-inex08artB	0.6514	0.6379	0.5901	0.5248	0.2261
p48-LIGMLFOCRI	0.7496	0.7209	0.5307	0.4440	0.1570	p56-VSMRIP02	0.7515	0.6333	0.4781	0.3667	0.1400
p78-FOER	0.7263	0.7089	0.6084	0.5485	0.2225	p5-GPX3COSFOC	0.6232	0.6220	0.5509	0.4626	0.2137
p5-GPX1COFOCe	0.7168	0.6972	0.6416	0.5616	0.2616	p25-RUCLLP08	0.5969	0.5969	0.5761	0.5545	0.2491
p29-LMnofb020	0.7193	0.6766	0.5926	0.5611	0.2951	p37-kulcaselem	0.6824	0.5626	0.3532	0.2720	0.1257
p10-TOPXCOallF	0.7482	0.6680	0.5555	0.4871	0.1925	p42-B2U0visith	0.6057	0.5364	0.4830	0.4449	0.1739
p25-weightedfi	0.6665	0.6634	0.5907	0.5646	0.2671	p16-001RunofUn	0.3111	0.2269	0.1675	0.1206	0.0365
p6-inex08artB	0.6689	0.6571	0.5570	0.4961	0.2104						
p9-UHelRun394	0.7024	0.6567	0.5602	0.5221	0.2255						
p72-UMDFocused	0.7259	0.6491	0.4947	0.3812	0.1115						

(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
p78-RICBest	0.4808	0.3818	0.2994	0.2274	0.2485	p6-inex08artB	0.3757	0.3113	0.2334	0.1847	0.1937
p5-GPX1CORICe	0.3946	0.3518	0.2670	0.2169	0.2166	p5-GPX3COSRIC	0.3482	0.3232	0.2381	0.1923	0.1764
p4-WHOLEDOC	0.4020	0.3534	0.2508	0.2009	0.2125	p56-VSMRIP05	0.3401	0.2796	0.2143	0.1616	0.1501
p10-TOPXCOallA	0.3892	0.3220	0.2366	0.1910	0.1967	p16-009RunofUn	0.0153	0.0156	0.0123	0.0095	0.0023
p92-manualQEin*	0.3818	0.3395	0.2515	0.1970	0.1933						
p6-inex08artB	0.3762	0.3140	0.2293	0.1790	0.1900						
p72-UMDRic2	0.3952	0.3434	0.2289	0.1868	0.1745						
p12-p8u3exp511	0.3229	0.2880	0.2245	0.1631	0.1680						
p48-LIGMLRIC4O	0.3818	0.3408	0.2461	0.1832	0.1583						
p56-VSMRIP04	0.2315	0.2031	0.1675	0.1368	0.1275						

(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
p78-BICER	0.3935	0.3386	0.2544	0.1956	0.2172	p5-GPX3COSBIC	0.3109	0.2883	0.2235	0.1780	0.1661
p25-weightedfi	0.3342	0.3065	0.2390	0.1958	0.2004	p56-VSMRIP08	0.2123	0.1911	0.1481	0.1214	0.1228
p5-GPX1COBICp	0.3663	0.3358	0.2504	0.1926	0.1983	p40-xfirmcos07	0.2381	0.1794	0.1348	0.1078	0.0908
p92-manualQEin*	0.3728	0.3383	0.2599	0.2082	0.1952	p55-KikoriBest	0.1817	0.1721	0.1422	0.1123	0.0803
p10-TOPXCOallB	0.2424	0.2419	0.1788	0.1457	0.1727	p16-006RunofUn	0.0307	0.0347	0.0307	0.0261	0.0128
p6-submitinex	0.3505	0.3062	0.2278	0.1713	0.1716						
p12-p8u3exp501	0.2586	0.2397	0.1934	0.1425	0.1448						
p72-UMDBIC1	0.3222	0.2751	0.1757	0.1377	0.1369						
p56-VSMRIP09	0.1562	0.1537	0.1377	0.1127	0.1038						
p40-xfirmbicco	0.1594	0.1546	0.1367	0.1137	0.0661						

elements. In total 86 topics had a non-trivial CAS query.[4] These CAS topics are numbered 544–550, 553–556, 564, 567, 568, 572, 574, 576–578, 580, 583, 584, 586–591, 597–605, 607, 608, 610, 615–625, 627, 629–633, 635–640, 646, 651–655, 658, 659, 661–670, 673, and 675–678. As it turned out, 39 of these CAS topics were assessed. The results presented here are restricted to the 39 CAS topics.

[4] Note that some of the wild-card topics (using the "*" target) in Table 11 had non-trivial about-predicates and hence have not been regarded as trivial CAS queries.

Table 13. Top 10 Participants in the Ad Hoc Track: Article retrieval.

Participant	P5	P10	1/rank	map	bpref
p78-BICER	0.6286	0.5343	0.8711	0.3789	0.3699
p92-manualQEin*	0.6429	0.5886	0.8322	0.3629	0.3924
p10-TOPXCOarti	0.5943	0.5443	0.8635	0.3516	0.3628
p5-GPX1COBICe	0.5743	0.5257	0.7868	0.3413	0.3588
p37-kulcoeleme	0.5286	0.4557	0.7468	0.3268	0.3341
p25-weightedfi	0.4971	0.4657	0.7192	0.3255	0.3355
p29-VSMfbElts0	0.5543	0.4857	0.7955	0.3195	0.3388
p60-JMUexpe136	0.5457	0.4857	0.7843	0.3192	0.3383
p9-UHelRun293	0.5829	0.5029	0.7766	0.3144	0.3323
p4-SWKL200	0.5714	0.5000	0.7950	0.3107	0.3297

Table 12 lists the top 10 participants measured using just the 39 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 CAS runs score lower than the CO query runs. For the Relevant in Context Task, the best CAS run would have ranked fifth among the CO runs. For the Best in Context Task, the best CAS run would rank seventh among the CO runs. Overall, we see the that teams submitting runs with both types of queries have higher scoring CO runs, with participant 6 as a notable exception for Relevant in Context.

6 Analysis of Article Retrieval

In this section, we will look in detail at the effectiveness of Ad Hoc Track submissions as article retrieval systems. We look first at the article rankings in terms of the Ad Hoc Track judgments—treating every article that contains highlighted text as relevant. Then, we look at the article rankings in terms of the clicked pages for the topics from the proxy log—treating every clicked article as relevant.

6.1 Article retrieval: Relevance Judgments

We will first look at the topics judged during INEX 2008, the same topics as in earlier sections, but now using the judgments to derive standard document-level relevance by regarding an article as relevant if some part of it is highlighted by the assessor. Throughout this section, we derive an article retrieval run from every submission using a first-come, first served mapping. That is, we simply keep every first occurrence of an article (retrieved indirectly through some element contained in it) and ignore further results from the same article.

We use `trec_eval` to evaluate the mapped runs and qrels, and use mean average precision (map) as the main measure. Since all runs are now article retrieval runs, the differences between the tasks disappear. Moreover, runs violating the task requirements—most notably non-overlapping results for all tasks, and having scattered results from the same article in relevant in context—are now also considered, and we work with all 163 runs submitted to the Ad Hoc Track.

Table 14. Top 10 Participants in the Ad Hoc Track: Article retrieval per task over judged topics (left) and clicked pages (right).

(a) Focused Task

Participant	P5	P10	1/rank	map	bpref	Participant	P5	P10	1/rank	map	bpref
p92-manualQEin*	0.6429	0.5886	0.8322	0.3629	0.3924	p5-Terrier	0.1594	0.0877	0.5904	0.5184	0.8266
p10-TOPXCOarti	0.5943	0.5443	0.8635	0.3516	0.3628	p6-inex08artB	0.1623	0.0870	0.5821	0.5140	0.8150
p5-GPX1COFOCp	0.5743	0.5257	0.7868	0.3413	0.3588	p92-autoindri0	0.1565	0.0884	0.5601	0.4853	0.8211
p37-kulcoeleme	0.5286	0.4557	0.7468	0.3268	0.3341	p60-JMUexpe142	0.1536	0.0862	0.5624	0.4853	0.8250
p78-FOER	0.5800	0.5043	0.7995	0.3259	0.3277	p48-LIGMLFOCRI	0.1449	0.0833	0.5191	0.4596	0.7153
p29-VSMfbElts0	0.5543	0.4857	0.7955	0.3195	0.3388	p10-TOPXCOarti	0.1522	0.0841	0.5164	0.4538	0.8167
p25-weightedfi	0.4971	0.4657	0.7192	0.3195	0.3324	p78-FOER	0.1304	0.0819	0.4979	0.4404	0.8136
p60-JMUexpe136	0.5457	0.4857	0.7843	0.3192	0.3383	p40-xfirmcos07	0.1217	0.0717	0.4301	0.3748	0.7184
p9-UHelRun293	0.5829	0.5029	0.7766	0.3144	0.3323	p55-KikoriFocu	0.1261	0.0732	0.4334	0.3727	0.7785
p6-inex08artB	0.5514	0.4800	0.7851	0.3010	0.3109	p22-EMSEFocuse*	0.1203	0.0783	0.4233	0.3704	0.8105

(b) Relevant in Context Task

Participant	P5	P10	1/rank	map	bpref	Participant	P5	P10	1/rank	map	bpref
p92-manualQEin*	0.6429	0.5886	0.8322	0.3629	0.3924	p5-Terrier	0.1594	0.0877	0.5904	0.5184	0.8266
p5-GPX1CORICp	0.5743	0.5257	0.7868	0.3413	0.3588	p6-inex08artB	0.1623	0.0870	0.5821	0.5140	0.8150
p78-RICBest	0.5886	0.5029	0.8161	0.3404	0.3422	p60-JMUexpe150	0.1536	0.0862	0.5624	0.4853	0.8167
p10-TOPXCOallA	0.5314	0.4843	0.8226	0.3122	0.3279	p92-autoindri0	0.1565	0.0884	0.5601	0.4853	0.8211
p60-JMUexpe150	0.5886	0.4900	0.8266	0.3119	0.3185	p48-LIGMLRIC4O	0.1464	0.0841	0.5238	0.4647	0.7081
p4-SWKL200	0.5714	0.5000	0.7950	0.3107	0.3297	p78-RICBest	0.1348	0.0812	0.4979	0.4422	0.8126
p6-inex08artB	0.5514	0.4800	0.7851	0.3010	0.3109	p10-TOPXCOallA	0.1333	0.0775	0.5139	0.4397	0.7863
p56-VSMRIP05	0.5486	0.4543	0.7752	0.2880	0.3045	p72-UMDRic2	0.1275	0.0717	0.4560	0.4088	0.7526
p72-UMDRic2	0.6000	0.5200	0.8579	0.2739	0.3048	p4-SWKL200	0.1159	0.0732	0.4168	0.3701	0.8007
p22-EMSERICStr*	0.5057	0.4543	0.7079	0.2728	0.3064	p55-KikoriRele	0.1232	0.0710	0.4125	0.3501	0.7712

(c) Best in Context Task

Participant	P5	P10	1/rank	map	bpref	Participant	P5	P10	1/rank	map	bpref
p78-BICER	0.6286	0.5343	0.8711	0.3789	0.3699	p5-Terrier	0.1594	0.0877	0.5904	0.5184	0.8266
p92-manualQEin*	0.6429	0.5886	0.8322	0.3629	0.3924	p6-submitinex	0.1594	0.0862	0.5673	0.4976	0.8164
p5-GPX1COBICe	0.5743	0.5257	0.7868	0.3413	0.3588	p92-autoindri0	0.1565	0.0884	0.5601	0.4853	0.8211
p10-TOPXCOallB	0.5314	0.4843	0.8226	0.3290	0.3344	p60-JMUexpe151	0.1536	0.0855	0.5624	0.4844	0.8214
p25-weightedfi	0.4971	0.4657	0.7192	0.3255	0.3355	p78-BICPRplus	0.1522	0.0841	0.5432	0.4673	0.7799
p60-JMUexpe157	0.5714	0.5000	0.8215	0.3098	0.3176	p10-TOPXCOallB	0.1333	0.0775	0.5139	0.4398	0.8205
p6-submitinex	0.5486	0.4757	0.7793	0.2984	0.3086	p72-UMDBIC1	0.1275	0.0710	0.4482	0.4011	0.7398
p56-VSMRIP08	0.5486	0.4543	0.7752	0.2880	0.3045	p40-xfirmcos07	0.1217	0.0717	0.4301	0.3748	0.7160
p72-UMDBIC2	0.5914	0.5171	0.8511	0.2761	0.3022	p55-KikoriBest	0.1261	0.0732	0.4334	0.3727	0.7785
p12-p8u3exp501	0.4829	0.4371	0.7044	0.2723	0.3061	p56-VSMRIP08	0.1130	0.0659	0.3943	0.3445	0.7258

Table 13 shows the best run of the top 10 participating groups. The first column gives the participant, see Table 5 for the full name of group. The second and third column give the precision at ranks 5 and 10, respectively. The fourth column gives the mean reciprocal rank. The fifth column gives mean average precision. The sixth column gives binary preference measures (using the top R judged non-relevant documents). Recall from the above that second ranked run (*p92-manualQEin*) is a manual article retrieval run submitted to all three tasks. Also the run ranked three (*p10-TOPXCOarti*) and the run ranked seven (*p60-JMUexpe136*) retrieve exclusively articles. The relative effectiveness of these article retrieval runs in terms of their article ranking is no surprise. Furthermore, we see submissions from all three ad hoc tasks. Most notably runs from the Best in Context task at ranks 1, 2, 4, and 6; runs from the Focused task at ranks 2, 3, 5, 7, 8, and 9; and runs from the Relevant in Context task at ranks 2 and 10.

If we break-down all runs over the original tasks, shown on the left-hand side of Table 14, we can compare the ranking to Section 4 above. We see some runs that are familiar from the earlier tables: three Focused runs correspond to

Table 15. Top 10 Participants in the Ad Hoc Track: Clicked articles.

Participant	P5	P10	1/rank	map	bpref
p5-Terrier	0.1594	0.0877	0.5904	0.5184	0.8266
p6-inex08artB	0.1623	0.0870	0.5821	0.5140	0.8150
p60-JMUexpe150	0.1536	0.0862	0.5624	0.4853	0.8167
p92-autoindri0	0.1565	0.0884	0.5601	0.4853	0.8211
p78-BICPRplus	0.1522	0.0841	0.5432	0.4673	0.7799
p48-LIGMLRIC4O	0.1464	0.0841	0.5238	0.4647	0.7081
p10-TOPXCOarti	0.1522	0.0841	0.5164	0.4538	0.8167
p72-UMDRic2	0.1275	0.0717	0.4560	0.4088	0.7526
p40-xfirmcos07	0.1217	0.0717	0.4301	0.3748	0.7184
p55-KikoriFocu	0.1261	0.0732	0.4334	0.3727	0.7785

Table 6, five Relevant in Context runs correspond to Table 7, and seven Best in Context runs correspond to Table 8. More formally, we looked at how the two system rankings correlate using Kendall's Tau.

- Over all 61 Focused task submissions the system rank correlation is 0.517 between iP[0.01] and map, and 0.568 between MAiP and map.
- Over all 40 Relevant in Context submissions the system rank correlation between MAgP and map is 0.792.
- Over all 35 Best in Context submissions the system rank correlation is 0.795

Overall, we see a reasonable correspondence between the rankings for the ad hoc tasks in Section 4 and the rankings for the derived article retrieval measures. The correlation with the Focused task runs is much lower than with the Relevant in Context and Best in Context tasks. This makes sense, since the ranking of articles is an important part of the two "in context" tasks.

6.2 Article retrieval: Clicked pages

In addition to the topics created and assessed by INEX participants, we also included 150 queries derived from a proxy log, and can also construct pseudo-relevance judgments by regarding every clicked Wikipedia article as relevant.

Table 15 shows the best run of the top 10 participating groups. The first column gives the participant, see Table 5 for the full name of group. The second and third column give the precision at ranks 5 and 10, respectively. The fourth column gives the mean reciprocal rank. The fifth column gives mean average precision. The sixth column gives binary preference measures (using the top R judged non-relevant documents). Compared to the judged topics, we immediately see much lower scores for the early precision measures (precision at 5 and 10, and reciprocal ranks), while at the same time higher scores for the overall measures (map and bpref). This is a result of the very low numbers of relevant documents, 1.8 on average, that make it impossible to get a grips on recall aspects. The runs ranked first (*p5-Terrier*), fourth (*p92-autoindri0*), and seventh (*p10-TOPXCOarti*) retrieve exclusively full articles. Again, it is no great surprise that these runs do well for the task of article retrieval.

The resulting ranking is quite different from the article ranking based on the judged ad hoc topics in Table 13. They have only one run in common, although they agree on five of the ten participants. Looking, more formally, at the system rank correlations between the two types of article retrieval we see the following.

- Over all 163 submissions, the system rank correlation is 0.357.
- Over the 76 Focused task submissions, the correlation is 0.356.
- Over the 49 Relevant in task submissions, the correlation is 0.366.
- Over the 38 Best in Context task submissions, the correlation is 0.388.

Hence the judged topics above and the topics derived from the proxy log vary considerable. A large part of the explanation is the dramatic difference between the numbers of relevant articles, with 70 on average for the judged topics and 1.8 on average for the proxy log topics.

7 Discussion and Conclusions

In this paper we provided an overview of the INEX 2008 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). We also look at effectiveness in term of article retrieval, both using the judged topics and using queries and clicks derived from a proxy log.

When examining the relative effectiveness of CO and CAS we found that for all tasks the best scoring runs used the CO query. This is in contrast with earlier results showing that structural hints can help promote initial precision [8]. Part of the explanation may be in the low number of CAS submissions (28) in comparison with the number of CO submissions (108). Only 39 of the 70 judged topics had a non-trivial CAS query, and the majority of those CAS queries made only reference to particular tags and not on their structural relations. This may have diminished the value of the CAS query in comparison with earlier years.

Given the efforts put into the fair comparison of element and passage retrieval approaches, the number of passage and FOL submissions was disappointing. Eighteen submissions used ranges of elements or FOL passage results, whereas 118 submissions used element results. In addition, many of the passage or FOL submissions used exclusively full articles as results. Although we received too few non-element runs to draw clear conclusions, we saw that the passage based approaches were competitive, but not superior to element based approaches. This outcome is consistent with earlier results in [4, 5].

As in earlier years, we saw that article retrieval is reasonably effective at XML-IR: for each of the ad hoc tasks there were three article-only runs among the best runs of the top 10 groups. When looking at the article rankings inherent

in all Ad Hoc Track submissions, we saw that again three of the best runs of the top 10 groups in terms of article ranking (across all three tasks) were in fact article-only runs. This suggests that element-level or passage-level evidence is still valuable for article retrieval. When comparing the system rankings in terms of article retrieval with the system rankings in terms of the ad hoc retrieval tasks, over the exact same topic set, we see a reasonable correlation especially for the two "in context" tasks. The systems with the best performance for the ad hoc tasks, also tend to have the best article rankings. Since finding the relevant articles can be considered a prerequisite for XML-IR, this should not come as a surprize. In addition, the Wikipedia's encyclopedic structure with relatively short articles covering a single topic results in relevant articles containing large fractions of relevant text (with a mean of 55% of text being highlighted). While it is straightforward to define tasks and measures that strongly favor precision over recall, a more natural route would be to try to ellicit more focused information needs that have natural answers in short excerpts of text.

When we look at a different topic set derived from a proxy log, and a shallow set of clicked pages rather than a full-blown IR test collection, we see notable differences. Given the low number of relevant articles (1.8 on average) compared to the ad hoc judgments (70 on average), the clicked pages focus exclusively on precision aspects. This leads to a different system ranking, although there is still some agreement on the best groups. The differences between these two sets of topics require further analysis.

Finally, the Ad Hoc Track had two main research questions. The first main research question was the comparative analysis of element and passage retrieval approaches, hoping to shed light on the value of the document structure as provided by the XML mark-up. We found that the best performing system used predominantly element results, although the number of non-element retrieval runs submitted is too low to draw any definite conclusions. The second main research question was to compare focused retrieval directly to traditional article retrieval. We found that the best scoring Ad Hoc Track submissions also tend to have the best article ranking, and that the best article rankings were generated using element-level evidence. For both main research questions, 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 Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO, grants 612.066.513, 639.072.601, and 640.001.501).

Bibliography

[1] J. P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference*, pages 302–310, 1994.

[2] C. L. A. Clarke. Range results in XML retrieval. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 4–5, 2005.

[3] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 40:64–69, 2006.

[4] W. Huang, A. Trotman, and R. A. O'Keefe. Element retrieval using a passage retrieval approach. In *Proceedings of the 11th Australasian Document Computing Symposium (ADCS 2006)*, pages 80–83, 2006.

[5] K. Y. Itakura and C. L. A. Clarke. From passages into elements in XML retrieval. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, pages 17–22, 2007.

[6] J. Kamps and M. Koolen. On the relation between relevant passages and XML document structure. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, pages 28–32, 2007.

[7] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. The importance of morphological normalization for XML retrieval. In *Proceedings of the First INEX Workshop*, pages 41–48, 2003.

[8] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Articulating information needs in XML query languages. *Transactions on Information Systems*, 24:407–436, 2006.

[9] J. Kekäläinen and K. Järvelin. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53:1120–1129, 2002.

[10] J. A. Thom and J. Pehcevski. How well does best in context reflect ad hoc XML retrieval. In *Pre-Proceedings of INEX 2007*, pages 124–125, 2007.

[11] A. Trotman and S. Geva. Passage retrieval and other XML-retrieval tasks. In *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, pages 43–50, 2006.

A Appendix: Full run names

Group	Run	Label	Task	Query	Results	Notes
4	151	p4-SWKL200	RiC	CO	Pas	
4	152	p4-WHOLEDOC	RiC	CO	Ele	Article-only
4	153	p4-WHOLEDOCPA	RiC	CO	Pas	Article-only
5	122	p5-Terrier	BiC	CO	Pas	Article-only
5	123	p5-Terrier	Foc	CO	Pas	Article-only
5	124	p5-Terrier	RiC	CO	Pas	Article-only
5	133	p5-GPX2COFOCp	Foc	CO	Pas	Article-only
5	138	p5-GPX1COBICe	BiC	CO	Ele	
5	139	p5-GPX1COFOCe	Foc	CO	Ele	
5	140	p5-GPX1CORICe	RiC	CO	Ele	
5	141	p5-GPX3COSBIC	BiC	CAS	Ele	
5	142	p5-GPX3COSFOC	Foc	CAS	Ele	
5	143	p5-GPX3COSRIC	RiC	CAS	Ele	
5	144	p5-GPX1COBICp	BiC	CO	Pas	Article-only
5	145	p5-GPX1COFOCp	Foc	CO	Pas	Article-only
5	146	p5-GPX1CORICp	RiC	CO	Pas	Article-only
6	255	p6-submitinex	BiC	CO	FOL	Article-only
6	264	p6-inex08artB	RiC	CAS	Ele	
6	265	p6-inex08artB	RiC	CO	Ele	
6	268	p6-inex08artB	RiC	CAS	Ele	
6	269	p6-inex08artB	RiC	CO	Ele	
6	270	p6-inex08artB	Foc	CAS	Ele	
6	271	p6-inex08artB	Foc	CO	Ele	
6	274	p6-inex08artB	Foc	CO	Ele	
6	276	p6-inex08artB	Foc	CO	Ele	
9	174	p9-UHelRun293	Foc	CO	Ele	
9	176	p9-UHelRun394	Foc	CO	Ele	
10	91	p10-TOPXCOallF	Foc	CO	Ele	
10	92	p10-TOPXCOallB	BiC	CO	Ele	
10	93	p10-TOPXCOallA	RiC	CO	Ele	
10	207	p10-TOPXCOarti	Foc	?	Ele	Article-only
12	97	p12-p8u3exp501	BiC	CO	Ele	
12	100	p12-p8u3exp511	RiC	CO	Ele	
14	205	p14-T2FBCOPARA	Foc	CO	Ele	
16	233	p16-009RunofUn	RiC	CAS	Ele	
16	234	p16-006RunofUn	BiC	CAS	Ele	
16	244	p16-001RunofUn	Foc	CAS	Ele	
22	62	p22-EMSEFocuse	Foc	CO	Ele	Manual Invalid
22	66	p22-EMSEFocuse	Foc	CO	FOL	Manual
22	68	p22-EMSERICStr	RiC	CO	Ele	Manual Invalid
25	30	p25-RUCLLP08	Foc	CAS	Ele	

Continued on Next Page...

Group	Run	Label	Task	Query	Results	Notes
25	278	p25-weightedfi	Foc	CO	Ele	
25	282	p25-weightedfi	BiC	CO	Ele	
29	238	p29-VSMfbElts0	Foc	CO	Ele	
29	253	p29-LMnofb020	Foc	CO	Ele	Article-only
37	227	p37-kulcaselem	Foc	CAS	Ele	
37	230	p37-kulcoeleme	Foc	CO	Ele	
40	54	p40-xfirmbicco	BiC	CO	Ele	
40	296	p40-xfirmcos07	BiC	CAS	Ele	
40	297	p40-xfirmcos07	Foc	CAS	Ele	Invalid
42	299	p42-B2U0visith	Foc	CAS	Ele	
48	59	p48-LIGMLFOCRI	Foc	CO	Ele	
48	72	p48-LIGMLRIC4O	RiC	CO	Ele	
55	279	p55-KikoriFocu	Foc	CAS	Ele	Invalid
55	280	p55-KikoriRele	RiC	CAS	Ele	Invalid
55	281	p55-KikoriBest	BiC	CAS	Ele	
56	190	p56-VSMRIP02	Foc	CAS	Ele	
56	197	p56-VSMRIP04	RiC	CO	Ele	Article-only
56	199	p56-VSMRIP05	RiC	CAS	Ele	Article-only
56	202	p56-VSMRIP08	BiC	CAS	Ele	
56	224	p56-VSMRIP09	BiC	CO	Ele	
60	11	p60-JMUexpe136	Foc	CO	Ele	Article-only
60	53	p60-JMUexpe142	Foc	CO	Ele	
60	81	p60-JMUexpe150	RiC	CO	Ele	Invalid
60	82	p60-JMUexpe151	BiC	CO	Ele	Invalid
60	175	p60-JMUexpe157	BiC	CO	Ele	Invalid
72	106	p72-UMDFocused	Foc	CO	Ele	
72	154	p72-UMDBIC1	BiC	CO	Ele	
72	155	p72-UMDBIC2	BiC	CO	Ele	
72	277	p72-UMDRic2	RiC	CO	Ele	
78	156	p78-FOER	Foc	CO	Ele	
78	157	p78-FOERStep	Foc	CO	Ele	
78	160	p78-BICER	BiC	CO	Ele	
78	163	p78-BICPRplus	BiC	CO	Pas	
78	164	p78-RICBest	RiC	CO	Ele	
92	177	p92-autoindri0	BiC	CO	Ele	Article-only
92	178	p92-autoindri0	Foc	CO	Ele	Article-only
92	179	p92-autoindri0	RiC	CO	Ele	Article-only
92	183	p92-manualQEin	BiC	CO	Ele	Manual Article-only
92	184	p92-manualQEin	Foc	CO	Ele	Manual Article-only
92	185	p92-manualQEin	RiC	CO	Ele	Manual Article-only