Overview of the INEX 2009 Ad Hoc Track

Shlomo Geva[1], Jaap Kamps[2], Miro Lethonen[3],
Ralf Schenkel[4], James A. Thom[5], and Andrew Trotman[6]

[1] Queensland University of Technology, Brisbane, Australia
s.geva@qut.edu.au
[2] University of Amsterdam, Amsterdam, The Netherlands
kamps@uva.nl
[3] University of Helsinki, Helsinki, Finland
miro.lehtonen@helsinki.fi
[4] Max-Planck-Institut für Informatik, Saarbrücken, Germany
schenkel@mpi-sb.mpg.de
[5] RMIT University, Melbourne, Australia
james.thom@rmit.edu.au
[6] University of Otago, Dunedin, New Zealand
andrew@cs.otago.ac.nz

Abstract. This paper gives an overview of the INEX 2009 Ad Hoc Track. The main goals of the Ad Hoc Track were three-fold. The first goal was to investigate the impact of the collection scale and markup, by using a new collection that is again based on a the Wikipedia but is over 4 times larger, with longer articles and additional semantic annotations. For this reason the Ad Hoc track tasks stayed unchanged, and the Thorough Task of INEX 2002–2006 returns. The second goal was to study the impact of more verbose queries on retrieval effectiveness, by using the available markup as structural constraints—now using both the Wikipedia's layout-based markup, as well as the enriched semantic markup—and by the use of phrases. The third goal was to compare different result granularities by allowing systems to retrieve XML elements, ranges of XML elements, or arbitrary passages of text. This investigates the value of the internal document structure (as provided by the XML mark-up) for retrieving relevant information. The INEX 2009 Ad Hoc Track featured four tasks: For the *Thorough Task* a ranked-list of results (elements or passages) by estimated relevance was needed. 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 setup of the track, the results for the four 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. In addition, we look at the effectiveness of systems using a reference run with a solid article ranking, and of systems using the phrase query. Finally, we look at the ability of focused retrieval techniques to rank articles.

1 Introduction

This paper gives an overview of the INEX 2009 Ad Hoc Track. There are three main research questions underlying the Ad Hoc Track. The first main research question is the impact of the new collection—four times the size, with longer articles, and additional semantic markup—on focused retrieval. That is, what is the impact of collection size? What is the impact of document length, and hence the complexity of the XML structure in the DOM tree? The second main research question is the impact of more verbose queries—using either the XML structure, or using multi-word phrases. That is, what is the impact of semantic annotation on both the submitted queries, and their retrieval effectiveness? What is the impact of explicitly annotated multi-word phrases? The third 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 to identify where the relevant information is within a document?

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 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 to arbitrary passages of text [11].

The INEX 2009 Ad Hoc Track featured four tasks:

1. For the *Thorough Task* a ranked-list of results (elements or passages) by estimated relevance must be returned. It is evaluated by mean average interpolated precision relative to the highlighted (or believed relevant) text retrieved.
2. 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.
3. 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.
4. 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 four tasks, giving results for the top 10 participating groups and discussing their 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 2009 ad hoc retrieval tasks and measures. Section 3 details the collection, topics, and assessments of the INEX 2009 Ad Hoc Track. In Section 4, we report the results for the Thorough Task (Section 4.2); the Focused Task (Section 4.3); the Relevant in Context Task (Section 4.4); and the Best in Context Task (Section 4.5). Section 5 details particular types of runs (such as element versus passage, using phrases or using the reference run), 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, treating any article with highlighted text as relevant. 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

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

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. At INEX 2009, we allowed the submission of three types of results: XML elements, file-offset-length (FOL) text passages, and ranges of XML elements. The submission format for all tasks is a variant of the familiar TREC format extended with two additional fields.

```
topic Q0 file rank rsv run_id column_7 column_8
```

Here:

- The first column is the topic number.
- The second column (the query number within that topic) is currently unused and should always be Q0.
- The third column is the file name (without .xml) from which a result is retrieved, which is identical to the ¡id¿ of the Wikipedia
- The fourth column is the rank the document is retrieved.
- The fifth column shows the retrieval status value (RSV) or score that generated the ranking.
- The sixth column is called the "run tag" identifying the group and for the method used.

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, and (with the .xml extension removed) identical to the ⟨id⟩ of the Wikipedia document. That is, file 9996.xml contains the article as the target document from the Wikipedia collection with ⟨id⟩ 9996.

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.

```
/article[1]/body[1]/section[1]/p[1]
```

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 may then be identified unambiguously using the combination of its file name (or ⟨id⟩) in column 3 and the element path in column 7. Column 8 will not be used. Example:

```
1 Q0 9996 1 0.9999 I09UniXRun1 /article[1]/bdy[1]/sec[1]
1 Q0 9996 2 0.9998 I09UniXRun1 /article[1]/bdy[1]/sec[2]
1 Q0 9996 3 0.9997 I09UniXRun1 /article[1]/bdy[1]/sec[3]/p[1]
```

Here the results are from 9996 and select the first section, the second section, and the first paragraph of the third section.

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

A result element may then be identified unambiguously using the combination of its file name (or ⟨id⟩) in column 3 and an offset in column 7 and a length in column 8. The following example is effectively equivalent to the example element result above:

```
1 Q0 9996 1 0.9999 I09UniXRun1 465 3426
1 Q0 9996 2 0.9998 I09UniXRun1 3892 960
1 Q0 9996 3 0.9997 I09UniXRun1 4865 496
```

The results are from article 9996, and the first section starts at the 466th character (so 465 characters beyond the first character which has offset 0), and has a length of 3,426 characters.

Ranges of Elements To support ranges of elements, elemental passages can be specified by their containing elements. We only allow elemental paths (ending in an element, not a text-node in the DOM tree) plus an optional offset.

A result element may then be identified unambiguously using the combination of its file name (or ⟨id⟩) in column 3, its start at the element path in column 7, and its end at the element path in column 8. Example:

```
1 Q0 9996 1 0.9999 I09UniRun1 /article[1]/bdy[1]/sec[1] /article[1]/bdy[1]/sec[1]
```

Here the result is again the first section from 9996. Note that the seventh column will refer to the beginning of an element (or its first content), and the eighth column will refer to the ending of an element (or its last content). Note that this format is very convenient for specifying ranges of elements, e.g., the first three sections:

```
1 Q0 9996 1 0.9999 I09UniXRun1 /article[1]/bdy[1]/sec[1] /article[1]/bdy[1]/sec[3]
```

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 2009 measures are solely based on the retrieval of highlighted text. We simplify all INEX tasks to highlighted text retrieval and assume that systems will try to 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.

Thorough Task Precision is measured as the fraction of retrieved text that was highlighted. Recall is measured as the fraction of all highlighted text that has been retrieved. Text seen before is automatically discounted. 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 overall performance, the main measure is mean average interpolated precision (MAiP), calculated over over 101 standard recall points (0.00, 0.01, 0.02, ..., 1.00). We also present interpolated precision at early recall points (iP[0.00], iP[0.01], iP[0.05], and iP[0.10]),

Focused Task As above, precision is measured as the fraction of retrieved text that was highlighted and recall is measured as the fraction of all highlighted text that has been retrieved. 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 main 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 [7] over articles, 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. 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. 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).

For further details on the INEX measures, we refer to [6]

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

Starting in 2009, INEX uses a new document collection based on the Wikipedia. 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. The annotation is enhanced with semantic markup of articles and outgoing links,

```
<article xmlns:xlink="http://www.w3.org/1999/xlink">
<holder confidence="0.9511911446218017" wordnetid="103525454">
<entity confidence="0.9511911446218017" wordnetid="100001740">
<musical_organization confidence="0.8" wordnetid="108246613">
<artist confidence="0.9511911446218017" wordnetid="109812338">
<group confidence="0.8" wordnetid="100031264">
<header>
<title>Queen (band)</title>
<id>42010</id>
...
</header>
<bdy>
...
<songwriter wordnetid="110624540" confidence="0.9173553029164789">
<person wordnetid="100007846" confidence="0.9508927676800064">
<manufacturer wordnetid="110292316" confidence="0.9173553029164789">
<musician wordnetid="110340312" confidence="0.9173553029164789">
<singer wordnetid="110599806" confidence="0.9173553029164789">
<artist wordnetid="109812338" confidence="0.9508927676800064">
<link xlink:type="simple" xlink:href="../068/42068.xml">
Freddie Mercury</link></artist>
</singer>
</musician>
</manufacturer>
</person>
</songwriter>
...
</bdy>
</group>
</artist>
</musical_organization>
</entity>
</holder>
</article>
```

Fig. 1. INEX 2009 Ad Hoc Track document `42010.xml` (in part).

based on the semantic knowledge base YAGO, explicitly labeling more than 5,800 classes of entities like persons, movies, cities, and many more. For a more technical description of a preliminary version of this collection, see [10].

The collection was created from the October 8, 2008 dump of the English Wikipedia articles and incorporates semantic annotations from the 2008-w40-2 version of YAGO. It contains 2,666,190 Wikipedia articles and has a total uncompressed size of 50.7 Gb. There are 101,917,424 XML elements of at least 50 characters (excluding white-space).

Figure 1 shows part of a document in the corpus. The whole article has been encapsulated with tags, such as the ⟨`group`⟩ tag added to the Queen page.

This allows us to find particular article types easily, e.g., instead of a query requesting articles about Freddie Mercury:

```
<topic id="2009114" ct_no="310">
  <title>self-portrait</title>
  <castitle>//painter//figure[about(.//caption, self-portrait)]</castitle>
  <phrasetitle>"self portrait"</phrasetitle>
  <description>Find self-portraits of painters.</description>
  <narrative>
    I am studying how painters visually depict themselves in their
    work.  Relevant document components are images of works of art, in
    combination with sufficient explanation (i.e., a reference to the
    artist and the fact that the artist him/herself is depicted in the
    work of art).  Also textual descriptions of these works, if
    sufficiently detailed, can be relevant.  Document components
    discussing the portrayal of artists in general are not relevant, as
    are artists that figure in painters of other artists.
  </narrative>
</topic>
```

Fig. 2. INEX 2009 Ad Hoc Track topic 2009114.

```
//article[about(., Freddie Mercury)]
```

we can specifically ask about a group about Freddie Mercury:

```
//group[about(., Freddie Mercury)]
```

which will return pages of (pop) groups mentioning Freddy Mercury. In fact, also all internal Wikipedia links have been annotated with the tags assigned to the page they link to, e.g., in the example about the link to Freddie Mercury gets the ⟨`singer`⟩ tag assigned. We can also use these tags to identify pages where certain types of links occur, and further refine the query as:

```
//group[about(.//singer, Freddie Mercury)]
```

The exact NEXI query format used to express the structural hints will be explained below.

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 phrase title, 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. For candidate topics without a ⟨`castitle`⟩ field, a default CAS-query was added based on the CO-query: `//*[about(., "`*CO-query*`")]`. Figure 2 presents an example of an ad hoc topic. Based on the submitted candidate topics, 115 topics were selected for use in the INEX 2009 Ad Hoc Track as topic numbers 2009001–2009115.

Each topic contains

title A short explanation of the information need using simple keywords, also known as the content only (CO) query. It serves as a summary of the content of the user's information need.

castitle A short explanation of the information need, specifying any structural requirements, also known as the content and structure (CAS) query. The castitle is optional but the majority of topics should include one.

phrasetitle A more verbose explanation of the information need given as a series of phrases, just as the ⟨`title`⟩ is given as a series of keywords.

description A brief description of the information need written in natural language, typically one or two sentences.

narrative A detailed explanation of the information need and the description of what makes an element relevant or not. The ⟨`narrative`⟩ should explain not only what information is being sought, but also the context and motivation of the information need, i.e., why the information is being sought and what work-task it might help to solve. Assessments will be made on compliance to the narrative alone; it is therefore important that this description is clear and precise.

The ⟨`castitle`⟩ contains the CAS query, an XPath expressions of the form: `A`[`B`] or `A`[`B`]`C`[`D`] where `A` and `C` are navigational XPath expressions using only the descendant axis. `B` and `D` are predicates using functions for text; the arithmetic operators <, <=, >, and >= for numbers; or the connectives `and` and `or`. For text, the `about` function has (nearly) the same syntax as the XPath function `contains`. Usage is restricted to the form `about`(*.path*, *query*) where *path* is empty or contains only tag-names and descendant axis; and *query* is an IR query having the same syntax as the CO titles (i.e. query terms). The about function denotes that the content of the element located by the path is about the information need expressed in the query. As with the title, the castitle is only a hint to the search engine and does not have definite semantics.

The purpose of the phrasetitle field is to explicate the order and grouping of the query terms in the title. The absence of a phrasetitle implies the absence of a phrase, e.g. a query with independent words. The title and phrasetitle together make the "phrase query" for phrase-aware search. Some topics come with quotations marks in the title, in which case the phrasetitle is at least partially redundant. However, we have made sure that the phrasetitle does not introduce words other than those in the title and that the identified phrases are encapsulated in quotation marks. This setting helps us study whether systems can improve their performance when given explicit phrases as opposed to individual words as implicit phrases.

3.3 Judgments

Topics were assessed by participants following precise instructions. The assessors used the 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 Thorough, 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.

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

	total		# per topic				
	topics	number	min	max	median	mean	st.dev
judged articles	68	50,725	380	766	754	746.0	49.0
articles with relevance	68	4,858	5	351	52	71.4	72.5
highlighted passages	68	7,957	5	594	75.5	117.0	121.5
highlighted characters	68	18,838,137	4,453	2,776,635	97,550.5	277,031.4	442,113.9

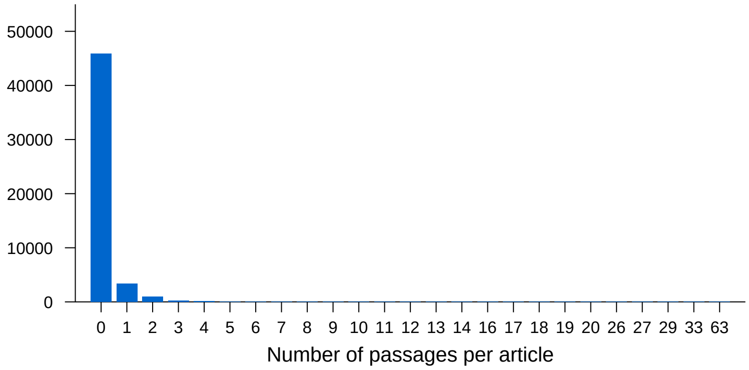

Fig. 3. Distribution of passages over articles.

The relevance judgments were frozen on November 10, 2009. At this time 68 topics had been fully assessed. Moreover, some topics were judged by two separate assessors, each without the knowledge of the other. All results in this paper refer to the 68 topics with the judgments of the first assigned assessor, which is typically the topic author.

– The 68 assessed topics were numbered 2009n with n: 001–006, 010–015, 020, 022, 023, 026, 028, 029, 033, 035, 036, 039–043, 046, 047, 051, 053–055, 061–071, 073, 074, 076–079, 082, 085, 087–089, 091–093, 095, 096, 104, 105, 108–113, and 115

Table 1 presents statistics of the number of judged and relevant articles, and passages. In total 50,725 articles were judged. Relevant passages were found in 4,858 articles. The mean number of relevant articles per topic is 71, but the distribution is skewed with a median of 52. There were 7,957 highlighted passages. The mean was 117 passages and the median was 76 passages per topic.[1]

Figure 3 presents the number of articles with the given number of passages. The vast majority of relevant articles (3,339 out of 4,858) had only a single highlighted passage, and the number of passages quickly tapers off.

[1] Recall from above that for the Focused Task the main effectiveness measures is precision at 1% recall. Given that the average topic has 117 relevant passages in 52 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 relevant articles.

	total		# per relevant article				
	topics	number	min	max	median	mean	st.dev
best entry point offset	68	4,858	2	86,545	311.5	2,493.2	6,481.8
first relevant character offset	68	4,858	2	86,545	295	2,463.0	6,375.6
length relevant documents	68	4,858	204	159,892	5,774.5	11,691.5	15,745.1
relevant characters	68	4,858	8	110,191	1,137	3,877.8	7,818.5
fraction highlighted text	68	4,858	0.00022	1.000	0.330	0.442	0.381

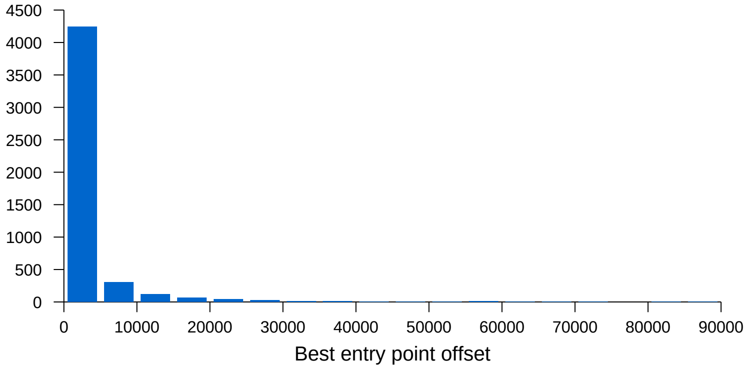

Fig. 4. Distribution of best entry point offsets.

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 2,493 but the distribution is very skewed with a median BEP offset of only 311. Figure 4 shows the distribution of the character offsets of the 4,858 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 2,463 but the median offset is only 295. This suggests a relation between the BEP offset and the FRC offset. Figure 5 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.

Table 2 also shows statistics on the length of relevant articles. Many articles are relatively short with a median length of 5,775 characters, the mean length is 11,691 characters. This is considerably longer than the INEX 2008 collection, where the relevant articles had a median length of 3,030 and a mean length of 6,793. The length of highlighted text in characters is on average 3,876 (mean

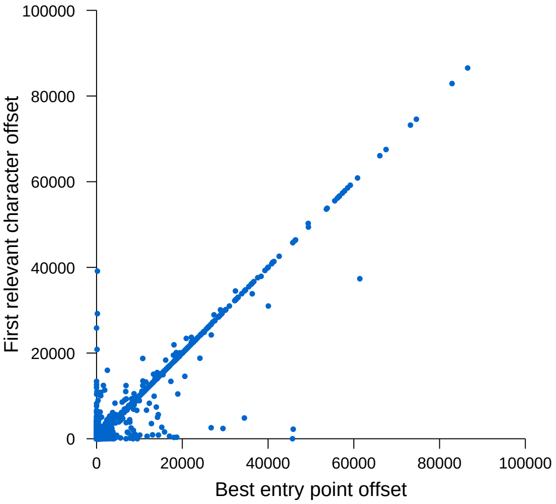

Fig. 5. Scatter plot of best entry point offsets versus the first relevant character.

1,137), in comparison to an average length of 2,338 (mean 838) in 2008. Table 2 also show that amount of relevant text varies from almost nothing to almost everything. The mean fraction is 0.44, and the median is 0.33, indicating that typically over one-third of the article is relevant. This is considerably less than the INEX 2008 collection, where over half of the text of articles was considered 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.

3.4 Questionnaires

At INEX 2009, 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 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 2009 Ad Hoc Track.

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)

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)

Table 5. Participants in the Ad Hoc Track.

Id	Participant	Thorough	Focused	Relevant in Context	Best in Context	CO query	CAS query	Phrase query	Reference run	Element results	Range of elements results	FOL results	# valid runs	# submitted runs
4	University of Otago	0	0	1	0	1	0	0	1	1	0	0	1	1
5	Queensland University of Technology	4	12	12	12	20	20	0	0	32	8	0	40	48
6	University of Amsterdam	4	2	2	2	7	3	0	0	10	0	0	10	10
10	Max-Planck-Institut Informatik	3	8	0	2	11	2	1	0	13	0	0	13	13
16	University of Frankfurt	0	2	0	0	0	2	0	0	2	0	0	2	2
22	ENSM-SE	0	4	0	0	4	0	4	0	4	0	0	4	4
25	Renmin University of China	1	3	3	2	7	2	0	0	9	0	0	9	9
29	INDIAN STATISTICAL INSTITUTE	0	2	0	0	2	0	0	0	2	0	0	2	2
36	University of Tampere	0	0	3	3	6	0	0	2	4	2	0	6	6
48	LIG	3	3	3	3	12	0	0	4	12	0	0	12	12
55	Doshisha University	0	1	0	0	0	1	0	0	1	0	0	1	1
60	Saint Etienne University	3	4	3	3	13	0	0	4	13	0	0	13	13
62	RMIT University	0	0	0	2	2	0	0	0	1	0	1	2	2
68	University Pierre et Marie Curie - LIP6	2	2	0	0	4	0	0	0	4	0	0	4	4
72	University of Minnesota Duluth	2	3	3	1	9	0	0	0	9	0	0	9	9
78	University of Waterloo	0	4	0	0	4	0	0	0	2	0	2	4	4
92	University of Lyon3	2	2	0	2	5	1	6	0	6	0	0	6	8
167	School of Electronic Engineering and Computer Science	3	3	1	3	10	0	0	4	10	0	0	10	12
346	University of Twente	3	2	2	2	0	9	0	4	9	0	0	9	12
	Total runs	30	57	33	37	117	40	11	19	144	10	3	157	172

4 Ad Hoc Retrieval Results

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

4.1 Participation

A total of 172 runs were submitted by 19 participating groups. Table 5 lists the participants and the number of runs they submitted, also broken down over the tasks (Thorough, Focused, Relevant in Context, or Best in Context); the used query (Content-Only or Content-And-Structure); whether it used the Phrase query or Reference run; and the used result type (Element, Range of elements, or FOL passage). Unfortunately, no less than 15 runs turned out to be invalid and will only be evaluated with respect to their “article retrieval” value in Section 6.

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

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p48-LIG-2009-thorough-3T	0.5967	0.5841	0.5444	0.5019	0.2855
p6-UAmsIN09article	0.5938	0.5880	0.5385	0.4981	0.2818
p5-BM25thorough	0.6168	0.5983	0.5360	0.4917	0.2585
p92-Lyon3LIAmanlmnt*	0.5196	0.4956	0.4761	0.4226	0.2496
p60-UJM_15494	0.5986	0.5789	0.5293	0.4813	0.2435
p346-utCASartT09	0.5461	0.5343	0.4929	0.4415	0.2350
p10-MPII-CASThBM	0.5860	0.5537	0.4821	0.4225	0.2133
p167-09RefT	0.3205	0.3199	0.2779	0.2437	0.1390
p68-I09LIP6OWATh	0.3975	0.3569	0.2468	0.1945	0.0630
p25-ruc-base-coT	0.5440	0.4583	0.3020	0.1898	0.0577

Participants were allowed to submit up to two element result-type runs per task and up to two passage result-type runs per task (for all four tasks). In addition, we allowed for an extra submission per task based on a reference run containing an article-level ranking using the BM25 model. This totaled to 20 runs per participant.[2] The submissions are spread well over the ad hoc retrieval tasks with 30 submissions for Thorough, 57 submissions for Focused, 33 submissions for Relevant in Context, and 37 submissions for Best in Context.

4.2 Thorough Task

We now discuss the results of the Thorough Task in which a ranked-list of non-overlapping results (elements or passages) was required. The official measure for the task was mean average interpolated precision (MAiP). Table 6 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 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, MAiP).

LIG Element retrieval run using the CO query. Description: Starting from 2K elements for each of the section types (sec, ss1, ss2, ss3, ss4) according to a multinomial language model with Dirichlet smoothing, we then interleave these five lists according to the score. We then group these results by the ranking of the reference run on articles, keeping within a document the element ranking. The run is based on the reference run.

University of Amsterdam Element retrieval run using the CO query. Description: A standard run on an article index, using a language model with a standard linear length prior. The run is retrieving only articles.

[2] As it turns out, one group submitted more runs than allowed: the *Queensland University of Technology* submitted 24 extra element runs. Some other groups submitted too many runs of a certain type or task. At this moment, we have not decided on any repercussions other than mentioning them in this footnote.

Queensland University of Technology Element retrieval run using the CO query. Description: Starting from a BM25 article retrieval run on an index of terms and tags-as-terms (produced by Otago), the top 50 retrieved articles are further processed by extracting the list of all (overlapping) elements which contained at least one of the search terms. The list is padded with the remaining articles, if needed.

University of Lyon3 A *manual* element retrieval run using the CO query. Description: Using Indri with Dirichlet smoothing and combining two language models: one of the full articles and one on the following tags: b, bdy, category, causal_agent, country, entry, group, image, it, list, location, p, person, physical_entity, sec, software, table, title. Special queries are created used NLP tools such as a summarizer and terminology extraction: the initial query based on the topic's phrase and CO title is expanded with related phrases extracted from the other topic fields and from an automatic summary of the top ranked documents by this initial query. In addition, standard query expansion are used, skip phrases are allowed, and occurrences in the title are extra weighted.

Saint Etienne University Element retrieval run using the CO query. Description: Using BM25 on an element index with element frequency statistics. The b and k parameters were tuned on the INEX 2008 collection, leading to value different from standard document retrieval. The resulting run is filtered for elements from articles in the reference run, while retaining the original element ranking. The run is based on the reference run.

Based on the information from these and other participants:

- All ten runs use retrieve element type results. Three out of ten runs retrieve only article elements: the second ranked *p6-UAmsIN09article*, sixth ranked *p346-utCASartT09*, and the eighth ranked *p167-09RefT*.
- Eight of the ten runs use the CO query, the runs ranked sixth, *p346-utCASartT09*, and seventh, *p10-MPII-CASThBM* use the structured CAS query.
- Three runs are based on the *reference run*: the first ranked *p48-LIG-2009-thorough-3T*, the fifth ranked *p60-UJM_15494*, and the eighth ranked *p167-09RefT*

4.3 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 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]).

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

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p78-UWatFERBM25F	0.6797	0.6333	0.5006	0.4095	0.1854
p68-I09LIP6Okapi	0.6244	0.6141	0.5823	0.5290	0.3001
p10-MPII-COFoBM	0.6740	0.6134	0.5222	0.4474	0.1973
p60-UJM_15525	0.6241	0.6060	0.5742	0.4920	0.2890
p6-UamsFSsec2docbi100	0.6328	0.5997	0.5140	0.4647	0.1928
p5-BM25BOTrangeFOC	0.6049	0.5992	0.5619	0.5057	0.2912
p16-Spirix09R001	0.6081	0.5903	0.5342	0.4979	0.2865
p48-LIG-2009-focused-1F	0.5861	0.5853	0.5431	0.5055	0.2702
p22-emse2009-150*	0.6671	0.5844	0.4396	0.3699	0.1470
p25-ruc-term-coF	0.6128	0.4973	0.3307	0.2414	0.0741

University of Waterloo FOL passage 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. It uses a fielded Okapi BM25F over two fields: a title composed of the concatenation of article and all ancestor's and current section titles, and a body field is the rest of the section. Training was done at element level and an average field length was used.

LIP6 Element retrieval run using the CO query. Description: A BM25 run with b=0.2 and k=2.0 and retrieving 1,500 articles for the CO queries, where negated words are removed from the query. For each document, the /article[1] element is retrieved. The run is retrieving only articles.

Max-Planck-Institut für Informatik Element retrieval run using the CO query. Description: Using EBM25, an XML-specific extension of BM25 using element frequencies of individual tag-term pairs, i.e., for each distinct tag and term, we precompute an individual element frequency, capturing the amount of tags under which the term appears in the entire collection. A static decay factor for the TF component is used to make the scoring function favor smaller elements rather than entire articles.

Saint Etienne University An element retrieval run using the CO query. Description: Using BM25 on an standard article index. The *b* and *k* parameters were tuned on the INEX 2008 collection. The run is retrieving only articles.

University of Amsterdam Element retrieval run using the CAS query. Description: Language model run on a non-overlapping section index with top 100 reranked using a link degree prior. The link degree prior is the indegree+outdegree using local links from the retrieved sections. The link degree prior is applied to the article level, thus all sections from the same article have the same link prior.

Based on the information from these and other participants:

- Seven runs use the CO query. Three runs, the fifth ranked *p6-UamsFSsec2docbi100*, the sixth ranked *p5-BM25BOTrangeFOC*, and the seventh ranked *p16-Spirix09R001* use the structured CAS query. The ninth run, *p22-emse2009-150*, uses a manually expanded query using words from the description and narrative fields.

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

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p5-BM25RangeRIC	0.3345	0.2980	0.2356	0.1786	0.1885
p4-Reference	0.3311	0.2936	0.2298	0.1716	0.1847
p6-UamsRSCMartCMdocbi100	0.3192	0.2794	0.2074	0.1660	0.1773
p48-LIG-2009-RIC-1R	0.3027	0.2604	0.2055	0.1548	0.1760
p36-utampere_given30_nolinks	0.3128	0.2802	0.2101	0.1592	0.1720
p346-utCASrefR09	0.2216	0.1904	0.1457	0.1095	0.1188
p60-UJM_15502	0.2003	0.1696	0.1311	0.0998	0.1075
p167-09RefR	0.1595	0.1454	0.1358	0.1205	0.1045
p25-ruc-base-casF	0.2113	0.1946	0.1566	0.1380	0.1028
p72-umd_ric_1	0.0943	0.0801	0.0574	0.0439	0.0424

- Eight runs retrieve elements as results. The top ranked *p78-UWatFERBM25F* retrieves FOL passages, and the sixth ranked *p5-BM25BOTrangeFOC* retrieves ranges of elements.
- The systems at rank second, (*p68-I09LIP6Okapi*), fourth (*p60-UJM_15525*), and seventh (*p16-Spirix09R001*) are retrieving only full articles.

4.4 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 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).

Queensland University of Technology Run retrieving ranges of elements using the CO query. Description: Starting from a BM25 article retrieval run on an index of terms and tags-as-terms (produced by Otago), the top 50 retrieved articles are further processed by identifying the first and last element in the article (in reading order) which contained any of the search terms. The focused result was then specified as a range of two elements (which could be one and the same). The list is padded with the remaining articles.

University of Otago Element retrieval run using the CO query. Description: the run uses the Okapi BM25 model on an article index, with parameters trained on the INEX 2008 collection. The run is retrieving only articles and is based on the reference run—in fact, it is the original reference run.

University of Amsterdam Element retrieval run using the CO query. Description: The results from section index are grouped and ranked based on the the article ranking from the article index. The section run is reranked using the Wikipedia categories as background models before we cut-off the section run at 1,500 results per topic. The article run is similarly reranked using the Wikipedia categories as background models and link degree priors using the local incoming and outgoing links at article level.

LIG Element retrieval run using the CO query. Description: First, separate lists of 2K elements are generated for the element types sec, ss1, ss2, ss3, and ss4, the five lists are merged according to score. Second, an article ranking is obtained using a mulinomial language model with Dirichlet smoothing. Third, the element results are group using the article ranking, by retaining with each article the reading order. Then we remove overlaps according to the reading order.

University of Tampere Element retrieval run using the CO query. Description: For each document the only retrieved passage was between the first and the last link to the top 30 documents. If there were no such links, the whole article was returned. The run is based on the reference run.

Based on the information from these and other participants:

- The runs ranked sixth (*p346-utCASrefR09*) and ninth (*p25-ruc-base-casF*) are using the CAS query. All other runs use only the CO query in the topic's title field.
- The top scoring run retrieves ranges of elements, all other runs retrieve elements as results.
- Solid article ranking seems a prerequisite for good overall performance, with second best run, *p4-Reference* and the eighth best run, *p167-09RefR*, retrieving only full articles.

4.5 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 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).

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

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p5-BM25bepBIC	0.2941	0.2690	0.2119	0.1657	0.1711
p62-RMIT09titleO	0.3112	0.2757	0.2156	0.1673	0.1710
p10-MPII-COBIBM	0.2903	0.2567	0.2053	0.1598	0.1662
p48-LIG-2009-BIC-3B	0.2778	0.2564	0.1969	0.1469	0.1571
p6-UamsBAfbCMdocbi100	0.2604	0.2298	0.1676	0.1478	0.1544
p92-Lyon3LIAmanBEP*	0.2887	0.2366	0.1815	0.1482	0.1483
p36-utampere_given30_nolinks	0.2141	0.1798	0.1462	0.1234	0.1207
p346-utCASrefB09	0.1993	0.1737	0.1248	0.0941	0.1056
p25-ruc-term-coB	0.1603	0.1610	0.1274	0.0976	0.1013
p167-09LrnRefB	0.1369	0.1250	0.1181	0.1049	0.0953

Queensland University of Technology Element retrieval run using the CO query. Description: Starting from a BM25 article retrieval run on an index of terms and tags-as-terms (produced by Otago), the top 50 retrieved articles are further processed by identifying the first element (in reading order) containing any of the search terms. The list is padded with the remaining articles.

RMIT University Element retrieval run using the CO query. Description: Using Zettair with Okapi BM25 on an article-level index. The BEP is assumed to be at the start of the article. The run is retrieving only articles.

Max-Planck-Institut für Informatik Element retrieval run using the CO query. Description: Using EBM25, an XML-specific extension of BM25 using element frequencies of individual tag-term pairs, i.e., for each distinct tag and term, we precompute an individual element frequency, capturing the amount of tags under which the term appears in the entire collection. A static decay factor for the TF component is used to make the scoring function favor smaller elements rather than entire articles, but the final run returns the start of the article as BEP. The run is retrieving only articles.

LIG Element retrieval run using the CO query. Description: First, separate lists of 2K elements are generated for the element types sec, ss1, ss2, ss3, and ss4, the five lists are merged according to score. Second, an article ranking is obtained from the reference run. Third, for each article the best scoring element is used as the entry point. The run is based on the reference run.

University of Amsterdam Element retrieval run using the CO query. Description: Article index run with standard pseudo-relevance feedback (using Indri), reranked with Wikipedia categories as background models and link degree priors using the local incoming and outgoing links at article level. The run is retrieving only articles.

Based on the information from these and other participants:

- The second best run (*p62-RMIT09titleO*) retrieves FOL passages, all other runs return elements as results. The FOL passage run is a degenerate case that always puts the BEP at the start of the article.
- 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

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

(a) Thorough Task

	1	2	3	4	5	6	7	8	9	10
p48		-	-	⋆	-	⋆	-	⋆	⋆	⋆
p6			-	⋆	-	⋆	-	⋆	⋆	⋆
p5				⋆	-	⋆	-	⋆	⋆	⋆
p92					-	-	-	⋆	⋆	-
p60						-	-	⋆	⋆	⋆
p346							-	⋆	⋆	⋆
p10								⋆	⋆	⋆
p167									-	-
p68										-
p25										

(b) Focused Task

	1	2	3	4	5	6	7	8	9	10
p78		-	-	-	-	-	-	-	-	⋆
p68			-	-	-	-	-	⋆	-	⋆
p10				-	-	-	-	-	-	⋆
p60					-	-	-	-	-	⋆
p6						-	-	-	-	⋆
p5							-	-	-	⋆
p16								-	-	⋆
p48									-	⋆
p22										⋆
p25										

(c) Relevant in Context Task

	1	2	3	4	5	6	7	8	9	10
p5		⋆	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆
p4			-	-	⋆	⋆	⋆	⋆	⋆	⋆
p6				-	-	⋆	⋆	⋆	⋆	⋆
p48					-	⋆	⋆	⋆	⋆	⋆
p36						⋆	⋆	⋆	⋆	⋆
p346							-	-	-	⋆
p60								-	-	⋆
p167									-	⋆
p25										⋆
p72										

(d) Best in Context Task

	1	2	3	4	5	6	7	8	9	10
p5		-	-	⋆	⋆	-	⋆	⋆	⋆	⋆
p62			-	⋆	-	-	⋆	⋆	⋆	⋆
p10				-	-	-	⋆	⋆	⋆	⋆
p48					-	-	⋆	⋆	⋆	⋆
p6						-	⋆	⋆	⋆	⋆
p92							-	⋆	⋆	⋆
p36								-	-	⋆
p346									-	-
p25										-
p167										

of all the retrieved articles at rank two (*p62-RMIT09titleO*), rank three (*p10-MPII-COBIBM*), rank five (*p6-UamsBAfbCMdocbi100*), and rank ten (*p167-09LrnRefB*).

– With the exception of the run ranked eight (*p346-utCASrefB09*), which used the CAS query, all the other best runs per group use the CO query.

4.6 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 the Thorough Task, we see that the performance (measured by MAiP) of the top scoring run is significantly better than the runs at rank 4, 6, 8, 9, and 10. The same holds for the second and third best run. The fourth best run is significantly better than the runs at rank 8 and 9. The fifth, sixth, and seventh ranked runs are all significantly better than the runs at rank 8, 9, and 10. Of the 45 possible pairs of runs, there are 26 (or 58%) significant differences. For the Focused Task, we see that the early precision (at 1% recall) is a rather unstable measure. All runs are significantly better than the run at rank 10, the second best run also is significantly better than the run at rank 8. Of the 45 possible pairs of runs,

Table 11. Ad Hoc Track: Runs with ranges of elements or FOL passages.

(a) Focused Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p78-UWatFERBM25F	0.6797	0.6333	0.5006	0.4095	0.1854
p5-BM25BOTrangeFOC	0.6049	0.5992	0.5619	0.5057	0.2912

(b) Relevant in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p5-BM25RangeRIC	0.3345	0.2980	0.2356	0.1786	0.1885
p36-utampere_auth_40_top30	0.2717	0.2509	0.2006	0.1583	0.1185

(c) Best in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p62-RMIT09titleO	0.3112	0.2757	0.2156	0.1673	0.1710

there are only 10 (or 22%) significant differences. 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 2 and 4 through 10. The second best run is significantly better than ranks 5 through 10. The third, fourth, and fifth ranked systems are significantly better than ranks 6 through 10. The sixth to ninth systems are significantly better than rank 10. Of the 45 possible pairs of runs, there are 33 (or 73%) significant differences, making MAgP a very discriminative measure. For the Best in Context Task, we see that the top run is significantly better than ranks 4 and 5, and 7 through 10. The second best run is significantly better than than ranks 4 and 7 to 10. The third, fourth, and fifth ranked runs are significantly better than than ranks 7 to 10. The seventh ranked system is better than the systems ranked 8 to 10, and the eighth ranked system better than rank 9 10. Of the 45 possible pairs of runs, there are 27 (or 60%) significant differences.

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 13 submissions using ranges of elements of FOL-passage results, from in total 4 participating groups. We will look at the relative effectiveness of element and passage runs.

As we saw above, in Section 4, for three tasks there were high ranking runs using FOL passages or ranges of elements in the top 10. Table 11 shows the best runs using ranges of elements or FOL passages for three ad hoc tasks, there were no such submissions for the Thorough Task. As it turns out, the best focused run retrieving FOL passages was the top ranked run in Table 7; the best relevant

Table 12. Ad Hoc Track: Runs using the phrase query.

(a) Thorough Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p92-Lyon3LIAmanlmnt*	0.5196	0.4956	0.4761	0.4226	0.2496

(b) Focused Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p22-emse2009-150*	0.6671	0.5844	0.4396	0.3699	0.1470
p10-MPII-COArBPP	0.5563	0.5477	0.5283	0.4681	0.2566
p92-Lyon3LIAmanQE*	0.4955	0.4861	0.4668	0.4271	0.2522

(c) Best in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p92-Lyon3LIAmanBEP*	0.2887	0.2366	0.1815	0.1482	0.1483

in context retrieving ranges of elements was the top scoring run in Table 8; and the best best in context run retrieving FOL passages was the second best run in Table 9. Given the low number of submissions using passages or ranges of elements, this is an impressive result. 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-BM25BOTrangeFOC* is an article retrieving run using ranges of elements, based on the CAS query. For Relevant in Context, *p5-BM25RangeRIC* is an article retrieving run using ranges of elements. For Best in Context, *p62-RMIT09titleO* is an article run using FOL passages. 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 Phrase queries

We received 10 submissions based on the phrase query. Table 12 shows the best runs using the phrase query for three of the ad hoc tasks, there were no valid submissions using the phrase title for Relevant in Context. The best phrase submission for the Thorough Task did rank 5th in the overall results. The best phrase submission for the Focused Task did rank 9th in the overall results. The best phrase submission for the Best in Context Task did rank 6th in the overall results.

Although few runs were submitted, the phrase title seems competitive, but not superior to the use of the CO query. The only participant submitting both types of runs, the *Max-Planck-Institute für Informatik* for the Focused Task, had marginally better performance for the CO query run over all 68 topics, and marginally better performance for the combined CO and Phrase title run over the 60 topics having a proper phrase in the Phrase title field. The differences between the query types are very small. A possible explanation for this is that all CO query have been expanded to contain the same terms as the more verbose phrase query. Hence the only difference is the explicit phrase markup, which

Table 13. Ad Hoc Track: Runs using the reference run.

(a) Thorough Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p48-LIG-2009-thorough-3T	0.5967	0.5841	0.5444	0.5019	0.2855
p60-UJM_15494	0.5986	0.5789	0.5293	0.4813	0.2435
p346-utCASrefF09	0.4834	0.4525	0.4150	0.3550	0.1982
p167-09RefT	0.3205	0.3199	0.2779	0.2437	0.1390

(b) Focused Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p48-LIG-2009-focused-3F	0.5946	0.5822	0.5344	0.5018	0.2732
p60-UJM_15518	0.5559	0.5136	0.4003	0.3104	0.1019
p346-utCASrefF09	0.4801	0.4508	0.4139	0.3547	0.1981
p167-09LrnRefF	0.3162	0.3072	0.2512	0.2223	0.1292

(c) Relevant in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p4-Reference	0.3311	0.2936	0.2298	0.1716	0.1847
p48-LIG-2009-RIC-3R	0.3119	0.2790	0.2193	0.1629	0.1757
p36-utampere_given30_nolinks	0.3128	0.2802	0.2101	0.1592	0.1720
p346-utCASrefR09	0.2216	0.1904	0.1457	0.1095	0.1188
p167-09RefR	0.1595	0.1454	0.1358	0.1205	0.1045
p60-UJM_15503	0.1825	0.1548	0.1196	0.0953	0.1020

(d) Best in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p48-LIG-2009-BIC-3B	0.2778	0.2564	0.1969	0.1469	0.1571
p36-utampere_given30_nolinks	0.2141	0.1798	0.1462	0.1234	0.1207
p346-utCASrefB09	0.1993	0.1737	0.1248	0.0941	0.1056
p167-09LrnRefB	0.1369	0.1250	0.1181	0.1049	0.0953
p60-UJM_15508	0.1274	0.1123	0.0878	0.0735	0.0795

requires special handling by the search engines. The available test collection with explicit phrases marked up in 60 topics is a valuable result of INEX 2009, and it can be studied in-depth in future experiments.

5.3 Reference run

There were 19 submissions using the reference run. Table 13 shows the best runs using the reference runs for the four ad hoc tasks. For the Thorough Task, the best submission based on the reference run ranked first. For the Focused Task, the best submission based on the reference run would have ranked tenth. For the Relevant in Context Task, the best submission based on the reference run—in fact, the actual reference run itself—ranked second. For the Best in Context Task, the best submission based on the reference run ranked fourth. The results show that the reference run indeed provides competitive article ranking that forms a good basis for retrieval.

There are also considerable differences in performance of the runs based on the same reference run. This suggests that the runs do not retrieve the exact

Table 14. Top 10 Participants in the Ad Hoc Track: Article retrieval based on the reference run.

Participant	P5	P10	1/rank	map	bpref
p4-Reference	0.6147	0.5294	0.8240	0.3477	0.3333
p36-utampere_given30_nolinks	0.6147	0.5294	0.8240	0.3477	0.3333
p48-LIG-2009-BIC-3B	0.6147	0.5294	0.8240	0.3463	0.3336
p60-UJM_15508	0.5324	0.4544	0.7020	0.2910	0.2925
p346-utCASrefB09	0.5441	0.4750	0.7494	0.2833	0.2768
p167-09RefT	0.3765	0.3603	0.5761	0.2443	0.2540

same set of articles. As explained later, in Section 6, we can look at the article rankings induced by the runs. Table 14 shows the best run of the top 10 participating groups, using the reference run. With the exception of *p36-utampere_given30_nolinks* the article rankings of the runs vary considerably.

5.4 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 15 show the distribution of target elements, with YAGO tags in emphatic. In total 81 topics had a non-trivial CAS query.[3] These CAS topics are numbered 2009n with n: 001–009, 011–013, 015–017, 020–025, 028–032, 036, 037, 039–045, 048–053, 057, 058, 060, 061, 064–072, 074, 080, 085–096, 098, 099, 102, 105, 106, and 108–115. As it turned out, 50 of these CAS topics were assessed. The results presented here are restricted to only these 50 CAS topics.

Table 16 lists the top 10 participants measured using just the 50 CAS topics and for the Thorough Task (a and b) and the Focused Task (c and d). For the Thorough Task the best CAS run, *p5-BM25BOTthorough*, would have ranked sixth amongst the CO runs on MAiP. The two participants submitting both CO and CAS runs had better MAiP scores for the CO runs. However, the best CAS run has higher scores on early precision, iP[0.00] through iP[0.05] than any of the CO submissions. For the Focused Task the best CAS run, *p6-UamsFSsec2docbi100*, would have ranked fifth amongst the CO runs. Two participants submitting both CO and CAS runs had better iP[0.01] scores for the CO runs, one participant had a better CAS run. For Relevant in Context Task (not shown), the best CAS run, *p5-BM25BOTrangeRIC*, would have ranked third among the CO runs. One participants submitting both CO and CAS runs had

[3] Note that some of the wild-card topics (using the "*" target) in Table 15 had non-trivial about-predicates and hence have not been regarded as trivial CAS queries.

Table 15. CAS query target elements over all 115 topics (YAGO tags slanted).

Target Element	Frequency
*	41
article	32
sec	9
group	5
p	4
music_genre	2
vehicles	1
theory	1
song	1
revolution	1
(p|sec|*person)*	1
(p|sec)	1
protest	1
(person|chemist|alchemist|scientist|physicist)	1
personality	1
museum	1
link	1
image	1
home	1
food	1
figure	1
facility	1
driver	1
dog	1
director	1
(classical_music|opera|orchestra|performer|singer)	1
bicycle	1
(article|sec|p)	1

better MAgP scores for a CO run, another participant had a better CAS run. For the Best in Context Task (not shown), the best CAS run, *p5-BM25BOTbepBIC*, would rank seventh among the CO runs. All three participants submitting both CO and CAS runs had better MAgP scores for their CO runs. Overall, we see the that teams submitting runs with both types of queries have higher scoring CO runs, with participant 5 as a notable exception for Focused.

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.

6.1 Article retrieval: Relevance Judgments

We will first look at the topics judged during INEX 2009, but now using the judgments to derive standard document-level relevance by regarding an article

Table 16. Ad Hoc Track CAS Topics: CO runs versus CAS runs.

(a) Thorough Task: CO runs

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p48-LIG-2009-thorough-1T	0.5781	0.5706	0.5315	0.4834	0.2729
p6-UAmsIN09article	0.5900	0.5821	0.5149	0.4613	0.2629
p92-Lyon3LIAmanlmnt*	0.5365	0.5039	0.4794	0.4330	0.2450
p5-BM25thorough	0.6273	0.6023	0.5191	0.4620	0.2389
p60-UJM_15494	0.6034	0.5766	0.5131	0.4612	0.2280
p10-MPII-COThBM	0.6436	0.5916	0.5135	0.3783	0.1909
p167-09RefT	0.3245	0.3237	0.2682	0.2392	0.1291
p68-I09LIP6OWATh	0.4146	0.3651	0.2512	0.1963	0.0608
p25-ruc-base-coT	0.5328	0.4333	0.2538	0.1653	0.0505
p72-umd_thorough_3	0.4073	0.2893	0.1697	0.0999	0.0494

(b) Thorough Task: CAS runs

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p5-BM25BOTthorough	0.6460	0.6169	0.5359	0.4472	0.2279
p346-utCASartT09	0.5541	0.5381	0.4819	0.4136	0.2227
p10-MPII-CASThBM	0.5747	0.5308	0.4406	0.3627	0.1651

(c) Focused Task: CO runs

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p78-UWatFERBM25F	0.6742	0.6222	0.4905	0.3758	0.1737
p60-UJM_15525	0.6373	0.6127	0.5696	0.4585	0.2811
p10-MPII-COArBM	0.6201	0.6060	0.5387	0.4648	0.2684
p68-I09LIP6Okapi	0.6130	0.6005	0.5660	0.5064	0.2798
p5-ANTbigramsRangeFOC	0.6089	0.5936	0.5331	0.4531	0.2597
p48-LIG-2009-focused-3F	0.5971	0.5802	0.5205	0.4775	0.2583
p22-emse2009-150*	0.6453	0.5598	0.4211	0.3471	0.1371
p92-Lyon3LIAmanQE*	0.5185	0.5058	0.4815	0.4339	0.2472
p25-ruc-term-coF	0.6277	0.4955	0.2900	0.2065	0.0668
p167-09LrnRefF	0.3357	0.3234	0.2536	0.2211	0.1216

(c) Focused Task: CAS runs

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p6-UamsFSsec2docbi100	0.6151	0.5974	0.4851	0.4230	0.1718
p16-Spirix09R001	0.6201	0.5958	0.5386	0.4920	0.2794
p5-BM25BOTrangeFOC	0.6031	0.5954	0.5470	0.4789	0.2713
p10-MPII-CASFoBM	0.5643	0.5161	0.4454	0.3634	0.1644
p25-ruc-base-casF	0.5114	0.4775	0.4077	0.3214	0.1666
p346-utCASrefF09	0.4353	0.3955	0.3477	0.2781	0.1471
p55-doshisha09f	0.1273	0.0651	0.0307	0.0227	0.0060

as relevant if some part of it is highlighted by the assessor. 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.

Table 17. Top 10 Participants in the Ad Hoc Track: Article retrieval.

Participant	P5	P10	1/rank	map	bpref
p6-UamsTAbi100	0.6500	0.5397	0.8555	0.3578	0.3481
p48-LIG-2009-BIC-1B	0.6059	0.5338	0.8206	0.3573	0.3510
p62-RMIT09title	0.6029	0.5279	0.8237	0.3540	0.3488
p5-BM25ArticleRIC	0.6147	0.5294	0.8240	0.3477	0.3333
p4-Reference	0.6147	0.5294	0.8240	0.3477	0.3333
p36-utampere_given30_nolinks	0.6147	0.5294	0.8240	0.3477	0.3333
p68-I09LIP6OWA	0.6118	0.5147	0.8602	0.3420	0.3258
p10-MPII-COArBP	0.6353	0.5471	0.8272	0.3371	0.3458
p92-Lyon3LIAmanQE*	0.6265	0.5265	0.7413	0.3335	0.3416
p78-UWatFERBase	0.5765	0.5088	0.8093	0.3267	0.3205

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 are now also considered, and we work with all 172 runs submitted to the Ad Hoc Track.

Table 17 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). No less than seven of the top 10 runs retrieve exclusively full articles: only rank two (*p48-LIG-2009-BIC-1B*), rank six (*p36-utampere_given30_nolinks*) and rank ten (*p78-UWatFERBase*) retrieve elements proper. The relative effectiveness of these article retrieval runs in terms of their article ranking is no surprise. Furthermore, we see submissions from all four ad hoc tasks. A run from the Thorough task at rank 1; runs from the Best in Context task at ranks 2 and 3; runs from the Relevant in Context task at ranks 4, 5 and 6; and runs from the Focused task at ranks 7, 8, 9 and 10.

If we break-down all runs over the original tasks, shown in Table 18), we can compare the ranking to Section 4 above. We see some runs that are familiar from the earlier tables: five Thorough runs correspond to Table 6, four Focused runs correspond to Table 7, six Relevant in Context runs correspond to Table 8, and five Best in Context runs correspond to Table 9. More formally, we looked at how the two system rankings correlate using kendall's tau.

- Over all 30 Thorough Task submissions the system rank correlation is 0.646 between MAiP and map.
- Over all 57 Focused task submissions the system rank correlation is 0.420 between iP[0.01] and map, and 0.638 between MAiP and map.
- Over all 33 Relevant in Context submissions the system rank correlation between MAgP and map is 0.598.
- Over all 37 Best in Context submissions the system rank correlation between MAgP and map is 0.517.

Table 18. Top 10 Participants in the Ad Hoc Track: Article retrieval per task.

(a)Thorough Task

Participant	P5	P10	1/rank	map	bpref
p6-UamsTAbi100	0.6500	0.5397	0.8555	0.3578	0.3481
p48-LIG-2009-thorough-1T	0.6118	0.5191	0.8042	0.3493	0.3392
p92-Lyon3LIAmanlmnt⋆	0.6382	0.5279	0.7706	0.3305	0.3374
p5-BM25thorough	0.6147	0.5294	0.8240	0.3188	0.3142
p10-MPII-COThBM	0.5853	0.5206	0.8084	0.3087	0.3138
p346-utCASartT09	0.5176	0.4588	0.7138	0.2913	0.2986
p60-UJM_15486	0.5647	0.4765	0.7149	0.2797	0.2884
p68-I09LIP6OWATh	0.4735	0.4353	0.7100	0.2665	0.2745
p72-umd_thorough_3	0.5382	0.4515	0.7406	0.2486	0.2674
p167-09RefT	0.3765	0.3603	0.5761	0.2443	0.2540

(b) Focused Task

Participant	P5	P10	1/rank	map	bpref
p48-LIG-2009-focused-1F	0.6059	0.5338	0.8206	0.3569	0.3506
p5-BM25ArticleFOC	0.6147	0.5294	0.8240	0.3477	0.3333
p68-I09LIP6OWA	0.6118	0.5147	0.8602	0.3420	0.3258
p10-MPII-COArBP	0.6353	0.5471	0.8272	0.3371	0.3458
p92-Lyon3LIAmanQE⋆	0.6265	0.5265	0.7413	0.3335	0.3416
p78-UWatFERBase	0.5765	0.5088	0.8093	0.3267	0.3205
p60-UJM_15525	0.5824	0.4926	0.8326	0.3256	0.3169
p16-Spirix09R002	0.5206	0.4588	0.7250	0.3133	0.3149
p6-UamsFSsec2docbi100	0.5941	0.4779	0.8958	0.2985	0.2994
p346-utCASartF09	0.5176	0.4588	0.7138	0.2913	0.2986

(c) Relevant in Context Task

Participant	P5	P10	1/rank	map	bpref
p48-LIG-2009-RIC-1R	0.6059	0.5338	0.8206	0.3569	0.3506
p6-UamsRSCMartCMdocbi100	0.6324	0.5309	0.9145	0.3523	0.3374
p5-BM25ArticleRIC	0.6147	0.5294	0.8240	0.3477	0.3333
p4-Reference	0.6147	0.5294	0.8240	0.3477	0.3333
p36-utampere_given30_nolinks	0.6147	0.5294	0.8240	0.3477	0.3333
p346-utCOartR09	0.5324	0.4882	0.7448	0.3120	0.3137
p72-umd_ric_2	0.5441	0.4544	0.7807	0.2708	0.2867
p167-09RefR	0.3765	0.3603	0.5761	0.2443	0.2540
p25-ruc-base-casF	0.4441	0.4176	0.6270	0.2243	0.2523
p60-UJM_15488	0.4382	0.3853	0.6043	0.2146	0.2343

(d) Best in Context Task

Participant	P5	P10	1/rank	map	bpref
p48-LIG-2009-BIC-1B	0.6059	0.5338	0.8206	0.3573	0.3510
p62-RMIT09title	0.6029	0.5279	0.8237	0.3540	0.3488
p5-BM25AncestorBIC	0.6147	0.5294	0.8240	0.3477	0.3333
p36-utampere_given30_nolinks	0.6147	0.5294	0.8240	0.3477	0.3333
p6-UamsBAfbCMdocbi100	0.6147	0.5118	0.8531	0.3361	0.3251
p10-MPII-COBIBM	0.5824	0.5191	0.8451	0.3325	0.3315
p92-Lyon3LIAmanBEP⋆	0.6382	0.5279	0.7706	0.3305	0.3374
p25-ruc-term-coB	0.5206	0.4779	0.7158	0.3197	0.3251
p346-utCOartB09	0.5324	0.4882	0.7448	0.3120	0.3137
p60-UJM_15508	0.5324	0.4544	0.7020	0.2910	0.2925

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 between article retrieval and the "in context" tasks was much higher (0.79) for the INEX 2008 collection. A likely effect of the increasing length of (relevant) Wikipedia articles.

7 Discussion and Conclusions

In this paper we provided an overview of the INEX 2009 Ad Hoc Track that contained four tasks: For the *Thorough Task* a ranked-list of results (elements or passages) by estimated relevance was required. 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 four tasks, and analysed the relative effectiveness of element and passage runs, of runs using phrases, of runs using the reference run, and of keyword (CO) queries and structured queries (CAS). We also look at effectiveness in term of article retrieval.

Given the efforts put into the fair comparison of element and passage retrieval approaches, the number submissions using FOL passages and range of elements was disappointing. Thirteen submissions used ranges of elements or FOL passage results, whereas 144 submissions used element results. In addition, several of the passage or FOL submissions used exclusively full articles as results. Still the non-element submissions were competitive with the top ranking runs for both the Focused and Relevant in Context Tasks, and the second ranking run for the Best in Context Task. There were too few submissions to draw any definite conclusions, but the outcome broadly confirms earlier results using passage-based element retrieval [3, 4].

There were also few submissions using the explicitly annotated phrases of the phrase query: ten in total. Phrase query runs were competitive with several of them in the overall top 10 results, but the impact of the phrases seemed marginal. Recall, that the exact same terms were present in the CO query, and the only difference was the phrase annotation. This is in line with earlier work. The use of phrases in queries has been studied extensively. In early publications, the usage of phrases and proximity operators showed improved retrieval results but rarely anything substantial [e.g., 2]. As retrieval models became more advanced, the usage of query operators was questioned. E.g., Mitra et al. [8] conclude that when using a good ranking algorithm, phrases have no effect on high precision retrieval (and sometimes a negative effect due to topic drift). Rasolofo and Savoy [9] combine term-proximity heuristics with an Okapi model, obtaining marginal improvements for early precision but with hardly observable impact on the MAP scores.

There were 19 submissions using the reference run providing a solid article ranking for further processing. These runs turned out to be competitive, with runs in the top 10 for all tasks. Hence the reference run was successful in helping participants to create high quality runs. However, run based on the reference run were not directly comparable, since participants used these runs in different ways leading to substantially different underlying article rankings.

When examining the relative effectiveness of CO and CAS we found that for all tasks the best scoring runs used the CO query but some CAS runs were in the top 10 for all four tasks. Part of the explanation may be in the low number of CAS submissions (40) in comparison with the number of CO submissions (117). Only 50 of the 68 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. The YAGO tags potentially expressing an information need naturally in terms of structural constraints, were popular: 36 CAS queries used them (21 of them judged). Over the 50 non-trivial CAS queries, most groups had a better performing run using the CO query. A notable exception was *QUT* who had better performance for CAS on the Focused Task. This is in accordance with earlier results showing that structural hints can help promote initial precision [5].

As in earlier years, we saw that article retrieval is a 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 also suggests that element-level or passage-level evidence is 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. The systems with the best performance for the ad hoc tasks, also tend to have the best article rankings.

Finally, the Ad Hoc Track had three main research questions. The first main research question was to study the effect of the new collection. We saw that the collection's size had little impact, but that the relevant articles were much longer (a mean length 3,030 in 2008 and 5,775 in 2009, a 52% increase), leading to a lower fraction of highlighted text per article (a mean of 58% in 2008 and 33% in 2009). This also reduced the correlation with article retrieval, e.g., from 79% for the "in context" tasks in 2008 to 51–58% in 2009. The second main research question was the impact of verbose queries using phrases or structural hints. The relatively few phase query submissions showed only marginal differences. The CAS query runs were in general less effective than the CO query runs, with one notable exception for the early precision measures of the Focused Task. The second 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. Despite the low number of non-element runs, we saw that some of the best performing system used FOL passages or ranges of elements. For all 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] C. L. A. Clarke. Range results in XML retrieval. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 4–5, Glasgow, UK, 2005.

[2] W. B. Croft, H. R. Turtle, and D. D. Lewis. The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 32–45, 1991.

[3] 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.

[4] 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. University of Otago, Dunedin New Zealand, 2007.

[5] 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.

[6] J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. Robertson. INEX 2007 evaluation measures. In *Focused access to XML documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, volume 4862 of *Lecture Notes in Computer Science*, pages 24–33. Springer Verlag, Heidelberg, 2008.

[7] 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.

[8] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases. In *Proceedings of RIAO-97*, 1997.

[9] Y. Rasolofo and J. Savoy. Term proximity scoring for keyword-based retrieval systems. In *Proceedings of the 25th European Conference on IR Research (ECIR 2003)*, pages 207–218, 2003.

[10] R. Schenkel, F. M. Suchanek, and G. Kasneci. YAWN: A semantically annotated Wikipedia XML corpus. In *12. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW 2007)*, pages 277–291, 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. University of Otago, Dunedin New Zealand, 2006.

A Appendix: Full run names

Group	Run	Label	Task	Query	Results	Notes
4	617	Reference	RiC	CO	Ele	Reference run Article-only
5	744	BM25AncestorBIC	BiC	CO	Ele	Article-only
5	757	BM25thorough	Tho	CO	Ele	
5	775	BM25ArticleFOC	Foc	CO	Ele	Article-only
5	781	BM25BOTrangeFOC	Foc	CAS	Ran	Article-only
5	792	ANTbigramsRangeFOC	Foc	CO	Ran	Article-only
5	796	BM25ArticleRIC	RiC	CO	Ele	Article-only
5	797	BM25RangeRIC	RiC	CO	Ran	Article-only
5	804	BM25BOTrangeRIC	RiC	CAS	Ran	Article-only
5	808	BM25BOTthorough	Tho	CAS	Ele	
5	824	BM25bepBIC	BiC	CO	Ele	Article-only
5	825	BM25BOTbepBIC	BiC	CAS	Ele	Article-only
6	634	UAmsIN09article	Tho	CO	Ele	Article-only
6	810	UamsTAbi100	Tho	CO	Ele	Article-only
6	813	UamsFSsec2docbi100	Foc	CAS	Ele	
6	814	UamsRSCMartCMdocbi100	RiC	CO	Ele	
6	816	UamsBAfbCMdocbi100	BiC	CO	Ele	Article-only
6	817	UamsBSfbCMsec2docbi100art1	BiC	CAS	Ele	Article-only
10	618	MPII-CASFoBM	Foc	CAS	Ele	
10	619	MPII-COFoBM	Foc	CO	Ele	
10	620	MPII-CASThBM	Tho	CAS	Ele	
10	621	MPII-COThBM	Tho	CO	Ele	
10	628	MPII-COArBM	Foc	CO	Ele	Article-only
10	632	MPII-COBIBM	BiC	CO	Ele	Article-only
10	700	MPII-COArBP	Foc	CO	Ele	Article-only
10	709	MPII-COArBPP	Foc	CO	Ele	Phrases Article-only
16	872	Spirix09R001	Foc	CAS	Ele	Article-only
16	873	Spirix09R002	Foc	CAS	Ele	Article-only
22	672	emse2009-150	Foc	CO	Ele	Phrases Manual
25	727	ruc-base-coT	Tho	CO	Ele	
25	737	ruc-term-coB	BiC	CO	Ele	
25	738	ruc-term-coF	RiC	CO	Ele	
25	739	ruc-term-coF	Foc	CO	Ele	
25	898	ruc-base-casF	Foc	CAS	Ele	
25	899	ruc-base-casF	RiC	CAS	Ele	
36	688	utampere_given30_nolinks	RiC	CO	Ele	Reference run
36	701	utampere_given30_nolinks	BiC	CO	Ele	Reference run
36	708	utampere_auth_40_top30	RiC	CO	Ran	
48	682	LIG-2009-thorough-1T	Tho	CO	Ele	
48	684	LIG-2009-thorough-3T	Tho	CO	Ele	Reference run
48	685	LIG-2009-focused-1F	Foc	CO	Ele	
48	686	LIG-2009-focused-3F	Foc	CO	Ele	Reference run
48	714	LIG-2009-RIC-1R	RiC	CO	Ele	
48	716	LIG-2009-RIC-3R	RiC	CO	Ele	Reference run
48	717	LIG-2009-BIC-1B	BiC	CO	Ele	

Continued on Next Page...

Group	Run	Label	Task	Query	Results	Notes
48	719	LIG-2009-BIC-3B	BiC	CO	Ele	Reference run
55	836	doshisha09f	Foc	CAS	Ele	
60	819	UJM_15518	Foc	CO	Ele	Reference run
60	820	UJM_15486	Tho	CO	Ele	
60	822	UJM_15494	Tho	CO	Ele	Reference run
60	827	UJM_15488	RiC	CO	Ele	
60	828	UJM_15502	RiC	CO	Ele	
60	829	UJM_15503	RiC	CO	Ele	Reference run
60	830	UJM_15490	BiC	CO	Ele	
60	832	UJM_15508	BiC	CO	Ele	Reference run
60	868	UJM_15525	Foc	CO	Ele	Article-only
62	895	RMIT09title	BiC	CO	Ele	Article-only
62	896	RMIT09titleO	BiC	CO	FOL	Article-only
68	679	I09LIP6Okapi	Foc	CO	Ele	Article-only
68	681	I09LIP6OWA	Foc	CO	Ele	Article-only
68	704	I09LIP6OWATh	Tho	CO	Ele	
72	666	umd_ric_1	RiC	CO	Ele	
72	667	umd_ric_2	RiC	CO	Ele	
72	870	umd_thorough_3	Tho	CO	Ele	
78	706	UWatFERBase	Foc	CO	FOL	
78	707	UWatFERBM25F	Foc	CO	FOL	
92	694	Lyon3LIAautoBEP	BiC	CAS	Ele	Phrases
92	695	Lyon3LIAmanBEP	BiC	CO	Ele	Phrases Manual Article-only
92	697	Lyon3LIAmanQE	Foc	CO	Ele	Phrases Manual Article-only
92	699	Lyon3LIAmanlmnt	Tho	CO	Ele	Phrases Manual
167	651	09RefT	Tho	CO	Ele	Reference run Article-only
167	654	09LrnRefF	Foc	CO	Ele	Reference run Article-only
167	657	09RefR	RiC	CO	Ele	Reference run Article-only
167	660	09LrnRefB	BiC	CO	Ele	Reference run Article-only
346	637	utCASartT09	Tho	CAS	Ele	Article-only
346	638	utCASartF09	Foc	CAS	Ele	Article-only Invalid
346	639	utCOartR09	RiC	CO	Ele	Article-only Invalid
346	640	utCOartB09	BiC	CO	Ele	Article-only Invalid
346	645	utCASrefF09	Tho	CAS	Ele	Reference run
346	646	utCASrefF09	Foc	CAS	Ele	Reference run
346	647	utCASrefR09	RiC	CAS	Ele	Reference run
346	648	utCASrefB09	BiC	CAS	Ele	Reference run