Overview of the INEX 2010 Ad Hoc Track

Paavo Arvola[1] Shlomo Geva[2], Jaap Kamps[3],
Ralf Schenkel[4], Andrew Trotman[5], and Johanna Vainio[1]

[1] University of Tampere, Tampere, Finland
paavo.arvola@uta.fi, s.johanna.vainio@uta.fi
[2] Queensland University of Technology, Brisbane, Australia
s.geva@qut.edu.au
[3] University of Amsterdam, Amsterdam, The Netherlands
kamps@uva.nl
[4] Max-Planck-Institut für Informatik, Saarbrücken, Germany
schenkel@mpi-sb.mpg.de
[5] University of Otago, Dunedin, New Zealand
andrew@cs.otago.ac.nz

Abstract. This paper gives an overview of the INEX 2010 Ad Hoc Track. The main goals of the Ad Hoc Track were three-fold. The first goal was to study focused retrieval under resource restricted conditions such as a small screen mobile device or a document summary on a hit-list. This leads to variants of the focused retrieval tasks that address the impact of result length/reading effort, thinking of focused retrieval as a form of "snippet" retrieval. The second goal was to extend the ad hoc retrieval test collection on the INEX 2009 Wikipedia Collection with additional topics and judgments. For this reason the Ad Hoc track topics and assessments stayed unchanged. The third goal was to examine the trade-off between effectiveness and efficiency by continuing the Efficiency Track as a task in the Ad Hoc Track. The INEX 2010 Ad Hoc Track featured four tasks: the *Relevant in Context* Task, the *Restricted Relevant in Context* Task, the *Restrict Focused* Task, and the *Efficiency* Task. We discuss the setup of the track, and the results for the four tasks.

1 Introduction

The main novelty of the Ad Hoc Track at INEX 2010 is its focus on retrieval under resource restricted conditions such as a small screen mobile device or a document summary on a hit-list. Here, retrieving full articles is no option, and we need to find the best elements/passages that convey the relevant information in the Wikipedia pages. So one can view the retrieved elements/passages as extensive result snippets, or as an on-the-fly document summary, that allow searchers to directly jump to the relevant document parts.

There are three main research questions underlying the Ad Hoc Track. The first goal is to study focused retrieval under resource restricted conditions, thinking of focused retrieval as a form of "snippet" retrieval, suggesting measures that factor in reading effort or by tasks that have restrictions on the length of results.

The second goal is to extend the ad hoc retrieval test collection on the INEX 2009 Wikipedia Collection—four times the size, with longer articles, and additional semantic markup than the collection used at INEX 2006–2008—with additional topics and judgments. For this reason the Ad Hoc track topics and assessments stayed unchanged, and the test collections of INEX 2009 and 2010 can be combined to form a valuable resource for future research. The third goal is to examine the trade-off between effectiveness and efficiency by continuing the Efficiency Track as a task in the Ad Hoc Track. After running as a separate track for two years, the Efficiency Track was merged into the Ad Hoc Track for 2010. For this new Efficiency Task, participants were asked to report efficiency-oriented statistics for their Ad Hoc-style runs on the 2010 Ad Hoc topics, enabling a systematic study of efficiency-effectiveness trade-offs with the different systems.

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 since INEX 2007. 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 [3] and to arbitrary passages of text [13].

The INEX 2010 Ad Hoc Track featured four tasks:

1. The *Relevant in Context* Task asks for non-overlapping results (elements or passages) grouped by the article from which they came, but is now evaluated with an effort-based measure.
2. The *Restricted Relevant in Context* Task is a variant in which we restrict results to maximally 500 characters per article, directly simulating the requirements of resource bounded conditions such as small screen mobile devices or summaries in a hitlist.
3. The *Restrict Focused* Task asks for a ranked-list of non-overlapping results (elements or passages) when restricted to maximally 1,000 chars per topic, simulating the summarization of all information available in the Wikipedia.
4. The *Efficiency* Task asks for a ranked-list of results (elements or passages) by estimated relevance and varying length (top 15, 150, or 1,500 results per topic), enabling a systematic study of efficiency-effectiveness trade-offs with the different systems.

Note that the resulting test collection also supports the INEX Ad Hoc tasks from earlier years: *Thorough*, *Focused*, and *Best in Context*. We discuss the results for the four tasks, giving results for the top 10 participating groups and discussing their best scoring approaches in detail.

The rest of the paper is organized as follows. First, Section 2 describes the INEX 2010 ad hoc retrieval tasks and measures. Section 3 details the collection, topics, and assessments of the INEX 2010 Ad Hoc Track. In Section 4, we report the results for the Relevant in Context Task (Section 4.2); the Restricted in Context Task (Section 4.3); the Restricted Focused Task (Section 4.4); and the Efficiency Task (Section 4.5). Section 5 discusses the differences between the measures that factor in result length and reading effort, and the old measures that were based on precision and recall of highlighted text retrieval. 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

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.

At INEX 2010, the task is interpreted as a form of "snippet" retrieval, and the evaluation will factor in result length/reading effort.

Restricted Relevant in Context Task The scenario underlying *Restricted Relevant in Context* addresses the requirements of resource bounded conditions, such as small screen mobile devices or summaries in a hitlist, directly by imposing a limit of maximally 500 characters per article.

Restricted 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.

At INEX 2010, we interpret the task as a form of summarization of all information available in the Wikipedia, and restrict results to exactly 1,000 chars per topic.

Efficiency Task The efficiency task is different in its focus on the trade-off between effectiveness and efficiency. Specifically, participants should create runs with the top-15, top-150, and top-1500 results for the Thorough task, a system-oriented task that has been used for many years in the Ad Hoc Track. Additionally, participants reported runtimes and I/O costs for evaluating each query as well as general statistics about the hard- and software environment used for generating the runs.

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.

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 2010, 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 only (hence 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 2010 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 the earlier Best in Context Task we used the distance between the best entry point in the run to that identified by an assessor.

Relevant in Context Task (INEX 2009) The evaluation of the Relevant in Context Task is based on the measures of generalized precision and recall [10] 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:

$$F_\beta = \frac{(1+\beta^2) \cdot \mathit{Precision} \cdot \mathit{Recall}}{(\beta^2 \cdot \mathit{Precision}) + \mathit{Recall}}.$$

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).

Relevant in Context Task (INEX 2010) The INEX 2010 version of the Relevant in Context Task is as before, but viewed as a form of snippet retrieval, and uses a different per-document score that takes reading effort into account. Specifically, the per document score is the character precision at a tolerance to irrelevance (T2I) point. In this measure, the user is expected to read the returned passages in document order. When result passages are read, the user is expected to continue reading from the beginning of the document and read the remaining parts in document order. The reading stops when the user's tolerance to irrelevance (i.e. the amount of irrelevant characters) is met, or all characters of a document are read. In other words, the reading/browsing is expected to end when the user has bypassed 300 (default) irrelevant characters. The T2I(300) score per document is again used in the measure based on generalized precision and 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).

Restricted Relevant in Context Task The evaluation of the Restricted Relevant in Context Task is the same as of the (unrestricted) Relevant in Context Task using T2I(300). So the main performance measure is mean average generalized precision (MAgP) based on T2I(300). We also present the generalized precision scores at early ranks (5, 10, 25, 50).

Restricted Focused Task We are interested in giving a quick overview of the relevant information in the whole Wikipedia. This is a variant of the Focused Task where we restrict the results to exactly 1,000 characters per topic. Evaluation will be in terms of set-based precision over the retrieved characters (char_prec). In addition, we will report on the earlier Focused measures such as 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]),

Efficiency 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. The Efficiency Task is evaluated as the INEX 2009 Thorough Task, which is basically identical to the Focused task. Since the Thorough Tasks allows for "overlapping" results, the evaluation will automatically discount text seen before in the ranked list. 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]),

For further details on the INEX measures, we refer to [1, 8].

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, 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 [12].

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:

```
//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

```xml
<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. Ad Hoc Track document `42010.xml` (in part).

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, 107 topics were selected for use in the INEX 2010 Ad Hoc Track as topic numbers 2010001–2010107.

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.

```
<topic id="2010048" ct_no="371">
  <title>Pacific navigators Australia explorers</title>
  <castitle>
    //explorer[about(., Pacific navigators Australia explorers)]
  </castitle>
  <phrasetitle>"Pacific navigators" "Australia explorers"</phrasetitle>
  <description>
    Find the navigators and explorers in the Pacific sea in search of
    Australia
  </description>
  <narrative>
    I am doing an essay on the explorers who discovered or charted
    Australia. I am already aware of Tasman, Cook and La Prouse and
    would like to get the full list of navigators who contributed to
    the discovery of Australia. Those for who there are disputes about
    their actual discovery of (parts of) Australia are still
    acceptable. I am mainly interested by the captains of the ships
    but other people who were on board with those navigators still
    relevant (naturalists or others). I am not interested in those
    who came later to settle in Australia.
</narrative>
</topic>
```

Fig. 2. INEX 2010 Ad Hoc Track topic 2010048.

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.

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

	total		# per topic				
	topics	number	min	max	median	mean	st.dev
judged articles	52	39,031	735	757	751	750.6	4.2
articles with relevance	52	5,471	5	506	65	105.2	112.8
highlighted passages	52	13,154	5	4,343	111	253.0	625.6
highlighted characters	52	17,641,119	3,841	2,624,502	129,440	339,252.3	527,349.0

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. All INEX 2010 tasks were evaluated against the text highlighted by the assessors, but the test collection does support the tasks of earlier years, such as the Thorough, Focused and Relevant in Context Tasks evaluated in terms of precision/recall, as well as the Best in Context Task evaluated against the best-entry-points.

The relevance judgments were frozen on November 3, 2010. At this time 52 topics had been fully assessed. Moreover, for 7 topics there is a second set of judgments by another assessor. All results in this paper refer to the 52 topics with the judgments of the first assigned assessor, which is typically the topic author.

- The 52 assessed topics were numbered 2010*n* with *n*: 003, 004, 006, 007, 010, 014, 016–021, 023, 025–027, 030–041, 043, 045–050, 054, 056, 057, 061, 068–070, 072, 075, 079, 095–097, 100, and 105–107.

Table 1 presents statistics of the number of judged and relevant articles, and passages. In total 39,031 articles were judged. Relevant passages were found in 5,471 articles. The mean number of relevant articles per topic is 105, but the distribution is skewed with a median of 65. There were 13,154 highlighted passages. The mean was 253 passages and the median was 111 passages per topic.

Figure 3 presents the number of articles with the given number of passages. The vast majority of relevant articles (3,388 out of 5,471) 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 3,166 but the distribution is very skewed with a median BEP offset of only 665. Figure 4 shows the distribution of the character offsets of the 5,471 best entry points. It is clear that the overwhelming majority of BEPs is at the beginning of the article.

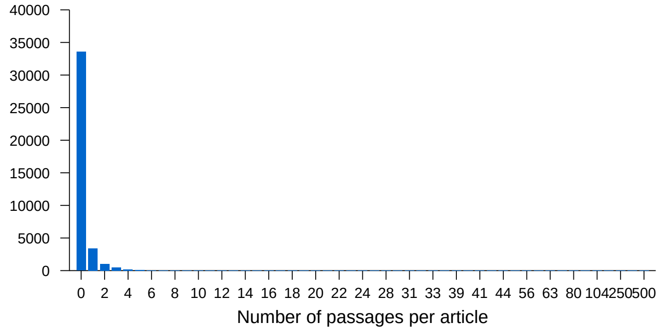

Fig. 3. Distribution of passages over articles.

Table 2. Statistics over relevant articles.

	total		# per relevant article				
	topics	number	min	max	median	mean	st.dev
best entry point offset	52	5,471	2	130,618	665	3,166.1	7,944.9
first relevant character offset	52	5,471	2	90,258	525	2,622.2	6,850.0
length relevant documents	52	5,471	249	179,200	5,545	12,084.9	17,274.5
relevant characters	52	5,471	4	179,166	897	3,224.5	7,326.1
fraction highlighted text	52	5,471	0.00036	1.000	0.239	0.358	0.332

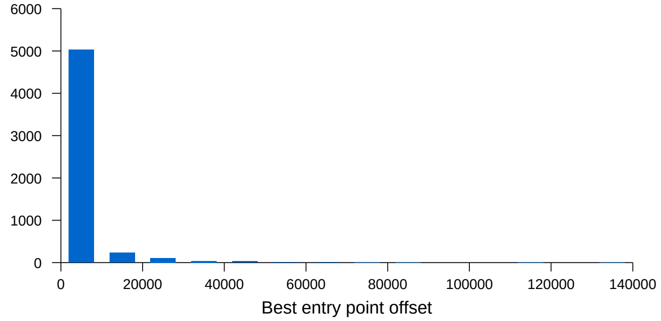

Fig. 4. Distribution of best entry point offsets.

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 2,662 but the median offset is only 525. 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.

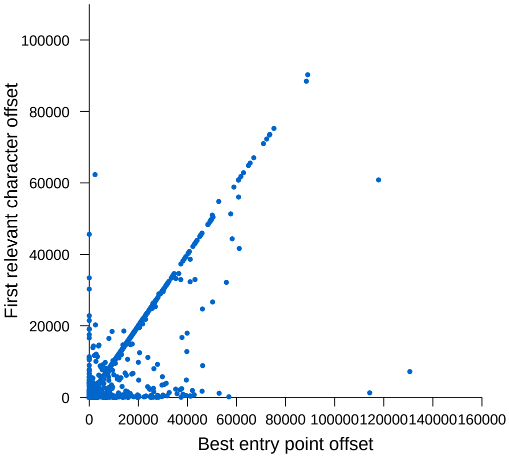

Fig. 5. Scatter plot of best entry point offsets versus the first relevant character.

Table 2 also shows statistics on the length of relevant articles. Many articles are relatively short with a median length of 5,545 characters, the mean length is 12,085 characters. The length of highlighted text in characters has a median of 897 (mean length is 3,225). Table 2 also show that amount of relevant text varies from almost nothing to almost everything. The mean fraction is 0.36, and the median is 0.24, indicating that typically one-third of 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.

3.4 Questionnaires

At INEX 2010, as in earlier years, 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 2010 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	Relevant in Context	Restricted Relevant in Context	Restricted Focused	Efficiency	CO query	CAS query	Phrase query	Reference run	Element results	Range of elements results	FOL results	# valid runs	# submitted runs
4	University of Otago	8	1	1	58	68	0	0	0	68	0	0	68	68
5	Queensland University of Technology	4	5	6	0	15	0	0	7	8	2	5	15	15
6	University of Amsterdam	2	2	2	0	6	0	0	0	0	0	6	6	6
9	University of Helsinki	0	0	4	0	4	0	0	0	0	0	4	4	8
22	ENSM-SE	4	0	0	0	4	0	4	2	4	0	0	4	4
25	Renmin University of China	2	0	0	0	2	0	0	0	2	0	0	2	2
29	INDIAN STATISTICAL INSTITUTE	2	2	3	3	10	0	0	1	3	0	7	10	12
55	Doshisha University	3	3	3	0	0	9	0	3	9	0	0	9	9
60	Saint Etienne University	1	0	0	0	1	0	0	1	1	0	0	1	2
62	RMIT University	2	0	0	0	2	0	0	0	2	0	0	2	2
65	Radboud University Nijmegen	1	1	3	0	4	1	0	3	0	0	5	5	9
68	University Pierre et Marie Curie - LIP6	0	0	3	3	6	0	0	2	6	0	0	6	6
72	University of Minnesota Duluth	1	1	1	0	0	3	0	0	3	0	0	3	0
78	University of Waterloo	1	1	1	0	3	0	0	0	0	0	3	3	3
98	LIA - University of Avignon	4	2	2	3	11	0	11	0	3	0	8	11	10
138	Kasetsart University	0	0	0	0	0	0	0	0	0	0	0	0	3
167	Peking University	12	9	2	17	40	0	0	0	40	0	0	40	45
557	Universitat Pompeu Fabra	0	0	3	0	3	0	0	1	0	0	3	3	9
	Total runs	47	27	34	84	179	13	15	20	149	2	41	192	213

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 213 runs were submitted by 18 participating groups. Table 5 lists the participants and the number of runs they submitted, also broken down over the tasks (Relevant in Context, Restricted Relevant in Context, Restricted Focused, or Efficiency); 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 21 runs turned out to be invalid.

Table 6. Top 10 Participants in the Ad Hoc Track Relevant in Context Task (INEX 2010 T2I-score).

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p22-Emse303R	0.3752	0.3273	0.2343	0.1902	0.1977
p167-36p167	0.2974	0.2536	0.1921	0.1636	0.1615
p98-I10LIA1FTri	0.2734	0.2607	0.2067	0.1692	0.1588
p5-Reference	0.2736	0.2372	0.1800	0.1535	0.1521
p4-Reference	0.2684	0.2322	0.1714	0.1442	0.1436
p65-runRiCORef	0.2642	0.2310	0.1694	0.1431	0.1377
p25-ruc-2010-base2	0.2447	0.2198	0.1744	0.1359	0.1372
p62-RMIT10titleO	0.2743	0.2487	0.1880	0.1495	0.1335
p55-DUR10atcl	0.1917	0.1484	0.1163	0.0982	0.1014
p6-0	0.1798	0.1614	0.1314	0.1183	0.0695

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. For the efficiency task, we allowed sets of runs with 15, 150, 1,500 results per topic. The submissions are spread well over the ad hoc retrieval tasks with 47 submissions for Relevant in Context, 27 submissions for Restricted Relevant in Context, 34 for Restricted Focused, and 84 submissions for Efficiency.

4.2 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, factoring reading effort with T2I(300). The official measure for the task was mean average generalized precision (MAgP).

Table 6 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 three groups (based on MAgP).

ENSM-SE An element run, using the keyword (CO) query, the phrase title and the reference run.

Description: The method for scoring one document/element is based on the proximity of query terms in the document [2]. In this basic method, the influence of query terms is modelized by triangular functions. For the Run Emse303R, the height of the triangle was enlarged proportionnally to a weight learnt with the 2009 queries and assessments [5]. In the final run the elements and the documents are sorted with many keys. The first documents returned are those that appear both in our list and in the reference run, then documents from our list. For each document, elements are returned according to their score.

Peking University An element run, using the keyword (CO) query.
Description: Starting from a BM25 article retrieval run, then according to the semantic query model MAXimal Lowest Common Ancestor (MAXLCA), candidate element results are extracted. These elements are further ranked by BM25 and Distribution Measurements.

LIA – **University of Avignon** A FOL run, using the keyword (CO) query, and the phrase query.
Description: Based on advanced query expansion. We first retrieve the 10 top documents with a baseline query. The queries of this baseline are generated by combining the words from the ⟨`title`⟩ and ⟨`phrasetitle`⟩ fields of the topics. The documents are ranked with a language modeling approach and the probabilities are estimated using Dirichlet smoothing. We select the 50 most frequent unigrams, 20 most frequent 2-grams and 10 most frequent 3-grams from these 10 top-ranked documents, and we use them to expand the baseline query, allowing term insertions within the 2-grams and 3-grams. Finally, we retrieve the 1000 top documents with this expanded query and we get the file offset lengths corresponding to the first ¡section¿ field of each document.

Based on the information from these and other participants:

- The runs ranked ninth (*p55-DUR10atcl*) is using the CAS query. All other runs use only the CO query in the topic's title field.
- The first (*p22-Emse303R*), second (*p167-36p167*) and fourth (*p5-Reference*) run retrieve elements; the second (*p167-36p167*) and tenth (*p6-0*) run use FOL passages.
- Solid article ranking seems a prerequisite for good overall performance, with fifth (*p4-Reference*) through ninth (*p55-DUR10atcl*) runs retrieving only full articles.

4.3 Restricted Relevant in Context Task

We now discuss the results of the Restricted Relevant in Context Task in which we allow for only 500 characters per article to be retrieved. The Restricted Relevant in Context Task was also evaluated using generalized precision with the generalized score per article based on T2I(300). 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 Restricted 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.

Table 7. Top 10 Participants in the Ad Hoc Track Restricted Relevant in Context Task (INEX 2010 T2I-score).

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p167-32p167	0.2910	0.2474	0.1872	0.1595	0.1580
p98-I10LIA2FTri	0.2631	0.2503	0.1972	0.1621	0.1541
p5-Reference	0.2722	0.2362	0.1785	0.1520	0.1508
p4-Reference	0.2684	0.2322	0.1714	0.1442	0.1436
p65-runReRiCORef	0.2641	0.2313	0.1686	0.1428	0.1375
p78-UWBOOKRRIC2010	0.1111	0.1001	0.0874	0.0671	0.0650
p55-DURR10atcl	0.1555	0.1300	0.1003	0.0822	0.0600
p6-categoryscore	0.1439	0.1191	0.1053	0.0980	0.0576
p29-ISI2010_rric_ro	0.1979	0.1673	0.1183	0.1008	0.0485
p72-1	0.0000	0.0000	0.0000	0.0000	0.0000

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

Peking University Element retrieval run using the CO query.
Description: This is a variant of the run for the Relevant in Context task. That is, starting from a BM25 article retrieval run, then according to the semantic query model MAXimal Lowest Common Ancestor (MAXLCA), candidate element results are extracted. These elements are further ranked by BM25 and Distribution Measurements. Here, the first 500 characters are returned for each element.

LIA – University of Avignon FOL passage retrieval using the CO query and phrases.
Description: Based on advanced query expansion. We first retrieve the 10 top documents with a baseline query. The queries of this baseline are generated by combining the words from the ⟨`title`⟩ and ⟨`phrasetitle`⟩ fields of the topics. The documents are ranked with a language modeling approach and the probabilities are estimated using Dirichlet smoothing. We select the 50 most frequent unigrams, 20 most frequent 2-grams and 10 most frequent 3-grams from these 10 top-ranked documents, and we use them to expand the baseline query, allowing term insertions within the 2-grams and 3-grams. Finally, we only select the 500 first characters of the first ⟨`section`⟩ field of each document (or less if the field contains less than 500 characters).

Queensland University of Technology Element retrieval run using the CO query, based on the reference run. 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.

Based on the information from these and other participants:

- The best run (*p167-32p167*), the third run (*p5-Reference*), and the tenth run (*p72-1*) retrieve elements. The fourth run (*p4-Reference*), seventh run

Table 8. Top 10 Participants in the Ad Hoc Track Restricted Focused Task.

Participant	char_prec	iP[.01]	iP[.05]	iP[.10]	MAiP
p68-LIP6-OWPCparentFo	0.4125	0.1012	0.0385	0.0000	0.0076
p55-DURF10SIXF*	0.3884	0.1822	0.0382	0.0000	0.0088
p9-yahRFT	0.3435	0.1186	0.0273	0.0000	0.0069
p98-LIAenertexTopic	0.3434	0.1500	0.0000	0.0000	0.0077
p167-40p167	0.3370	0.1105	0.0384	0.0000	0.0067
p65-runFocCORef	0.3361	0.0964	0.0435	0.0000	0.0067
p5-Reference	0.3199	0.1170	0.0431	0.0000	0.0070
p557-UPFpLM45co	0.3066	0.1129	0.0264	0.0000	0.0070
p4-Reference	0.3036	0.0951	0.0429	0.0000	0.0063
p29-ISI2010_rfcs_ref	0.2451	0.1528	0.0192	0.0000	0.0072

(*p55-DURR10atcl*), eighth run (*p6-categoryscore*) retrieve full articles, and the remaining four runs retrieve FOL passages.

– With the exception of the runs ranked seventh (*p55-DURR10atcl*) and tenth (*p72-1*), which used the CAS query, all the other best runs per group use the CO query.

4.4 Restricted Focused Task

We now discuss the results of the Restricted Focused Task in which a ranked-list of non-overlapping results (elements or passages) was required, totalling maximally 1,000 characters per topic.

The official measure for the task was the set-based character precision over the 1,000 characters retrieved (runs were restricted or padded to retrieve exactly 1,000 characters if needed). Table 8 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 column gives the character-based precision over 1,000 characters retrieved, the third to fifth column give the interpolated precision at 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 three groups (based on official measure for the task, char_prec).

LIP6 An element retrieval run using the CO query.
Description: A learning to rank run that is retrieving elements for the CO queries (negated words are removed and words are not stemmed). We limit the domain of elements to the tag-types: {sec, ss, ss1, ss2, ss3, ss4, p}.

Doshisha University A manual element retrieval run, using the CAS query.
Description: We used the result reconstruction method from earlier years. In this method, we aim to extract more relevant fragments without irrelevant parts to return appropriate granular fragments as search results. We considered: 1) which granular fragments are more appropriate in overlapped fragments, and 2) what size is more suitable for search results. Our method

Table 9. Participants in the Ad Hoc Track Efficiency Task.

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p167-18P167	0.4561	0.4432	0.4215	0.3936	0.2354
p4-OTAGO-2010-10topk-18	0.4425	0.4272	0.4033	0.3697	0.2304
p68-LIP6-OWPCRefRunTh	0.4790	0.4651	0.4343	0.3985	0.2196
p29-ISI2010_thorough.1500	0.2931	0.2930	0.2480	0.2145	0.0846
p98-I10LIA4FBas	0.5234	0.4215	0.2500	0.1677	0.0417

combines neighbor relevant fragments to satisfy these views, by using the initial fragments obtained by a well-known scoring technique: BM25E as a basic scoring method for scoring each fragment, and ITF (inverse tag frequency) instead of IPF (inverse path frequency) because there are a number of tags in the test collection.

University of Helsinki A passage retrieval run using the CO query.
Description: The result list for each topic consists of a total of 1,000 characters from the beginning of the top two articles as ranked by the Yahoo! search-engine. Retrieving the passages from the beginning of the article is based on the assumption that the best entry point is in the beginning of the article. Because Yahoo! does not suggest any other entry point to the article, retrieving the beginning of the article is also what Yahoo! provides to users. Only the title field of the topic was used in the query.

Based on the information from these and other participants:

- Nine runs use the CO query. Only the second run (*p55-DURF10SIXF*) is a manual run using the CAS query.
- Only the ninth ranked system, (*p4-Reference*), retrieves full articles. The runs ranked first (*p68-LIP6-OWPCparentFo*), second (*p55-DURF10SIXF*⋆), and fifth (*p167-40p167*), and seventh (*p5-Reference*), retrieve elements. The remaining five runs retrieve FOL passages.

4.5 Efficiency Task

We now discuss the results of the Efficiency Task focusing on efficiency rather than effectiveness, and especially the trade-off between efficiency and effectiveness. Participants were asked to submit ranked-lists of 15 results, or 150 results, or 1,500 results per topic. The official measure for the task was mean average interpolated precision (MAiP). Table 9 shows the best run of the 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 three groups (based on official measure for the task, MAiP).

Peking University An element retrieval run using the CO query.

Description: This is again a variant of the runs for (Restricted) Relevant in Context. That is, starting from a BM25 article retrieval run, then according to the semantic query model MAXimal Lowest Common Ancestor (MAXLCA), candidate element results are extracted. These elements are further ranked by BM25 and Distribution Measurements. Here, the parameters in ranking functions are tuned by a learning method.

University of Otago An article retrieval run using the CO query.

Description: The goal of the Otago runs was sub-millisecond per query. This was achieved using three techniques: impact ordered indexes, static pruning, and the use of a top-k ranking algorithm. Run p4-OTAGO-2010-10topk-18 scored the best in precision because it did the least pruning and least top-k restriction. It used BM25 and index-time S-stripper stemming. The fastest runs were, indeed, sub-millisecond, but at a reduced precision.

LIP6 An article retrieval run using the CO query.

Description: A learning to rank run that is retrieving top 1,500 documents for the CO queries (negated words are removed and words are not stemmed). For each document, the /`article`[1] element is retrieved.

Figure 6 shows the effectiveness, in terms of either iP[0.01] or MAiP, against the run-time efficiency. There is a vague diagonal trend—the best scoring runs tend to be the least efficient—but the trend is weak at best. Only the *University of Otago* submitted provided a large set of runs with all details. The MAiP scores tend to improve with longer runs, other things being equal this is no surprise. For the iP[0.01] scores, this is hardly the case.

Based on the information from these and other participants:

- The top scoring run (*p167-18P167*) uses elements, and the fifth run (*p98-I10LIA4FBas*) uses FOL passages. The other three runs retrieve articles.
- All runs 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 “$\star$”) than lower ranked runs. For the Relevant in Context Task, we see that the top run is significantly better than ranks 2 through 10. The second best run is significantly better than ranks 4 through 10. The third run better than ranks 6–10, the fourth run better than ranks 5-10, the fifth run better than runs 6 and 9–10, the sixth through eighth run better than runs 9–10. Of the 45 possible pairs of runs, there are 36 (or 80%) significant differences, making MAgP a very discriminative measure. For the Restricted Relevant in Context Task, we see that the top run is significantly better than ranks 2 through 10. The second best run is significantly better than ranks 6 through 10. The third run better than ranks 4–10, the fourth run better than ranks 5–10, the fifth run better than runs 6–10, the sixth run better than 9–10, and the seventh through ninth run better than runs 10. Of the 45 possible

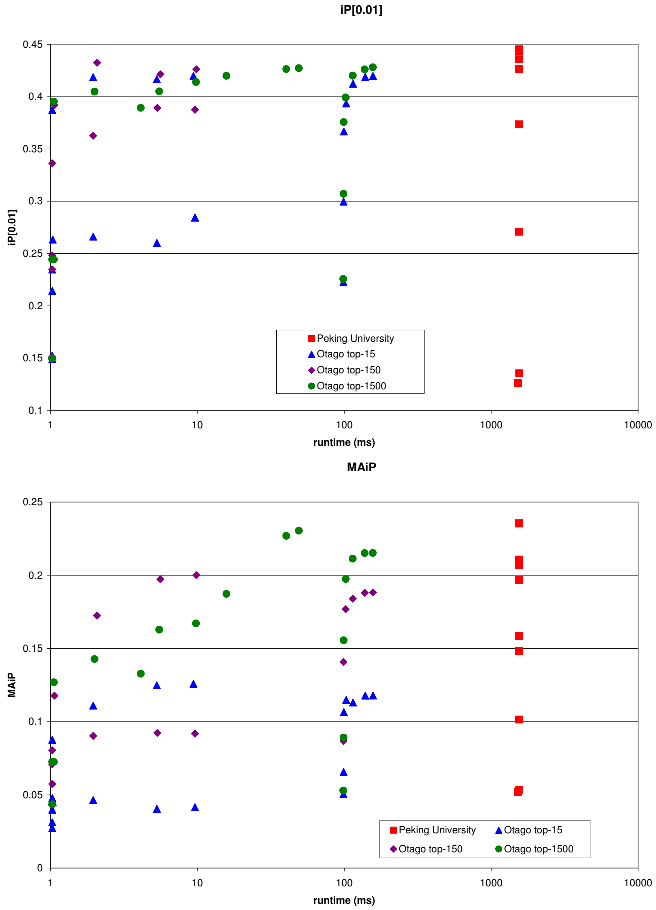

Fig. 6. Trade-off between Effectiveness and Efficiency: iP[0.01] (top) and MAiP (bottom).

pairs of runs, there are again 36 (or 80%) significant differences, confirming that MAgP is a very discriminative measure. For the Restricted Focused Task, we see that character precision at 1,000 characters is a rather unstable measure. The best run is significantly better than runs 7–10, and the runs ranked 2–5 and significantly better than the run ranked 10. Of the 45 possible pairs of runs, there are only 8 (or 18%) significant differences. Hence we should be careful when drawing conclusions based on the Focused Task results. For the Efficiency Task, we see that the performance (measured by MAiP) of the top scoring run is significantly better than the runs at rank 4 and 5. The same holds for the second and third best run. The fourth best run is significantly better than the run at rank 5. Of the 10 possible pairs of runs, there are 7 (or 70%) significant differences.

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

(a) Relevant in Context Task

	1	2	3	4	5	6	7	8	9	10
p22		⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
p167			-	⋆	⋆	⋆	⋆	⋆	⋆	⋆
p98				-	-	⋆	⋆	⋆	⋆	⋆
p5					⋆	⋆	⋆	⋆	⋆	⋆
p4						⋆	-	-	⋆	⋆
p65							-	-	⋆	⋆
p25								-	⋆	⋆
p62									⋆	⋆
p55										-
p6										

(b) Restricted Relevant in Context Task

	1	2	3	4	5	6	7	8	9	10
p167		-	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
p98			-	-	-	⋆	⋆	⋆	⋆	⋆
p5				⋆	⋆	⋆	⋆	⋆	⋆	⋆
p4					⋆	⋆	⋆	⋆	⋆	⋆
p65						⋆	⋆	⋆	⋆	⋆
p78							-	-	⋆	⋆
p55								-	-	⋆
p6									-	⋆
p29										⋆
p72										

(c) Restricted Focused Task

	1	2	3	4	5	6	7	8	9	10
p68		-	-	-	-	-	⋆	⋆	⋆	⋆
p55			-	-	-	-	-	-	-	⋆
p9				-	-	-	-	-	-	⋆
p98					-	-	-	-	-	⋆
p167						-	-	-	-	⋆
p65							-	-	-	-
p5								-	-	-
p557									-	-
p4										-
p29										

(d) Efficiency Task

	1	2	3	4	5
p167		-	-	⋆	⋆
p4			-	⋆	⋆
p68				⋆	⋆
p29					⋆
p98					

5 Analysis of Reading Effort

In this section, we will look in detail at the impact of the reading effort measures on the effectiveness of Ad Hoc Track submissions, by comparing them to the INEX 2009 measures based on precision and recall.

5.1 Relevant in Context

Table 11 shows the top 10 participating groups (only the best run per group is shown) in the Relevant in Context Task evaluated using the INEX 2009 measures based on a per article F-score. 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.

Comparing Table 11 using the F-score and Table 6 using the T2I-score, we see some agreement. There are six runs in both tables, and some variant of the runs. There are however, notable upsets in the system rankings:

- Over all 47 Relevant in Context submissions, the system rank correlation is 0.488 between the F-score based and the T2I-score based evaluation.

Table 11. Top 10 Participants in the Ad Hoc Track Relevant in Context Task (INEX 2009 F-score).

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p22-Emse301R	0.3467	0.3034	0.2396	0.1928	0.1970
p167-21p167	0.3231	0.2729	0.2107	0.1767	0.1726
p4-Reference	0.3217	0.2715	0.2095	0.1751	0.1710
p25-ruc-2010-base2	0.2761	0.2627	0.2128	0.1686	0.1671
p65-runRiCORef	0.3190	0.2700	0.2078	0.1735	0.1623
p62-RMIT10title	0.2869	0.2585	0.1958	0.1573	0.1541
p98-I10LIA1FTri	0.2230	0.2048	0.1725	0.1421	0.1298
p55-DUR10atcl	0.2031	0.1663	0.1339	0.1096	0.1122
p29-ISI2010_ric_ro	0.2082	0.1874	0.1429	0.1250	0.0693
p5-Reference	0.0978	0.0879	0.0698	0.0640	0.0634

Table 12. Top 10 Participants in the Ad Hoc Track Restricted Relevant in Context Task (INEX 2009 F-score).

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p5-Reference	0.1815	0.1717	0.1368	0.1206	0.1064
p98-I10LIA2FTri	0.1639	0.1571	0.1340	0.1130	0.1053
p167-27p167	0.1622	0.1570	0.1217	0.1061	0.1030
p4-Reference	0.1521	0.1469	0.1119	0.0968	0.0953
p65-runReRiCORef	0.1610	0.1508	0.1138	0.0986	0.0945
p55-DURR10atcl	0.1369	0.1102	0.0870	0.0727	0.0537
p78-UWBOOKRRIC2010	0.0760	0.0777	0.0711	0.0544	0.0497
p6-0	0.0996	0.0880	0.0816	0.0782	0.0462
p29-ISI2010_rric_ro	0.1276	0.1189	0.0820	0.0759	0.0327
p72-1	0.0000	0.0000	0.0000	0.0000	0.0000

– Taking the top 10 systems based on the T2I-score, their system ranks on the F-score have a correlation of 0.467.
– Taking the top 10 systems based on the F-score, their system ranks on the T2I-scores have a correlation of 0.956.

The overall system rank correlation is fairly low: the reading effort measure significantly affects the ranking. There is an interesting unbalance between the top 10 rankings. On the one hand, systems scoring well on the F-score tend to get very similar rankings based on the T2I-score. This makes sense since systems with a high F-score will tend to retrieve a lot of relevant text, and hence are to some degree immune to the T2I conditions. On the other hand, systems that score well on the T2I-score tend to have fairly different rankings based on the F-score. This can be explained by the high emphasis on precision of the T2I measures, and the relative importance of recall for the F-score.

Restricted Relevant in Context Table 12 shows the top 10 participating groups (only the best run per group is shown) in the Restricted Relevant in Context Task evaluated using the INEX 2009 measures based on a per article

F-score. 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.

Comparing Table 12 using the F-score and Table 7 using the T2I-score, we see some agreement.

- Over all 27 Restricted Relevant in Context submissions, the system rank correlation is 0.761 between the F-score based and the T2I-score based evaluation.
- Taking the top 10 systems based on the T2I-score, their system ranks on the F-score have a correlation of 0.022.
- Taking the top 10 systems based on the F-score, their system ranks on the T2I-scores have a correlation of 0.156.

The overall system rank correlation is higher than for the Relevant in Context task above, but the system rank correlations between the top 10's however are substantially lower.

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 2010, 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. 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 are now also considered, and we work with all 213 runs submitted to the Ad Hoc Track.

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).

No less than five of the top 10 runs retrieved exclusively full articles: the three runs at rank one (*p22-Emse301R*), rank two (*p167-38P167*), and rank six (*p5-Reference*) retrieved elements proper, and the two runs at rank four

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

Participant	P5	P10	1/rank	map	bpref
p22-Emse301R	0.6962	0.6423	0.8506	0.4294	0.4257
p167-38P167	0.7115	0.6173	0.8371	0.3909	0.3863
p25-ruc-2010-base2	0.6077	0.5846	0.7970	0.3885	0.3985
p98-I10LIA2FTri	0.6192	0.5827	0.7469	0.3845	0.3866
p4-Reference	0.6423	0.5750	0.7774	0.3805	0.3765
p5-Reference	0.6423	0.5750	0.7774	0.3805	0.3765
p62-RMIT10title	0.6346	0.5712	0.8087	0.3653	0.3683
p68-LIP6-OWPCRefRunTh	0.6115	0.5673	0.7765	0.3310	0.3480
p78-UWBOOKRRIC2010	0.5615	0.5115	0.7281	0.3237	0.3395
p65-runRiCORef	0.5808	0.5346	0.7529	0.3177	0.3382

(*p98-I10LIA2FTri*) and rank nine (*p78-UWBOOKRRIC2010*) retrieved FOL passages. 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. Runs from the Relevant in Context task at ranks 1, 3, 7; runs from the Restricted Relevant in Context task at ranks 4, 5, 9, 10; runs from the Restricted Focused task at ranks 6; and runs from the Efficiency task at ranks 2, 8

If we break-down all runs over the original tasks, shown in Table 14, we can compare the ranking to Section 4 above. We see some runs that are familiar from the earlier tables: five Relevant in Context runs correspond to Table 6, seven Restricted in Context runs correspond to Table 7, seven Restricted Focused runs correspond to Table 8, and five Efficiency runs correspond to Table 9. More formally, we looked at how the two system rankings correlate using kendall's tau.

- Over all 47 Relevant in Context submissions the system rank correlation between MAgP and map is 0.674.
- Over all 27 Restricted Relevant in Context submissions the system rank correlation between MAgP and map is 0.647.
- Over all 34 Restricted Focused task submissions the system rank correlation is 0.134 between char_prec and map, and 0.194 between MAiP and map.
- Over all 84 Efficiency Task submissions the system rank correlation is 0.697 between MAiP and map.

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 only exception is the correlation between article retrieval and the Restricted Focused task. This is a likely effect of the evaluation over the bag of all retrieved text, regardless of the internal ranking.

7 Discussion and Conclusions

The Ad Hoc Track at INEX 2010 studied focused retrieval under resource restricted conditions such as a small screen mobile device or a document summary

Table 14. Top 10 Participants in the Ad Hoc Track: Article retrieval per task.

(a) Relevant in Context Task

Participant	P5	P10	1/rank	map	bpref
p22-Emse301R	0.6962	0.6423	0.8506	0.4294	0.4257
p25-ruc-2010-base2	0.6077	0.5846	0.7970	0.3885	0.3985
p98-I10LIA1ElTri	0.6192	0.5827	0.7469	0.3845	0.3866
p167-21p167	0.6423	0.5750	0.7774	0.3805	0.3765
p4-Reference	0.6423	0.5750	0.7774	0.3805	0.3765
p5-Reference	0.6423	0.5750	0.7774	0.3805	0.3765
p62-RMIT10title	0.6346	0.5712	0.8087	0.3653	0.3683
p78-UWBOOKRIC2010	0.5615	0.5115	0.7281	0.3237	0.3395
p65-runRiCORef	0.5808	0.5346	0.7529	0.3177	0.3382
p557-UPFpLM45co	0.5885	0.5423	0.7623	0.3041	0.3210

(b) Restricted Relevant in Context Task

Participant	P5	P10	1/rank	map	bpref
p98-I10LIA2FTri	0.6192	0.5827	0.7469	0.3845	0.3866
p4-Reference	0.6423	0.5750	0.7774	0.3805	0.3765
p167-29p167	0.6423	0.5750	0.7774	0.3805	0.3765
p5-Reference	0.6423	0.5750	0.7774	0.3805	0.3765
p78-UWBOOKRRIC2010	0.5615	0.5115	0.7281	0.3237	0.3395
p65-runReRiCORef	0.5808	0.5346	0.7529	0.3177	0.3382
p557-UPFsecLM45co	0.5846	0.5212	0.7904	0.2684	0.2919
p9-goo100RRIC	0.6423	0.5712	0.8830	0.2180	0.2503
p6-categoryscore	0.3115	0.2981	0.4319	0.1395	0.2566
p55-DURR10atcl	0.3269	0.2769	0.4465	0.1243	0.1540

(c) Restricted Focused Task

Participant	P5	P10	1/rank	map	bpref
p4-Reference	0.6423	0.5750	0.7774	0.3805	0.3765
p5-Reference	0.6423	0.5750	0.7774	0.3805	0.3765
p65-runFocCORef	0.5808	0.5346	0.7529	0.3177	0.3382
p98-LIAenertexDoc	0.5654	0.3192	0.7388	0.0636	0.0759
p55-DURF10SIXF*	0.4000	0.2442	0.7186	0.0531	0.0603
p557-UPFpLM45co	0.3769	0.2038	0.7308	0.0492	0.0531
p167-40p167	0.3038	0.1519	0.8462	0.0474	0.0484
p6-0	0.3154	0.3096	0.4230	0.0384	0.0591
p9-goo100RFT	0.3038	0.1519	0.8654	0.0382	0.0399
p29-ISI2010_rfcs_ref	0.2577	0.1308	0.5689	0.0300	0.0346

(d) Thorough Task

Participant	P5	P10	1/rank	map	bpref
p167-38P167	0.7115	0.6173	0.8371	0.3909	0.3863
p4-OTAGO-2010-10topk-18	0.6115	0.5654	0.7632	0.3738	0.3752
p98-I10LIA4FBas	0.6115	0.5673	0.7984	0.3648	0.3671
p68-LIP6-OWPCRefRunTh	0.6115	0.5673	0.7765	0.3310	0.3480
p29-ISI2010_thorough.1500	0.3731	0.2865	0.7294	0.0886	0.1804

on a hit-list. Here, retrieving full articles is no option, and we need to find the best elements/passages that convey the relevant information in the Wikipedia pages. So one can view the retrieved elements/passages as extensive result snippets, or as an on-the-fly document summary, that allow searchers to directly jump to the relevant document parts.

In this paper we provided an overview of the INEX 2010 Ad Hoc Track that contained four tasks: The *Relevant in Context* Task asked for non-overlapping results (elements or passages) grouped by the article from which they came, but evaluated with an effort-based measure. The *Restricted Relevant in Context* Task is a variant in which we restricted results to maximally 500 characters per article, directly simulating the requirements of resource bounded conditions such as small screen mobile devices or summaries in a hitlist. The *Restrict Focused* Task asked for a ranked-list of non-overlapping results (elements or passages) restricted to maximally 1,000 chars per topic, simulating the summarization of all information available in the Wikipedia. The *Efficiency* Task asked for a ranked-list of results (elements or passages) by estimated relevance and varying length (top 15, 150, or 1,500 results per topic), enabling a systematic study of efficiency-effectiveness trade-offs with the different systems. We discussed the results for the four tasks.

The Ad Hoc Track had three main research questions. The first goal was to study focused retrieval under resource restricted conditions such as a small screen mobile device or a document summary on a hit-list. That is, to think of focused retrieval as a form of "snippet" retrieval. This leads to variants of the focused retrieval tasks that address the impact of result length/reading effort, either by measures that factor in reading effort or by tasks that have restrictions on the length of results. The results of the effort based measures are a welcome addition to the earlier recall/precision measures. It addresses the counter-intuitive effectiveness of article-level retrieval—given that ensuring good recall is much easier than ensuring good precision [7]. As a result there are significant shifts in the effectiveness of systems that attempt to pinpoint the exact relevant text, and are effective enough at it. Having said that, even here locating the right articles remains a prerequisite for obtaining good performance, and finding a set of measures that resonate closely with the perception of the searchers remains an ongoing quest in focused retrieval.

The second goal was to extend the ad hoc retrieval test collection on the INEX 2009 Wikipedia Collection—four times the size, with longer articles, and additional semantic markup—with additional topics and judgments. For this reason the Ad Hoc track topics and assessments stayed unchanged, and the test collections of INEX 2009 and 2010 combined form a valuable resource for future research. INEX 2010 added 52 topics to the test collection on the INEX Wikipedia Corpus, making it a total of 120 topics. In addition there are seven double judged topics. This results in an impressive test collection, with a large topic set and highly complete judgments [11]. There are many ways of (re)using the resulting test collection for passage retrieval, XML element retrieval, or article retrieval.

The third goal was to examine the trade-off between effectiveness and efficiency by continuing the Efficiency Track as a task in the Ad Hoc Track. After running as a separate track for two years, the Efficiency Track was merged into the Ad Hoc Track for 2010. For this new Efficiency Task, participants were asked to report efficiency-oriented statistics for their Ad Hoc-style runs on the 2010 Ad Hoc topics, enabling a systematic study of efficiency-effectiveness trade-offs with the different systems. The Efficiency task received more runs than at INEX 2009 but of a smaller number of participants. Regarding efficiency, average running times per topic varied from 1ms to 1.5 seconds, where the fastest runs where run on indexes kept in memory. This is again almost an order of magnitude faster than the fastest system from INEX 2009, and the low absolute response times clearly demonstrate that the current Wikipedia-based collection is not large enough to be a true challenge for current systems. Result quality was comparable to other runs submitted to other tasks in the AdHoc Track.

This is the fifth year that INEX has studied ad hoc retrieval against the Wikipedia. In 2006–2008 the English Wikipedia of early 2006 transformed into XML was used covering 659,338 Wikipedia articles [4]. Over the three years a combined test collection of 291 topics was created. In 2009–2010 a new collection was created based on a late 2008 dump of the English Wikipedia, containing 2,666,190 Wikipedia articles and incorporating semantic annotations from YAGO [based on 12]. Over the last two years a combined test collection of 120 topics was created. The test collections on Wikipedia have large sets of topics, 291 for the 2006–2008 Wikipedia and 120 for the 2009–2010 Wikipedia. There are relevance judgments at the passage level (both best-entry-points as well as the exact relevant text) plus derived article-level judgments. The resulting judgments are relatively "complete" due to the varied pools and especially the encyclopedic corpus [11]. There is a range of evaluation measures for evaluating the various retrieval tasks [1, 8], in addition to the standard measures that can be used for article-level retrieval. In addition, there is rich information on topic authors and assessors, and their topics and judgments based on extensive questionnaire, allowing for detailed further analysis and reusing topics that satisfy particular conditions [6, 9]. After five years, there seems little additional benefit in continuing with focused retrieval against the Wikipedia corpus, given the available test collections that are reusable in various ways. It is time for a new challenge, and other tracks have started already addressing other aspects of ad hoc retrieval: the INEX 2010 Book Track using a corpus of scanned books, the INEX 2010 Data Centric Track using a corpus of IMDb data, and the INEX 2010 Interactive Track using a corpus of Amazon and Library Thing data.

Acknowledgments

Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO, grants 612.066.513, 639.072.601, and 640.001.501). Paavo Arvola and Johanna Vainio were supported by the Academy of Finland (grants #115480 and #130482).

Bibliography

[1] P. Arvola, J. Kekäläinen, and M. Junkkari. Expected reading effort in focused retrieval evaluation. *Information Retrieval*, 13:460–484, 2010.

[2] M. Beigbeder. Focused retrieval with proximity scoring. In *Proceedings of the 2010 ACM Symposium on Applied Computing (SAC'10)*, pages 1755–1759. ACM Press, New York NY, USA, 2010.

[3] 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.

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

[5] M. Géry, C. Largeron, and F. Thollard. Integrating structure in the probabilistic model for information retrieval. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 763–769. IEEE Computer Society, 2008.

[6] J. Kamps and B. Larsen. Understanding differences between search requests in XML element retrieval. In A. Trotman and S. Geva, editors, *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, pages 13–19, 2006.

[7] J. Kamps, M. Koolen, and M. Lalmas. Locating relevant text within XML documents. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 847–849. ACM Press, New York NY, USA, 2008.

[8] 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.

[9] J. Kamps, M. Lalmas, and B. Larsen. Evaluation in context. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas, editors, *Proceedings of the 13th European Conferences on Digital Libraries (ECDL 2009)*, volume 5714 of *LNCS*, pages 339–351. Springer Verlag, Berlin, Heidelberg, 2009.

[10] 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.

[11] S. Pal, M. Mitra, and J. Kamps. Evaluation effort, reliability and reusability in XML retrieval. *Journal of the American Society for Information Science and Technology*, 62:375–394, 2011.

[12] 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.

[13] 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	1019	Reference	RiC	CO	Ele	Article-only
4	1020	Reference	RRiC	CO	Ele	Article-only
4	1021	Reference	RFoc	CO	Ele	Article-only
4	1138	OTAGO-2010-10topk-18	Eff	CO	Ele	Article-only
5	1205	Reference	RiC	CO	Ele	Reference run
5	1206	Reference	RRiC	CO	Ele	Reference run
5	1207	Reference	RFoc	CO	Ele	Reference run
5	1208	Reference	RiC	CO	Ran	Reference run Invalid
5	1212	Reference	RRiC	CO	Ele	Reference run
5	1213	Reference	RFoc	CO	Ele	Reference run
6	1261	0	RiC	CO	FOL	
6	1265	categoryscore	RRiC	CO	FOL	Article-only
6	1266	0	RRiC	CO	FOL	
6	1268	0	RFoc	CO	FOL	
9	1287	goo100RRIC	RRiC	CO	FOL	Invalid
9	1294	goo100RFT	RFoc	CO	FOL	
9	1295	yahRFT	RFoc	CO	FOL	
22	1249	Emse301R	RiC	CO	Ele	Phrases Reference run
22	1251	Emse303R	RiC	CO	Ele	Phrases Reference run
25	1282	ruc-2010-base2	RiC	CO	Ele	Article-only
29	1067	ISI2010_thorough.1500	Eff	CO	Ele	Article-only
29	1073	ISI2010_rric_ro	RRiC	CO	FOL	
29	1094	ISI2010_ric_ro	RiC	CO	FOL	
29	1096	ISI2010_ref_ric_aggr	RiC	CO	FOL	Reference run Invalid
29	1098	ISI2010_rfcs_ref	RFoc	CO	FOL	Reference run
55	1163	DUR10atcl	RiC	CAS	Ele	Reference run Article-only
55	1164	DURF10SIXF	RFoc	CAS	Ele	Manual
55	1169	DURR10atcl	RRiC	CAS	Ele	Reference run Article-only
60	1289	UJM_33456	RiC	CO	Ele	Reference run
62	1290	RMIT10title	RiC	CO	Ele	Article-only
62	1291	RMIT10titleO	RiC	CO	Ele	Article-only
65	1273	runRiCORef	RiC	CO	FOL	Reference run Article-only
65	1274	runReRiCORef	RRiC	CO	FOL	Reference run
65	1275	runFocCORef	RFoc	CO	FOL	Reference run
68	1170	LIP6-OWPCparentFo	RFoc	CO	Ele	
68	1181	LIP6-OWPCRefRunTh	Eff	CO	Ele	Reference run Article-only
72	1031	1	RRiC	CAS	Ele	
78	1024	UWBOOKRIC2010	RiC	CO	FOL	
78	1025	UWBOOKRRIC2010	RRiC	CO	FOL	
98	1255	I10LIA4FBas	Eff	CO	FOL	Phrases
98	1258	I10LIA1ElTri	RiC	CO	Ele	Phrases
98	1260	I10LIA1FTri	RiC	CO	FOL	Phrases
98	1270	I10LIA2FTri	RRiC	CO	FOL	Phrases
98	1284	LIAenertexTopic	RFoc	CO	FOL	Phrases
98	1285	LIAenertexDoc	RFoc	CO	FOL	Phrases

Continued on Next Page...

Group	Run	Label	Task	Query	Results	Notes
167	1049	21p167	RiC	CO	Ele	
167	1076	32p167	RRiC	CO	Ele	
167	1079	29p167	RRiC	CO	Ele	
167	1081	27p167	RRiC	CO	Ele	
167	1092	36p167	RiC	CO	Ele	
167	1219	40p167	RFoc	CO	Ele	
167	1241	18P167	Eff	CO	Ele	
167	1242	38P167	Eff	CO	Ele	
557	1313	UPFpLM45co	RiC	CO	FOL	Reference run Invalid
557	1316	UPFsecLM45co	RRiC	CO	FOL	Reference run Invalid
557	1319	UPFpLM45co	RFoc	CO	FOL	Reference run