Overview of INEX 2013

Patrice Bellot, Antoine Doucet, Shlomo Geva[1], Sairam Gurajada, Jaap Kamps[2], Gabriella Kazai, Marijn Koolen, Arunav Mishra, Véronique Moriceau, Josiane Mothe, Michael Preminger, Eric SanJuan, Ralf Schenkel[3], Xavier Tannier, Martin Theobald, Matthew Trappett, and Qiuyue Wang

[1] INEX co-chair & QUT, Australia
[2] INEX co-chair & University of Amsterdam, The Netherlands
[3] INEX co-chair & University of Passau, Germany

Abstract. INEX investigates focused retrieval from structured documents by providing large test collections of structured documents, uniform evaluation measures, and a forum for organizations to compare their results. This paper reports on the INEX 2013 evaluation campaign, which consisted of four activities addressing three themes: *searching professional and user generated data* (Social Book Search track); *searching structured or semantic data* (Linked Data track); and *focused retrieval* (Snippet Retrieval and Tweet Contextualization tracks). INEX 2013 was an exciting year for INEX in which we consolidated the collaboration with (other activities in) CLEF and for the second time ran our workshop as part of the CLEF labs in order to facilitate knowledge transfer between the evaluation forums. This paper gives an overview of all the INEX 2013 tracks, their aims and task, the built test-collections, and gives an initial analysis of the results.

1 Introduction

Traditional IR focuses on pure text retrieval over "bags of words" but the use of structure—such as document structure, semantic metadata, entities, or genre/topical structure—is of increasing importance on the Web and in professional search. INEX has been pioneering the use of structure for focused retrieval since 2002, by providing large test collections of structured documents, uniform evaluation measures, and a forum for organizations to compare their results.

INEX 2013 was an exciting year for INEX in which we joined forces with CLEF and ran our workshop as part of the CLEF labs in order to foster further collaboration and facilitate knowledge transfer between the evaluation forums. In total four research tracks were included, which studied different aspects of focused information access:

Social Book Search Track investigating techniques to support users in searching and navigating collections of digitised or digital books, metadata and complementary social media. The *Social Book Search Task* studies the relative value of authoritative metadata and user-generated content using a test collection with data from Amazon and LibraryThing. The *Prove It Task* asks

for pages confirming or refuting a factual statement, using a corpus of the full texts of 50k digitized books.

Linked Data Track investigating retrieval over a strongly structured collection of documents based on DBpedia and Wikipedia. The *Ad Hoc Search Task* has informational requests to be answered by the entities in DBpedia/Wikipedia. The *Jeopardy Task* asks for the (manual) formulation of effective SPARQL queries with additional keyword filters, aiming to express natural language search cues more effectively.

Tweet Contextualization Track investigating tweet contextualization, helping a user to understand a tweet by providing him with a short background summary generated from relevant Wikipedia passages aggregated into a coherent summary.

Snippet Retrieval Track investigate how to generate informative snippets for search results. Such snippets should provide sufficient information to allow the user to determine the relevance of each document, without needing to view the document itself.

Both Tweet Contextualization and Snippet retrieval use the same XML'ified corpus of Wikipedia, and address focused retrieval in the form of constructing some concise selection of information in a form that is of interest to NLP researchers (tweet contextualization) and to IR researchers (snippet retrieval).

In the rest of this paper, we discuss the aims and results of the INEX 2013 tracks in relatively self-contained sections: the Social Books Search track (Section 2), the Linked Data track (Section 3), and the paired Tweet Contextualization (Section 4) and Snippet Retrieval (Section 5) tracks.

2 Social Book Search Track

In this section, we will briefly discuss the INEX 2013 Social Book Search Track (addressing the searching professional and user generated data theme). Further details are in [7].

2.1 Aims and Tasks

Prompted by the availability of large collections of digitized books, the Social Book Search Track aims to promote research into techniques for supporting users in searching, navigating and reading full texts of digitized books and associated metadata. This year, the track ran two tasks: the Social Book Search task and the Prove It task:

1. The *Social Book Search* (SBS) task, framed within the scenario of a user searching a large online book catalogue for a given topic of interest, aims at exploring techniques to deal with both complex information needs of searchers—which go beyond topical relevance and can include aspects such as genre, recency, engagement, interestingness, quality and how well-written a book is—and heterogeneous information sources including user profiles, personal catalogues, professional metadata and user-generated content.

2. The *Prove It* (PI) task aims to test focused retrieval approaches on collections of books, where users expect to be pointed directly at relevant book parts that may help to confirm or refute a factual claim;

In addition to these task, the *Structure Extraction* (SE) task runs at ICDAR in 2013 [3] and aims at evaluating automatic techniques for deriving structure from OCR and building hyperlinked table of contents. The extracted structure could then be used to aid navigation inside the books.

2.2 Test Collections

For the Social Book Search task a new type of test collection has been developed. Unlike traditional collections of topics and topical relevance judgements, the task is based on rich, real-world information needs from the LibraryThing (LT) discussion forums and user profiles. The collection consists of 2.8 million book descriptions from Amazon, including user reviews, and is enriched with user-generated content from LT. For the information needs we used the LT discussion forums. We selected 386 discussion threads which focus on members asking for book recommendations on a certain topic. The initial messages in these threads often contain detailed descriptions of what they are looking for. The relevance judgements come in the form of suggestions from other LT members in the same discussion thread. We paid trained annotators to indicate for each book suggestion in the thread whether the person suggesting the book has read it and whether they are positive, neutral or negative to it. These opinions are used to derive relevance values for the books. Opinions from the topic creator are the most important, then those of others who have read the book and finally those of members who have not. The final set of judgements contain suggestions for 380 topics with an average of 16 judgements per topic. The judgements are independent of the submitted runs, which avoids pooling bias. Previously we investigated the reliability of using forum suggestions for evaluation and found they are complete enough, but different in nature from editorial judgements based on topical relevance [6].

The PI task builds on a collection of over 50,000 digitised out-of-copyright books (about 17 million pages) of different genre (e.g., history books, text books, reference works, novels and poetry) marked up in XML. The task was first run in 2010 and was kept the same for 2011 and 2012. This year the aim is to evaluate book-pages not only on whether they contain information confirming or refuting a statement, but also whether the book is authoritative and of an appropriate genre and subject matter such that a reader would trust the confirming or refuting information.

The SE task relies on a subset of the 50,000 digitized books of the PI task. In 2013, the participants were to extract the tables of contents of 1,000 books extracted from the whole PI book collection. In previous years, the ground truth was constructed collaboratively by participating institutions. For the first time in 2013, the ground truth production was performed by an external provider, and partly funded by the Seventh Framework Program (FP7) of the EU Commission.

This centralized construction granted better consistency. In addition, it also validated the collaborative process used since 2009, as the results this year were in line with those of the previous rounds.

2.3 Results

Eight teams together submitted 33 runs to the SBS task and two teams submitted 12 runs to the Prove It! task. The *Social Book Search* task evaluation has shown that the most effective systems use all available book information—professional metadata and user-generated content— and incorporate either the full topic statement, which includes the title of the topic thread, the name of the discussion group, the full first message that elaborates on the request and the query generate by annotators, or a combination of the title and the query. None of the groups used user profile information for the runs they submitted. The best performing run is *run3.all-plus-query.all-doc-fields* by **RSLIS**, which used all topic fields combined against an index containing all available document fields. The second best group is **UAms (ILLC)** with run *inex13SBS.tiqu.bayes_avg.LT_rating*, which uses only the topic titles and moderated query ran against an index containing the title information fields (title, author, edition, publisher, year), user-generated content fields (tags, reviews and awards) and the subject headings and Dewey decimal classification titles from the British Library and Library of Congress. The retrieval score of each book was then multiplied by a prior probability based on the Bayesian average of LT ratings for that book. The third group is **ISMD**, with manual run *run_ss_bsqstw_stop_words_free....* This run is generated after removing Book Search Query Stop Words (bsqstw), standard stopwords and the member field from the topics and running against an index where stopwords are removed and the remaining terms are stemmed with the Krovetz stemmer. If we ignore the manual runs, ismd is still the third group with the fully automatic run ism run ss free text 2013, which is generated using free text queries on Krovetz stemmed and stopwords removed index.

For the *Prove It* task, we expect to have relevance judgments from Mechanical Turk with book appropriateness and evaluation results in time for the INEX proceedings. Evaluation results with relevance judgments for the statements split into their atomic aspects indicate that performance increases when matching named entities (persons and locations) from the statements with named entities in the pages.

The Structure Extraction task is conjoint with ICDAR and therefore ran a bit earlier than the other tasks, with a run submission deadline in May. A total of 9 organizations signed up, 6 of which submitted runs. This increase in active participants is probably a direct result of both 1) the availability of training data and 2) the removal of the requirement for participating organizations to create part of the ground truth. This round of the competition further provided rejoicing results, as for the first time since the competition started, one organization has beaten the baseline BookML format provided by MDCS (Microsoft Development Center Serbia) in 2008. The University of Innsbruck indeed performed best in terms of link-based evaluation.

2.4 Outlook

Next year, we continue with the SBS task to further investigate the role of user information. We plan to run an additional pilot task for which we have a few options. One option is to investigate how we can use the interactivity in the forum thread to simulate interactive sessions. Another is to extend the original task by requiring systems to not only determine which book ISBNs to return, but also what information about those books to return. Book descriptions contain a mixture of professional metadata, user tags and up to 100 user reviews. A new challenge could be to determine which tags and reviews are relevant to the user in determining whether she wants to read a book or not.

The Prove It task attracted no new participants in the last two years and will not continue next year. We are considering a new task centered around entity recognition, such as identifying and mapping characters in novels.

The structure extraction task has reached a record high number of active participants, and has for the first time witnessed an improvement of the state the art. In future years, we aim to investigate the usability of the extracted ToCs, both for readers in navigating books and systems that index and search parts of books. To be able to build even larger evaluation sets, we hope to experiment with crowdsourcing methods. This may offer a natural solution to the evaluation challenge posed by the massive data sets handled in digitized libraries

3 Linked Data Track

In this section, we will briefly discuss the INEX 2013 Linked Data Track (addressing the searching structured or semantic data theme). Further details are in [4].

3.1 Aims and Tasks

The goal of the Linked Data track was to investigate retrieval techniques over a combination of textual and highly structured data, where RDF properties carry additional key information about semantic relations among data objects that cannot be captured by keywords alone. We intend to investigate if and how structural information could be exploited to improve ad-hoc retrieval performance, and how it could be used in combination with structured queries to help users navigate or explore large result sets via Ad-hoc queries, or to address Jeopardy-style natural-language queries which are translated into a SPARQL-based query format. The Linked Data track thus aims to close the gap between IR-style keyword search and Semantic-Web-style reasoning techniques. Our goal is to bring together different communities and to foster research at the intersection of Information Retrieval, Databases, and the Semantic Web.

For INEX 2013, we explored two different retrieval tasks that continue from INEX 2012:

- The classic Ad-hoc Retrieval task investigates informational queries to be answered mainly by the textual contents of the Wikipedia articles.

– The Jeopardy task employs natural-language Jeopardy clues which are manually translated into a semi-structured query format based on SPARQL with keyword conditions.

3.2 Test Collection

The Linked Data track used a subset of DBpedia 3.8 and YAGO2s together with a recent dump of Wikipedia core articles (dump of June 1st, 2012). Valid results are entities occurring in both Wikipedia and DBpedia (and hence in YAGO), hence we provided a complete list of valid URIs to the participants. In addition to these reference collections, we will also provide two supplementary collections: 1) to lower the participation threshold for participants with IR engines, a fusion of XML'ified Wikipedia articles with RDF properties from both DBpedia and YAGO2s, and 2) to lower the participation threshold for participants with RDF engines, a dump of the textual content of Wikipedia articles in RDF. Participants are explicitly encouraged to make use of more RDF facts available from DBpedia and YAGO2s, in particular for processing the reasoning-related Jeopardy topics.

The goal of the Ad-hoc Task is to return a ranked list of results in response to a search topic that is formulated as a keyword query. Results had to be represented by their Wikipedia page ID's, which in turn had to be linked to the set of valid DBpedia URI's. A set of 144 Ad-hoc task search topics for the INEX 2013 Linked Data track had been released in March 2013 and was made available for download from the Linked Data Track homepage. In addition, the set of QRels from the 2012 Ad-Hoc Task topics was provided for training.

These are familiar IR topics, an example is:

```
<topic id="2009002">
  <title>best movie</title>
  <description>information of classical movies</description>
  <narrative>
    I spend most of my free time seeing movies. Recently, I want to
    retrospect some classical movies. Therefore, I need information about
    the awarded movies or movies with good reputation. Any information,
    such as the description or comments of the awarded movies on famous
    filmfests or movies with good fame, is in demand.
  </narrative>
</topic>
```

As in 2012, the Jeopardy task continued to investigate retrieval techniques over a set of natural-language Jeopardy clues, which were manually translated into SPARQL query patterns with additional keyword-based filter conditions. A set of 105 Jeopardy task search topics, out of which 74 topics were taken over from 2012 and 31 topics were newly added to the 2013 setting. 72 single-entity topics (with one query variable) were also included into the set of 144 Ad-hoc topics. All topics were made available for download in March 2013 from the Linked Data Track homepage. In analogy to the Ad-hoc Task, the set of topics from 2012 was provided together with their QRels for training. An example topic is:

```
<topic id="2013301" category="Falls">
  <jeopardy_clue>
    This river's 350-foot drop at the Zambia-Zimbabwe border creates this
    water falls.
  </jeopardy_clue>
  <keyword_title>
    river's 350-foot drop Zambia-Zimbabwe Victoria Falls
  </keyword_title>
  <sparql_ft>
    SELECT DISTINCT ?x ?o WHERE {
      ?x <http://dbpedia.org/property/watercourse>
      ?o . FILTER FTContains (?x, "Victoria Falls") .
      FILTER FTContains (?o, "river water course Victoria 350-foot drop
        Zimbabwe") .
    }
  </sparql_ft>
</topic>
```

3.3 Results

In total, 4 ad-hoc search runs were submitted by 3 participants and 2 valid Jeopardy! runs were submitted by 1 participant. Assessments for the Ad-hoc Task were done on Amazon Mechanical Turk by pooling the top-100 ranks from the 4 submitted runs in a round-robin fashion. Conversely, the top-10 results were pooled from the 3 Jeopardy submissions for the single-entity topics and by pooling the top-20 for the multi-entity topics, respectively, again in a round-robin fashion. A total of 72 Ad-hoc topics and 77 Jeopardy topics were assessed.

The TREC-eval tool was adapted to calculate the following well-known metrics (see [1, 5]) used in ad-hoc and entity ranking settings: Precision, Recall, Average-Precision (AP), Mean-Average-Precision (MAP), Mean-Reciprocal-Rank (MRR), and Normalized-Discounted-Cumulated-Gain (NDCG). The best scoring submission for the ad hoc task was *ruc-all-2200-paragraph-80* by the **Renmin University of China (RUC)** with a MAiP of 0.3880. The best scoring submission for the Jeopardy task was *MPIUltimatum_Phrase* by the **Max-Planck Institute for Informatics (MPI)** with a MRR of 0.7671.

Given the low number of submissions, it is difficult to draw general conclusions from the runs, but individual participants found various interesting results demonstrating the value of the build test collection for research in this important emerging area. We hope and expect that the test collection will be (re)used by researchers for future experiments in this active area of research.

3.4 Outlook

The Linked Data Track was organized towards our goal to close the gap between IR-style keyword search and Semantic-Web-style reasoning techniques. The track thus continues one of the earliest guiding themes of INEX, namely to investigate whether structure may help to improve the results of ah-hoc keyword search. A

key contribution is the introduction of a new and much larger supplementary XML collection, coined *Wikipedia-LOD v2.0*, with XML-ified Wikipedia articles which were additionally annotated with RDF properties from both DBpedia 3.8 and YAGO2. However, due to the very low number of participating groups, in particular for the Jeopardy, detailed comparisons of the underlying ranking and evaluation techniques can only be drawn very cautiously.

4 Tweet Contextualization Track

In this section, we will briefly discuss the INEX 2013 Tweet Contextualization Track (one of the two tracks addressing the focused retrieval theme). Further details are in [2].

4.1 Aims and Tasks

Twitter is increasingly used for on-line client and audience fishing, this motivated the proposal of a new track addressing tweet contextualization. The objective of this task is to help a user to understand a tweet by providing him with a short summary (500 words). This summary should be built automatically using local resources like the Wikipedia and generated by extracting relevant passages and aggregating them into a coherent summary. The task is evaluated considering informativeness which is computed using a variant Kullback-Leibler divergence and passage pooling. Meanwhile effective readability in context of summaries is checked using binary questionnaires on small samples of results. Running since 2010 as a complex QA track at INEX, the results showed that only systems that efficiently combine passage retrieval, sentence segmentation and scoring, named entity recognition, text POS analysis, anaphora detection, diversity content measure as well as sentence reordering are effective.

4.2 Test Collection

The document collection has been built based on a recent dump of the English Wikipedia from November 2011. This date is anterior to all selected topics. Since we target a plain XML corpus for an easy extraction of plain text answers, we removed all notes and bibliographic references that are difficult to handle and kept only non empty Wikipedia pages (pages having at least one section). Resulting documents consist of a title (`title`), an abstract (`a`) and sections (`s`). Each section has a sub-title (`h`). Abstract and sections are made of paragraphs (`p`) and each paragraph can contain entities (`t`) that refer to other Wikipedia pages.

In 2012, topics were made of 53 tweets from New York Times (NYT). In 2013, the task was enriched and evaluated topics were made of 120 tweets manually collected by organizers. These tweets were selected and checked, in order to make sure that:

– They contained "informative content" (in particular, no purely personal messages); Only non-personal accounts were considered (*i.e.* @CNN, @TennisTweets, @PeopleMag, @science. . .).
– The document collection from Wikipedia contained related content, so that a contextualization was possible.

From the same set of accounts, more than 1,800 tweets were then collected automatically. These tweets were added to the evaluation set, in order to avoid that fully manual, or not robust enough systems could achieve the task. All tweets were then to be processed by participants, but only the 120 short list was used for evaluation. Participants did not know which topics were selected for evaluation. These tweets were provided in a text-only format without metadata and in a JSON format with all associated metadata.

4.3 Measures

Tweet contextualization is evaluated on both informativeness and readability. Informativeness aims at measuring how well the summary explains the tweet or how well the summary helps a user to understand the tweet content. On the other hand, readability aims at measuring how clear and easy to understand the summary is. Informativeness measure is based on lexical overlap between a pool of relevant passages (RPs) and participant summaries. Once the pool of RPs is constituted, the process is automatic and can be applied to unofficial runs. The release of these pools is one of the main contributions of Tweet Contextualization tracks at INEX [8]. By contrast, readability is evaluated manually and cannot be reproduced on unofficial runs. In this evaluation the assessor indicates where he misses the point of the answers because of highly incoherent grammatical structures, unsolved anaphora, or redundant passages.

Three metrics were used: **Relevancy (or Relaxed) metric**, counting passages where the T box has not been checked (*Trash* box if the passage does not make any sense in the context of the previous passages); **Syntax**, counting passages where the S box was not checked either (i.e, the passage has no syntactic problems), and the **Structure (or Strict) metric** counting passages where no box was checked at all. In all cases, participant runs were ranked according to the average, normalized number of words in valid passages.

4.4 Results

A total number of 13 teams from 9 countries (Brasil, Canada, France, India, Ireland, Mexico, Russia, Spain, USA) submitted runs to the Tweet Contextualization track in 2013. This year, the best participating system *256* from **Université de Nantes** used hashtag preprocessing. The best run by this participant used all available tweet features including web links which was not allowed by organizers. However their second best run *258* without using linked web pages is ranked first among official runs. Second best participant on informativeness was run *275* from **IRIT, Toulouse** which score best in readability and used state

of the art NLP tools. Third best participant was run *254* from **University of Minnesota Duluth** was first in 2012, suggesting that their system performs well on a more diversify set of tweets in 2013.

All participants but two used language models, however informativeness of runs that only used passage retrieval is under 5%. Terminology extraction and reformulation applied to tweets was also used in 2011 and 2012. Appropriate stemming and robust parsing of both tweets and wikipedia pages are an important issue. All systems having a run among the top five in informativeness used the Stanford Core NLP tool or the TreeTagger. Automatic readability evaluation and anaphora detection helps improving readability scores, but also informativeness density in summaries. State of the art summarization methods based on sentence scoring proved to be helpful on this task. Best runs on both measures used them. Best run in 2013 also experimented a tweet tag scoring technique while generating the summary. Finally, this time the state-of-the-art system proposed by organizers since 2011 combining LM indexation, terminology graph extraction and summarization based on shallow parsing was not ranked among the ten best runs which shows that participant systems improved on this task over the three editions.

4.5 Outlook

The discussion on next year's track is only starting, and there are links to related activities in other CLEF labs that need to be further explored. The use case and the topic selection should remain stable in 2014 TC Track, so that 2013 topics can be used as a training set. Nevertheless, we will consider more diverse types of tweets, so that participants could better measure the impact of hashtag processing on their approaches.

5 Snippet Retrieval Track

In this section, we will briefly discuss the INEX 2013 Snippet Retrieval Track (one of the tracks addressing the focused retrieval theme). Further details are in [9].

5.1 Aims and Task

The goal of the snippet retrieval track is to determine how to generate informative snippets for search results. Such snippets should provide sufficient information to allow the user to determine the relevance of each document, without needing to view the document itself, allowing the user to quickly find what they are looking for.

The task was to return a ranked list of documents for the requested topic to the user, and with each document, a corresponding text snippet describing the document. Each run had to return 20 documents per topic, with a maximum of 180 characters per snippet. The snippets may be created in any way – they may consist of summaries, passages from the document, or any other text at all.

5.2 Collection

The Snippet Retrieval Track uses the exact same collection as the Tweet Contextualization track—an XML version of the English Wikipedia, based on a dump taken on November 2012. Since the task is to generate snippets for the documents given in the reference run, a link to an archive containing only those 700 documents (as well as the reference run submission file itself) was provided.

There were 35 topics in total—10 taken from the INEX 2010 Ad Hoc Track, and 25 created specifically for this track, with the goal being to create topics requesting more specific information than is likely to be found in the first few paragraphs of a document. Each topic contains a short content only (CO) query, a phrase title, a one line description of the search request, and a narrative with a detailed explanation of the information need, the context and motivation of the information need, and a description of what makes a document relevant or not.

5.3 Assessment and Evaluation

To determine the effectiveness of the returned snippets at their goal of allowing a user to determine the relevance of the underlying document, manual assessment is being used. Both snippet-based and document-based assessment are being used. The documents will first be assessed for relevance based on the snippets alone, as the goal is to determine the snippet's ability to provide sufficient information about the document. The documents will then be assessed for relevance based on the full document text, with evaluation based on comparing these two sets of assessments.

We created snippet assessment packages (the size of a single submission) to assess, each participating organization will receive as many packages as they have submitted runs. For each topic, the assessor will read through the details of the topic, after which they will read through each snippet, and determine whether or not the underlying document is relevant to the topic. This is expected to take around 1-2 hours per package. Ideally, each package should be assessed by a different person if feasible. Additionally, it will be required to perform one assessment of the document assessment package. For each of the 35 topics, the assessor is shown the full text of each of the 20 documents. They must read through enough of the document to determine whether or not it is relevant to the topic. This is expected to take around 3-7 hours, depending on the assessor.

Submissions are evaluated by comparing the snippet-based relevance judgements with the document-based relevance judgements, which are treated as a ground truth. The primary evaluation metric used is the geometric mean of recall and negative recall (GM). A high value of GM requires a high value in recall and negative recall—i.e., the snippets must help the user to accurately predict both relevant and irrelevant documents. If a submission has high recall but zero negative recall (e.g. in the case that everything is judged relevant), GM will be zero. Likewise, if a submission has high negative recall but zero recall (e.g. in the case that everything is judged irrelevant), GM will be zero. Details of additional metrics used are given in [9].

5.4 Results

As of this writing, only preliminary results are available. The best scoring system is *snippets_2013_knapsack* of **IRIT, Toulouse**, with a GM score of 0.5352. The second scoring run is *QUT_2013_Focused* of **Queensland University of Technology (QUT)** with a GM score of 0.4774. Further discussion of the results will be available in [9].

5.5 Outlook

We have discussed the setup of the track, and presented the preliminary results of the track. The preliminary results show that in all submitted runs, poor snippets are causing users to miss over half of all relevant results, indicating that a lot of work remains to be done in this area. Final results will be released at a later date, once further document assessment has been completed.

6 Envoi

This complete our walk-through of INEX 2013. INEX 2013 focused on three themes: *searching professional and user generated data* (Social Book Search track); *searching structured or semantic data* (Linked Data track); and *focused retrieval* (Snippet Retrieval and Tweet Contextualization tracks). The last two tracks use the same Wikipedia corpus and both address focused retrieval in the form of constructing some concise selection of information in a form that is of interest to NLP researchers (tweet contextualization) and to IR researchers (snippet retrieval) The INEX tracks cover various aspects of focused retrieval in a wide range of information retrieval tasks. This overview has only touched upon the various approaches applied to these tasks, and their effectiveness. The online proceedings of CLEF 2013 contains both the track overview papers, as well as the papers of the participating groups. The main result of INEX 2013, however, is a great number of test collections that can be used for future experiments, and the discussion amongst the participants that happens at the CLEF 2013 conference in Valencia and throughout the year on the discussion lists.

References

[1] S. Amer-Yahia and M. Lalmas. XML search: languages, INEX and scoring. *SIGMOD Record*, 35, 2006.

[2] P. Bellot, V. Moriceau, J. Mothe, E. Sanjuan, and X. Tannier. Overview of the INEX 2013 tweet contextualization track. In *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*, 2013.

[3] A. Doucet, G. Kazai, S. Colutto, and G. Muehlberger. Overview of the ICDAR 2013 Competition on Book Structure Extraction. In *Proceedings of the Twelfth International Conference on Document Analysis and Recognition (ICDAR'2013)*, Washington, USA, September 2013.

[4] S. Gurajada, J. Kamps, A. Mishra, R. Schenkel, M. Theobald, and Q. Wang. Overview of the INEX 2013 linked data track. In *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*, 2013.

[5] J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. Robertson. INEX 2007 evaluation measures. In *Focused access to XML documents (INEX 2007)*, volume 4862 of *LNCS*, pages 24–33. Springer Verlag, 2008.

[6] M. Koolen, J. Kamps, and G. Kazai. Social Book Search: The Impact of Professional and User-Generated Content on Book Suggestions. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM 2012)*. ACM, 2012.

[7] M. Koolen, G. Kazai, M. Preminger, and A. Doucet. Overview of the INEX 2013 social book search track. In *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*, 2013.

[8] E. SanJuan, P. Bellot, V. Moriceau, and X. Tannier. Overview of the inex 2010 question answering track (qa@inex). In *Comparative Evaluation of Focused Retrieval, INEX 2010*, volume 6932 of *Lecture Notes in Computer Science*, pages 269–281. Springer, 2010.

[9] M. Trappett, S. Geva, A. Trotman, F. Scholer, and M. Sanderson. Overview of the INEX 2013 snippet retrieval track. In *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*, 2013.