Report on INEX 2013

Patrice Bellot, Antoine Doucet, Shlomo Geva, Sairam Gurajada, Jaap Kamps, Gabriella Kazai, Marijn Koolen, Arunav Mishra, Véronique Moriceau, Josiane Mothe, et al.

Report on INEX 2013

P. Bellot **A. Doucet** **S. Geva** **S. Gurajada** **J. Kamps**
G. Kazai **M. Koolen** **A. Mishra** **V. Moriceau** **J. Mothe**
M. Preminger **E. SanJuan** **R. Schenkel** **X. Tannier** **M. Theobald**
M. Trappett **A. Trotman** **M. Sanderson** **F. Scholer** **Q. Wang**

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 continued as part of CLEF for the second year, motivated by the desire 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 4), the Linked Data track (Section 2), and the paired Tweet Contextualization (Section 5) and Snippet Retrieval (Section 3) tracks. We end with an outline of the plans for INEX 2014, again held as part of CLEF 2013 in Sheffield, UK.

2 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].

2.1 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.

2.2 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>
```

2.3 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 [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).

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. The track will team up with the CLEF QA track, and continue the Jeopardy task focusing on complex questions.

2.4 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.

3 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].

3.1 Aims and Task

The goal of the snippet retrieval track is to determine how best 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. This allows the user to quickly find what they are looking for.

A set of topics (or queries) was provided, each with a corresponding set of search results. Participants were then tasked with automatic generation of text snippets describing each document returned by a query. Returned snippets were limited to a maximum of 180 characters. The snippets may be created in any way—they may consist of summaries, passages from the document, or any other text at all.

3.2 Collection

The topics for the 2013 track have been reused from the 2012 Snippet Retrieval track. There were 35 topics in total — 10 originally taken from the INEX 2010 Ad Hoc Track, and 25 created specifically for the Snippet Retrieval track in 2012, 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 irrelevant.

For each topic, there is a corresponding set of twenty documents — the search results for the topics. These XML documents are based on a dump of the English language Wikipedia, from November 2012 — the same collection as the Tweet Contextualisation track.

3.3 Assessment and Evaluation

To determine the effectiveness of the returned snippets at allowing a user to determine the relevance of the underlying document, manual assessment is used. Both snippet-based and document-based assessment are used.

The documents are first 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. Each topic within a submission is assigned an assessor. The assessor, after reading the details of the topic, reads through the top 100 returned snippets, and judges which of the underlying documents seem relevant based on the snippets alone.

To avoid bias introduced by assessing the same topic more than once in a short period of time, and to ensure that each submission is assessed by the same assessors, the runs are shuffled in such a way that topics from each submission are spread evenly amongst all assessors.

Additionally, each of the 20 documents returned for each of the 35 topics is assessed for relevance based on the full document text. This full set of 700 documents is assessed multiple times, by separate assessors. The majority judgment formed by all of the document assessments is treated as a ground truth.

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].

3.4 Results

In the 2013 Snippet Retrieval track, 4 runs were submitted, from 2 participating groups. In addition, a baseline run was generated and evaluated, consisting of the first 180 characters of each document.

As of this writing, only preliminary results have been released. While each of the submissions have had their snippets assessed, the set of full-text documents has been assessed only once. The final set of results will use the majority opinion of multiple document assessors as its ground truth relevance judgments. This will be released at a later date, once further document assessment has been completed.

The initial results have all runs scoring highly on recall, but poorly on negative recall. This indicates that while users are able to easily identify most irrelevant results based on snippets alone, poor snippets are causing users to miss over half of all relevant results.

4 Social Book Search Track

In this section, we briefly discuss the INEX 2013 Social Book Search (SBS) Track. Further details can be found in [7].

4.1 Aims and Tasks

Prompted by the availability of large collections of digitized books, the SBS 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, and quality—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 ran at ICDAR 2013 [3], with the aim of 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.

4.2 Test Collections

For the SBS 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 associated user profiles of LT members. The collection consists of 2.8 million book descriptions from Amazon, including reviews, and is enriched with user-generated content from LT. For the information needs we used the LT discussion forums. We selected 380 discussion threads which focus on members asking for book recommendations on various topics. The initial messages in these threads often contain detailed descriptions of what the members are looking for. The relevance judgements come in the form of suggestions from other LT members in the same discussion thread. We hired trained annotators to indicate whether the person suggesting the book has read it and whether their recommendations had a positive, neutral or negative sentiment. From these we then derived relevance values for the suggested books, combining the topic creator's feedback on the recommendations with the opinions of those who recommended them. The final set of judgements contain an average of 16 book suggestions per topic. Note that 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 (leading to different system ranking) 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 marked up in XML. The task was first run in 2010 and was kept the same for 2011 and 2012. Since the topic statements are generally complex, relevance judgements in 2012 were collected on the atomized statements, meaning that assessors were asked to judge each aspect of each statement, from which the final judgments (weather a book page confirms or refutes the statement) were compiled. This year the aim was to evaluate book-pages not only on whether they contained 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 information within.

The SE task at ICDAR2013 used a subset of the PI task's collection and required participants to extract the tables of contents for 1,000 digitized books. 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. In previous years this was done by the task participants.

4.3 Results

Eight teams together submitted 34 runs to the SBS task and two teams submitted 12 runs to the PI task. The SBS 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. Analysis of the individual topics revealed that there is a large set of topics on which all systems fail to retrieve relevant results. This is probably due to the incompleteness of book suggestions and the vagueness and complexity of the topic statements. Next year, we will select only topics that are well-defined and have

enough discussion in the thread to signal relative completeness of the suggestions.

For the PI task, we expect to have crowdsourced relevance judgments from Mechanical Turk with book appropriateness and evaluation results in time for the INEX proceedings. Evaluation results so far 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.

A total of 9 organizations signed up to the SE task, 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 demonstrated encouraging progress, 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.

4.4 Outlook

Next year, we continue with the SBS task to further investigate the role of professional and user-generated information, but focus more on the recommendation aspect of the task. A larger set of user profiles will be incorporated to allow both content-based recommendation and collaborative filtering techniques. In addition, there will be the new Interactive SBS task. The CLEF-CHiC Lab will not continue next year, but the Interactive task of CHiC will move to the SBS Track and use the same A/LT collection. The organisers of the Interactive task feel the A/LT collection and LT forum topics offer a better use case for Interactive IR experiments dealing with heterogeneous metadata.

The PI task attracted no new participants in the last two years and will not continue next year.

The SE task will continue exclusively at ICDAR, where it has been jointly running for several years. With the discontinuation of the PI task and the introduction of the Interactive task, the focus of the track is shifting more towards the A/LT collection, leaving little connection with the SE task.

5 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].

5.1 Aims and Tasks

Text Contextualization differs from text expansion in that it aims at helping a human to understand a text rather than a system to better perform its task. For example, in the case of query expansion in IR, the idea is to add terms to the initial query that will help the system to better select the documents to be retrieved. Text Contextualization on the contrary can be viewed as a way to provide more information on the corresponding text in the objective to make it understandable and to relate this text to information that explains it.

In the context of micro-blogging, which is increasingly used for many purposes such as for on-line client and audience fishing, contextualization is specifically important since 140 characters long messages are rarely self-content. This motivated the proposal in 2011 of a new track at Clef INEX lab of Tweet Contextualization.

The use case is as follows: given a tweet, the user wants to be able to understand the tweet by reading a short textual summary; this summary should be readable on a mobile device without having to scroll too much. In addition, the user should not have to query any system and the system should use a resource freely available. More specifically, the guideline specified the summary should be 500 words long and built from sentences extracted from a dump of WikiPedia. WikiPedia has been chosen both for evaluation purpose and because this is an increasing popular resource while being generally trustable.

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.

5.2 Test Collection

The document collection has been built based on a recent dump of the English Wikipedia from November 2012. All collections are made available on the task website as well as an API to use online indexes powered by Indri. Programs to automatically generate the collections from WikiPedia repositories are also available. Generated documents from Wikipedia dump, 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.

Over 2012 and 2013 editions, evaluated topics were made of 120 (60 topics each year) 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 treated 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, an example is:

```
<topic id="303481535074549763">
  <title>007 in #SKYFALL's Floating Dragon casino. #SKYFALL IS OUT ON BLU-RAY/DVD
      TODAY IN THE UK! http://t.co/kTj5d0Ux</title>
  <txt>
    "to_user_id":0,
     "source":"&lt;a href=&quot;http://www.hootsuite.com&quot;&gt;HootSuite&lt;/a&gt;",
     "profile_image_url":
       "http://a0.twimg.com/profile_images/1620182823/Wall_Icon_normal.png",
     "profile_image_url_https":
       "https://si0.twimg.com/profile_images/1620182823/Wall_Icon_normal.png",
     "created_at":"Mon, 18 Feb 2013 12:30:11 +0000",
```

```
    "text":"007 in #SKYFALL's Floating Dragon casino. #SKYFALL IS OUT ON BLU-RAY/DVD
        TODAY IN THE UK! http://t.co/kTj5d0Ux",
    "metadata":"result_type":"recent",
    "id":303481535074549763,
    "from_user_id_str":"389229444",
    "from_user":"007",
    "geo":null,
    "from_user_id":389229444,
    "id_str":"303481535074549763",
    "iso_language_code":"en",
    "to_user_id_str":"0",
    "to_user":null,
    "from_user_name":"James Bond",
    "to_user_name":null
  </txt>
</topic>
```


5.3 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. Since 2012, readability is evaluated in the context of the tweet. Passages not related to the tweet are considered as unreadable.

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.

5.4 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. Overall, informativity and readability scores are this year strongly correlated (Kendall test: $\tau > 90\%$, $p < 10^{-3}$) which shows that all systems have integrated these two constrains.

This year, the best participating systems used advanced hashtag preprocessing algorithms. The best run on informativeness used all available tweet features, and was ranked second on readability. The best scoring run on readability used state of the art NLP tools, and was second best on informativeness. The best scoring approach of 2012 rank now third on informativeness, signaling the importance of hashtag processing for the tweets of 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.

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.

5.5 Outlook

In 2014 Tweet Contextualization will reuse data released by RepLab (CLEF e-reputaion lab). These are tweets in two languages (English and Spanish) with extra tags: target entity, opinion and priority. The user case could be the following. Given a tweet x that has been labeled as carrying a strong negative opinion about some entity y described by a WikiPedia page, "explain why tweet x could damage the e-reputation of y".

A pilot task in Spanish involving six participants just started to try to deal with humorous tweets and is under evaluation. Spanish has almost as many NLP resources than English so it is easy for a participant to deal with both languages, however evaluating readability in Spanish appears to be different due to the average sentence length and strong grammatical structure.

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.

INEX 2014 will continue the successful collaboration with CLEF, further integrating its activities into the CLEF labs organization. The Social Book Search Track will continue strong, and will focus on the interface between search and recommendation using the profiles

of LibraryThing users. the CHIC@CLEF lab will team up with INEX and promote user-centric studies into book search and exploration. INEX has a long history of "interactive" tracks in 2004–2010, and we are very excited about a return of this user-centered focus in close coupling with the system-centered evaluation tasks. The Tweet Contextualization Track will continue with the main task as well as the tweets and resources of the online reputation management (CLEF RepLab) lab. Finally, the Linked Data Track will continue with its Jeopardy task in close collaboration with the CLEF QA lab.

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.