# The Heterogeneous Collection Track at INEX 2006

Ingo Frommholz[1] and Ray Larson[2]

[1] University of Duisburg-Essen
Duisburg, Germany
ingo.frommholz@uni-due.de
[2] University of California
Berkeley, California 94720-4600
ray@sims.berkeley.edu

**Abstract.** While the primary INEX test collection is based on a single DTD, it is realistic to assume that most XML collections consist of documents from different sources. This leads to a heterogeneity of syntax, semantics and document genre. In order to cope with the challenges posed by such a diverse environment, the heterogeneous track was offered at INEX 2006. Within this track, we set up a collection consisting of several different and diverse collections. We defined retrieval tasks and identified a set of topics. These are the foundations for future run submissions, relevance assessments and proper evaluation of the proposed methods dealing with a heterogeneous collection.

## 1 Introduction

The primary INEX test collection has been based on a single DTD. In practical environments, such a restriction will hold in rare cases only. Instead, most XML collections will consist of documents from different sources, and thus with different DTDs or Schemas. In addition, distributed systems (federations or peer-to-peer systems), where each node manages a different type of collection, will need to be searched and the results combined. If there is a semantic diversity between the collections, not every collection will be suitable to satisfy the user's information need. On the other hand, querying each collection separately is expensive in terms of communication costs and result post-processing, therefore it has been suggested in the distributed IR literature that preselection of appropriate collections should be performed. Given these conditions and requirements, heterogeneous collections pose a number of challenges for XML retrieval, which is the primary motivation for including a heterogeneous track in INEX 2006.

## 2 Research Questions

Dealing with a set of heterogeneous collections that are syntactically and semantically diverse poses a number of challenges. Among these are:

N. Fuhr, M. Lalmas, and A. Trotman (Eds.): INEX 2006, LNCS 4518, pp. 312–317, 2007.

- For content-oriented queries, most current approaches use the DTD or Schema for defining elements that would form reasonable answers. In heterogeneous collections, DTD-independent methods need to be developed.
- For content-and-structure queries, there is the added problem of mapping structural conditions from one DTD or Schema onto other (possibly unknown) DTDs and Schemas. Methods from federated databases could be applied here, where schema mappings between the different DTDs are defined manually. However, for a larger number of DTDs, automatic methods must be developed, e.g. based on ontologies. One goal of an INEX track on heterogeneous collections is to set up such a test collection, and investigate the new challenges posed by its structural diversity.
- Both content-only and content-and-structure approaches should be able to preselect suitable collections. This way, retrieval costs can be minimised by neglecting collections which would probably not yield valuable answers but are expensive to query in terms of time, communication costs, and processing.

The Heterogeneous track aims to answer, among others, the following research questions:

- For content-oriented queries, what methods are possible for determining which elements contain reasonable answers? Are pure statistical methods appropriate, or are ontology-based approaches also helpful?
- For content-and-structure queries, what methods can be used to map structural criteria onto other DTDs? Should mappings focus on element names only, or also deal with element content or semantics?
- For all types of queries, how can suitable collections be preselected in order to improve retrieval efficiency and without corrupting retrieval effectiveness?
- What are appropriate evaluation criteria for heterogeneous collections?

In order to cope with above questions, we need collections which are both heterogeneous syntactically (based on different DTDs) and semantically (dealing with different topics, in this case from computer science research to IT business to non-IT related issues). As in the previous years, the main focus of effort for the track was on the construction of an appropriate testbed, consisting of different single collections, and on appropriate tools for evaluation of heterogeneous retrieval. The testbed provides a basis for the exploration of the research questions outlined above.

## 3 Testbed Creation

In order to create a testbed for heterogeneous retrieval, we had to find suitable collections first. Subsequently, corresponding topics had to be found and relevance assessments to be performed.

### 3.1 Collection Creation

We reused several subcollections offered in the last years' and the current INEX runs. Most of the collections from previous years of the Heterogeneous track were restructured so that each document was a separate file embedded within a new directory structure, in order to be able to use the normal INEX evaluation tools. Additionally, we downloaded and prepared new collections like ZDNet News (IT related articles and discussion) and IDEAlliance. A specific DTD was defined for every subcollection, if not already given, ensuring syntactic heterogeneity. Table 1 shows some statistics about the subcollections.

**Table 1.** Components of the heterogeneous collection. *Element counts estimated for large collections.*

| Collection | Size | SubColl. | Documents | Elements | Mean Elements per Document |
|---|---|---|---|---|---|
| Berkeley | 52M | | 12800 | 1182062 | 92.3 |
| bibdb Duisburg | 14M | | 3465 | 36652 | 10.6 |
| CompuScience | 993M | | 250986 | 6803978 | 27.1 |
| DBLP | 2.0G | | 501102 | 4509918 | 9.0 |
| hcibib | 107M | | 26390 | 282112 | 10.7 |
| IEEE (2.2) | 764M | | 16820 | 11394362 | 677.4 |
| IDEAlliance | 58M | eml | 156 | 66591 | 426.9 |
| | | xml1 | 301 | 45559 | 151.4 |
| | | xml2 | 264 | 58367 | 221.1 |
| | | xmle | 193 | 32901 | 170.5 |
| | | xtech | 71 | 14183 | 199.8 |
| Lonely Planet | 16M | | 462 | 203270 | 440.0 |
| qmulcsdbpub | 8.8M | | 2024 | 23435 | 11.6 |
| Wikipedia | 4.9G | | 659385 | 1193488685 | 1810.0 |
| ZDNet | 339M | Articles | 4704 | 242753 | 51.6 |
| | | Comments | 91590 | 1433429 | 15.7 |
| **Totals** | 9.25G | | 1570713 | 1219818257 | 776.6 |

The subcollections serve different domains, ranging from computer science (e.g. bibdb Duisburg, IEEE, DBLP) through technology news (ZDNet) to travel advice (Lonely Planet) and general purpose information (Wikipedia). We find several document genres like metadata records, fulltexts of scientific papers, articles and web sites as well as textual annotations which form discussion threads. The bottom line is that we have subcollections which differ with respect to their syntax (DTD), semantic (domains served) and document genre.

### 3.2 Topic Creation

The topic creation phase resulted in 61 topics. Among these are selected topics of the adhoc track as well as 36 new topics created especially for the heterogeneous

track. In order to develop the latter topics we used the Chesire II[1] system. Converting the subcollections into a unified internal representation turned out to be a very time-consuming task as new collections had to be incorporated.

Appendix A shows the topic DTD used. Besides keywords and titles, also content-and-structure titles (`<castitle>` in our DTD) were given for most topics in order to make them suitable for the CAS tasks. Content-and-structure titles do not only contain keywords, but also a NEXI [1] path expression for the desired structural elements. For instance, the `castitle` expression

```
//article[about(.,user interface)]//section[about(.,design)]
```

requests sections about design in articles about user interfaces. Whenever given, the scope of a topic provides a hint about the collection used to identify the topic (which in fact does not necessarily mean that a topic is not relevant for other subcollections as well).

## 4 Tasks and Run Submissions

The following tasks were proposed for this year's heterogeneous track:

**Adhoc CO Task.** Here, content-oriented queries are implied. The systems return a ranked list of documents from all collections.

**CAS Task 1.** The system should return only elements specified in `<castitle>`.

**CAS Task 2.** The system should basically return the elements specified in `<castitle>` , but also similar elements. As an example, `<doctitle>` in ZDNet and `<title>` in other collections are most probably equivalent. The `<description>` in ZDNet, which is the description grabbed from RSS feeds, is similar, but not equivalent, to the `<about>` tag elsewhere. A possible scenario for both CAS tasks would be a system which likes to present the user only the title and a representative summary of the content so that she could decide if a document is relevant or not without higher cognitive overload (the need for reading the whole article). The system should thus return only the title and the summary of relevant documents, but might base the relevance decision on, e.g., the whole document fulltext.

**Resource Selection.** The goal here is to select the most relevant resources (i.e., collections) for a given topic. The system should return a ranked list of collections for this task. The scenario is that a system should identify relevant collections beforehand and query them, instead of querying all resources (which might be expensive when it comes to communication or access costs).

For run submissions we defined a DTD which can viewed in Appendix B. This DTD covers rankings of elements as well as rankings of subcollections. Note that this DTD allows those submitting runs to specify the collections actually used in resolving the topics. Thus it permits users to submit runs for only a subset of the collections, and in principle such runs could be scored without counting the ignored collections.

[1] `http://cheshire.berkeley.edu/`

## 5 Pooling and Assessment

Having set up the heterogeneous collection with tasks and topics, next steps include the actual submission of runs. The once submitted runs are the basis for a *pooling* procedure to extract the set of relevant elements for each query and task. This step can also provide us with with new insights whether the pooling procedure can be applied to a heterogeneous scenario or if there is the need for suitable adaptations.

We plan to use the XRai system for relevance assessments based on the pooled elements. Part of the motivation in the restructuring of the collections so that each record or document was a separate file was to be able to use XRai.

## 6 Conclusion and Outlook

In this year's heterogeneous track, we managed to set up a collection whose subcollections have heterogeneous syntax, semantics and document genres. We also set up a number of test topics for evaluation. We have, therefore, laid the foundations for a new heterogeneous track which may now concentrate on submitting runs, creating a pooled result set and providing relevance assessments, and these in turn will be used to evaluate the submitted runs.

## Acknowledgement

Special thanks go to Miro Lehtonen for providing us with the IDEAlliance collection, and to Saadia Malik and Benjamin Piwowarski for technical support.

## Reference

1. Trotman, A., Sigurbjornsson, B.: Narrowed extended XPath I (NEXI). In: Fuhr, N., Lalmas, M., Malik, S., Szlavik, Z. (eds.) Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004, Revised Selected Papers, vol. 3493. Springer-Verlag GmbH, (May 2005) http://www.springeronline.com/3-540-26166-4

## A Topic DTD

```
<?xml version="1.0" encoding="UTF-8"?>
<!ENTITY   lt      "&#38;#60;">
<!ENTITY   gt      "&#62;">
<!ENTITY   amp     "&#38;#38;">

<!ELEMENT inex_het_topic
  (title,castitle?,description,narrative,ontopic_keywords,scope?)>
```

```
<!ATTLIST inex_het_topic
  topic_id     CDATA     #REQUIRED
>

<!ELEMENT title (#PCDATA)>
<!ELEMENT castitle     (#PCDATA)>
<!ELEMENT description   (#PCDATA)>
<!ELEMENT narrative     (#PCDATA)>
<!ELEMENT ontopic_keywords     (#PCDATA)>
<!ELEMENT scope (collection+) >
<!ELEMENT collection (#PCDATA) >
```

## B Run Submission DTD

```
<!ENTITY % collection-ids "berkeley | bibdbpub | compuscience | dblp | hcibib |
                        qmulcsdbpub | ieee | zdnetart | zdnetcom | wikipedia |
                        lp | idea_eml | idea_xml1 | idea_xml2 | idea_xmle |
                        idea_xtech">
<!ELEMENT inex-het-submission (topic-fields, description, collections, topic+)>
<!ATTLIST inex-het-submission
  participant-id CDATA #REQUIRED
  run-id CDATA #REQUIRED
  task (CO | CAS1 | CAS2 | RS) #REQUIRED
  query (automatic | manual) #REQUIRED
>
<!ELEMENT topic-fields EMPTY>
<!ATTLIST topic-fields
  title (yes|no) #REQUIRED
  castitle (yes|no) #REQUIRED
  description (yes|no) #REQUIRED
  narrative (yes|no) #REQUIRED
  ontopic_keywords (yes|no) #REQUIRED
>
<!ELEMENT description (#PCDATA)>
<!ELEMENT topic (result*|collections)>
<!ATTLIST topic topic-id CDATA #REQUIRED >
<!ELEMENT collections (collection*)>
<!ELEMENT collection (rank?, rsv? )>
<!ATTLIST collection collectionid (%collection-ids;) #REQUIRED>
<!ELEMENT result (file, path, rank?, rsv?)>
<!ATTLIST result collectionid (%collection-ids;) #IMPLIED>
<!ELEMENT file (#PCDATA)>
<!ELEMENT path (#PCDATA)>
<!ELEMENT rank (#PCDATA)>
<!ELEMENT rsv (#PCDATA)>
```