

# Toward meaningful test collections for information integration benchmarking

Peter Bailey  
CSIRO ICT Centre  
GPO Box 664  
Canberra ACT 2601 AUSTRALIA  
+61 2 6216 7055

Peter.Bailey@csiro.au

David Hawking  
CSIRO ICT Centre  
GPO Box 664  
Canberra ACT 2601 AUSTRALIA  
+61 2 6216 7060

David.Hawking@csiro.au

Alexander Krumpholz  
CSIRO ICT Centre  
GPO Box 664  
Canberra ACT 2601 AUSTRALIA  
+61 2 6216 7735

Alexander.Krumpholz@csiro.au

## ABSTRACT

Meaningful comparison of algorithms for search, extraction and aggregation across heterogeneous sources will require well-designed benchmarks, preferably based on freely available test collections. In this paper, we discuss issues which will inevitably arise during the construction of such benchmarks. We argue that the creation of benchmarks requires careful consideration of the evaluation methods. These go hand-in-hand with the selection of an appropriate test collection. Conversely, selection of inappropriate tasks or test collections can lead to false or incomplete conclusions about the effectiveness of a retrieval system which addresses information integration challenges. We draw on our experience with the development of collections for the evaluation of information retrieval systems focused on Web data and tasks. Our particular interest is in effective retrieval over heterogeneous data sources.

## Categories and Subject Descriptors

H.1.2 [User/Machine Systems] – *human information processing*.  
H.3.3 [Information search and retrieval] – *retrieval models, search process*. H.3.4 [Systems and Software] – *performance evaluation (efficiency and effectiveness)*.

## General Terms

Measurement, Performance, Design, Experimentation.

## Keywords

Test collections, evaluation methods, user models.

## 1. INTRODUCTION

We suggest that the primary purpose of information integration is ultimately to enable users (or other agents) to discover, search, navigate, and manipulate the information as if it was entirely, or at least reasonably, consistent. To enable this, many hard problems have to be solved, all of which present their own challenges.

The Information Retrieval group at CSIRO has had many years of interest and activity in the development of effective collections for carrying out comparative experiments in the context of large scale and Web-oriented data. These activities have included the development of a number of collections (e.g. VLC2, WT2g, WT10g [2]), which have been used in various TREC tracks [8]

(<http://trec.nist.org>) and by research organizations.

Collections are understood in the information retrieval community to consist of three components: a corpus of data, a set of tasks, and a set of judgements of relevant answers for each task. The judgements enable the researcher to assess how well their system performs over the aggregate of tasks for the particular corpus. To conduct experiments, and to draw statistically valid conclusions, a non-trivial number of tasks are required.

We note that the term “collection”, or occasionally “test collection”, appears to be used more loosely in the arena of information integration research, and often synonymously with a dataset; what would be referred to as a corpus in information retrieval terminology. We try to use the more specific information retrieval definition of collection throughout this paper.

The typical context in which these collections were initially used was TREC ad hoc retrieval style tasks. A progressive reformulation of the tasks to reflect more accurately the activities of typical Web search engine users emerged over the years of the TREC Web tracks. In parallel, understanding grew that the relevance metrics that were appropriate for TREC ad hoc were not necessarily appropriate in the Web context. For example, it is typical in web search that some relevant documents are of far greater utility to the searcher than others. Consider when a searcher is trying to navigate to the Intel site by typing the query “intel” in a web search engine, approximately 170 million pages are at least faintly relevant, but one (the site entry page) is far more useful than all the others and effectiveness of this search can be measured by the reciprocal rank of its appearance in the results list. On homepage finding and named page finding tasks, mean reciprocal rank of the desired answer is a measure which more accurately reflects utility to the searcher than e.g. precision at 10 or average precision based on the full set of relevant documents.

Throughout the development of these collections, and follow-on collections used at TREC in the new Terabyte track, including the .GOV and .GOV2 collections, two concerns have typically been stressed: that the collections are of large scale and that they are broadly representative of general properties of real Web data [14]. In doing so, tradeoffs were made in trying to avoid properties that were perceived not to contribute to the requirements of the retrieval tasks. For example, in the WT10g collection, efforts were made to remove sites containing only a single document, even though many sites do exist with this property. The likelihood of such sites returning relevant results for any query is low, and therefore the value of having these sites represented is small, even if representative of the underlying data (and the Web at large).

New benchmarks will be needed for search tools operating over integrated heterogeneous environments. There is also a great opportunity to apply quality evaluation methodology to aggregation and extraction algorithms and systems. Our particular interest is in effective retrieval over heterogeneous data sources.

## 2. RELATED WORK

A small number of information integration test collections have been created to date. An impetus for these came from a meeting of database researchers in 2003, producing the Lowell Report, which specifically called for the creation of a test bed and defining a set of integration tasks [6]. The report acknowledges the difficulty of obtaining representative data, with a suggestion to use computer science course descriptions from a number of universities.

This suggestion was taken up in THALIA, a publicly available test bed and benchmark, which uses course catalogs from 40 computer science departments at universities around the world as the corpus [7]. The THALIA data has been extracted to produce valid and well-formed XML, with corresponding schemas. The benchmark activities consist of 12 query tasks, split into 3 groups – attribute heterogeneities, missing data, and structural heterogeneities. The essence of the tasks is to be able to provide answers to information needs when the underlying data is not represented uniformly. In each task, the particular integration problems differ. For example, one task is to list all instructors for courses on software systems. One underlying schema may represent multiple instructor information as a single attribute “instructors”, while another represents it by multiple “section-instructor” attribute pairs. The benchmark tasks appear to be aimed towards addressing systems which can address information extraction problems, primarily by the creation of an integrated data schema. The individual tasks have been synthetically derived to require resolution of each heterogeneity problem, and they make use of two data sources (course catalogs) per benchmark query. The strength of this approach is that there are known answers for each benchmark query, and thus systems can be easily judged as to their success in creating an integrated schema that can be queried to provide the correct answers from the data.

A different effort has been underway at the University of Illinois, Urbana-Champaign, with the UIUC Web Integration Repository [1]. A series of datasets have been collected, from as early as at least 2002. At the current time, 4 of the 5 datasets consist of query interfaces (or query interface schemas) to web sites in various domains (such as real estate, travel, jobs, etc). The exception is the OntoBuilder dataset, which is a collection of more than 100 ontologies spanning 14 domains, represented by DTDs. These datasets, in comparison to those of THALIA, are thus effectively at a meta-level. In addition, certainly with the query interface datasets, there has been no stated effort to clean up the underlying data to be well-formed – they are “real” in the sense they have been extracted from live web sites. There are two kinds of integration tasks that have been associated with these datasets – schema matching and query capabilities extraction. The schema matching task entails attempting to discover matching attributes across the sets of query interfaces, thereby enabling query mediation in otherwise heterogenous information sources, thought most likely in the same domain. The second task, query capability extraction, attempts to understand how an information source may be queried based on the query interface representation. Solving this problem is part of the challenge in providing access to “deep

web” data sources, particularly when considered as distributed information retrieval.

In 2004, the Heterogeneous track at INEX [11] used the existing INEX collection (based on IEEE journal proceedings), plus a set of computer science publication and bibliography databases. In 2006, it added additional data sources from other INEX tracks, attempting to broaden the scope from the computer science only focus to provide more realistic mixed data. In a review of the heterogeneous track goals for 2005 [13], Larson wrote in conclusion that:

“You can either have heterogeneous retrieval, or precise element specifications in queries, but you cannot have both simultaneously.”

The Heterogeneous track in 2006 at INEX is using a corpus built from conference proceedings provided by the IDEAlliance. Obviously the corpus consists of XML data, since the INEX conference series is solely focused on XML retrieval. The conference proceedings provide text data from different DTDs. The purpose of the track is to consider issues relating to how to best handle content-oriented retrieval tasks, how to map structural criteria from one DTD onto others, what kind of mappings to perform, retrieval efficiency versus effectiveness tradeoffs, and evaluation criteria.

The primary focus of all the INEX Heterogeneous track collections is on information retrieval, and is thus most closely aligned to our interests. That said, to do so, they have needed to address fundamental information integration issues, such as schema mapping. The THALIA testbed primarily supports investigations into integrating data heterogeneities. The UIUC Web Integration Repository primarily supports investigations into extracting or inducing data schemas from corresponding query interfaces, and also query capability induction.

## 3. MOTIVATING EXAMPLE

As a practical example, consider the following application. A person wishes to search for information relating to a planned vacation. They are interested in obtaining: cheap airfares from Italy to Mexico in August, car hire and hotel accommodation for 2 in Mexico and California in August/September, and general information about visa requirements for Europeans into Mexico and the USA.

A series of information needs corresponding to these different aspects of the combined travel plans can be formulated. Individual queries for each need can be constructed; some of these queries may be structured (e.g. From: Italy To: Mexico By: airline Date: August) (and converted to an appropriate structured query language such as XQuery) and some may be free text (e.g. Mexican visa requirements).

A large scale and heterogeneous corpus is required. The UIUC Web Integration Repository could provide the basis for this – a collection of query interfaces spanning a number of domains. In addition, the underlying databases that these query interfaces expose need to be available, and corresponding query processing systems that respond to each query interface. As a baseline, an SQL database could be assumed. Finally, additional unstructured data from the Web could be included, such as that found in the .GOV collection. Such a corpus obviously consists of more than static data, but is analogous to the testbeds created for conducting distributed information retrieval research experiments, such as reported in [9].

An information retrieval system that hopes to address all of these information needs successfully must be able to solve various interesting information integration problems. For example: how to select appropriate servers to answer each query and to map the user query to each server's query interface; how to determine the query type and the retrieval approach most appropriate to it.

Individual information integration aspects may also be tackled independently using the same collection.

## 4. COLLECTION CONSIDERATIONS

In considering the development of useful collections and benchmarks for semi-structured data, we believe that the lessons we have learned in unstructured Web data retrieval experiments apply even more strongly. These lessons may be considered along the three axes of: collection design, selection of representative tasks, and evaluation methods. Making matters more complex, these tend to be inter-dependent.

### 4.1 Selecting properties of a collection

Any collection represents a specific set of choices made in the properties by which data has been included or excluded. In the Web collections we have created, and in collections developed for information integration experiments (for example, the UIUC Web Integration Repository [1]), one of their characteristics is that they represent an aggregation of multiple data sources. In Web collections, these arise from individual Web servers; the unifying property is that all data is HTML, but the quality of the HTML and specific version of the (X)HTML standard used has always differed markedly among and even within servers. In collections such as the one we have used as a motivating example, one of the purposes for their existence is to explore integration across multiple data formats, and we can imagine a wide variety of data sources.

Nevertheless, it may make sense to restrict data to be represented in an XML format [3], albeit with multiple schemas or DTDs, simply to make the parsing aspect more uniform and accessible, given the wide availability of XML parsing toolkits.

Collections with highly varied data are valuable since they provide opportunities for investigation across a number of interesting axes and may reflect real-world data. But they are more problematic for a number of reasons: they are difficult to process effectively by a range of research retrieval systems without extensive engineering; data properties may be mostly inappropriate for individual experiments; and they tend to be either too large in scale to be readily processed or too small to contain sufficient numbers of relevant documents.

Collections which are developed with specific properties are valuable since they support targeted experiments more appropriately and can facilitate the investigation of important problems without having to deal with as many data engineering anomalies. However they suffer from: the expense of creating a specific collection is similar to that of a general collection; and they may cause researchers to pursue solutions to problems that do not exist in the world at large.

For the purposes of retrieval experiments in the context of information integration, we believe that creating a corpus from rich and varied (but well-formed) datasets would have great value. This would more accurately reflect the challenges faced for retrieval systems working with Web-accessible data, while balancing the engineering effort to deal with data quality, which

we discuss in the next section. Such a corpus could be used for the kind of motivating example we discussed.

### 4.2 Engineering to cope with data messiness

A frequent challenge observed in any collection that has been created from real-world data is dealing with poor data quality. For example, in processing the Web collections, we progressively dealt with issues such as duplicate documents, invalid HTML, URLs that were split over line breaks, and so on. A substantial amount of work in modern browser technology entails coping with invalid HTML and representing it meaningfully nevertheless.

Semi-structured data, particularly if represented as XML, surely should be more manageable. By definition, XML documents are only XML if they are well formed as defined in [3], and preferably they are also valid (that is, complying with a particular schema for the document). In practice, this is not always the case. For example, in carrying out experiments with the INEX collection from 2004 [11], we found that the breakdown of the document collection into folder structures and files, with volume files declaring individual article files as external entities, led to an inability to open the individual article files, since they included entities declared only in the volume file. However, current XML tools do not automatically open external entities, so it is not possible to open the volume file either. This is not to criticise the INEX organisers; merely to remark that to participate in research activities with this collection, basic engineering issues have to be solved.

As collections grow in size and/or in the number of subsidiary data sources, the number and variety of potential errors also increases. As a research organisation, spending time solving these problems provides limited or no value, unless the research goal is to handle messy data. Wherever possible, the creator of the collection should pre-process the data to eliminate these errors as much as possible, and maximise data integrity. This issue demonstrates the interplay between collection design and the selection of the benchmark – if you provide messy data, you will force people to address that problem, even if it is not the problem you would like them to be addressing.

Commercial XML data integration systems almost certainly have to solve these problems. As a research issue however, we believe that there are more fundamental problems to be addressing.

### 4.3 Engineering to cope with scale

A common challenge posed by the Web collections addressed in the TREC Web tracks was one of sheer scale. For example, in 1998, the VLC2 collection was 100GB in size. In itself, large scale collections are not a bad thing to use, as they force the system designers to implement efficient algorithms and address issues such as data compression.

The second challenge of large scale collections is that classic approaches to retrieval effectiveness evaluation which rely on being exhaustive with respect to judgements may simply not apply [15]. It is not possible for human judges in a sufficiently short timeframe to read all possible documents that are presented, or all possible variations of the integration of information, when the numbers run into the thousands or tens of thousands. This issue demonstrates the interplay between collection design and evaluation methods.

At present, none of the information integration datasets referred to in the related work section deal with scale on this level. Yet one of the main motives posited for addressing information integration tasks is that the scale of the information in the “deep web” vastly exceeds that on the “visible” web [12]. A corpus to support the kind of motivating example we gave will need to be substantially larger than existing testbeds.

#### 4.4 Availability of data

Another issue which affects the selection of a collection is the availability of data. For the development of a Web collection, public Web data is more than sufficient, and the only issue of substance is actually crawling it and some legal issues surrounding copyright and privacy. For the development of a large scale collection of semi-structured data sources over which to apply information integration, the obvious candidates are large enterprises which have this kind of data in abundance. However, large enterprises are naturally very wary of sharing their private data. Indicative of this problem, the TREC Enterprise track has made use of public W3C websites, which include mailing lists. While this represents one form of enterprise, there are relatively few organisations in the world which carry out their activities in such an open forum. Thus it is arguable how applicable the results for systems addressing this collection would be if transferred to more typical enterprise data sets.

Similarly, the INEX and THALIA datasets have used mostly public computer science publications and bibliography databases. Do the results achieved over these datasets extend to other domains such as biology or social sciences? Do they apply in contexts completely outside of publications and bibliographies? The challenge for us is to obtain not just the query interfaces from web sites such as those in the UIUC Web Integration Repository datasets, but also the underlying databases as well.

### 5. BENCHMARK CONSIDERATIONS

#### 5.1 Modelling a user task

Many retrieval experiments attempt to capture a benchmark of interest by modelling a user’s information need, and then represent solving this need by a task description. The task description incorporates some system that interacts between the user (possibly an automated agent) and the test collection to address the information need. For experiments with large scale collections, typically a large number of queries are posed (50 or more are common in most TREC tracks which involve automated agents), to allow the application system to produce sufficient results to be analysed for statistical significance.

The main issue here is to identify meaningful user tasks that can be represented and then subsequently evaluated once the system has processed them. For example, in our motivating example, should the results for the travel need be judged in order of cheapest airfares returned that meet the travel itinerary or ones which minimise the travel time? Once again, this demonstrates the interplay between benchmark selection and the evaluation method and the test collection.

#### 5.2 Lessons learned from TREC Web tracks

One of the lessons learned from running the Web tracks at TREC is that sometimes it is easier to build a collection or three than it is to understand that the focus of the application is wrong or at least

the mapping of the application to the task description does not reflect what you wanted to explore, *and* how to fix it. The problem is that you may only discover the application is not understood properly until several groups have run the experiments and evaluation has taken place, possibly over several years. In parallel, it may also be discovered that the selection properties of the collection have led to a test collection that is not amenable to supporting the application effectively.

Reading through the TREC Web track summaries from 1999 [10] to 2004 [5] (the final year of the TREC Web track) illuminates how the benchmarks changed, reflecting a growing understanding of how user information needs over Web data are different from classic TREC ad hoc retrieval information needs. In TREC-8, standard ad hoc topic descriptions were used, and standard ad hoc evaluation methods were used. By TREC-2004, three user tasks were considered – homepage finding, named page finding, and topic distillation. These represented a far more realistic model of how users carry out retrieval activities over Web data.

### 6. EVALUATION CONSIDERATIONS

#### 6.1 One size does not fit all

Unsurprisingly, there is no one evaluation method that meets all information retrieval experimentation needs [4]. Nor is there one evaluation metric which is appropriate in all circumstances. New approaches may be required such as that suggested by Thomas [16], particularly when contrasting systems that provide different result presentations as can be expected with retrieval experiments involving information integration.

#### 6.2 Lessons learned from TREC

Our experience with the Web tracks and the lessons learned in the TREC community more generally indicate that the choice of evaluation metrics may evolve over time in accordance with changes in application focus. Similarly, even if an evaluation metric remains unchanged, for example the question answering track at TREC has used the MRR of the first correct answer since inception, the degree of latitude allowed in how that answer is displayed or justified may change [17].

Hard problems arise as the nature of the test collections in use change. For example, with the development of TREC ad hoc collections as large as a few hundred megabytes in size, pooling methods were used to identify potentially relevant documents. Subsequently, the pooling methodology was assessed and found to be a valid one, for collections of that scale. When collections grow yet further, to of the order of 100 million documents, pooling methods and other measures such as recall are simply impractical to apply, and new evaluation methods must be developed [15].

### 7. CONCLUSIONS

In conclusion, we believe that developing effective test collections for retrieval experiments involving information integration over heterogeneous datasets will be challenging, and may take place over a period of years if our experiences with the TREC Web tracks are a good indicator.

Although seemingly counter-intuitive, it could speed up the process to spend more time understanding the benchmarks of interest, and appropriate evaluation methods, rather than focusing on the identification of a suitable (or just available) test collection.

In particular, the application benchmarks may dramatically affect the properties of the collection to be developed. Time should also be spent maximising data integrity once it has been selected.

A final recommendation is to consider iterative prototyping of the benchmark-evaluation-collection development process. This may prove impractical from a logistical perspective, but more rapid feedback than provided by yearly cycles could prove extremely beneficial, in a manner similar to the development of software.

## 8. REFERENCES

- [1] The UIUC web integration repository. Computer Science Department, University of Illinois at Urbana-Champaign. <http://metaquerier.cs.uiuc.edu/repository>. (2003)
- [2] Bailey, P., Craswell, N., and Hawking, D. 2003. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management* 39, 6 (November).
- [3] Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F. 2004. Extensible markup language (XML) 1.0 (third edition). W3C recommendation, W3C.
- [4] Craswell, N., Bailey, P., and Hawking, D. 1999. Is it fair to evaluate web systems using TREC ad hoc methods? In *ACM SIGIR '99 Workshop on Web Retrieval* (1999).
- [5] Craswell, N. and Hawking, D. 2004. Overview of the TREC-2004 web track. In *Proc. TREC 2004* (2004).
- [6] Gray, J., Schek, H., Stonebraker, M., and Ullman, J. The Lowell Report. In *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*. (New York, USA, 2003), p. 680.
- [7] Hammer, J., Stonebraker, M., and Topsakal, O. Thalia: Test harness for the assessment of legacy information integration approaches. In *21<sup>st</sup> International Conference on Data Engineering (ICDE'05)*. (April 2005), pp. 485-486.
- [8] Hawking, D. and Craswell, N. 2005. Very Large Scale Retrieval and Web Search. In *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press. (2005)
- [9] Hawking, D. and Thomas, P. Server selection methods in hybrid portal search. In *Proceedings of ACM SIGIR 2005*. pp. 75-82. Salvador, Brazil (2005).
- [10] Hawking, D., Voorhees, E., Craswell, N., and Bailey, P. Overview of TREC-8 web track. In *Proc. TREC-9* (1998).
- [11] Initiative for the Evaluation of XML Retrieval.. INEX collection. <http://inex.is.informatik.uni-duisburg.de/>.
- [12] Larson, R. Distributed IR for digital libraries. In *Research and Advanced Technology for Digital Libraries (ECDL 2003)*, pp. 487-498. Springer (LNCS #2769) (2003).
- [13] Larson, R. XML element retrieval and heterogeneous retrieval: In pursuit of the impossible? In *Proceedings of the INEX 2005. Workshop on Element Retrieval Methodology, Second Edition*, (pp. 43-46) (2005).
- [14] Soboroff, I. Do TREC web collections look like the web? *SIGIR Forum* 36, 2 (September 2002).
- [15] Soboroff, I., Voorhees, E., and Craswell, N. Summary of the SIGIR 2003 workshop on evaluation methodologies for terabyte-scale test collections. *SIGIR Forum* 37, 2, pp. 55-58 (September 2003)
- [16] Thomas, P. Personal Information Retrieval. Poster at *HCSNet*. Sydney, Australia (December 2005)
- [17] Voorhees, E. Question answering in TREC. In *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press. (2005)