Project no.507618

DELOS

A Network of Excellence on Digital Libraries

Instrument: Network of Excellence

Thematic Priority: IST-2002-2.3.1.12
Technology-enhanced Learning and Access to Cultural Heritage

D 7.4.2: Progress Report I on DL evaluation

Due date of deliverable: 31/12/2005
Actual submission date: 25/1/2006

Start Date of Project: 01 January 2004
Duration: 48 Months

University of Duisburg-Essen, Germany

Project co-funded by the European Commission within the Sixth Framework Programme
(2002-2006)

Executive Summary

This report summarizes the results of the DELOS Evaluation Cluster (WP7) activities in the second half of 2005. During this period, three of the WP7 tasks were active:

- T7.3: INEX – Initiative for the Evaluation of XML Retrieval
- T7.4: CLEF – Cross-language Evaluation Forum
- T7.5: A Digital Library Testbed Framework for the Evaluation of Architectures, Services and Execution Dynamics

Here we first give a brief survey of the two evaluation initiatives and then summarize the results of T7.5

T7.3: INEX – Initiative for the Evaluation of XML Retrieval

The Initiative for the Evaluation of XML Retrieval (Task 7.3) supports research in content-oriented access to structured documents, where XML is used as standard document format. INEX provides an infrastructure for testing XML retrieval methods and a discussion forum for researchers and developers working in this area. The major results achieved by INEX during this reporting period are the following:

- stimulation of research activity in new, previously unexplored areas, such as interactive aspects of XML information access and multimedia XML IR,
- study and implementation of evaluation methodologies for diverse types of XML access systems
- novel evaluation methodology where user aspects can be formalised in the evaluation measures (XCG metrics)
- development of test collections and evaluation methodologies for XML access system evaluation
- building of a strong, multidisciplinary research community).

T7.4: CLEF – Cross-language Evaluation Forum

The Cross Language Evaluation Forum (Task 7.4) promotes research in multilingual information access and retrieval. It does this by providing an infrastructure for multilingual and cross-language system testing and evaluation and a forum where researchers and system developers can discuss and compare different approaches and share their experiences. The results achieved by CLEF in this reporting period can be summarised in the following main points:

- creation of a new technical infrastructure managed by the DIRECT system (U.Padua) which provides tools and an easy to use Web interface for relevance assessment, run submission and validation.
- study and implementation of evaluation methodologies for diverse types of cross-language IR systems
- stimulation of research activity in new, previously unexplored areas, such as cross-language question answering, image and geographic information retrieval
- documented improvement in system performance for cross-language text retrieval systems
- creation of important, reusable test collections for system benchmarking
- building of a strong, multidisciplinary research community
- production of a large number of publications reporting research and evaluation activities in this field.

T7.5: A Digital Library Testbed Framework for the Evaluation of Architectures, Services and Execution Dynamics

This task is working on the development of a logging standard and a testbed framework; in order to demonstrate these concepts, specific evaluations are carried out.

The goal of the logging standard development is to specify a standard for a normal and comparative evaluation of digital libraries. The following results were achieved here:

- A first draft for a general log event classification model was prepared; it is planned to submit the first version to an international conference in early 2006.
- For the Daffodil system that will form the basis of the testbed framework to be developed, we verified and implemented new and missing user- and system-triggered DL events, and the corresponding logging facilities were integrated into Daffodil.
- A log schema converter for the Daffodil logs is under development; this converter will allow for exploitation of the large volumes of Daffodil log data that are already available, by converting this data into the new standard logging format.

In order to set up a standard testbed framework for comparative evaluations, we are working on a theoretical framework and extensions of the Daffodil framework.

For the theoretical framework, a whitepaper on digital library evaluation has been completed. This paper gives an overview on the state of the art in DL evaluation and describes the merging of two evaluation frameworks, namely the Evaluation Computer and Interaction Triptych Framework; in addition, the paper gives recommendations for further work in this field. This paper has been submitted for publication in a journal.

On the practical side, the development of a layered help function for the Daffodil system has started, and several new services were developed and integrated into this system:

- A multi-level hypertext browser allows for clustering and browsing of search results (joint work with DELOS task 4.7).
- An annotation tool was integrated that allows for both *out-of-line* and *in-line* annotation of documents, including sharing and searching of annotations and supporting of discussion threads. This part will be extended and evaluated along with DELOS task 4.10 (DILAS).
- For supporting synchronous collaboration, a chat facility and a whiteboard tool with several functions for discussion and topic structuring has been integrated.

Concerning specific evaluations, we were involved in two activities:

- As joint work with the INEX task, the baseline system for the INEX interactive track was developed, which required the implementation of new services for retrieving and browsing of XML documents; a large volume of logging data was collected here.
- We started the preparation of a comparative evaluation between "The European Library" and a Daffodil-based system. This evaluation will follow both an analytical and an empirical approach.

Table of Contents

1 Introduction

Digital libraries (DL) need to be evaluated to determine how useful, usable, and economical they are, and whether they achieve reasonable cost-benefit ratios. Results of evaluation studies can provide strategic guidance for the design and deployment of future systems, can assist in determining whether digital libraries address the appropriate social, cultural, and economic problems, and whether they are as maintainable as possible.

In order to promote research in this area, the DELOS WP7 (Evaluation) currently supports three tasks. T7.3 (INEX) and T7.4 (CLEF) are evaluation initiatives, each with more than 50 participating groups from all over the world (most of them from outside of DELOS); they both focus on specific problems of information access to digital libraries: while INEX deals with structured documents, CLEF addresses multilingual and cross-language systems. In contrast to these two initiatives, T7.5 (Digital Library Testbed Framework) is a more standard task, which is working on the development of a logging standard and a testbed framework.

Due to the different nature of the three tasks, the tasks proceeded in parallel. For this reason, we describe each task separately in the following.

2 INEX: An evaluation campaign for XML information access in digital libraries

2.1 Introduction

The DELOS task 7.3 is concerned with the evaluation of content-oriented access to XML documents, where XML stands for "extensible Markup Language". XML is increasingly being used in digital libraries and similar systems or platforms (e.g. XML is becoming the W3C standard for representing documents). The provision of effective access to XML-based content has become a key research issue, and is the focal point of XML retrieval research. XML retrieval systems aim to exploit the logical structure of documents to retrieve document components, the so-called XML elements, instead of whole documents in response to a user's query. According to this retrieval paradigm, an XML retrieval system needs not only to find relevant information in the XML documents, but also to determine the appropriate level of component granularity to return to the user. Evaluating how good these systems are, hence, requires test-beds where the evaluation paradigms are provided according to criteria that take into account the imposed structural aspects.

In 2002, the Initiative for the Evaluation of XML Retrieval (INEX[1]) started to address these issues. INEX has a strong international character; participants from over 50 organisations, distributed across Europe, North America, Australia, New Zealand and Asia, have participated in this year's fourth INEX run. The aim of the INEX initiative is to establish an infrastructure and to provide means, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of content-oriented XML retrieval systems.

2.2 Main Results

The results achieved by INEX are summarised in the following main points, and discussed in details in the next sub-sections.

- stimulation of research activity in new, previously unexplored areas, such as interactive aspects of XML information access and multimedia XML IR,
- study and implementation of evaluation methodologies for diverse types of XML access systems
- novel evaluation methodology where user aspects can be formalised in the evaluation measures (XCG metrics)
- development of test collections and evaluation methodologies for XML access system evaluation
- building of a strong, multidisciplinary research community.

2.2.1 Expanding research directions

In total seven research tracks were included in INEX 2005, that studied different aspects of XML information access: Ad-hoc, Interactive, Multimedia, Relevance Feedback, Heterogeneous, Document Mining and Natural Language (NLP). Two (multimedia and document mining) are new for the 2005 campaign; the rest reached their second year. The interactive track expanded in the numbers of tasks offered and in the number of participating groups; the track tries to answer some fundamental questions of XML IR. The heterogeneous track expanded by studying new collections with different DTDs and their effect on XML IR system effectiveness. The relevance feedback track investigated approaches for queries that also include structural hints (rather than content-only queries in 2004). The NLP track included a new task in 2005 that allows new participants with NLP expertise to join the INEX workshop without the need to develop a search engine, and thus encouraged wider accessibility. The

---

[1] http://inex.is.informatik.uni-duisburg.de/

consolidation of the existing tracks, and the expansion to new areas offered by the two new tracks, allows the INEX initiative to grow in reach.

Regarding the two new tracks:

- The aim of the document mining track, done in collaboration with the PASCAL network of Excellence (http://www.pascal-network.org/), is to develop machine learning methods for structured data mining and to evaluate these methods for XML document mining tasks. The track in 2005 focused on classification and clustering for XML documents.
- The main objective of the INEX 2005 multimedia track was to provide an evaluation platform/forum for structured document access systems that do not only include text in the retrieval process, but also other types of media, such as images, speech, and video.

2.2.2 Construction of evaluation methodologies and infrastructure

In INEX 2005 the main effort on developing an evaluation infrastructure and methodology for various aspects of XML information access systems continued with significant results in the following areas.
.
Test Collections

Test collections consist of sets of documents, relevance assessments and topics (or queries)

Document Collections

- The INEX corpus until 2004 was composed of the full-texts, marked up in XML, of 12,107 articles of the IEEE Computer Society's publications from 12 magazines and 6 transactions, covering the period of 1995-2002, and totalling 494 megabytes in size. The collection has a suitably complex XML structure (192 different content models in DTD) and contains scientific articles of varying length. On average an article contains 1,532 XML nodes, where the average depth of a node is 6.9. An addition of 4,712 articles totalling 241 megabytes in size from 2002-2004 have been added to the collection, thus reaching a total of **16,819 articles**.
- A new corpus was acquired for the multimedia track in 2005. This corpus was made available by the **Lonely Planet** organization. The Lonely Planet collection consists of 462 XML documents with information about destinations that is particularly useful for travellers that want to find interesting details for their next holiday or business trip. This particular collection is referred to as the "WorldGuide" and can be viewed online at: http://www.lonelyplanet.com/worldguide.
- As part of the heterogeneous track the following document collections have been made available:
  - Berkeley (Library catalog entries for CS literature): 12,800 XML items
  - CompuScience (Bibliographic entries from the Computer Science database of FIZ Karlsruhe): 250,987 XML items.
  - bibdbpub (BibTeX converted to XML by the IS group at Univ. of Duisburg-Essen): 3465 XML items.
  - dblp (Bibliographic entries from the Digital Bibliography & Library Project in Trier): 501,101 XML items.
  - hcibib (Human-Computer Interaction Resources, bibliography from www.hcibib.org): 26,402 XML items.
  - qmuldcsdpub (Publications database of QMUL Department of Computer Science): 2024 XML items.
  - ZDNet (Articles and Comments) provided by ZDNet.com to the INEX evaluation: 96,351 items (4734 Articles and 91,617 comments on those articles). This sub-collection was added in 2005.
- For the document mining track, 2 new collections were developed: The **WIPO** corpus is composed of 75,250 XML documents, and the **MovieDB** corpus (based on the Internet Movie Database) consists of 9463 XML documents.

Topics/Queries
- A total of 139 topics were submitted by 51 INEX participants for the ad-hoc track. Of these, 40 CO+S and 27 CAS (see below) were selected for use in the evaluation campaign. To date, INEX has a total of 288 topics, most of them with associated relevance assessments.
- For the multimedia track 25 new topics were created that capture information needs for multimedia items using both content and structural constraints.
- For the heterogeneous track, the topics were selected from the ad hoc track, thus minimizing effort.

Evaluation Methodology and Measures

Since its launch in 2002, INEX has been challenged by the issue of how to measure an XML information access system's effectiveness. In 2005, INEX adopted a new set of metrics, the **eXtended Cumulated Gain (XCG) metrics** to support the evaluation of XML access systems. These new metrics aim to provide an evaluation framework that allows to consider the dependency that exists among XML document components and, in particular, incorporate mechanisms to reward the retrieval of so-called near-misses and to address issues of overlap. With these measures, both system and user-oriented evaluation aspects are considered and both recall and precision-like qualities are measured. **User-oriented measures** allow to reason about a system's ability to satisfy users, and typically focus on the early ranks of a system's output as users are more likely to limit their search to these results. System-oriented measures allow system developers to obtain an overall picture of performance. The metrics have been implemented in a Java package, EvalJ, which has been made available to all INEX participants as **open source code** on sourceforge.net.

The main retrieval task to be performed in INEX was defined as the ad-hoc retrieval of XML documents. In information retrieval literature, ad-hoc retrieval is described as a simulation of how a library might be used, and it involves the searching of a static set of documents using a new set of topics. While the principle is the same, the difference for INEX is that the library consists of XML documents, the queries may contain both content and structural conditions and, in response to a query, arbitrary XML elements may be retrieved from the library. Within the main ad-hoc retrieval task, three sub-tasks have been identified depending on how structural constraints are expressed in queries.

1. In the Content-Only (**CO**) sub-task, queries ignore the document structure and contain only content-related conditions.
2. An extension of the CO sub-task that includes structural hints is the **+S** sub-task, where a user may decide to add structural hints to his query to narrow down the number of returned documents resulting from a CO query.
3. In the Content and Structure (**CAS**) sub-task, structural constraints are explicitly stated in the query and they can refer both to where to look for the relevant elements (i.e. support elements), and what type of elements to return (i.e. target elements). A structural constraint can also be interpreted as strict (i.e. the structural requirements must be followed strictly) or vague (i.e. the structural constraints are interpreted as hints and the main goal is to satisfy the overall information need). Strict and vague interpretations can be applied to both support and target elements, giving a total of four strategies for the CAS subtask.

With regards to evaluation methodology for the ad-hoc track, depending on how we assume that a user would want the output of an XML retrieval system to be, **three different strategies** are defined and used. In a *focussed* strategy, we assume that a user prefers a single element that most exhaustively discusses the topic of the query (most exhaustive element), while at the same time it is most specific only to that topic (most specific element). In a *thorough* strategy, we assume that a user prefers all highly exhaustive and specific elements, and in a *fetch and browse* strategy we assume that a user is interested in highly exhaustive and specific elements that are contained only within highly relevant articles.

Evaluation Infrastructure

To facilitate the laborious process of assessing the relevance of XML components to topics, an interface has been made available to participants. The **relevance assessment interface** allows assessors to view the pooled set of results, to browse the INEX document collection and to record the

relevance assessments. The interface is available as **open source code** at https://developer.berlios.de/projects/x-rai/. In 2005, a new assessment procedure is employed by the interface, by using interactive highlighting of relevant document elements by the assessor for one dimension of relevance as defined by INEX (relevance in INEX is two dimensional) and interactive querying by the system to obtain relevance values for the other dimension. With the extensive use of this on-line tool, large amounts of data about **user behaviour** when assessing the relevance of XML components has been collected, and will be analysed after the INEX workshop.

For the **interactive track**, a version of the Daffodil system was made available to INEX participants (http://www.is.informatik.uni-duisburg.de/projects/daffodil/). This version of Daffodil was enhanced to manage the INEX data collection, to enable the logging of all user actions with the interface, and to facilitate the experimental procedures used in the interactive track (e.g. automatic recording of experimental conditions). The logging facilities used correspond to the Logging Standard developed by UNIDE as part of DELOS Task 7.5. The interactive system was used by 76 test persons recruited by 11 institutions world wide.

The **multimedia track** has provided an evaluation platform for structured document retrieval systems that do not only include text in the retrieval process. The track acquired and exploited the Lonely Planet collection, created search topics for the track, and provided an experimental system (see http://contentlab.cs.uu.nl/) for studies. This system was also used in one of the tasks of the Interactive track in INEX 2005.

2.2.3 Community building and Research Integration activities

The following Institutions have contributed to the organisation of the different tracks of INEX 2005. These institutions all worked on a voluntary basis, without any DELOS funding, finding local funding to cover their INEX participation and activities:

1. Centre for Mathematics and Computer Science, CWI, Amsterdam, the Netherlands
2. Informatics Institute, University of Amsterdam, the Netherlands
3. Centre for Web Research, University of Chile, Chile
4. Department of Computer Science, University of Otago, New Zealand
5. Royal School of Library and Information Science, Copenhagen, Denmark
6. Language Technologies Institute, School of Computer Science, Carnegie Mellon University, USA
7. Queensland University of Technology, Australia
8. Laboratoire d'Informatique de Paris 6, LIP6 Paris, France
9. University of Minnesota-Duluth, USA
10. IBM research Labs, Haifa, Israel
11. Center for Content and Knowledge Engineering, Utrecht University
12. School of Information Management & Systems, University of California Berkeley, USA
13. INRIA-Rocquencourt, France
14. University of Twente, the Netherlands
15. University of Helsinki, Finland
16. RMIT, Melbourne, Australia

The discussion lists during the campaign and the workshop at the end give the different groups the opportunity to come together exchanging and sharing ideas, experiences, tools and methodologies. INEX puts a strong emphasis on resource building and sharing. Many groups that originally met through INEX have continued to work together and collaborate. The result is a strong, well-connected and enthusiastic community.

There have been more than 40 publications on methodologies and approaches using the INEX test bed in 2004 (this number was 13 in 2003). We do not have final number for publication up to 2005, but we estimate this to be in the order of 80.

2.3 Events and Meetings

The main event was the INEX 2005 Workshop, 28-30 November, Schloss Dagstuhl, Germany; this was attended by approximately 50 participants from over 25 institutions – see http://inex.is.informatik.uni-duisburg.de. A second event (INEX 2005 Workshop on Element Retrieval Methodology – see http://www.cs.otago.ac.nz/inexmw/) was organised as part of the Information Retrieval festival held and hosted by the University of Glasgow in July 2005. About 30 people participated in the event, including leading non-INEX researchers in evaluation and structured document retrieval (Prof van Rijsbergen from the University of Glasgow, Prof David Hawkins and Prof Ross Wilkinson both from CSIRO, Australia).

2.4 Synergies and Dissemination

INEX activities are widely known and disseminated in the relevant international research communities (in particular information retrieval and information science). With the addition of new tracks, INEX activities are also becoming known in new communities (NLP, DB and Data Mining, Interactivity, Multimedia), leading to a high profile and awareness of DELOS activities in all these areas.

In addition, the data mining track in 2005 is a joint effort between INEX and the PASCAL network of excellence (http://www.pascal-network.org/). A second round of evaluation for the track is being organized to run between January and March 2006 with a presentation of the participants' results at a later workshop. A joint INEX-Pascal workshop is currently planned for 2006.

2.5 Publications

The main forum for publication of INEX activities is the INEX 2004 Proceedings of revised selected papers:

N. Fuhr, M. Lalmas, S. Malik and Z. Szlavik (eds). Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2004), Springer, 2005.

Note that formal proceedings for INEX 2005 will also be published by LNCS, Springer, in 2006.

Publications related to methodological issues:

A. Trotman, M. Lalmas, N. Fuhr. Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Information Retrieval Festival, University of Glasgow, 30 July 2005.

A. Trotman, M. Lalmas. Report on the INEX 2005 Workshop on Element Retrieval Methodology. ACM SIGIR Forum, 39(2), December 2005.

G. Kazai and M. Lalmas. Notes on what to measure in INEX, Proceedings of the INEX 2005 Workshop on

Element Retrieval Methodology, Glasgow, July 2005.

N. Fuhr and M. Lalmas (eds). Special issue on INEX, Special Issue of Information Retrieval, 8(4), December 2005.

M. Lalmas. INEX: Evaluating XML Retrieval Effectiveness. ERCIM News N0 63, October 2005.

M. Lalmas. INEX: Evaluating Content-Oriented XML Retrieval Effectiveness. Feature Article of the Informer, BCS IRSG News Letter. Number 16, Autumn 2005.

A. Tombros, S. Malik, B. Larsen. Report on the INEX 2004 interactive track. ACM SIGIR Forum, 39(1):43-49, June 2005.

A. Tombros, B. Larsen, S. Malik. The interactive track at INEX 2004. Proceedings of the 3rd INEX Workshop (INEX 2004), Lecture Notes in Computer Science, Volume 3493.

G. Kazai, M. Lalmas and A. de Vries. Reliability Tests for the XCG and inex-2002 Metrics. Proceedings of the 3rd INEX Workshop (INEX 2004), Lecture Notes in Computer Science, Volume 3493.

Z. Szlavik and T. Roelleke. Building and Experimenting with a Heterogeneous Collection. In Proceedings of the 3rd INEX Workshop (INEX 2004), Lecture Notes in Computer Science, Volume 3493.

S. Malik, M. Lalmas, Norbert Fuhr (2005). Overview of INEX 2004. In Proceedings of the 3rd INEX Workshop (INEX 2004), Lecture Notes in Computer Science, Volume 3493.

S. Betsi (2005). XML Retrieval: User expectations, MSc in Information Management, Queen Mary University of London.

The following publications consist of work related to the INEX initiative (e.g. making use of the INEX data, or investigating issues related to those raised within INEX):

Z. Szlavik, A. Tombros, M. Lalmas. Investigating the use of summarisation for interactive XML retrieval. To appear in the Proceedings of the 21st ACM Symposium on Applied Computing, Information Access and Retrieval Track (SAC-IARS'06), 2006.

M.-R. Amini, A. Tombros, N. Usunier, M. Lalmas and P. Gallinari. Learning to Summarize XML Documents by Combining Content and Structure Features, 14th ACM Conference on Information and Knowledge Management (CIKM 2005), Bremen, Germany, 297-298, October 2005.

Z. Kong and M. Lalmas. XML Multimedia Retrieval, Symposium on String Processing and Information Retrieval (SPIRE 2005), Buenos Aires, Argentina, pp 218-223, November 2005.

J. Reid, M. Lalmas, K. Finesilver and M. Hertzum. Best Entry Points for Structured Document Retrieval – Part I: Characteristics, Information Processing & Management, 42(1):74-88, 2006.

J. Reid, M. Lalmas, K. Finesilver and M. Hertzum. Best Entry Points for Structured Document Retrieval – Part II: Types, Usage and Effectiveness, Information Processing & Management, 42(1):89-105, 2006.

Tutorials, invited lectures and presentations that have a strong emphasis on the INEX initiative:

M. Lalmas. Accessing XML Content: From DB and IR Perspectives, Tutorial at the 14th Conference on Information and Knowledge Management (CIKM), Bremen, Germany, 31 October - 5 November, 2005 (with Sihem Amer-Yahia).

M. Lalmas. Structure/XML retrieval. Lecture at the 5th European Summer School in Information Retrieval, ESSIR2005, Dublin, 5-9 September 2005.

M. Lalmas. XML retrieval and evaluation: Where are we? Invited Talk at the 5th Dutch-Belgian Information Retrieval Workshop (DIR'5), Utrecht, The Netherlands, 10-11 January 2005.

M. Lalmas. XML Retrieval, University of Twente, October 2005.

M. Lalmas. XML retrieval and evaluation: the INEX experience, The Open University, UK, November 2005.

M. Lalmas. XML retrieval and evaluation: the INEX experience, University of Edinburgh, UK, December 2005.

In addition, The INEX web site has accumulated publications describing work making use of the INEX data. The publication list is currently been updated. See http://inex.is.informatik.uni-duisburg.de/2005/index.html under Publications heading.

2.6   Conclusions

With the inclusion of the various research tracks, INEX is expanding in scope and in the number of participating organisations. INEX is also in the process of acquiring new collections of XML documents in an effort to enhance the evaluation environment. INEX has shown that XML retrieval is a challenging field within IR and DL research. In addition to learning more about XML retrieval approaches, INEX is making steps in the evaluation methodology for accessing and retrieving XML documents.

3 CLEF: Cross-Language Evaluation Forum

The provision of functionality to store, represent, access and retrieve information in any language is a key issue in the digital library research domain. However, in practice, few operational DL systems offer anything more than very limited multilingual services. The CLEF activity within DELOS aims at advancing the state-of-the-art in this area. It does this by providing an infrastructure for multilingual and cross-language system testing and evaluation and a forum where researchers and system developers can discuss and compare different approaches and share their experiences. Over the years, CLEF has stimulated research work through the design of appropriate tasks. In many cases, the problems studied are new. The aim has been to encourage was a shift of focus of research from textual document retrieval to information extraction and multimedia retrieval over languages. For this reason, new tasks have been added each year: an interactive task in 2001; cross-language spoken document retrieval in 2002, question-answering and image retrieval in the multilingual context in 2003, and a web track and cross-language geographical information retrieval in the 2005 campaign (see Figure 1). In the rest of this section we describe this year's campaign and the main results achieved.
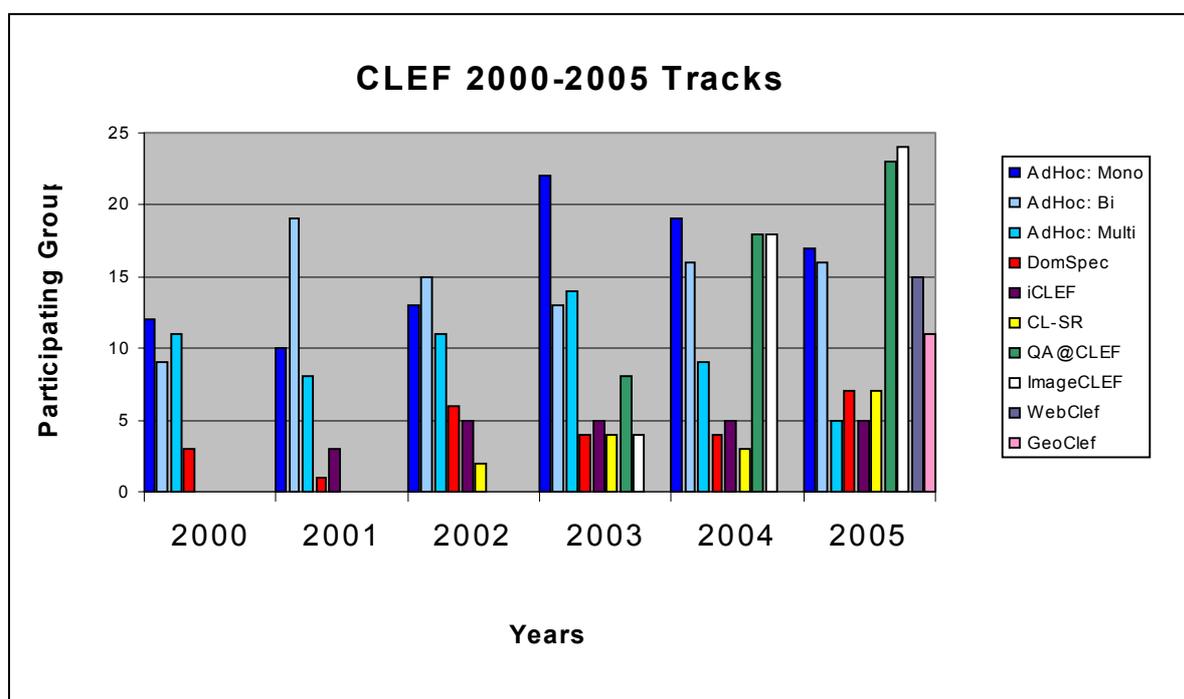

Figure 1: Shift in R&D focus at CLEF over the years

3.1 The CLEF 2005 Campaign

There were eight evaluation tracks in CLEF 2005:
- Ad-Hoc: the ad-hoc track offered a task aimed at studying in-depth the problem of results merging over collections/over languages and at measuring progress in multilingual IR system development over time;
- Domain-Specific: this track studied mono- and cross-language retrieval on structured scientific data;
- iCLEF: the interactive track was devoted to the comparative study of user-inclusive cross-language search strategies in two contexts: cross-language question answering and retrieval of annotated images;
- QA@CLEF: the question-answering track focused on building up a common and replicable evaluation framework to test both mono- and cross-language QA systems. New types of natural

language questions and new evaluation measures – namely the K1 value and r coefficient – were introduced in order to build more challenging test sets and the explore system self-scoring ability;

- ImageCLEF: the image retrieval track explored the use of both text and content-based retrieval methods for cross-language image retrieval; a major goal was to investigate the effectiveness of combining text and image for retrieval;
- CL-SR: the speech track focused on a very difficult task: searching spontaneous speech from oral historical interviews (Shoah archives) rather than the more commonly used news broadcasts. The aim is to encourage the development of technologies to facilitate access to spontaneous speech
- WebCLEF: the web track constructed a multilingual web corpus, with web content in many languages, as an important first step towards a cross-lingual web retrieval test collection. This will serve as an important resource to better understand the challenges of multilingual web retrieval. The creation of such a corpus raised many issues including copyright and linguistic balance, among others
- GeoCLEF: the geographic information retrieval track was run as a pilot task with the aim of building an evaluation infrastructure to evaluate the retrieval of multilingual documents with an emphasis on geographic search; this was the first time that GIR systems ha been evaluated in a multingual context. The interest in this initial work, especially from industry, was encouraging.

3.2 Main Results

The results achieved by CLEF in this period can be summarised in the following main points:

- development of a new technical infrastructure designed to support innovative research on on the evolution of evaluation information access methods and techniques (see Di Nunzio & ferro, 2005a, 2005b)
- study and implementation of evaluation methodologies for diverse types of cross-language IR systems
- documented improvement in system performance for cross-language text retrieval systems (see Gonzalo and Peters, 2005)
- quantitative and qualitative evidence with respect to best practice in cross-language system development (see Gonzalo and Peters, 2005)
- R&D activity in new areas such as cross-language question answering, multilingual retrieval for mixed media, and cross-language geographic information retrieval (see, for example, CLEF 2005 Working Notes and CLEF 2004 Proceedings)
- creation of important, reusable test collections for system benchmarking
- building of a strong, multidisciplinary research community.

Further details are given in the following.

3.2.1 New CLEF Infrastructure

This year the University of Padua took over responsibility for the technical coordination of CLEF. They designed and developed a new CLEF technical infrastructure to support innovative research on the evolution of evaluation information access methods and techniques. A prototype software system, the Distributed Information Retrieval Evaluation Campaign Tool (DIRECT) was designed and developed to support the new infrastructure; DIRECT managed the CLEF infrastructure and provided the CLEF assessors and participants with both tools and an easy to use Web interface for relevance assessment, run submission and validation. This system is fully documented in Di Nunzio & Ferro, 2005a, 2005b.

3.2.2 Test Collections

One of the objectives of CLEF is to build reusable test collections for system benchmarking. Each collection consists of documents, topics/queries, relevance assessments. They are built up and expanded over the years. Here below we list the current CLEF document collections, new additions in 2005 are evidenced. Each track created sets of appropriate topics or questions (simulations of

information needs) in a number of languages for use with these document collections. The form of the topics and the size and languages of the topic sets varied according to the requirements of the different evaluation tasks.

CLEF Document Collections
- CLEF multilingual comparable corpus of more than 2M news docs in 12 languages: **Bulgarian** (new in 2005); Dutch; English; Finnish; French; German; **Hungarian** (new in 2005); Italian; Russian; Spanish; Swedish; Portuguese (used for Ad Hoc, QA@CLEF, iCLEF, GeoCLEF tracks).
- Domain Specific Scientific Collections:
  - GIRT-4 social science database in English and German: more that 300,000 docs
  - **Russian Social Science Corpus**: almost 100,000 docs (used for the Domain-Specific track)
- Historical and Medical images collection:
  - St Andrews historical photographic archive: 28,000 images
  - CasImage radiological medical database with case notes in French and English: 9,000 images;
  - **PEIR 33,000 images, MIR 2,000 images, PathoPic 9,000 images;**
  - **IRMA collection** in English and German for automatic medical image annotation: 10,000 (Used for ImageCLEF)
- Collections of spoken transcripts:
  - **Malach** collection of spontaneous conversational speech derived from the Shoah archives: 589 hours (Used for CL-SR)
- Collection of Web pages:
  - **EuroGOV,** a multilingual collection of more than 2M webpages crawled from European governmental sites (Used for WebCLEF).

3.2.3 Community Building
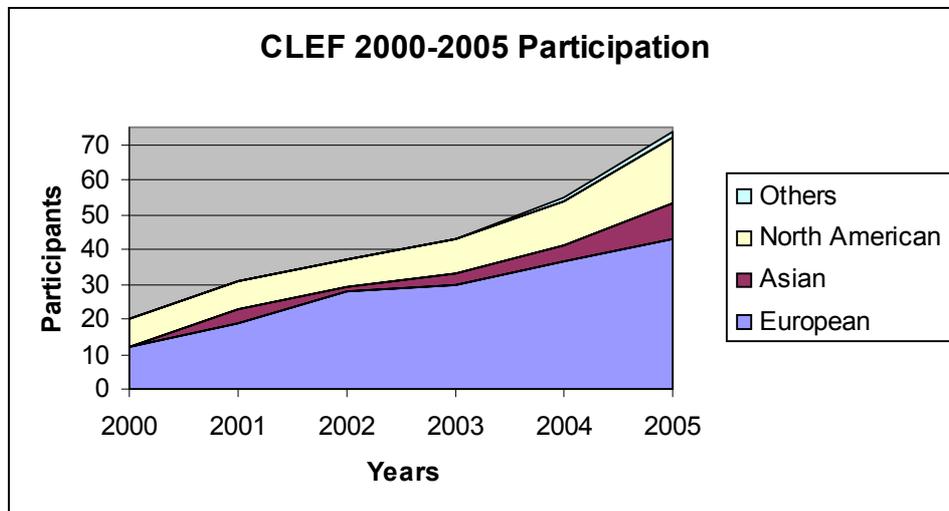
Figure 2: Growth in participation at CLEF over the years

Participation in CLEF has increased enormously over the years. A total of 74 groups from 26 nations and 5 continents submitted results in CLEF 2005. Although the majority of groups are from academia there was a small but strong contingent from industry. As new areas are covered, CLEF is becoming an increasingly multidisciplinary forum with participants from the Information Retrieval (IR), Natural Language Processing (NLP), image and speech processing, and Geographic Information System (GIS) communities, and a consequent synergy of diverse expertise. The discussion lists during the campaign and the workshop at the end give the different groups the opportunity to come together exchanging and

sharing ideas, experiences, tools and methodologies. CLEF puts a strong emphasis on resource building and sharing. Many groups that originally met at a CLEF workshop have continued to work together and collaborate. The result is a strong, well-connected and enthusiastic community. The figure shows the growth in participating groups over the years.

The CLEF workshop, held in Vienna, 21-23 September in conjunction with ECDL was attended by 110 participants. A joint workshop was organised between the ImageCLEF group and the MUSCLE Network of Excellence on Image and Video Retrieval Evaluation on the previous day.

The CLEF activities are widely known and disseminated in the relevant international research communities (in particular IR and NLP) , thus helping to contribute to a high profile and awareness of DELOS activities in these areas.

3.3   Publications

The main publications of CLEF in the reporting period are the CLEF 2005 Working Notes and the CLEF2004 post-campaign proceedings:

Peters, C., (eds.) Cross Language Evaluation Forum: Results of the CLEF 2005 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria. http://www.clef-campaign.org/2005/working_notes/CLEF2005WN-Contents1.htm

Peters, C., Clough, P.D., Gonzalo, J, Jones, G., Kluck, M., Magnini, B.(eds.) "Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign", CLEF 2004, Bath, UK, 2004. Revised papers". Lecture Notes in Computer Science 3491, Springer 2005, 845 p.

In addition the DELOS partners in CLEF (ISTI-CNR, U.Padua, SICS) have published the following papers:

Peters, C. (2005). "Comparative Evaluation of Cross-language Information Retrieval Systems". In Matthias Hemmje, Claudia Niederee, Thomas Risse (eds.) From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments, Springer LNCS.

Braschler, M., Di Nunzio, G., Ferro, N., Peters, C. (2005). "CLEF 2004: Ad Hoc Track Overview and Results Analysis." In Peters, C., Clough, P., Gonzalo, J, Jones, G.J.F., Kluck, M., Magnini, B. (eds.) "Multilingual Information Access for Text, Speech and Images". Fifth Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK. Revised papers". Lecture Notes in Computer Science 3491, Springer 2005, 10-26.

Bacchin, M., Ferro, N. and Melucci, M. "A Probabilistic Model for Stemmer Generation." Information Processing and Management," 2005, vol. 41, n. 1, pp. 121–137.

Di Nunzio, G. M. and Ferro, N. (2005a). DIRECT: a System for Evaluating Information Access Components of Digital Libraries. In Rauber, A., Christodoulakis, S., and Min Tjoa, A., editors, Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005), pages 483–484. Lecture Notes in Computer Science (LNCS) 3652, Springer, Heidelberg, Germany.

Di Nunzio, G. M., and Ferro, N. (2005b). DIRECT: a Distributed Tool for Information Retrieval Evaluation Campaigns. In Ioannidis, Y., Schek, H.-J., and Weikum, G., editors, Proc. 8th International Workshop of the DELOS Network of Excellence on Digital Libraries on Future Digital Library Management Systems (System Architecture & Information Access), 2005.

J. Gonzalo, C. Peters, (2005). The impact of evaluation on multilingual text retrieval. Proceedings of SIGIR 2005, ACM Press, NY, USA, pp 603-604.

Hansen, Preben and Karlgren, Jussi (2005) Effects of Foreign Language and Task Scenario on Relevance Assessment. Journal of Documentation 61:5.

A recent survey of CLEF 2004 and 2005 participants asking for information on "Publications by CLEF participants that cite CLEF results" has produced the following results
- International Journal Articles : 2005 = 7, 2004 = 13
- Articles or Chapters in Monographs : 2005 = 6, 2004 = 1
- International Conferences: 2006 = 2 (accepted) 2005 = 27 2004 = 24
- Theses: 2005 = 1, 2004 = 2

It is to be presumed that these results are partial as not all groups replied. A complete list can be found at http://clef.isti.cnr.it/clef-bibliography.pdf.

4 TASK 7.5: A Digital Library Testbed Framework for the Evaluation of Architectures, Services and Execution Dynamics

4.1 Objectives

Most DL evaluations use specific systems, which are difficult to compare. The aim of this effort is to provide a standard testbed framework for comparative evaluation of DL systems. Based on a theoretical framework for DL evaluation, we develop a framework system that can be easily adopted for new application domains or extended by new services. As technical basis for this framework, the Daffodil system is used. For analyzing applications of this framework as well as for comparison with other systems, we develop a standard event model of DL services along with a logging standard and corresponding evaluation tools.

In the following sections the achieved results are presented. First, a short introduction to Daffodil is given, since it is referred to in all sections. Section 4.3 presents our work on the logging standard. In Section 4.4, we describe the development of the testbed. Results and preparations on evaluations with INEX and "The European Library" are presented in Section 4.5.

4.2 The Daffodil Framework

The Daffodil system forms the technical basis for the evaluation framework to be developed. Here we give a brief description of the current state of this system. Daffodil (see http://www.daffodil.de) is a virtual digital library, targeted at strategic support of users during the information search process. For searching, exploring, and managing digital library objects it provides user-customisable information seeking patterns over a federation of heterogeneous digital libraries. Searching with Daffodil makes a broad range of information sources easily accessible and enables quick access to a rich information space. The Daffodil framework consists of two major parts, the graphical frontend client and the backend agent-based services.

4.2.1 The Graphical Client

The graphical client combines a set of high-level search activities as integrated tools. The current Daffodil prototype for the domain of computer science provides seven main tools:

- **Search tool**, to specify the search domain, set filters and compose queries. The queries are broadcast to a set of distributed information services (via agents and wrappers). Integrated result lists are displayed for navigation and detail inspection.
- **Classification Browser**, to allow hierarchical topic-driven access to the information space. It enables browsing of classification schemes like e.g. the ACM Computing Classification System.
- **Thesaurus Browser**, to transform search terms to broader or narrower terms. Subject-specific or Web-based thesauri, like e.g. WordNet, are used for finding related terms. Items can be used via Drag&Drop to another tool.
- **Author Network Browser**, to compute and browse co-author networks for a list of given authors. The list can be either typed or given by dropping a document item on the tool.
- **Journal / Conference Browser**, to search for a journal/conference title and browse many directories, often with direct access to meta-data or the full-text of articles.
- **Personal Library** which stores DL objects in personal or group folders, along with the possibility of enabling awareness for these items.

The client is deployed to the user via Java Webstart technology. There are no limitations on integration of new tools, as long as they are written in Java.

4.2.2 The backend services

On the backend side each frontend tool is represented by one or more agent-based services, which provide the actually functionality. Currently more than 30 services and 15 wrapper agents are running. The services communicate (for efficiency reasons) via CORBA, but can also be accessed via SOAP. For communication XML messages are used. The agent framework itself is very simplisticly modelled for high performance and provides parallel threads for each user.

4.3 Logging standard and corresponding evaluation tools

The goal of this subtask is to specify a logging standard for a normal and comparative evaluation of digital libraries. The following results are archived:

- A first version of a general log event classification model was prepared and described in a paper submitted to an international conference in January 2006 [Klas et al 06]. In this paper, we are proposing a new logging schema that will account for all kinds of data about users, systems and the user-system interactions. We present a novel, multi-level logging framework that will provide complete coverage of the different aspects of DL usage. The main focus is the user behaviour level, for which we describe the logging scheme in some detail. Based on this specification, we show how the various DL stakeholders can analyze the logging data according to their specific interests. In addition, specific analysis tools and a freely accessible log data repository will yield synergies and sustainablity in DL evaluation.
- In parallel to the ongoing discussion about the logging standard, we verified and implemented new and missing user and system triggered digital library events in the Daffodil system. The neccessary logging facilities were integrated into the Daffodil system and the evaluation tools are under preparation.
- A log schema converter for the the Daffodil logs is under development. Since we have collected already large volumes of Daffodil log data, including events and event content (e. g. queries, documents), we can provide rich event logs in the standard format, thus allowing other research groups to analyze this data

4.4 Standard testbed framework for comparative evaluations

In order to setup a standard testbed framework for comparative evaluations we are working on a theoretical framework and extensions of the Daffodil framework.

4.4.1 Evaluation Framework

For the theoretical framework, we have finished a whitepaper [Fuhr et al. 06] on digital library evaluation, including state of the art overview and recommendations. This whitepaper contains a first version of the merging of two frameworks, namely the Evaluation Computer and Interaction Triptych Framework. This paper has been submitted for publication in a journal.
In parallel with this work, new modeling techniques are being considered, like e.g. the ontology-based model described in [Kovacs & Miszik 06].

The Interaction Triptych Framework aims to identify and describe interaction processes within a DL and to provide a basis on which interaction evaluation elements (methods, criteria, etc.) are developed. As a user-centered model it explores and emphasizes on the dialogue elements that are mutually expressed and transacted during an interaction period between the user and the DL and recognizes the random and iterative nature of interaction events.
The Evaluation Computer (EC) is a systematic approach for the description and analysis of DL evaluation activities. This model is able to provide insights about the distance between different evaluation procedures and to locate white spots in the domain. According to this model, the evaluation process is defined as a selection of a point from a multidimensional space. The space consists of a set of aspects, namely evaluation, user, organizational, content and system. EC is able to answer the major

questions regarding the evaluation process (evaluation aspect), the DL components (user, content, system aspects) and the contextual conditions (organizational aspects).

4.4.2 Practical Results

On the practical side, new services were developed and integrated into the Daffodil system, and a layered help function of the system is under development. The following new services were integrated:
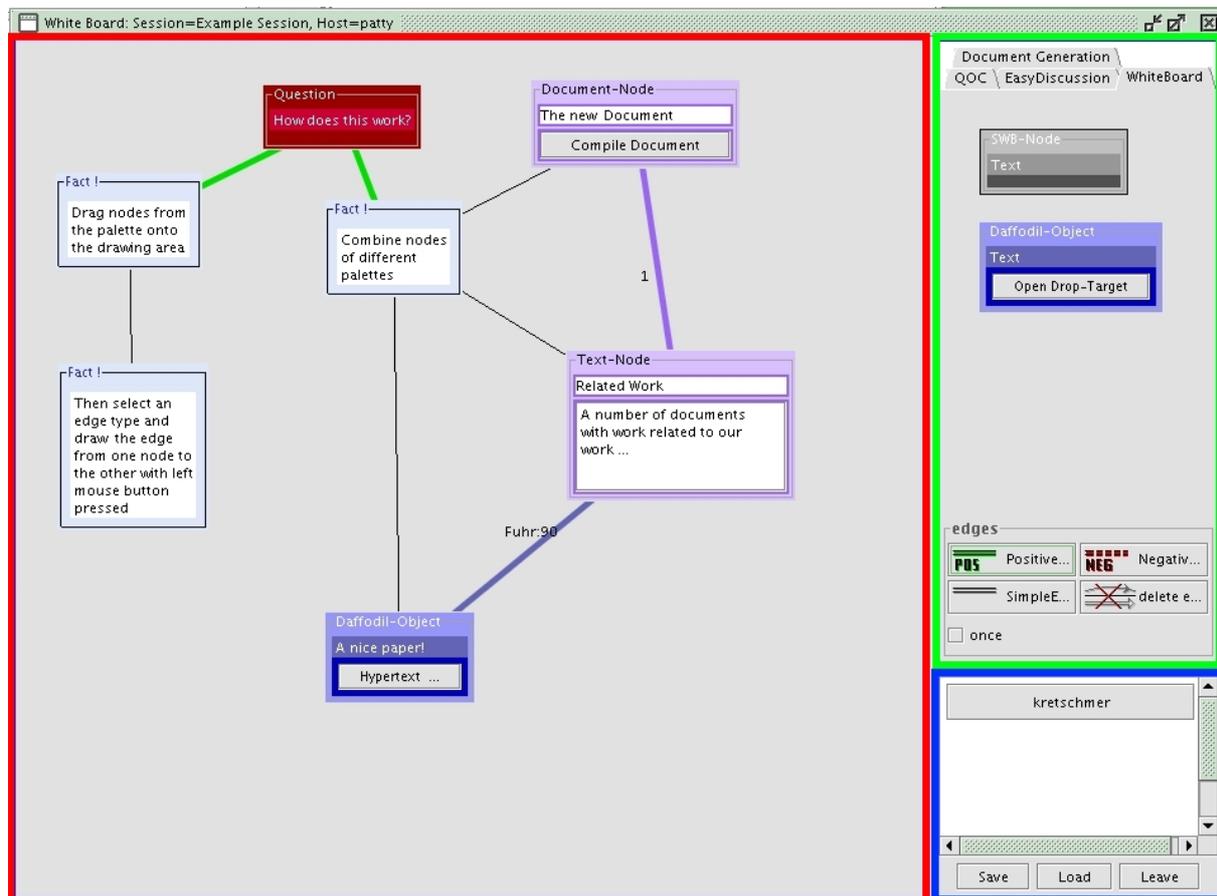

Figure 3: Whiteboard for Collaboration

- A multi-level hypertext browser allows for clustering and browsing of search results (joint work with task 4.7).

- For searching XML document collections, several services had to be extended appropriately (joint work with task 7.3 (INEX) - see below).

- As an extension to the XML services, a summarization tool for XML fulltext originally developed at the Queen Mary University of London was rewritten and integrated.

- An annotation tool was integrated that allows for both *out-of-line* and *in-line* annotation of documents, including sharing and searching of annotations and supporting of discussion threads. This part will be extended and evaluated along with task 4.10 (DILAS).

- For supporting synchronous collaboration, a chat facility and a whiteboard (see Figure 3) with several functionalities for discussion and topic structuring has been integrated.

4.5 Evaluation Activities

Besides the two standardize tasks, our group started, prepared and ran two major activities, including implementation of new services for applying the testbed framework in specific evaluations.

The first application in the interactive track of INEX was performed in fall 2005, whereas the evaluation in the context of the European Library is still in the planning stage.

4.5.1 INEX Interactive Track Evaluation
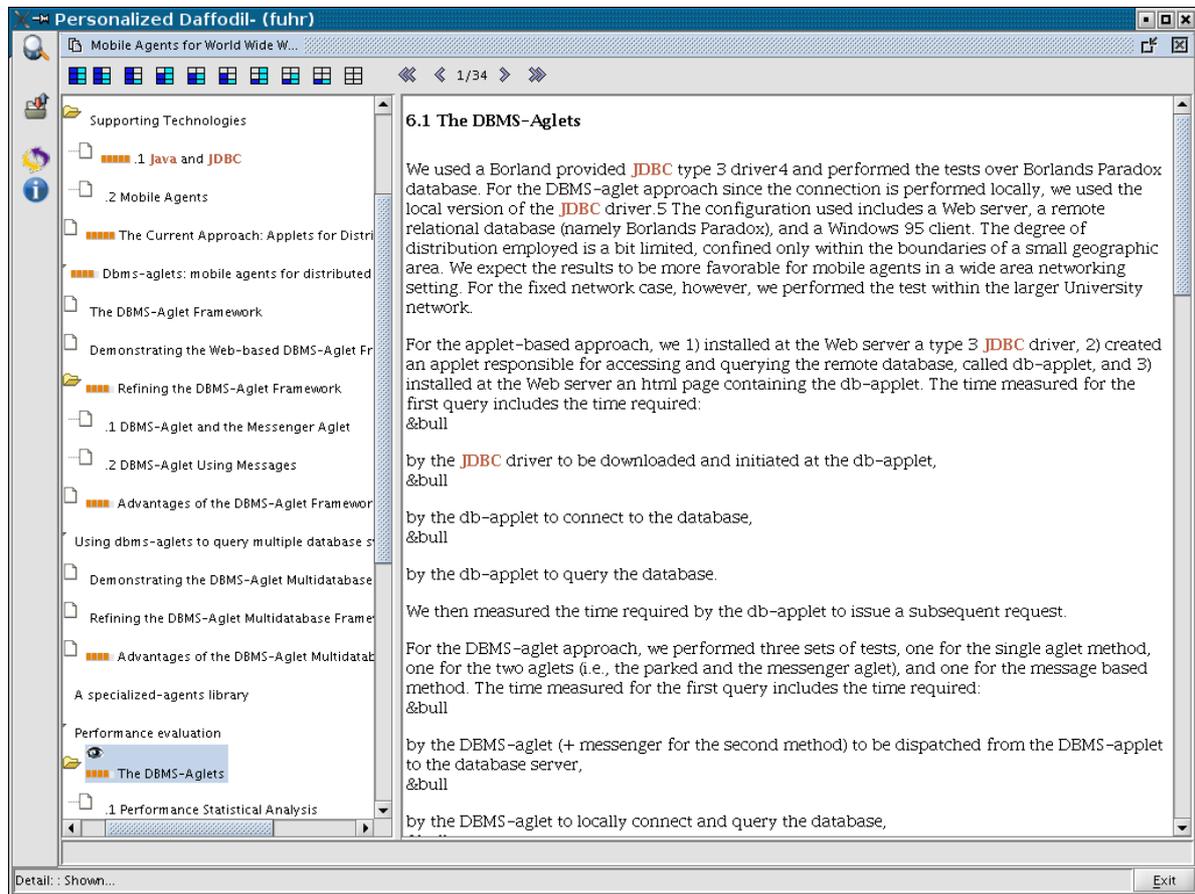
Figure 4: New detail view in Daffodil for the INEX interactive track

In the Interactive Track of INEX 2005, Daffodil, as joint work with the INEX task 7.3, was used as baseline system for interactive XML retrieval. There were 11 groups participating in this track, each performing experiments with at least 6 subjects. Both relevance assessments and interaction logs were collected for later analysis.

For this evaluation, the Daffodil system was extended to meet the necessary requirements, which included:

1. Performance and stability issues,
2. Integration of a new search-tool with special result list handling and view for XML documents,
3. Integration of a new detail viewer with fulltext ability and table of content browsing (see Figure 4)
4. Integration of services and sources to access the XML retrieval backend,
5. Implementation of new user events, like TOC browsing according to the standard log schema.

The questionnaires, compiled log events and user statements will be made accessible at the testbed website.

4.5.2 Evaluation of „The European Library"

Following the DELOS/TEL meeting in London, we discussed with members of the TEL consortium about a comparative evaluation between the current TEL interface and an appropriate variant of Daffodil. The TEL currently provides searching in national libraries.

The TEL system uses JavaScript as technology and access the national libraries via the protocols Z39.50, SRW/SRU and OAI; in addition, several of the catalogs are mirrored in a central database.

Figure 5: Current search interface of TEL

On the technical side, Daffodil needs to be extended to
1. access the same information sources, i.e. connect to the national digital libraries or the TEL gateway and
2. modify the existing search-tool, so that it allows for incremental display of result lists, since TEL follows the same strategy.

On the conceptual side, group members work on the design of the comparative evaluation. This evaluation will follow both an analytical and an empirical approach. The goal of the analytical evaluation is to assess the functional similarities and differences between the existing TEL search tool and the Daffodil-based search interface. The goal of the empirical evaluation is to evaluate how well each tool supports the users needs.

The analytical evaluation considers the following use-oriented characteristics of the two search tools:
- Usability
- Search and browsing functions
- Display functions
- Feedback/help functions

The analytical evaluation will be conducted through expert walkthrough of systems/prototypes as well as a desk study of the corresponding design documentation. The expert walkthrough of the systems/prototypes will access similar information sources.
The empirical evaluation will be based on a user-centered and qualitative approach. Its focus is on the users experience with the tools. It considers the following perspectives:
- User characteristics, preferences and strategies (i.e. their experience),
- Types of activities that users do/tasks they perform,

- The environment in which the search tool is used, either in the natural work setting or in a controlled laboratory setting.

The empirical evaluation will be conducted in natural settings as well as laboratory settings, and enroll representatives of two major categories of users:
- stakeholders of TEL (eg., librarians)
- end-users of TEL.

4.6   Publications

Norbert Fuhr; Giannis Tsakonas; Trond Aalberg; Maristella Agosti; Preben Hansen; Sarantos Kapidakis; Claus-Peter Klas; László Kovács; Monica Landoni; András Micsik; Christos Papatheodorou; Carol Peters; Ingeborg Solvberg. Evaluation of Digital Libraries. (Submitted for publication). 2006. http://www.is.informatik.uni-duisburg.de/bib/docs/Fuhr_etal_06.html

An Experimental Framework for Comparative Digital Library Evaluation: The Logging Schema. C.-P. Klas, N. Fuhr, S. Kriewel, H. Albrechtsen, L. Kovacs, A. Micsik, P. Hansen, G. Tsakonas, S. Kapidakis, Ch. Papatheodorou, E. Jacob (Submitted for publication). 2006. http://www.is.informatik.uni-duisburg.de/bib/docs/Klas_etal_06.html

Laszo Kovacs, Andras Micsik, An Ontology-Based Model of Digital Libraries, Proceedings ICADL 2005, Lecture Notes in Computer Science, Volume 3815, Jan 2006, Pages 38 – 43. Springer, Berlin et al.