DELOS NoE
on Digital Libraries

WP7: Evaluation

WP Leader: Norbert Fuhr
University of Duisburg-Essen, Germany
WP7 tasks

- 7.3: INEX - Initiative for the Evaluation of XML Retrieval
- 7.4 CLEF – Cross-Language Evaluation Forum
- 7.5 A Digital Library Testbed Framework for the Evaluation of Architectures, Services and Execution Dynamics
Task 7.3: INEX - Initiative for the Evaluation of XML Retrieval

Task leader: Mounia Lalmas
Queen Mary University of London, UK
Objectives of INEX
Promote research and stimulate development of XML information access and retrieval, through

- Creation of evaluation infrastructure and organisation of regular evaluation campaigns for system testing
- Building of an XML information access and retrieval research community
- Construction of test-suites

INEX 2005 has allowed a new community in XML information access to emerge, as shown by the number of publications (64+ in 2005, 37 in 2004 and 13 in 2003). INEX is allowing research in XML information access to thrive.
INEX 2005 Results: Expansion of Test Collections

• Ad-hoc Track:
  – fulltexts in XML from 18 IEEE-CS journals
  – 12,107 +4,712 articles (1995-2004), 494 +241 MB

• Multimedia track: Lonely Planet World Guide
  – 462 XML documents with 1800 photos, 452 maps

• Document mining
  – WIPO corpus: 75,250 XML documents
  – MovieDB corpus (based on the Internet Movie Database): 9463 XML documents

• Heterogeneous track
  – 7 corpora with 900,000+ XML documents
INEX 2005: Results
Research in New Directions I

- **Interactive track**
  - 3 different tasks
  - 11 participating groups with ~100 test persons
  - How do users interact with the search system (basis for the definition of metrics)
  - How can interactive retrieval become more effective? (functionality, UI design)

- **Heterogeneous track**
  - new collections with different DTDs and their effect on XML IR system effectiveness.

- **Relevance feedback**
  - queries that also include structural hints (rather than content-only queries in 2004).
INEX 2005: Results
Research in New Directions II

- **Natural language query processing track**
  - New task allowing new participants with NLP expertise to join INEX without the need to develop a search engine, and thus encouraging wider accessibility.

- **Document mining track**
  - Collaboration with the PASCAL network of Excellence
  - Development of machine learning methods for structured data mining
  - This year: classification and clustering for XML documents.

- **Multimedia track**
  - Access methods for structured multimedia documents
  - This year: text and image retrieval
**Ad-hoc track** a simulation of how a XML content-based library might be used. Assumptions about user requirements led to three different strategies:

- **focussed**: user prefers a single element that most exhaustively and specifically discusses the topic of the query.
- **thorough**: user prefers all highly exhaustive and specific elements.
- **fetch and browse**: user is interested in highly exhaustive and specific elements that are contained only within highly relevant articles.
INEX 2005: Results
Research in New Directions V

• New set of metrics:
eXtended Cumulated Gain (XCG)
  – evaluation framework that considers dependency among XML document components
  – incorporate mechanisms to reward the retrieval of near-misses and to address issues of overlap.
  – system and user-oriented evaluation aspects and recall and precision-like qualities are measured.
INEX 2005: Overall Results

- stimulation of research activity in new, previously unexplored areas, such as interactive aspects of XML information access and multimedia XML access,
- 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
- participation is up: 64 groups in 2005
Plans for INEX 2006

INEX 2005 Tracks

- XML retrieval (ad-hoc)
- Interactive track
- Heterogeneous track
- Relevance feedback track
- Natural language query processing track
- Multimedia XML retrieval track
- Document mining track

- New tracks
  - XML entity search track (under discussion)

Details of new tracks and new tasks and evaluation methodologies under discussion

Further development of the infrastructure for results analysis and testing of methodology.

N. Fuhr: DELOS WP7 Report: Advisory Board 12/05
Task 7.4: CLEF – Cross language Evaluation Forum

Task Leader: Carol Peters
CNR-ISTI, Italy
Objectives of CLEF
Promote research and stimulate development of multilingual IR systems for European languages, through

- Creation of evaluation infrastructure and organisation of regular evaluation campaigns for system testing
- Building of an MLIA/CLIR research community
- Construction of publicly available test-suites

CLEF 2005
- continued expansion of participation, tasks, and test collections
- Focus shifted from cross-language textual document retrieval to include various types of information extraction in multilingual/multimedia context.
CLEF 2005 Results:
Research in New Directions I

- **Ad-hoc track:**
  - results merging over collections/over languages and at measuring progress in multilingual IR system development over time;

- **Interactive track:**
  - comparative study of user-inclusive cross-language search strategies

- **Question-answering track:**
  - mono- and cross-language QA systems.
  - new types of natural language questions
  - new evaluation measures – namely the K1 value and r coefficient

- **Image retrieval track:**
  - combining text and content-based retrieval for cross-language image retrieval
  - effectiveness of combining text and image for retrieval
CLEF 2005: Results
Research in New Directions II

- **Speech track**
  - searching spontaneous speech from oral historical interviews (Shoah archives)

- **Web track**
  - Construction of a multilingual web corpus
  - Resource for investigating multilingual web retrieval.

- **Geographic information retrieval track**
  - retrieval of multilingual documents with an emphasis on geographic search
  - first time that GIR systems have been evaluated in a multilingual context
  - strong interest, especially from industry, was encouraging.
CLEF 2005: Overall Results

- Stimulation of research activity in new, previously unexplored areas, such as cross-language question answering, image and geographic information retrieval
- 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
- Creation of a large set of empirical data about multilingual information access from the user perspective;
- Quantitative and qualitative evidence with respect to best practice in cross-language system development
- R&D activity in new areas such as cross-language question answering, multilingual retrieval for mixed media, and cross-language geographic information retrieval
- Creation of important, reusable test collections for system benchmarking building of a strong, multidisciplinary research community

N. Fuhr: DELOS WP7 Report: Advisory Board 12/05
Plans for CLEF 2006

CLEF 2006 Tracks (same tracks – new tasks)

- Multilingual, Bilingual and Monolingual (non-English) Text Retrieval on News Collections
- Domain-Specific Retrieval on Social Science Docs
- Interactive Cross-Language IR
- Multilingual Question Answering
- Cross-Language Image Retrieval
- Cross-Language Speech Retrieval
- Multilingual Web Track
- Cross-Language Geographical IR

Details of new tasks and evaluation methodologies now under discussion

Further development of the infrastructure for results analysis and testing of pooling methodology.
Task 7.5: A Digital Library Testbed Framework for the Evaluation of Architectures, Services and Execution Dynamics

Task Leader: Claus-Peter Klas
University of Duisburg-Essen
DL Testbed: Objectives

• Standard Testbed Framework for DL Evaluation
  – Guide, Repository, Testbed-Framework Daffodil
• Standard Event Model of DL users and services
  – Logging standard schema and corresponding evaluation tools
• Evaluation Actions:
  – INEX Baseline System
  – The European Library
  – Collaboration in Computer Science DLs
Standard Event Model for DL Usage and Services

- Digital Library Life Cycle
  Discover - Retrieve - Collate - Interpret - Re-Present

- Logging levels
  1. **User behavior level**: context, task, user info, etc.
  2. **Concept level**: browse - store - annotate - help - navigate - inspect – communicate – author
  3. **Service level**: metadata search, detail search, journal browse, ....
  4. **HCI (keystroke) level**: mouse movement/clicking, keyboard input, eye tracking
  5. **System level**: server load, resource, network traffic etc.

N. Fuhr: DELOS WP7 Report: Advisory Board 12/05
Comparative Evaluation:
The European Library

N. Fuhr: DELOS WP7 Report: Advisory Board 12/05
Comparative Evaluation:
TEL standard interface vs. DAFFODIL

Analytical evaluation of functional similarities & differences

- Usability
- Search & Browse
- Display
- Feedback / Help Functions

• Empirical evaluation
  - How well does each tool support the users’ needs?
  - Experts and End-Users
INEX Interactive

- Daffodil as baseline system for interactive XML retrieval
  - Integration of services and sources into Daffodil
  - Integration of new search tool with special result list handling and display of XML data
INTERSYSTEM HANDOFF

Our description of the intersystem handoff follows IS-41 [2] (GSM follows similar procedures) and we assume network-controlled handoff. Figure 3 illustrates the trunk (voice or data circuit) connection before and after the handoff. A communicating mobile user moves out of the base station served by MSC1 and enters the area covered by MSC2. The handoff follows these steps:

- MSC1 requests MSC2 to perform handoff measurement. MSC2 then selects a candidate base station, BS2, for handoff. That is, MSC2 finds a base station that covers the mobile phone and has a free radio channel to cover the call. MSC2 returns the signal-quality parameter values and other information to MSC1.
- MSC1 checks if the mobile phone has made too many handoffs or if intersystemtrunks are not available. If so, MSC1 exits the procedure. Otherwise, MSC1 asks MSC2 to set up a voice channel. Suppose that a voice channel is available in BS2. MSC2 asks MSC1 to start the radio link transfer.
- MSC1 sends the mobile phone a handoff order. The mobile phone tries to synchronize to BS2. After the mobile phone connects to BS2, MSC2 informs MSC1 that the handoff is successful. MSC1 then connects the call path (trunk) to MSC2 and completes the handoff.

Figure 3: Before (a) and after (b) an intersystem handoff.

In this process, MSC1 is called the anchor MSC, and is always in the call path after the handoff (see Figure 4a). If the mobile phone moves back to MSC1 again, the connection between MSC2 and MSC1 terminates.
Next Activities

• Define standard for logging schema
  – January 2006

• Setup evaluation testbed framework
  – January 2006

• TEL Evaluation
Next Activities

• Evaluation of cooperation in DL usage (computer science domain)
  – Evaluation of existing tools
    • Personal Library
    • Annotations/Discussion threads
    • Chat tool (text/voice)
    • Whiteboard
  – Study on tools required for collaboration

• Evaluation of Annotation-Service (Joint work with task 4.10)
WP7 - Conclusion

• Evaluation initiatives:
  – Community building: growing participation, strong support from volunteering groups
  – Foci of international research on XML IR / CLIR
  – New tasks addressing more application-oriented settings

• DL Testbed Framework:
  – Testbed framework: new applications & services
  – Development of logging schema
  – Prototype evaluations: INEX, TEL, Cooperation in DLs