



DELOS WP7: Evaluation

Norbert Fuhr

Univ. of Duisburg-Essen, Germany

WP Objectives

Digital Library Evaluation (DLE):

- Enable communication between evaluation experts and DL researchers/developers
- Continue existing evaluation initiatives relevant for the DL area
- Develop new evaluation models, methods and testbeds



WP7 Activities

- I. DLE Infrastructure
- II. DLE research and development

I. DLE Infrastructure

Testbed Metalibrary

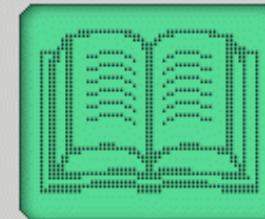
- Developed in 1st DELOS NoE
http://www.sztaki.hu/delos_wg21/metalibrary
- Describes 62 testbeds by the following groups of criteria:
 - general data
 - users and usage
 - applied technologies
 - data collection

Testbed Metalibray



DELOS Working Group 2.1

Evaluation and test environments for digital library research.



MetaLibrary

If you are searching for a digital library testbed or system:

You may browse MetaLibrary entries according the criteria in the following major dimensions:

- [general data](#)
- [about users and usage](#)
- [about applied technologies](#)
- [about data collection](#)

You may also search for patterns in Metalibrary entries:

[List all digital libraries and collections entered into MetaLibrary!](#)

Metalibrary contains mainly collections

→Development of new testbeds in current DELOS NoE

DLE Infrastructure in current DELOS NoE



- Collection of DLE resources (literature, testbeds and toolkits)
<http://dlib.ionio.gr/WP7>
- Communication forum
<http://dlib.ionio.gr/delosforum>
- Support prototype evaluations
- Organization of evaluation campaigns:
CLEF, INEX

Cross-Language Evaluation Forum



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 2004 has seen shift in focus from cross-language document retrieval to include information extraction in multilingual multimedia context

CLEF 2004: Evaluation Tracks



CLEF 2004 offered six tracks designed to evaluate the performance of systems for:

- mono-, bi- and multilingual document retrieval on news collections (Ad-hoc)
- mono- and cross-language domain-specific retrieval (GIRT)
- interactive cross-language retrieval (iCLEF)
- multiple language question answering (QA@CLEF)
- cross-language retrieval on image collections (ImageCLEF)
- cross-language spoken document retrieval (CL-SDR)

CLEF 2004: Results

- Participation is up: 55 groups in 2004 (42 in 2003)
- Expansion of test-suite
- Great success of QA@CLEF and ImageCLEF
- Synergy of diverse expertise partly consequence of new tracks – IR, NLP, Image Processing, Medical Informatics..



- **Background:**

- Increased use of XML as document format on the Web and in digital libraries
- Development of retrieval systems to store and access XML documents

- **Objectives:**

- Creation of evaluation infrastructure and organisation of regular evaluation campaigns for system testing
- Building of an XML-IR research community
- Construction of test beds + appropriate scoring methods for evaluating content-oriented XML retrieval

INEX - 4 main tasks



- Evaluation of **retrieval effectiveness**, especially by refining the evaluation criteria, in order to consider how XML elements satisfy information needs in the context of digital libraries.
- Evaluation of **efficiency**, taking into account the larger number of possible answers (XML elements) and their possible overlap.
- Prototype evaluation of **usability**, considering various types of information-seeking activities in an interactive setting.
- Investigation of **new testbeds** for heterogeneous and multimedia documents in the context of XML.

Effectiveness



Evaluation of retrieval effectiveness:

- Develop/refine evaluation criteria: how do XML elements satisfy information needs in the context of digital libraries.
- Ad hoc retrieval + 4 tracks
 - Relevance feedback track
 - Heterogeneous data track
 - Natural language track
 - Interactive track
- Development of evaluation methodologies including metrics

Usability



Prototype evaluation of usability, considering various types of information-seeking activities in an interactive setting.

- Done as part of the interactive track
 - Investigate user behaviour when interacting with XML documents
 - Develop and investigate retrieval approaches that are effective in interactive settings

INEX 2004



- 57 participants:
 - 55% Europe
 - 22% USA
 - 13% Asia + Australia/Oceania
- Strong involvement of the participants in the various tasks and evaluation methodologies
- INEX 2004 workshop: December 6-8 in Dagstuhl/Germany

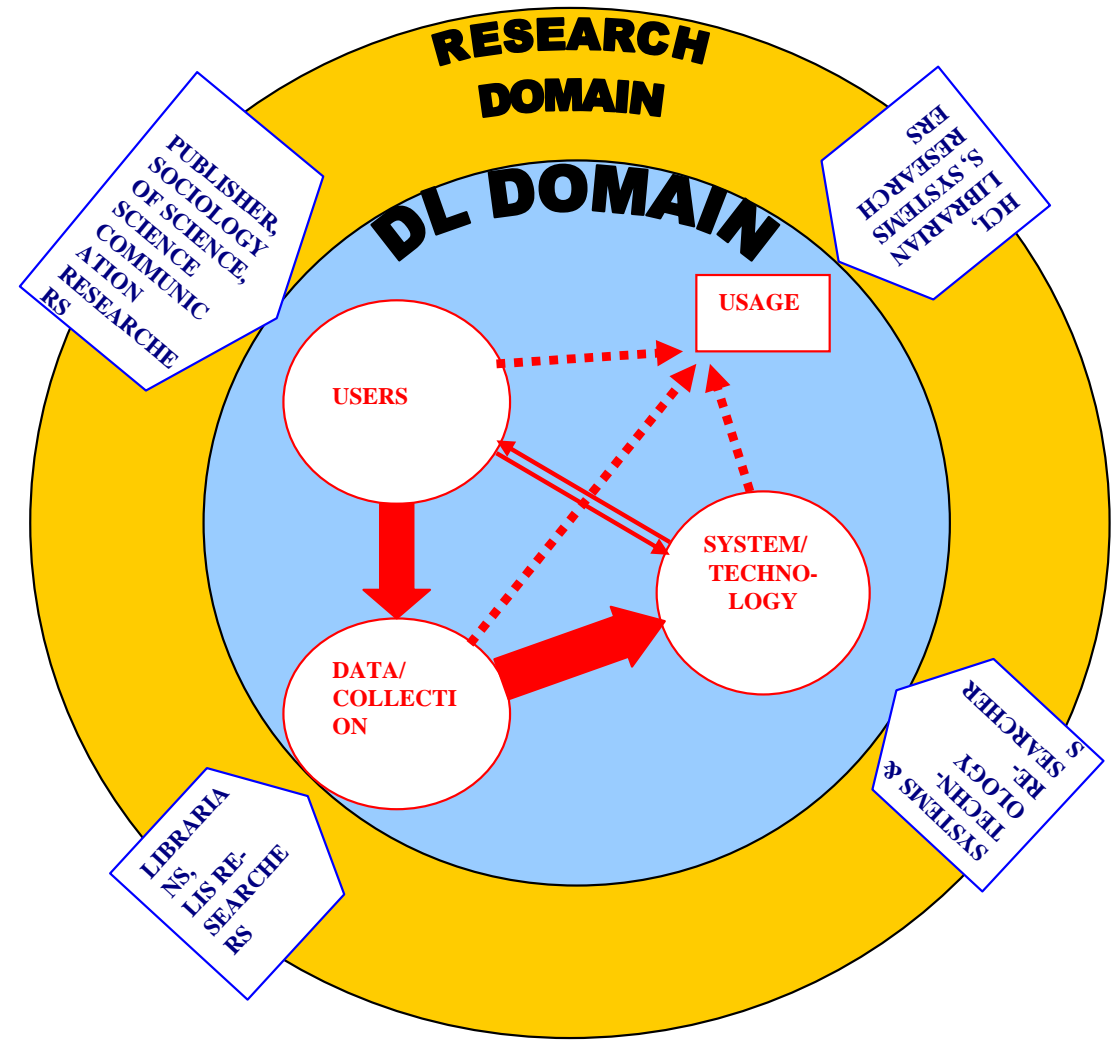
II. DLE research and development

- A conceptual model for Digital libraries and their evaluation
- Evaluation approaches, models and methods
- DLE testbeds

II.1 Conceptual model



Conceptual DL
Model
developed in
1st DELOS
NoE

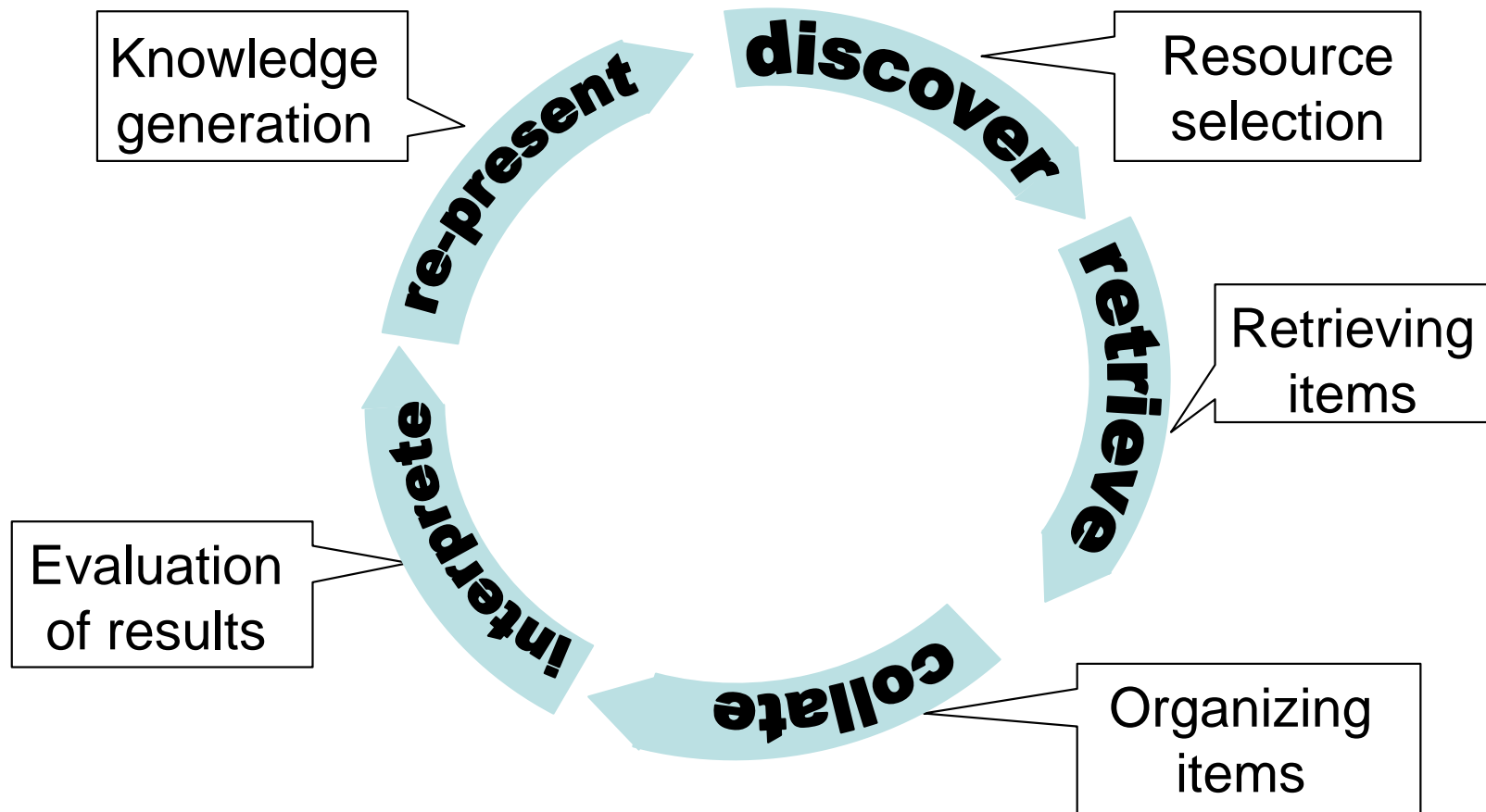




DL usage

- DL life cycle
- Levels of search activities

Usage: DL life cycle



Many evaluations restricted to discover+retrieve stages!

Usage: Levels of search activities

(Bates 1990):

1. Move: Low-level search function
(e.g. type in search term, view retrieved document)
2. Tactic: several moves to further a search
(e.g. broaden/narrow a query)
3. Stratagem: set of actions on a single domain
(citation database, tables of contents of journals)
4. Strategy: complete plan for satisfying an information need
(e.g. subject search, browse relevant journals, find referenced articles)

Little support for higher levels in current systems!

II.2 Evaluation approaches, models and methods

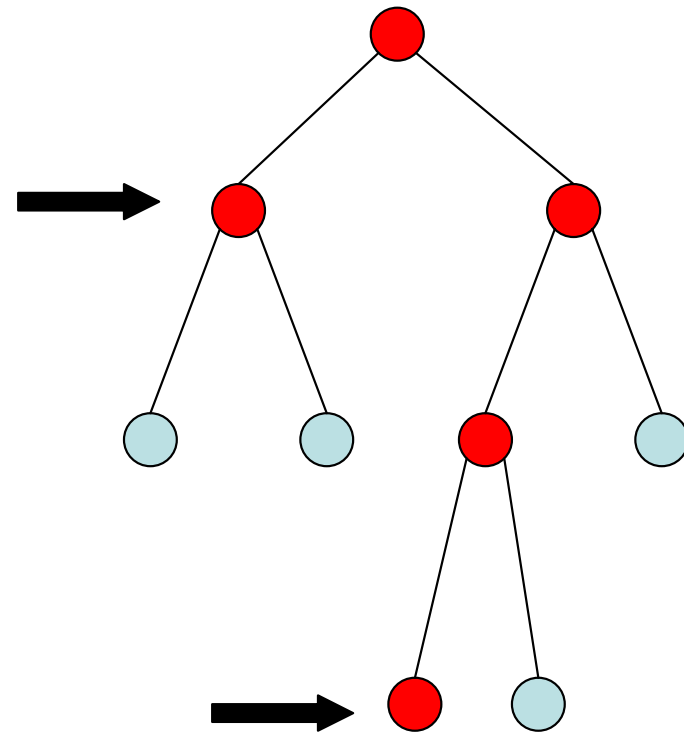


- Participation of users in the evaluation cycle
- Meta-analysis of existing evaluation studies
- Comparison and evaluation of DLE techniques
- Development of new DLE approaches, models and methods

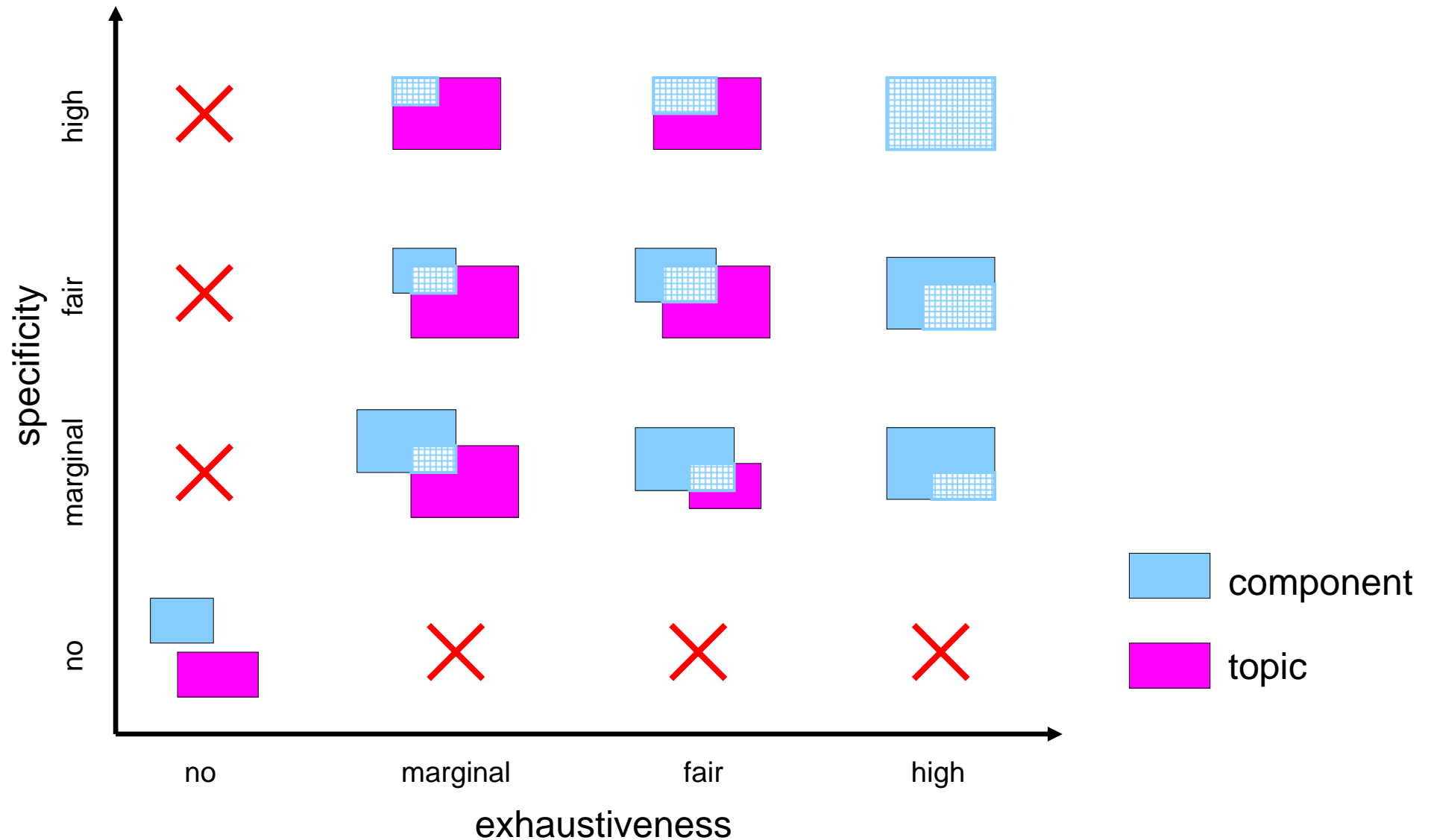
Example: INEX metrics



- Content-oriented retrieval of XML elements
- Retrieval strategy: retrieve most specific elements satisfying the query



INEX relevance scale

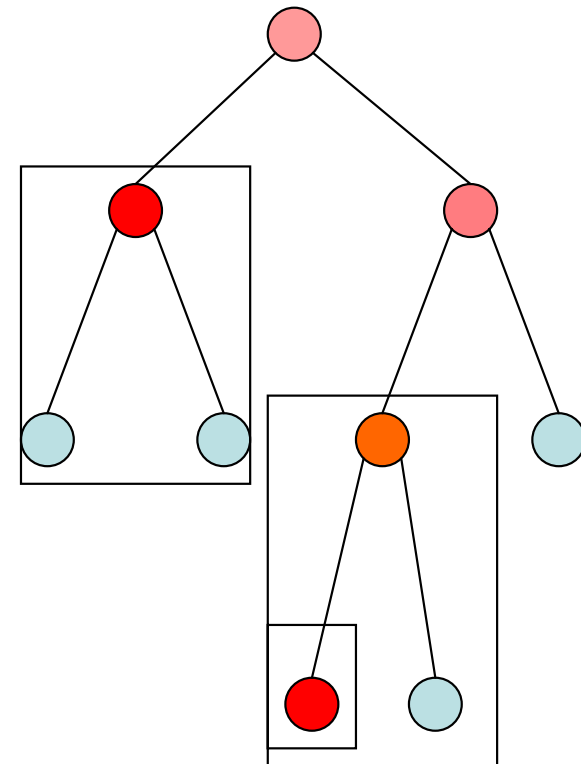


Multiple answers in a document



Recall/precision metrics must consider:

- Relevance scale
- Overlap of elements
- Size of elements
- Assumptions about user behaviour?
-> INEX interactive track



II.3 Testbeds

Characteristics of existing testbeds:

- **Collection** properties
 - media
 - structure
 - heterogeneity
- **Usage**
- Research groups apply their own **systems**

Information Media



- Text
- Facts
- 2D: graphics, images
- Speech
- Video
- 3D





TREC



TREC

TREC

Information structure

- Unstructured  TREC
- Semi-structured (XML) 
- Fully structured (standard databases)
- Hyperlinked (Web) TREC

Heterogeneity



- Language: multilingual
- Media: multimedia
- Heterogeneous structures
- Heterogeneous services



Usage



- ad-hoc (batch retrieval)
- filtering (relevance feedback)
- interactive retrieval
- question answering



TREC



TREC



TREC



TREC

New Testbeds

- Multimedia
 - MPEG-7 collection?
 - Images in the INEX collection?
- Usage-oriented
 - Test-bed of user interactions with DLs?
 - Test-bed framework:
Daffodil

The Daffodil framework



The screenshot displays the Daffodil framework interface, which is a personalized search environment. The main window is titled "Personalized Daffodil- (klas) # db.-4a0bdca6:fb5e0649a6:-7f26@UA". It is divided into several panes:

- Search:** Shows the query "Author=Edward Fox AND Title=digital libraries" and the results. The results are sorted by "Ranked" and show 21 items. The first six results are listed, with the first one being "Thanos (eds.) Costantino; Chris Khoo; Ee-Peng Lim; Schubert Foo; Hsinchun".
- Personal Lib:** A browsing pane showing a tree structure of folders and files. The selected item is "Digital library evaluation by analysis of user retrieval patterns.".
- Results:** A detailed view of the selected paper, showing the title, author(s), journal, and keywords.
- Related Term List:** A list of related terms such as "data mining", "database selection", "digital library", "formal model", and "human computer interaction".
- Author List:** A list of authors associated with the search results, including "Fox", "Douglas Foxvog", "Edward A. Fox", "Eric Foxley", and "Eric Foxlin".

The interface also includes a toolbar at the bottom with various icons for navigation and actions, and a status bar at the bottom showing "ONLINE" and "Attributes: Opening view, done."

Conclusion

- Started activities:
 - DLE infrastructure
 - DLE testbeds and evaluation campaigns
- Workshop as starting point for
 - Survey on existing DL approaches
 - Development of new evaluation model and methods