



# 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 1<sup>st</sup> DELOS NoE  
[http://www.sztaki.hu/delos\\_wg21/metalibrary](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



The screenshot shows the website for DELOS Working Group 2.1. At the top left is a green square logo with 'DELOS' and 'WG-2.1' in a digital font. To its right is the text 'DELOS Working Group 2.1' and 'Evaluation and test environments for digital library research.' At the top right is a green square icon of an open book. Below this is the heading 'MetaLibrary' and the instruction 'If you are searching for a digital library testbed or system:'. This is followed by a list of dimensions: 'general data', 'about users and usage', 'about applied technologies', and 'about data collection'. Below the list is a search box with a 'Search' button and a link to '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..

## Initiative for the Evaluation of XML Retrieval



- **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

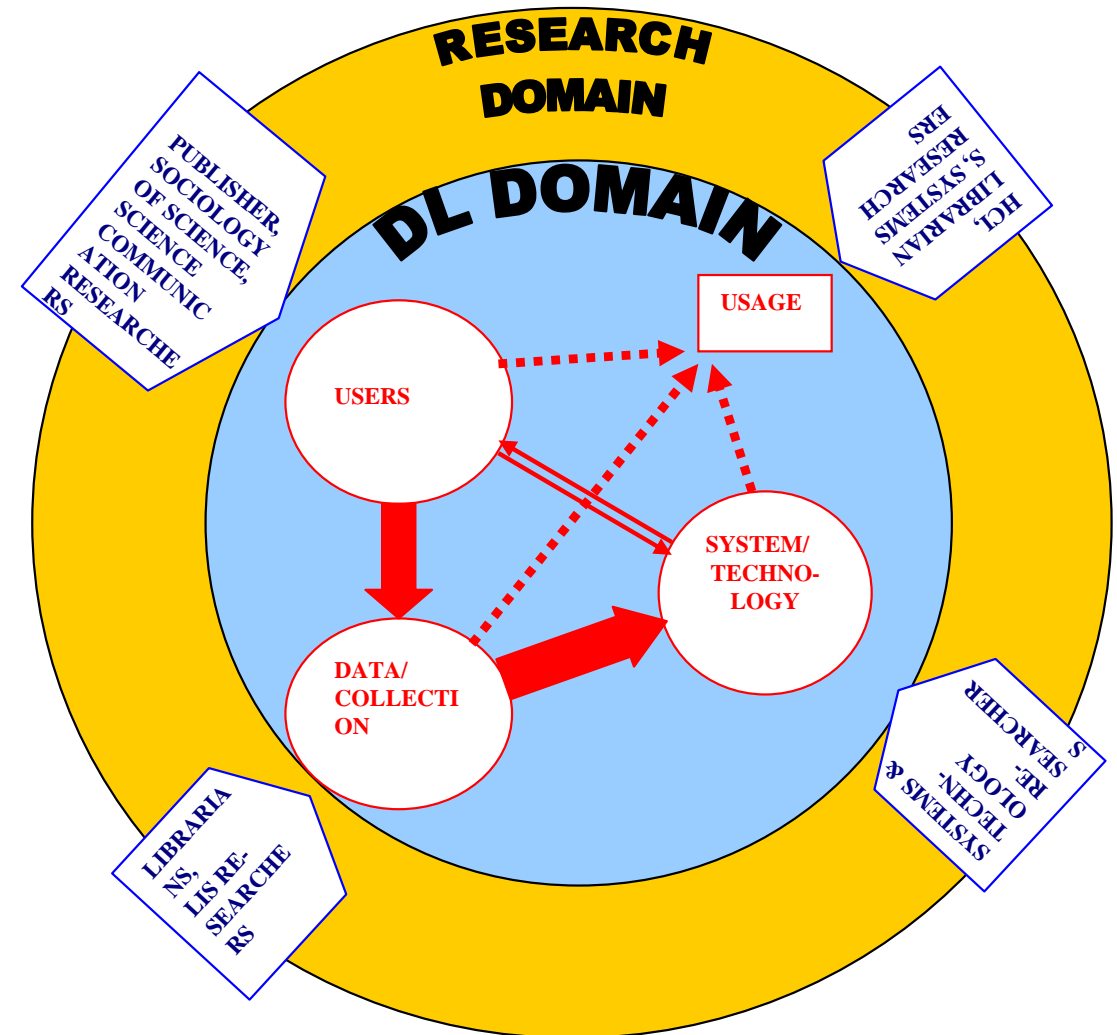
## II. DLE research and development

- 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  
1<sup>st</sup> DELOS  
NoE



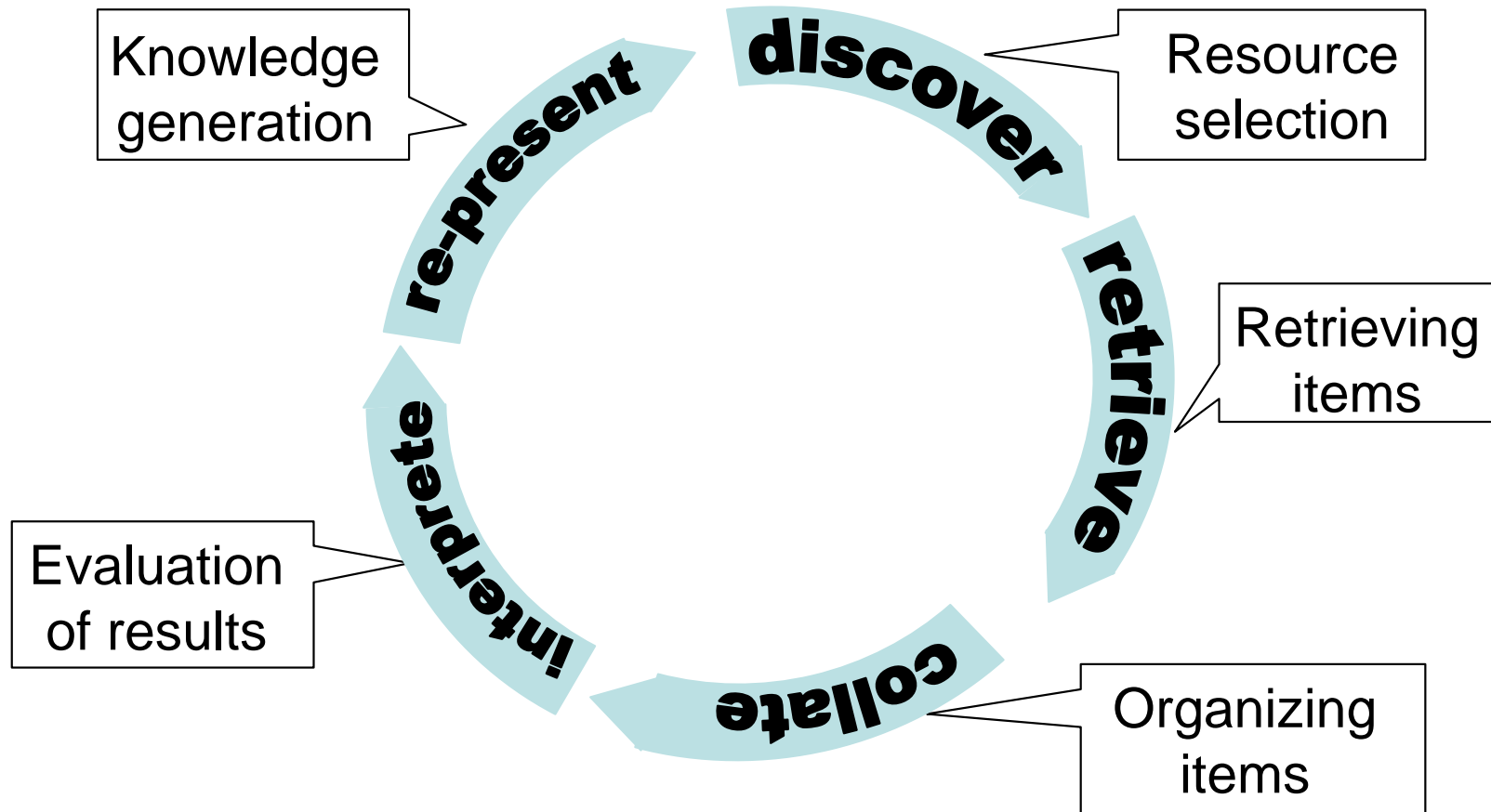




# DL usage

- 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

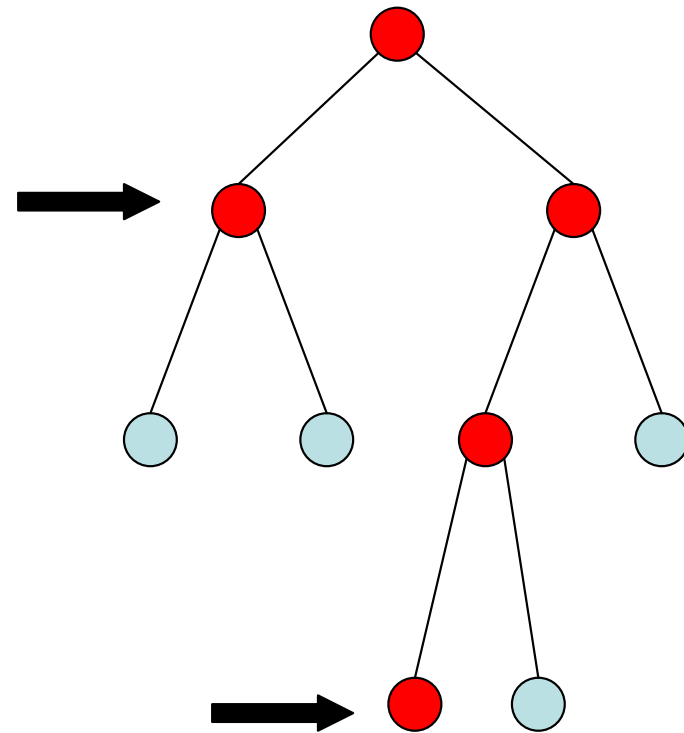


- Participation of users in the evaluation cycle
- Meta-analysis of existing evaluation studies
- Comparison and evaluation of DLE techniques
- Development of new DLE approaches, models and methods

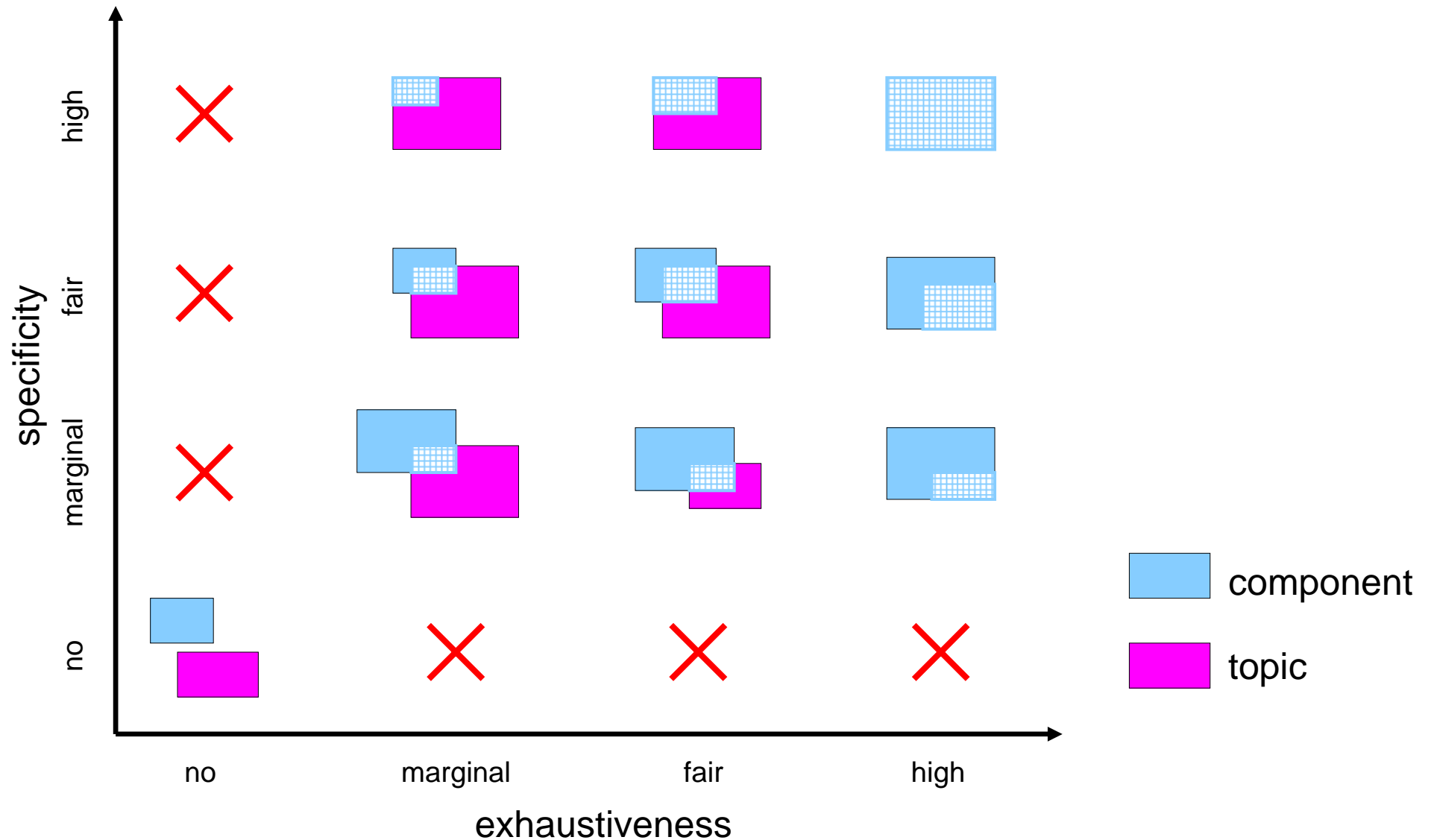
# Example: INEX metrics



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



# INEX relevance scale

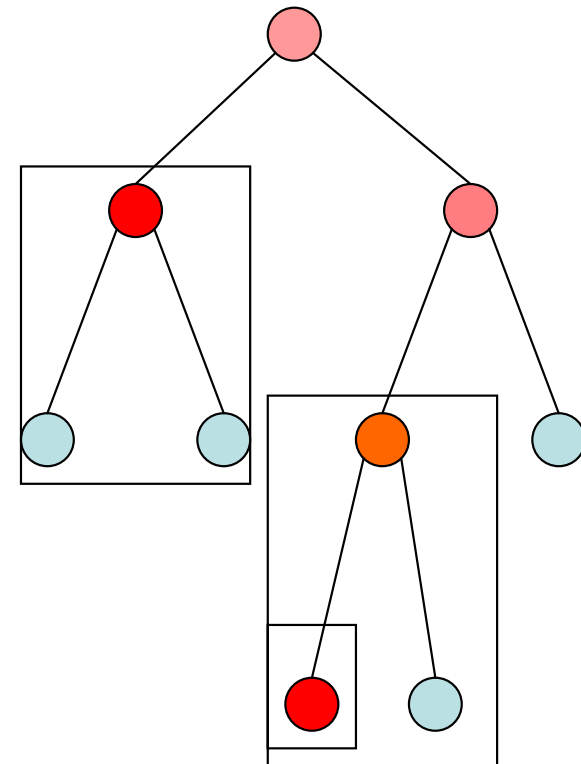


# Multiple answers in a document



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 with the following components:

- Search Window:** Contains the query "Author=Edward Fox AND Title=digital libraries" and shows 21 results. The top results are:
  - Thanos (eds.) Costantino; Chris Khoo; Ee-Peng Lim; Schubert Foo; Hsinchun... *Digital libraries*. (1995) from CompuScience; BibDB; Ieee; ACM.
  - Edward A. Fox. *Digital Libraries of the Future* (1993) from DBLP; ACM.
  - Edward A. Fox; Gary Marchionini. *Digital libraries: Introduction* (2001) from DBLP; Ieee; Achilles; ACM.
  - Edward A. Fox; Constantinos Phanouriou; Neill A. Kipp; Ohm Sornil; Paul Mat... A *digital library* for authors: recent progress of the networked *digital library* (1999) from BibDB; HCIBIB.
  - Pablo A. Roberto; Edward A. Fox; Altigran S. da Silva; Pável P. Calado; Alber... Tools for building *digital libraries*: The Web-DL environment for building *digita* (2003) from ACM.
  - Edward A. Fox. Tutorials: Overview of *digital libraries*
- Personal Lib Window:** Shows a tree view of folders including "DL Evaluation" and "Digital library evaluation by analysis of user retrieval patterns." (highlighted).
- Related Term List:** Lists terms like "data mining", "database selection", "digital library", "formal model", and "human computer interaction".
- Author List:** Lists authors including "Fox", "Douglas Foxvog", "Edward A. Fox", "Eric Foxley", and "Eric Foxlin".
- Document View:** Displays the title "Digital library evaluation by analysis of user retrieval patterns." and author information: "Author(s): Somasekhar Vemulapalli (Query for Name) (Search for Homepage), Johan Bollen (Query for Name) (Search for Homepage), Weining Xu (Query for Name) (Search for Homepage)". It also shows journal information: "Journal: Annals of Mathematics and Artificial Intelligence. In: Agosti, Maristella; Thanos, Costantino (eds.). Lecture Notes in Computer Science. Berlin (Germany): Springer. v. 2458, xvi, 664~p., 432-447. Year: 2002. Number: 1-3. Pages: 121-136." and keywords.
- Taskbar:** Includes icons for Search, Personal Lib, Progress, Networks, Graph, Thesaurus, Conferences, Journals, References, Classification, History, Recommender, Related Terms, Attributes, Export, and Help.

# 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