

# Searching XML Documents Preliminary Work

INEX 2005 Workshop  
28.— 30. November 2005

*Marcus Hassler*  
Abdelhamid Bouchachia

ALPEN-ADRIA  
UNIVERSITÄT  
KLAGENFURT 

# [ Overview ]

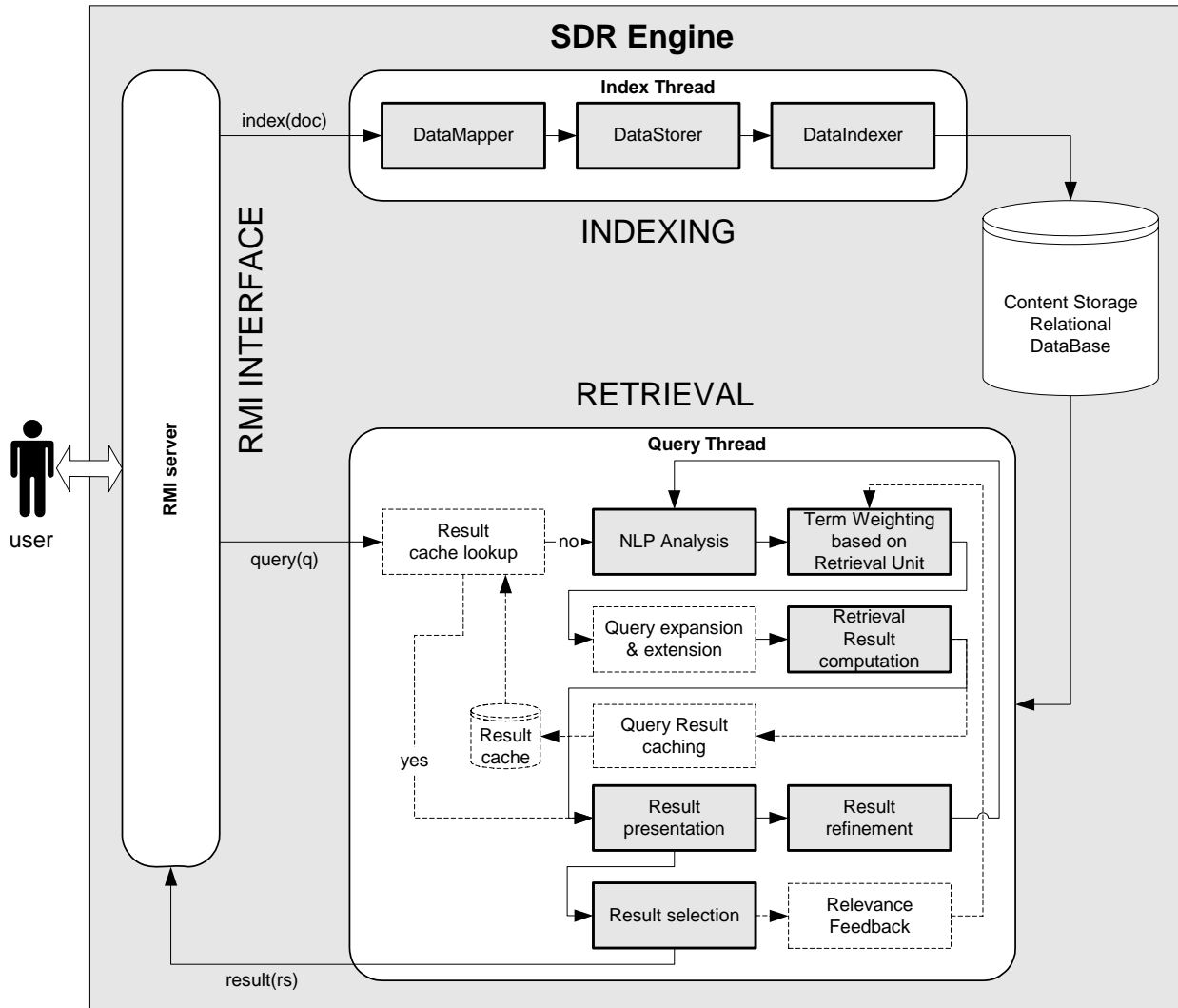
- Introduction
- Architecture
  - Representation
  - Storage
  - Indexing
  - Retrieval
- First Results
- Conclusion

# Introduction

- Context of this work
  - eLearning
  - Infrastructure PlaNet-ET
  
- What is PlaNet-ET
  - Online eLearning – portal [ <http://planet-et.at/> ]
  - HyperWave Information Server
    - courses
    - chunks (learning objects)
  - ➔ structured content + metadata (attributes of objects, LOM)
  
- Relevance of XML representation
  - Extract XML documents from HIS
  - ➔ independent search engine



# Architecture

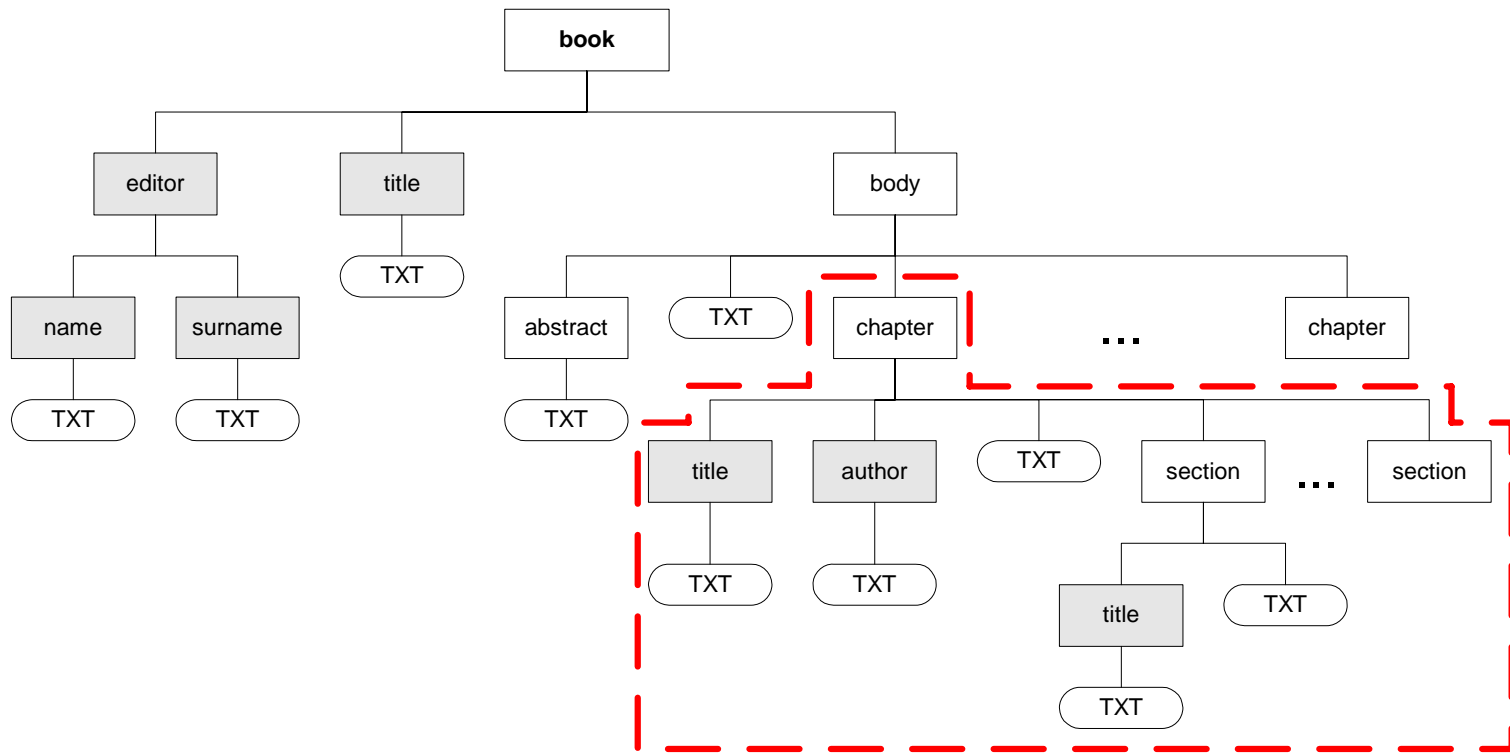


# [ Representation ]

- XSLT transformation → general schema
- Consists of 3 main elements
  - **DOC**
  - **SEC**
  - **FRA** (smallest retrievable unit)
- Each Element
  - **Metadata** - block
  - **Content** - block
- Pros
  - no mixed-content nodes (txt + structure)
  - Semantic relativism (**sec=ss1=ss2=ss3, p=p1=ip=ilrj**)
  - Efficient storage & retrieval

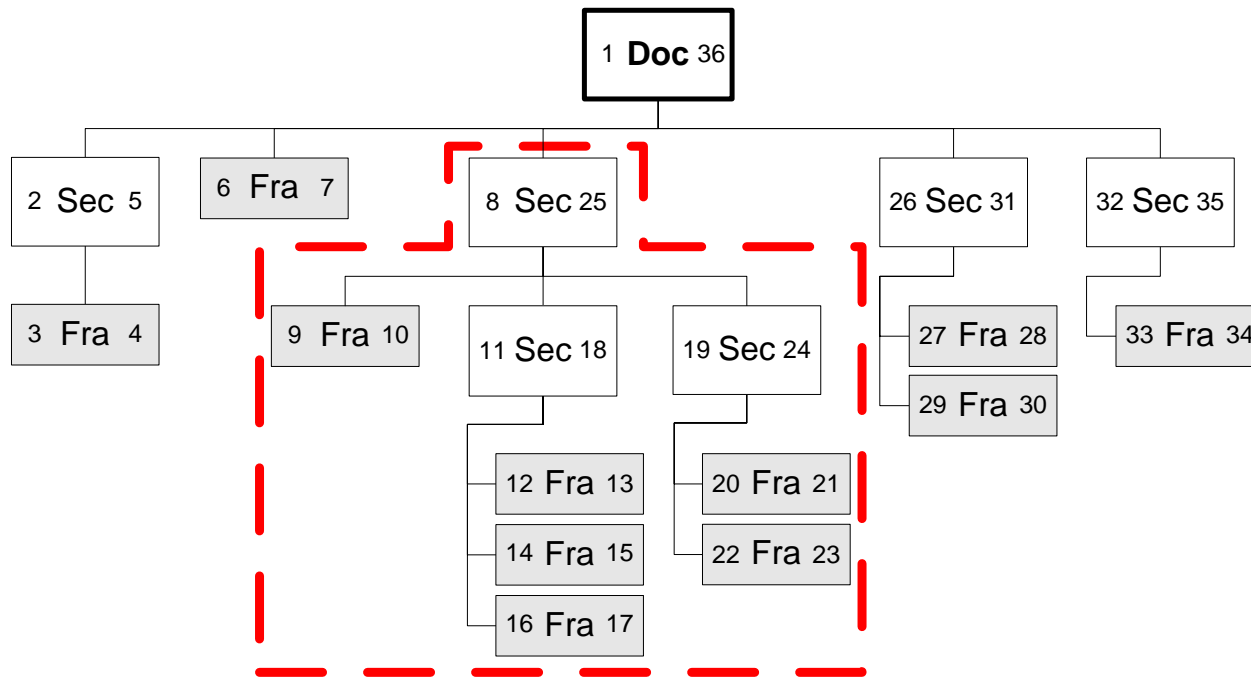
# [ Representation ]

## ■ Example



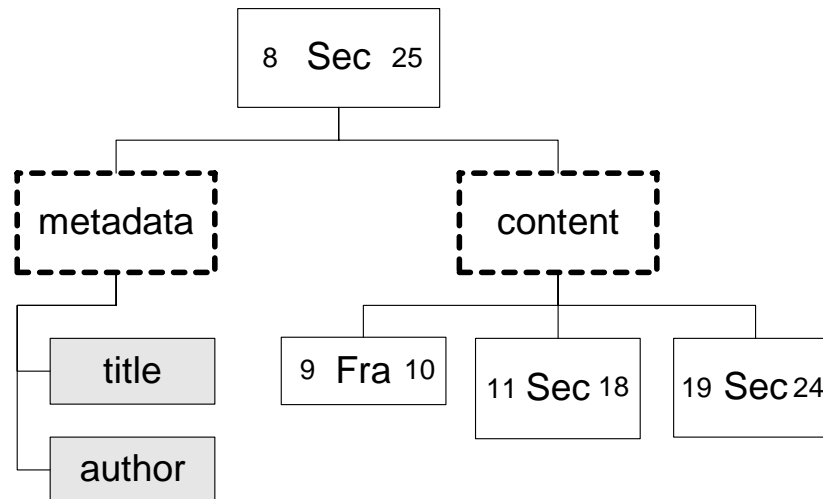
# [ Representation ]

## ■ Example



# [ Representation ]

- Example





# [ Storage ]

- Based on Grust's work

[T. Grust, Accelerating XPath location steps, ACM SIGMOD 2002, pp. 109--120]

- preorder and postorder

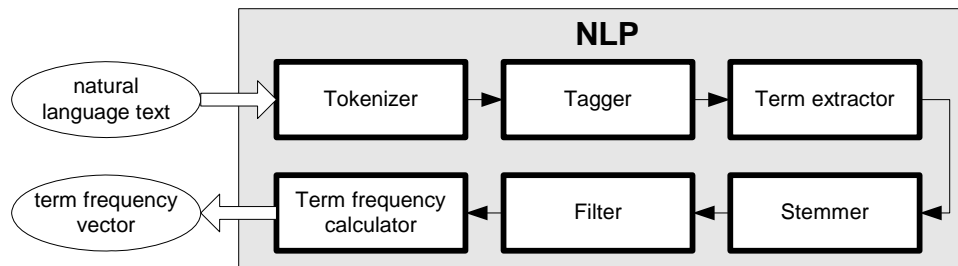
docID	pre	post	parentID	tag	path
1	1	36	0	DOC	/DOC
1	2	5	1	SEC	/DOC/SEC
1	3	4	2	FRA	/DOC/SEC/FRA
1	6	7	1	FRA	/DOC/FRA
1	8	25	1	SEC	/DOC/SEC
1	9	10	8	FRA	/DOC/SEC/FRA
1	11	18	8	FRA	/DOC/SEC/SEC
1	12	13	11	FRA	/DOC/SEC/SEC/FRA
1	14	15	11	FRA	/DOC/SEC/SEC/FRA
...	...	...	...	...	...

docID	pre	cdata
1	3	In this Paper we describe the ...
1	6	In Sec. 1 the architecture is ...
1	9	In our approach we distinguish ...
...	...	...

docID	pre	sec_title	sec_author
1	2	Abstract	R. Smith
1	8	Introduction	J. Alf
1	11	Architecture	NULL
...	...	...	...

# [ Indexing ]

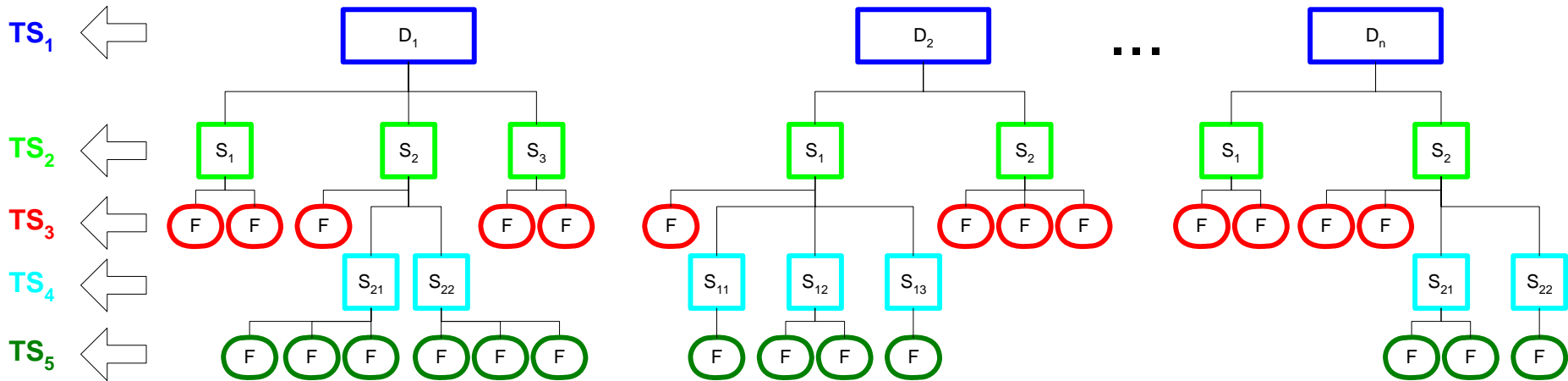
- Content only in leaf nodes (**FRA** element)
- Representation: term-frequency vector



- Pros:
  - Dynamic granularity (dyn. TS + dyn. IDF)
  - Dynamic document environment
    - add, remove, re-index → on-the-fly

# [ Retrieval ]

- Term weighting (VSM)
    - Context of a node:
      - all nodes in the same path
- e.g., /DOC, /DOC/SEC/SEC



- Dynamic term space & *idf* calculation

# [ Retrieval ]

Topic  
No. 265

The screenshot displays the 'Query Interface for SDR' window. At the top, there are three tabs: 'INEX Query' (selected), 'Keyword Query', and 'NL Query'. Below the tabs is the 'General Settings' section with 'Query type' set to 'INEX' and 'Language' set to 'english'. The main area is divided into 'INEX Query' and 'Examples'. The 'INEX Query' section contains five text input fields with 'Assist' buttons to their right. The first two fields contain the queries: '//DOCUMENT[meta(.,documentMeta.title like "%digital libraries%")]' and '//SECTION[about(.,"information retrieval")]'. The 'Examples' section on the right shows three buttons: '01 (~ 5 sec)', '02 (~50 sec)', and '03 (~55 sec)'. Below these is the 'Parameters' section, which includes: 'Maximum results' (1500), 'Minimum similarity (%)' (0), 'Generality (%)' (20), and 'Content importance (%)' (30). Each of these parameters has a slider control. The 'Result type' section at the bottom has two radio buttons: 'unfocussed' (selected) and 'focussed'. At the very bottom are 'OK', 'Reset', and 'Cancel' buttons. A red box highlights the INEX Query input fields and the 'Assist' buttons. Three red arrows point from the 'Assist' buttons to the 'Generality (%)' and 'Content importance (%)' sliders. The 'Generality (%)' and 'Content importance (%)' sliders and their corresponding input fields are circled in red.

# [ Retrieval ]

- Result Computation

- Metadata similarity
- Content similarity
- ➔ Combined to calculate RSV

$$rsv = contSim * ci + metaSim * (1-ci)$$

- User-defined Granularity

$$sim_{new} = sim_{old} * gf + sim_{new} * (1-gf)$$

- Result element: (docID, pre, post, simMeta, simCont)

- Result filtering (Focused task)

- exploit preorder and postorder

# Retrieval

FOCUS @ Alps-Adria University Klagenfurt

Document

Query Result Viewer #2 //DOCUMENT[meta(.,documentMeta.title like "%digital libraries%")]//SECTION[about(.,"information retrieval")]

Result

Query  
 //DOCUMENT[meta(.,documentMeta.title like "%digital libraries%")]  
 //SECTION[about(.,"information retrieval")]

Results: 109 Time: 28329 ms

ID: 0 VecType: static Type: kwd Lang: english MaxRes: 1500 ResType: GenFac: 0.2 ContImp: 0.3 MinSim: 0.0

Document	Element	M sim	C sim	RSV	ID/Pre/Post
File: /root/XMLSearch_cfg/./xml/co/1998/_r1093.xml Title: Content-Based Retrieval in Digital Libraries	Path: /DOCUMENT/SECTION Title:	0.2	0.3092	0.2328	ID: 1861 Pre: 2 Post: 7
File: /root/XMLSearch_cfg/./xml/co/1998/_r1093.xml Title: Content-Based Retrieval in Digital Libraries	Path: /DOCUMENT/SECTION/SECTION Title:	0.2	0.2876	0.2263	ID: 1861 Pre: 41 Post: 50
File: /root/XMLSearch_cfg/./xml/ex/1996/_x3008.xml Title: The role of AI in digital libraries	Path: /DOCUMENT/SECTION/SECTION Title:	0.2	0.2849	0.2255	ID: 4153 Pre: 95 Post: 98
File: /root/XMLSearch_cfg/./xml/co/1998/_r1093.xml Title: Content-Based Retrieval in Digital Libraries	Path: /DOCUMENT/SECTION Title: VIDEO RETRIEVAL	0.2	0.2819	0.2246	ID: 1861 Pre: 34 Post: 57
File: /root/XMLSearch_cfg/./xml/tp/1996/_j0769.xml Title: Introduction to the Special Section on Digital Libraries: Representation and Retrieval	Path: /DOCUMENT/SECTION Title: PAPERS IN THIS	0.2	0.2765	0.223	ID: 10719 Pre: 8 Post: 43
File: /root/XMLSearch_cfg/./xml/co/1998/_r1093.xml Title: Content-Based Retrieval in Digital Libraries	Path: /DOCUMENT/SECTION Title: OBJECT RETRIEVAL	0.2	0.2419	0.2126	ID: 1861 Pre: 8 Post: 17

Preview

Subdocument search and indexing

Libraries index entire documents, not sections or passages within. However, in **information retrieval** there is significant work on finer-grain indexing, which should be a natural part of a digital library to provide higher-fidelity information access.

Previous 20 results    Viewing results 0 - 19    Next 20 results    View document

# Retrieval

The screenshot displays the FOCUS @ Alps-Adria University Klagenfurt interface. The main window title is "Document - /root/XMLSearch\_cfgf../xmlVex1996f\_x3008.xml". The interface is divided into several panels:

- Left Panel (Tree View):** A hierarchical tree of document sections. The selected section is "SEC: Subdocument search and indexing". Other visible sections include "DOC: The role of AI in digital libraries", "SEC: Information agents: A new challenge", "SEC: The digital library as a community", "SEC: Digital librarians: beyond the digital library", "SEC: The digital reference librarian", "SEC: The digital acquisition librarian", "SEC: Beyond traditional library functions", "FRA: text", "SEC: More sophisticated information", "SEC: On-demand document summarization", "SEC: Multidocument summarization", "SEC: Multimedia search and indexing", "SEC: Access to live or near-live information", "SEC: Active information sources", "SEC: Symbiotic human-machine systems", "SEC: Information on tap, anywhere", "SEC: The universal library", "SEC: References", "SEC: APPENDIX -- Interesting URLs", "SEC: APPENDIX -- Coming in August", and "SEC: Curricula vitae".
- Top Panel (Element Meta Info):** Displays metadata for the selected section:
  - InternalID:** 4153
  - Preorder:** 95
  - Postorder:** 98
  - Title:** Subdocument search and indexing
  - Type:** SECTION
  - Sourcepath:** /article[1]/bdy[1]/sec[6]/ss1[1]
- Right Panel (Sub-Element Info):** Shows a summary of the section's content:
  - SECTIONS (1):** sections (0), subsections (1), subsubsections (0), subsubsubsections (0)
  - FRAGMENTS (1):** texts (1), text lists (0), figures (0), tables (0), formulas (0), codes (0), definitions (0), theorems (0), proofs (0), references (0)
- Main Content Area:** Displays the text of the selected section: "Subdocument search and indexing". The text reads: "Libraries index entire documents, not sections or passages within. However, in information retrieval there is significant work on finer-grain indexing, which should be a natural part of a digital library to provide higher-fidelity information access." The phrase "information retrieval" is highlighted in blue.
- Bottom Panel:** A yellow area containing a series of vertical bars of varying heights, representing a visualization of the document's structure or content. The second bar from the left is highlighted with a red border.

# First Results

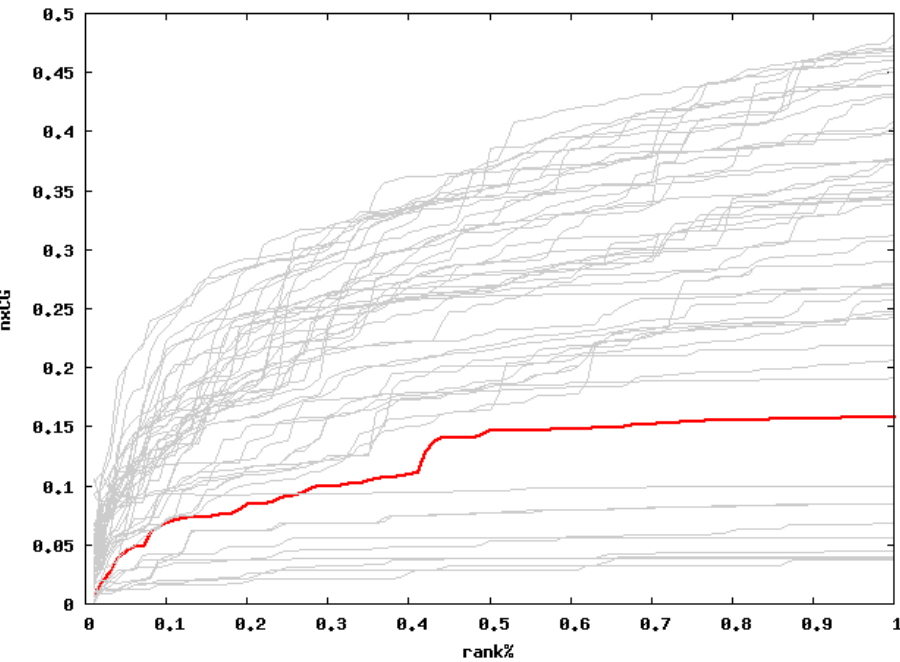
- Used collection v1.4 (12107 docs)
  - missing 4712 docs from v1.8 (16819 docs)
- XML transformation (XSLT)
  - ~ 0,42 % errors (51 docs) → manually fixed
- Topic transformation
  - exchange element/path names  
e.g., `/article` → `/DOC`
  - map conditions for metadata  
e.g., `about(../atl,"information retrieval")` →  
`meta(/DOC,title like '%information retrieval%')`
- Difficulties:
  - Some topics (215, 231, 241) contain “and”, “in”, “for”, “+”, ...
  - Path mapping:  
`/article` → `/DOC`  
`/article/bdy` → `/DOC`



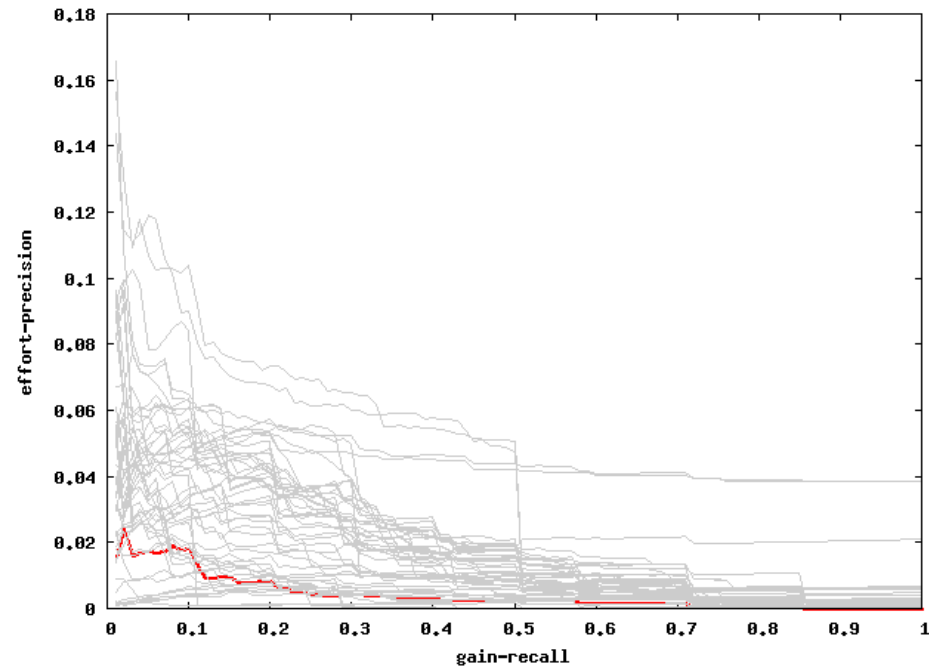
# [ First Results ]

## ■ CO.Thorough

ncxG (overlap=off,strict)

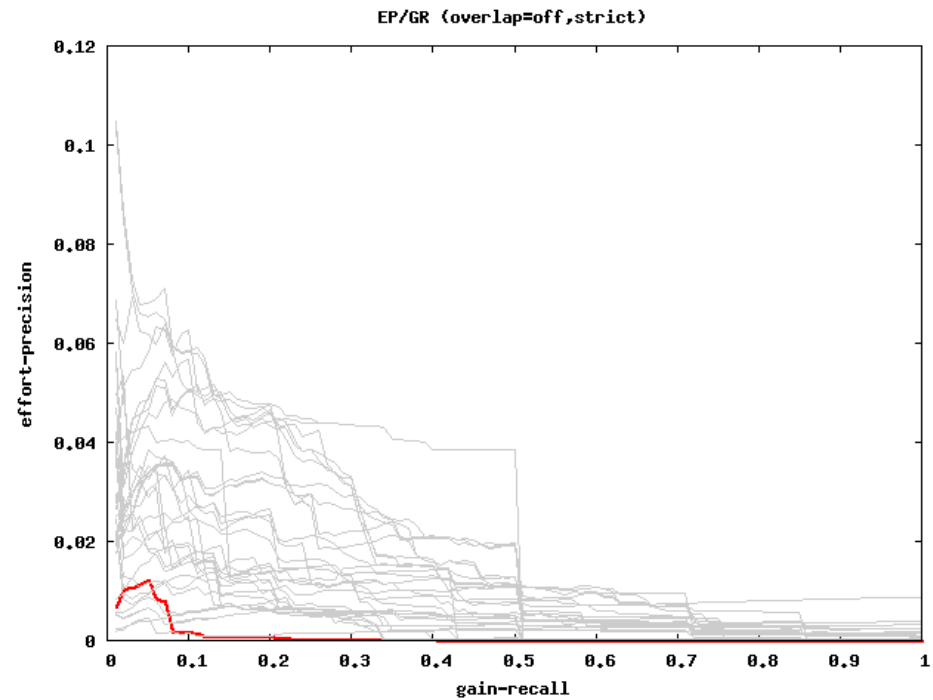
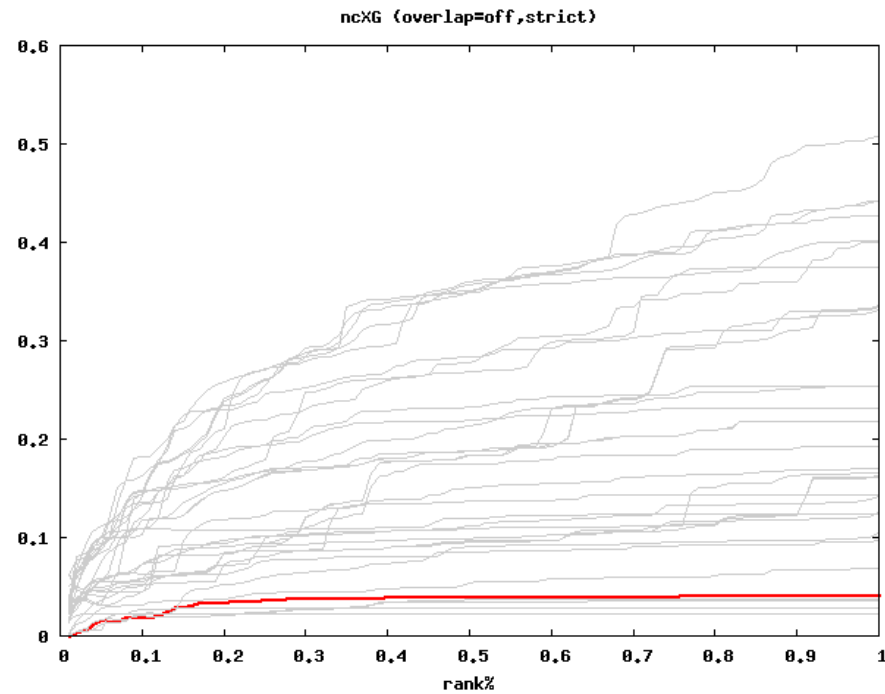


EP/GR (overlap=off,strict)



# [ First Results ]

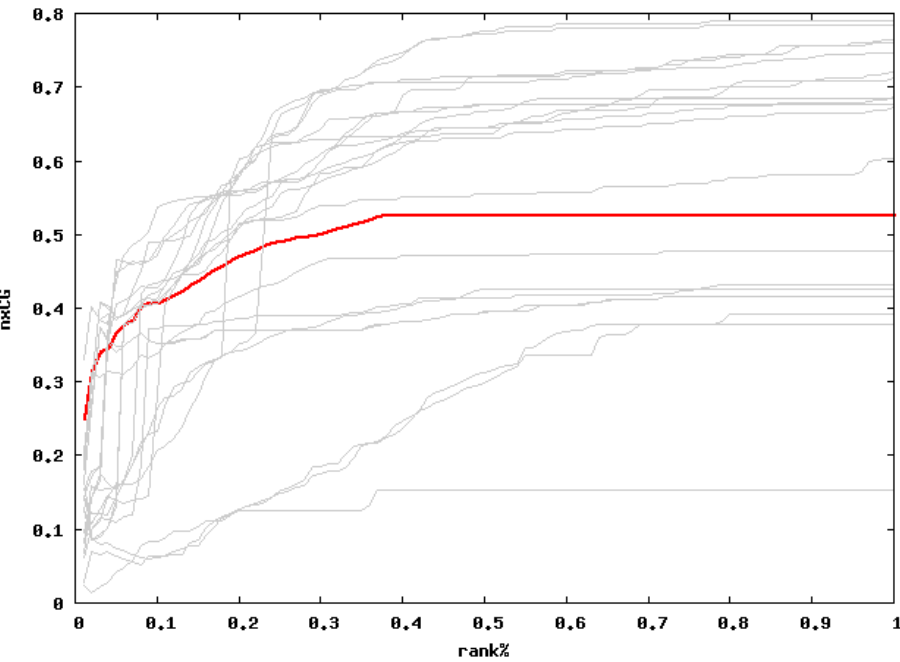
## ■ COS.Thorough



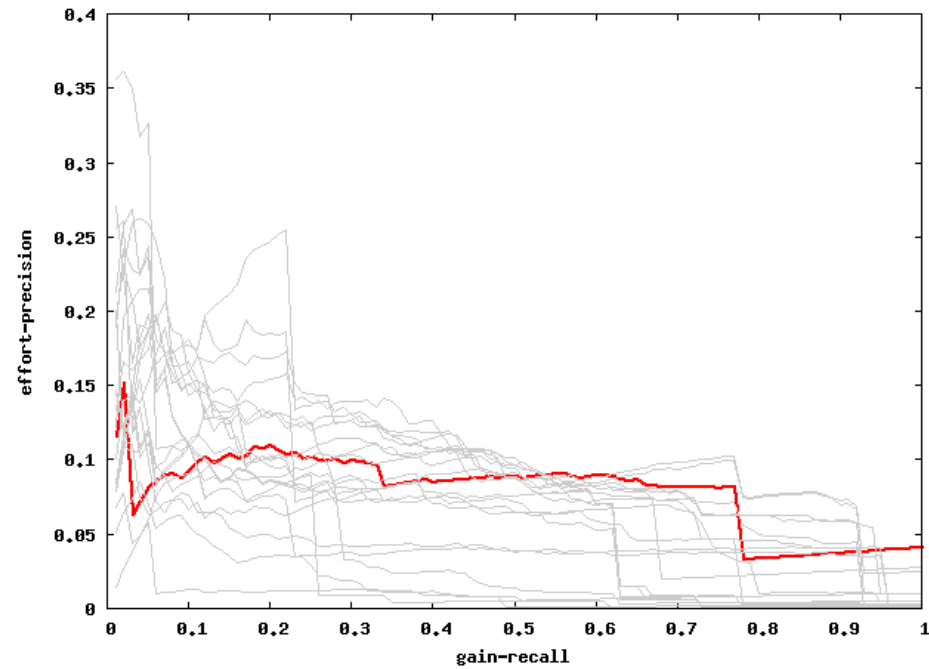
# [ First Results ]

## ■ SSCAS

ncXG (overlap=off,strict)



EP/GR (overlap=off,strict)

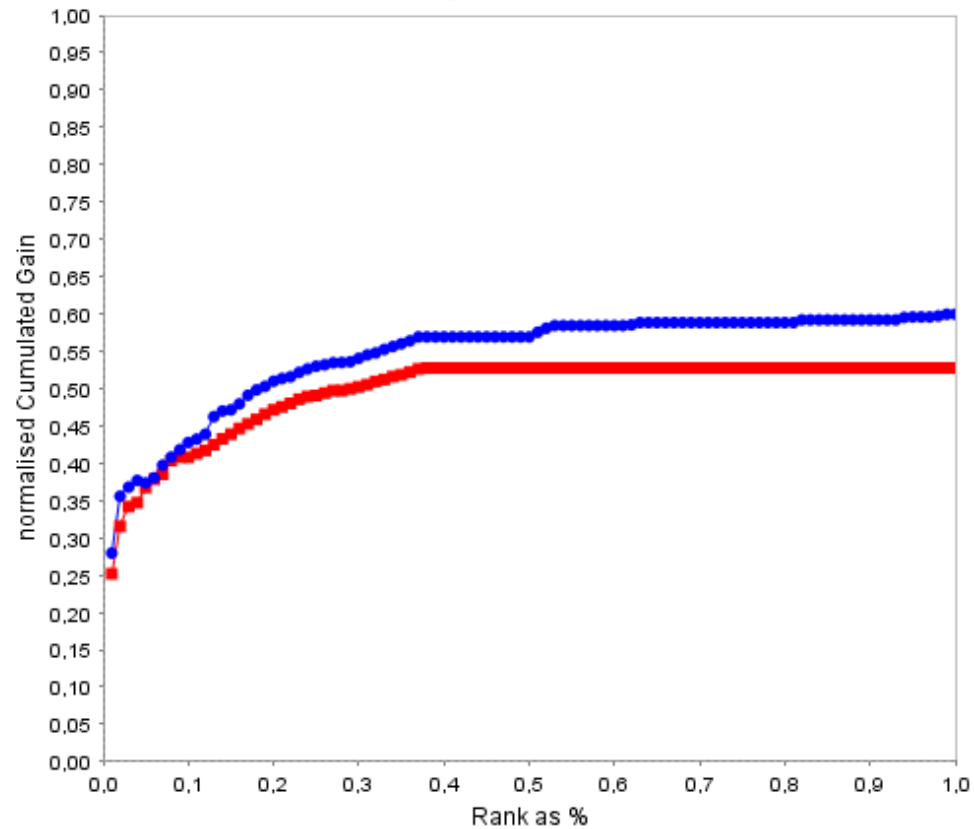


# [ First Progress ]

## ■ SSCAS

Metric: nRnxCg Overlap: off Quantisation: strict

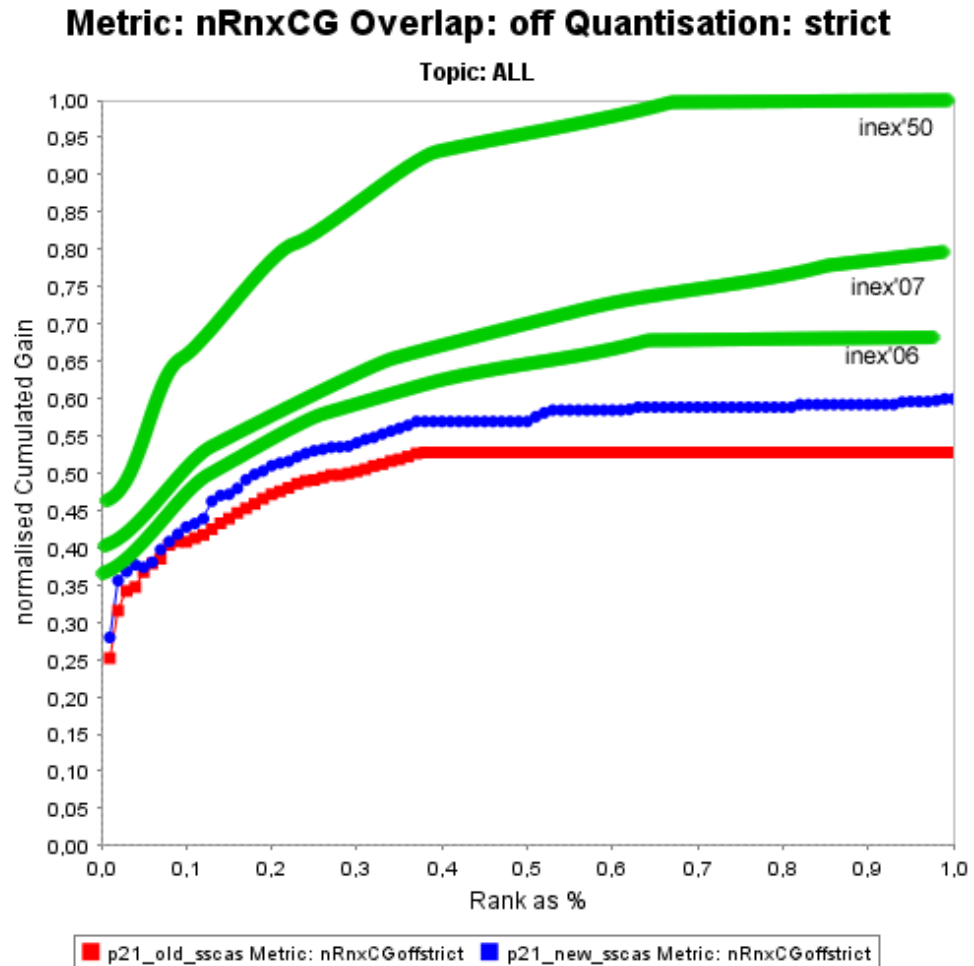
Topic: ALL



■ p21\_old\_sscas Metric: nRnxCgoffstrict ■ p21\_new\_sscas Metric: nRnxCgoffstrict

# First Progress

## SSCAS



# [ Conclusion ]

- Independent system architecture
- Dynamic aspects
  - Term frequency propagation
  - Term space
  - *idf* calculation
- Initial experiments
  - Only 3 runs (CO.Thorough, COS.Thorough, SSCAS)
  - Not too bad results (SSCAS)
- Further work
  - Adapt INEX corpus 1.8
  - Improve metadata typing and matching
  - Include multi-term index (e.g., information retrieval)
  - Complete Evaluation