Short Presentation of my Previous Research Work & Future Directions

Information Retrieval of Text, Structure and Sequential Data in Heterogeneous XML Document Collections

Talk @ The Leesklub INS2 group research meeting @ CWI

Eugen.Popovici@cwi.nl

March 12th 2009

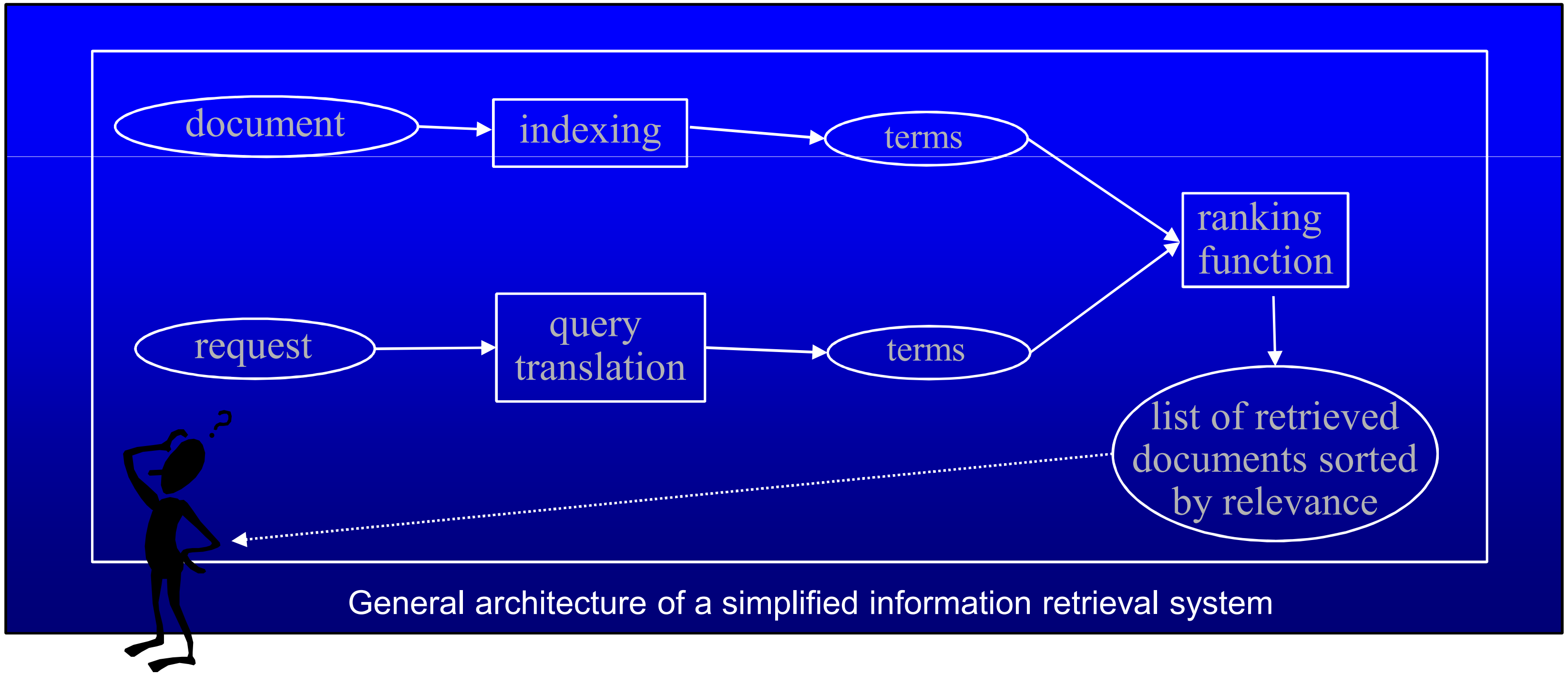
UBS UNIVERSITÉ DE BRETAGNE-SUD

VALORIA

CWI Centrum Wiskunde & Informatica

Information Retrieval

- **Information Retrieval (IR)** deals with the representation, storage, organization of and access to information items [Baeza-Yates 99].



General architecture of a simplified information retrieval system

XML Multimedia Documents

- Nowadays digital documents represent a complex mixture of meta-data and multimedia information
  - Textual and sequential/time series data are an ubiquitous form of representation in many scientific, medical, financial applications…
  - XML = *de facto* standard for the data-exchange and presentation of documents
- XML Multimedia Documents = complex documents integrating **structure**, **text**, and ... **sequential/time series data**

  Library of Congress Collection, Scientific articles (INEX IEEE collection)

  Medical records, Annotated biological databases (SwissProt DB)

  Musical pieces (MusicXML, MidiXML), Multimedia Descriptions (MPEG7-DLL)
- Increasing volumes of data

Challenges…

(Among the) challenges in indexing and searching collections of XML documents with heterogeneous structures and multimedia content

***(C1)* Answer multi-criteria approximate requests**

- having an incomplete, imprecise or even erroneous knowledge about both the structure and the content (text and sequential data) of the documents

***(C2)* Provide focused access to relevant information**

- by pointing the user to the **appropriate locations** within the documents and within the multimedia parts of these documents

***(C3)* Process large volumes of documents**

- by using specialized hardware to accelerate the access and processing of data

Outline

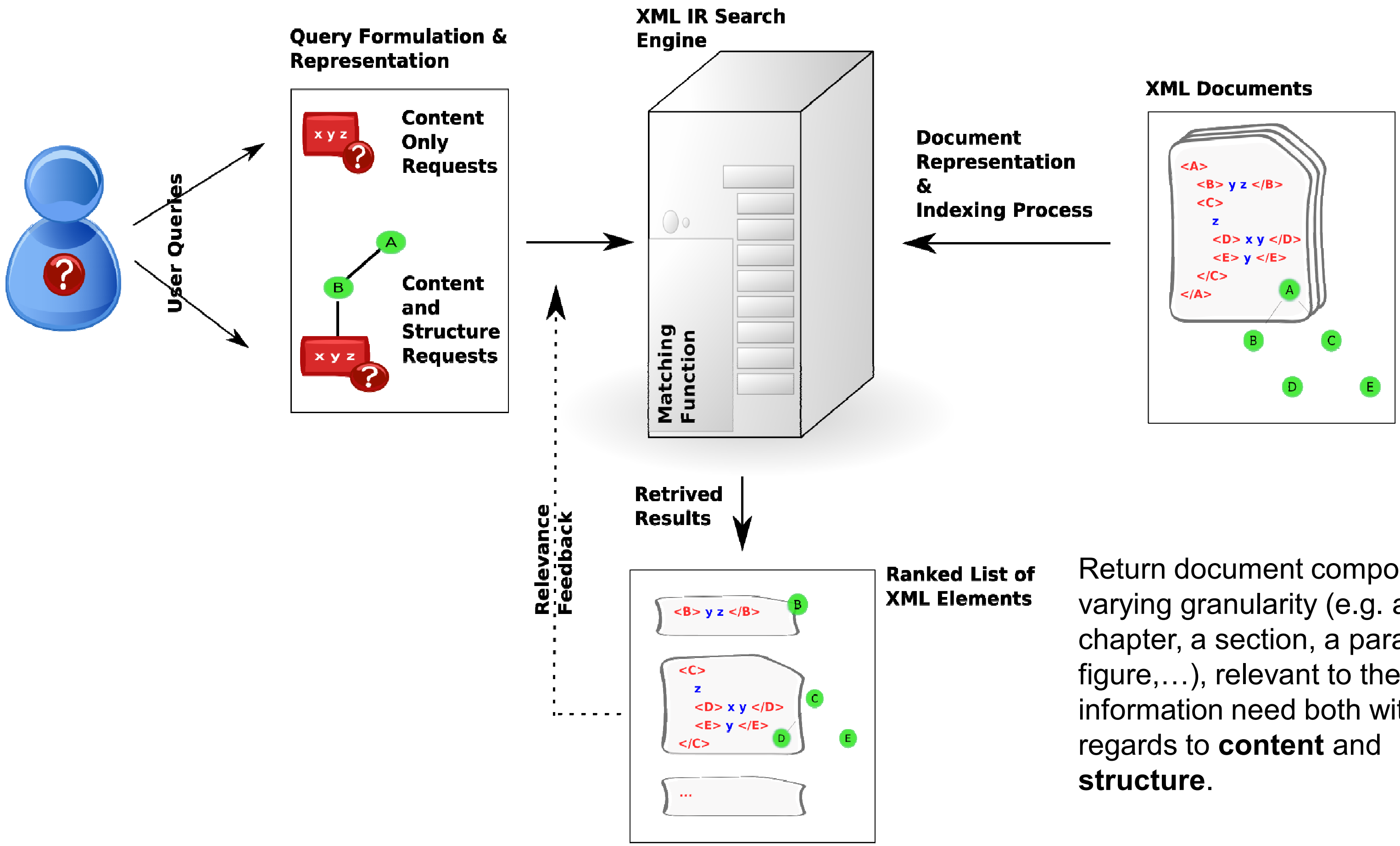
1. **XML Information Retrieval**
   Structure management, Focused access

2. **XML Multimedia IR**
   Sequential data

3. **XML IR on Specialized Hardware**
   Hardware accelerator

4. **Summary & Future Work**

Conceptual Model for XML IR



Return document components of varying granularity (e.g. a book, a chapter, a section, a paragraph, a figure,…), relevant to the user's information need both with regards to **content** and **structure**.

Content Oriented XML IR & Structural Constraints Interpretation

A heterogeneous XML Database
- Different sources
- Different structures
- Different uses

document-centric
view of the db,
XML IR perspective

- ranked answers
- no predefined unit of retrieval
- overlapping results
- the user may have an incomplete or imprecise knowledge about both
  - the structure and
  - the content

  of the XML documents. [XIRQL, NEXI]

data-centric
view of the db
- no ranking
- strict answers on both structure and content

[XQuery, XPath]

data-centric view & ***IR*** techniques for ranking the textual content
[XQuery & XPath Full Text]

We present & evaluate an XML retrieval scheme that manages two levels of approximation:
- On the XML structure
- On the textual content

Path Approximate Matching

Structural constraints are interpreted vaguely as “structural hints”

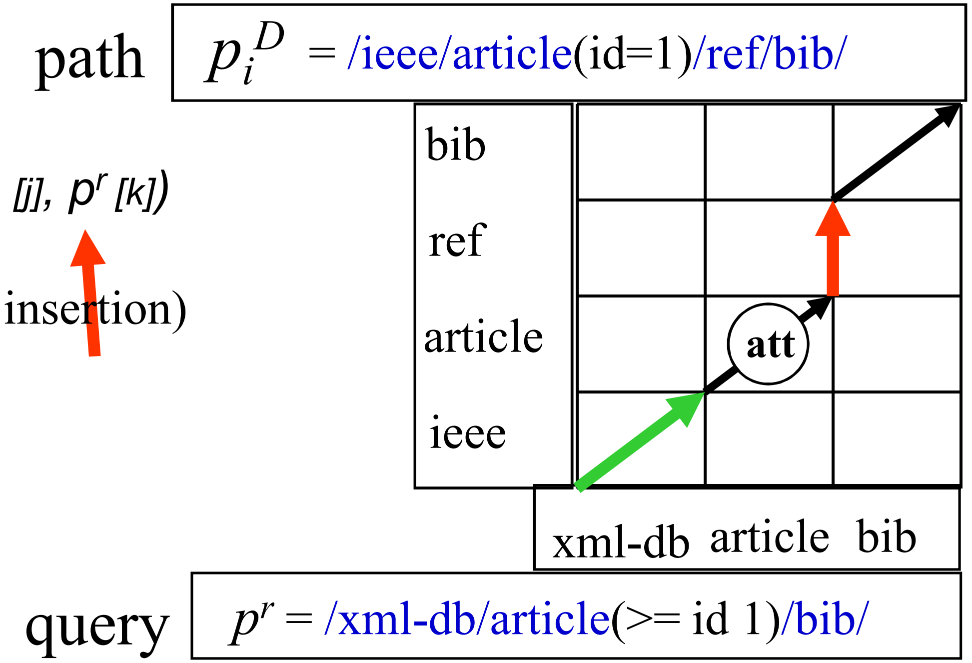
[Trotman and Lalmas, SIGIR06]

Levenshtein Editing Distance: compares two Strings S_1 and S_2 and finds the minimal set of transformations (substitution, insertion, deletion) to get from S_1 to S_2: the result is the sum of the transformation cost. $\delta_L(\mathit{life}, \mathit{likes}) = 2$

Wagner&Fisher Algorithm: O(nm), n and m respectively lengths of S_1 and S_2.

$$\delta_L(p_i^D, p^r) = \min GC\ \Phi\ (p_i^D[j], p^r[k])$$

$$= (1\ \text{substitution} + 1\ \text{insertion})$$



Data Model: XML tree represented by the set of its root-to-leaves paths.

Motivation: reducing the complexity of the tree alignment algorithm to a path alignment problem.

Weighting Strategies

$C_{subst}(elm^R, elm) = \xi$, (att)

$C_{insert}(elm) = \xi$

$C_{delete}(elm) = 0$

(att) $C_{AttCond}(att) = 0.5 \cdot \xi$

Models an end user having *precise* but *incomplete* information about:

- the xml tags,
- their attributes conditions and
- their ancestor-descendant relationships.

Example of distances between the indexed path p_i^D and the request path p^R:

δ_L(/book /sec/template/p,
//**article** //template) = ξ

δ_L(/**article**/sec/template**(name='book')**/p,
//**article** //template**(OR (== @name book)** (== @name reference) **)** = 0

δ_L(/**article**/sec/template**(name='book')**/p,
//**article** //template(OR (== @name author) (== @name reference)) = 0.5·ξ

Normalizing and Aggregating Matching Scores

- **Normalized structural similarity**

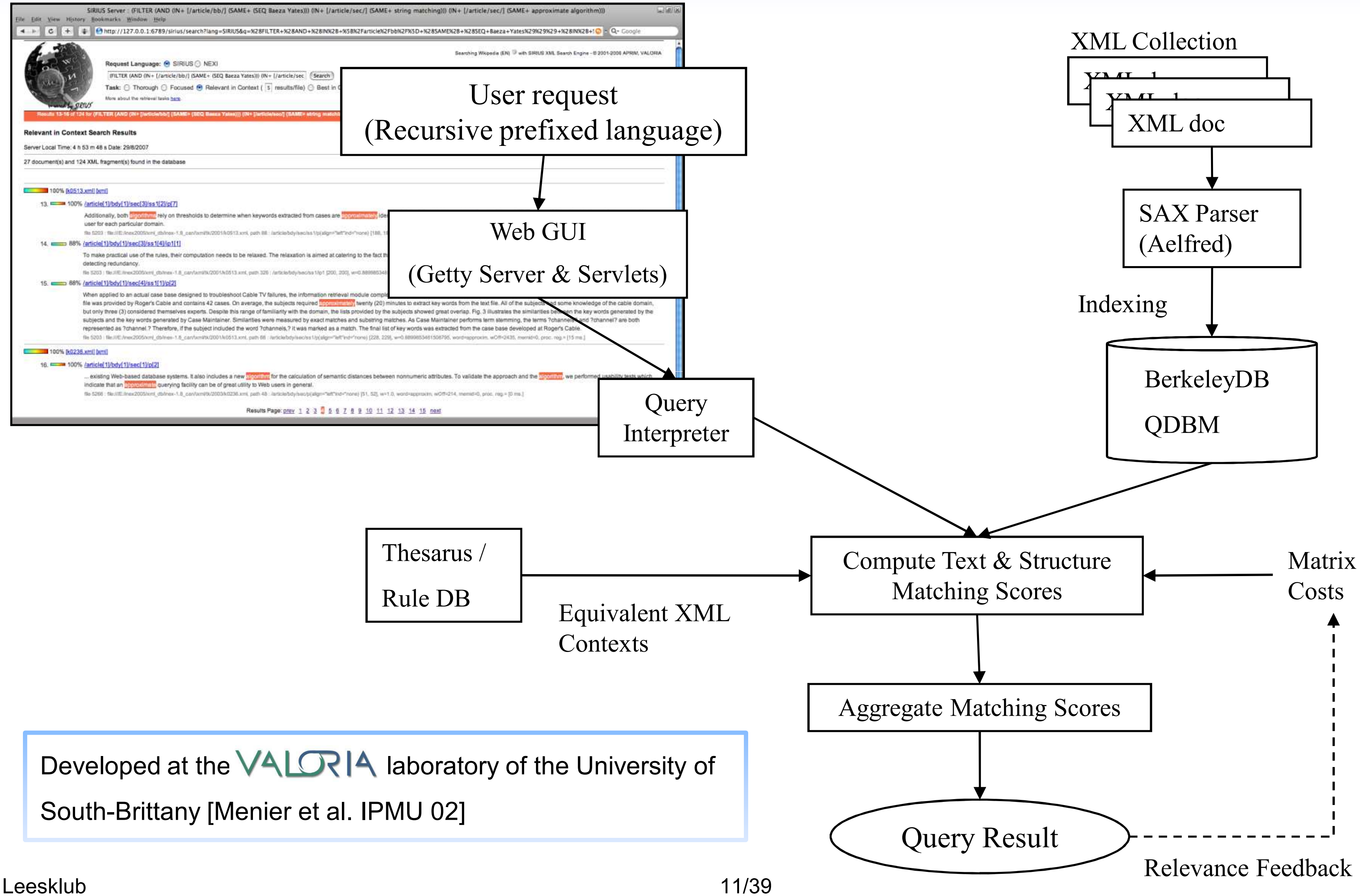
$$\sigma_{struct}(p_i^D, p^r) = 1/(1+ \delta_L(p_i^D, p^r))$$

takes values between 1 for perfect match and → 0 for lowest similarity scores

- **Textual Ranking**
  - Indexing disjoint elements [XFIRM, GPX]
  - Returning the most focused elements (i.e. leaf elm.)
  - Ranking function based on the vector space model

- **Merging Structure and Content Matching Scores**
  - weighted linear aggregation between the conditions on structure and content match.

SIRIUS XML IR System Architecture



Developed at the VALORIA laboratory of the University of South-Brittany [Menier et al. IPMU 02]

INEX Evaluation Campaigns

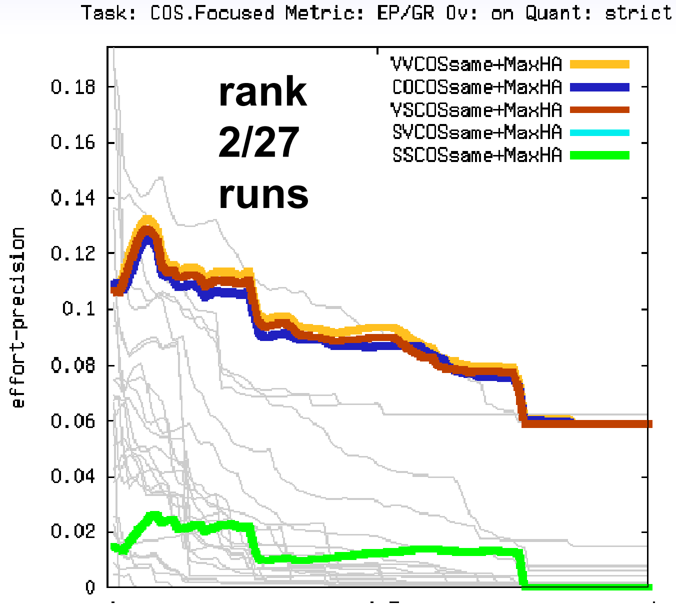
Initiative for the evaluation of XML Retrieval

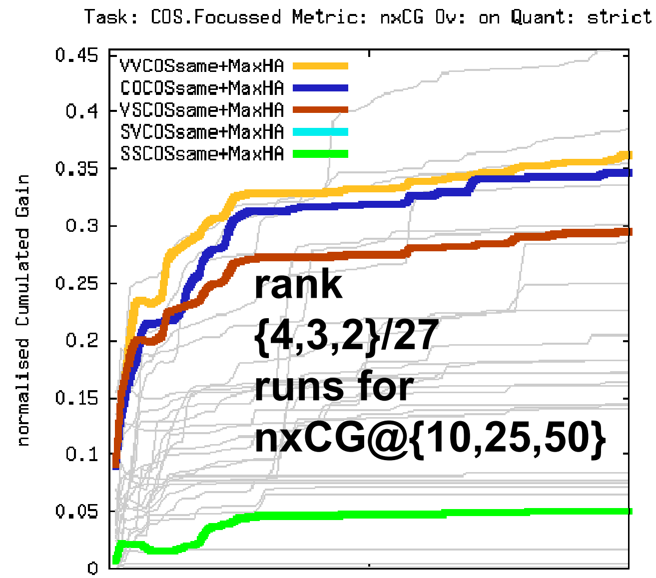
- **Datasets**
  - inex-1.8 IEEE collection: 16819 doc, 11M elem, 748 MB
  - Wikipedia XML collection: 659,388 doc, 30M elem, 4.6 GB
- **Requests (content only-CO & content and structure-CAS)**
  - 40 CO + 47 CAS topics (INEX 2005)
  - 110 CO & CAS topics (INEX 2006)
- **Tracks & Retrieval Tasks**
  - **Tracks: Ad Hoc, Multimedia, Heterogeneous, Passage Retrieval…**
  - Ad HocTasks: **VVCAS** (Thorough)**, Focused, BestInContext**
- **Pertinence Judgments**
- **Evaluation measures**
  - nxCG (user oriented)
  - ep/gr (system oriented)
    - overlap = off/on, quantization = strict (only fully specific & fully exhaustive elem.)

Approximate Structural Match for XML IR

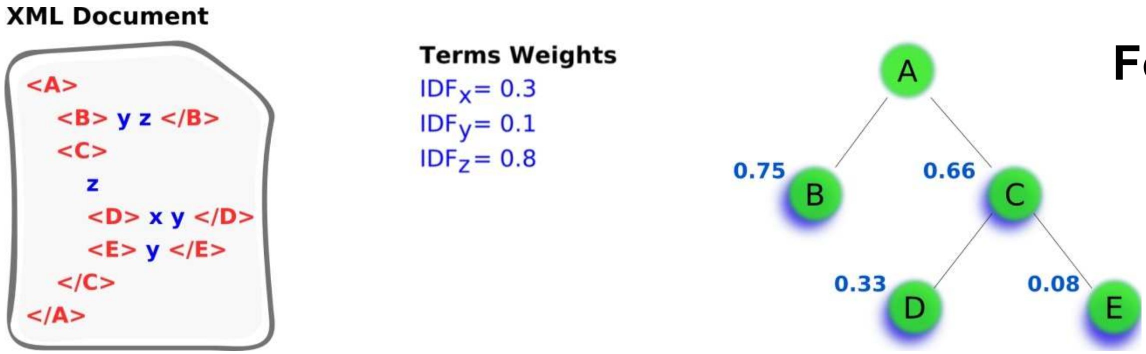
INEX 2005 Focused Task (ov=on, quant.=strict)

- A vague interpretation of the structural constraints can highly improve the quality of the retrieved results versus a strict interpretation
- Taking the structural hints into account may increase the system retrieval performances
  - This was not confirmed (in average) by the results on the INEX 2006 Wikipedia collection…
- Encouraging results relative to current state of the art XML IR systems
  - good quality evaluation results for the top 50 first ranked answers





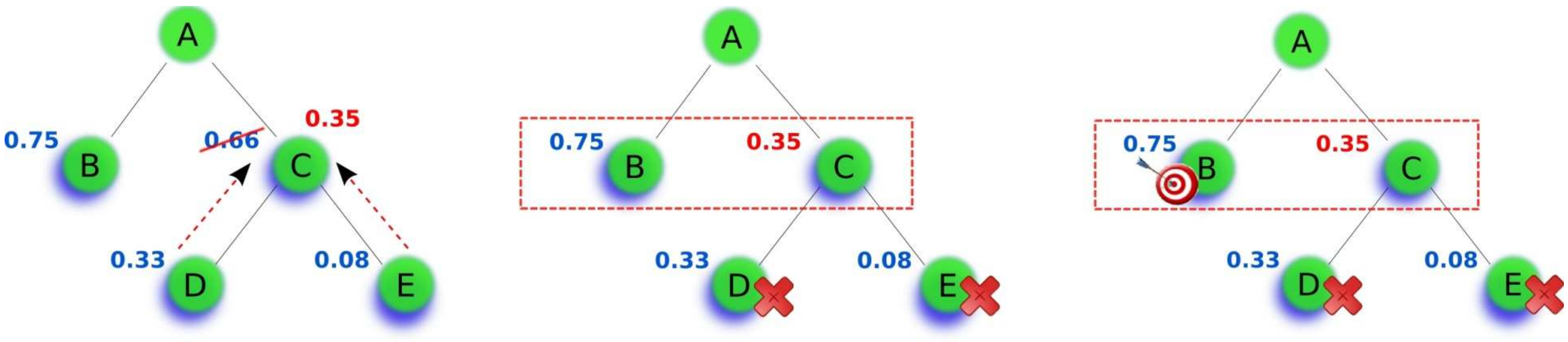
Detecting Best Entry Points in XML Documents



Focused Retrieval Strategy

– no overlapping elm. allowed

XML document with mixed content and term weights



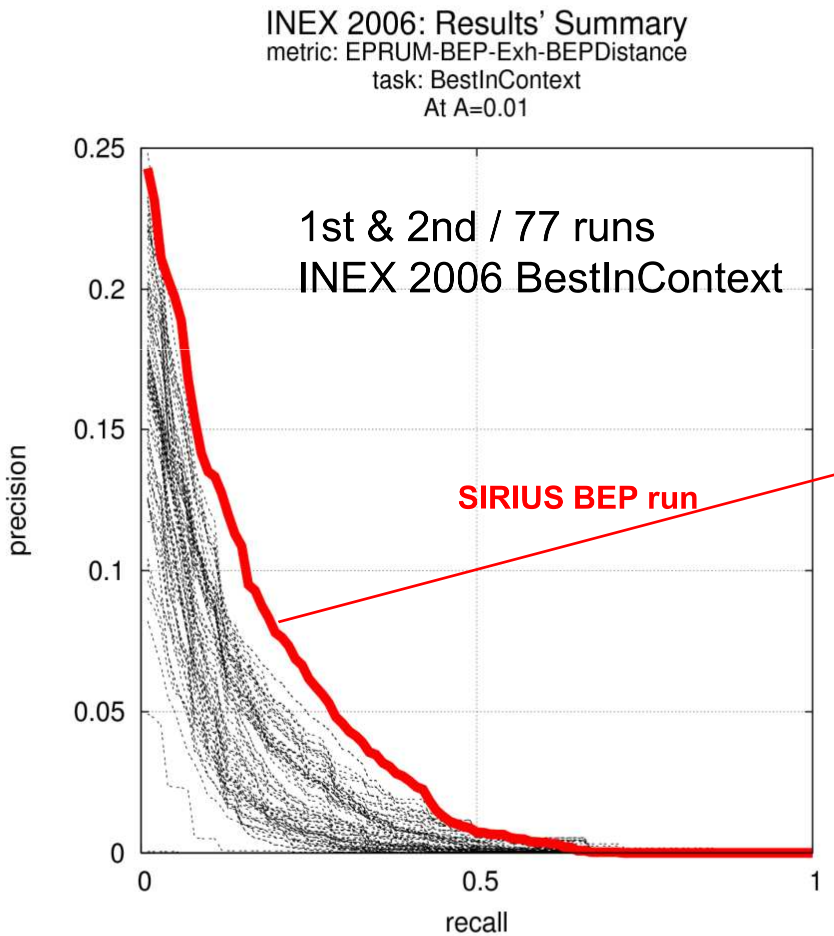
a) Bottom-up Elm. Score Aggregation (AVG)

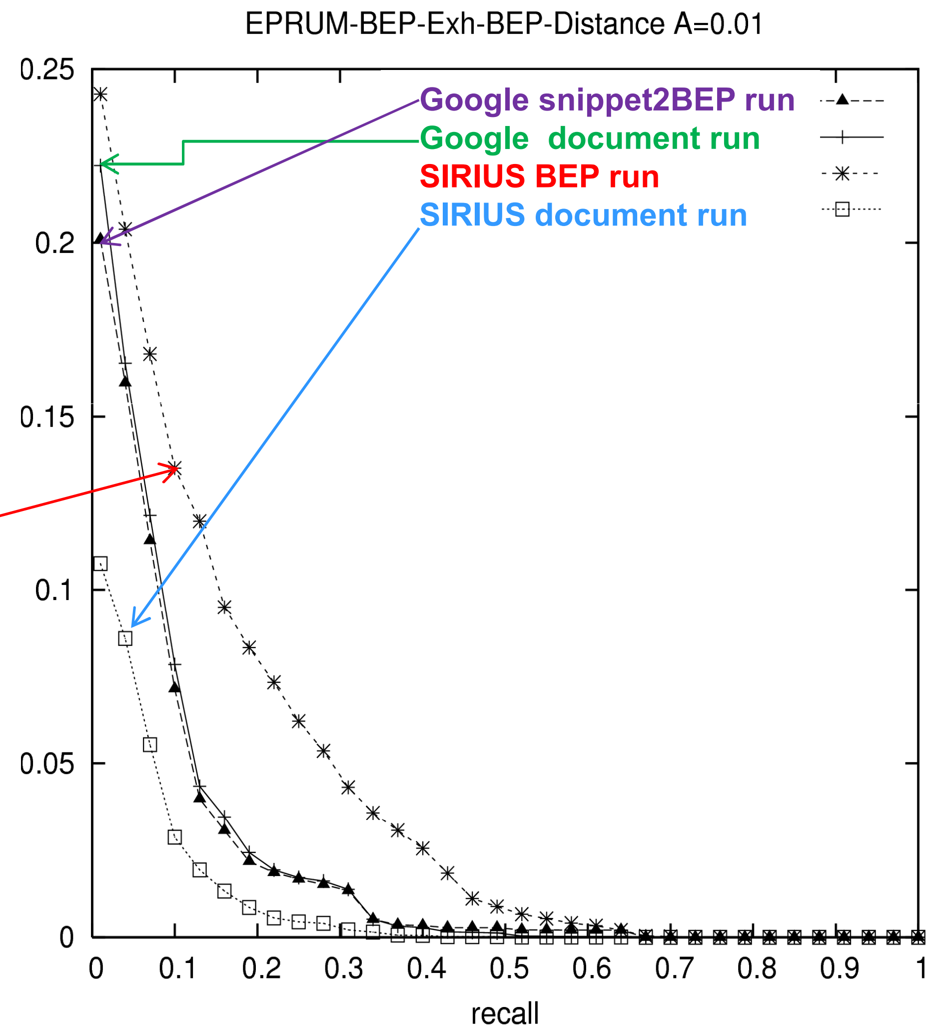
b) Removing Overlapping Elements (MRD) [BruteForce Filtering]

c) BEPs Selection Heuristic

Sort articles by the highest scoring elm.

BEPs Retrieval Strategy vs. Document Retrieval & Current (XML) Search Engines Technology





CWI

Experimental Results

- Effective strategy for detecting BEPs within a document, but less effective for document ranking
  - Emphasizing the weight of the most specific non overlapping elements with relevant content
    - ~ start of relevant textual content [Trotman and Lalmas, SIGIR07] and "Start Reading Here" BEP type [Kazai and Ashoori, 06]
  - Encouraging results: 1st & 2nd/77 runs for the INEX 2006 *BestInContext* task…(evaluated with A=0.01).
  - Compared with current 'flat' Web search engine technology and *document snippets* approximated to BEPs approach.

Outline

1. **XML Information Retrieval**
   Structure management, Focused access

2. **XML Multimedia IR**
   Sequential data

3. **XML IR on Specialized Hardware**
   Hardware accelerator

4. **Summary & Future Work**

Motivation

- Sequential data = an ubiquitous form of representation in scientific, medical, financial applications…and in XML document collections

Different **structures**
Different types of **sequential data**
Different uses

→ ***document-centric*** view
XML IR perspective

⇩

the user may have an incomplete or imprecise knowledge about both the **structure** and the **content** of the XML documents [Fuhr,TIS04].

⇩

Extract, index and retrieve heterogeneous types of sequential data embedded in heterogeneous collections of XML documents.

By using two levels of approximation:
1) on the structural organization of the sequential data (XML IR)
2) on the sequential data content

Sequential Data & XML Context

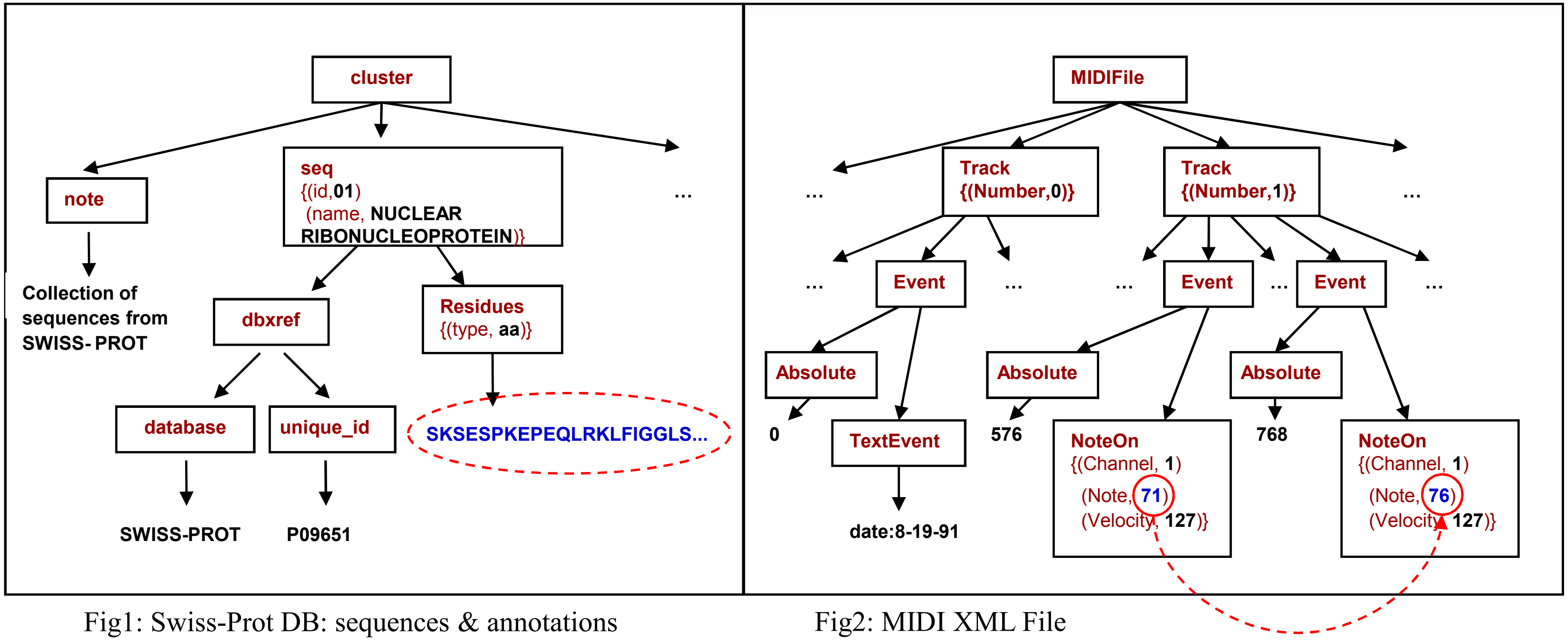


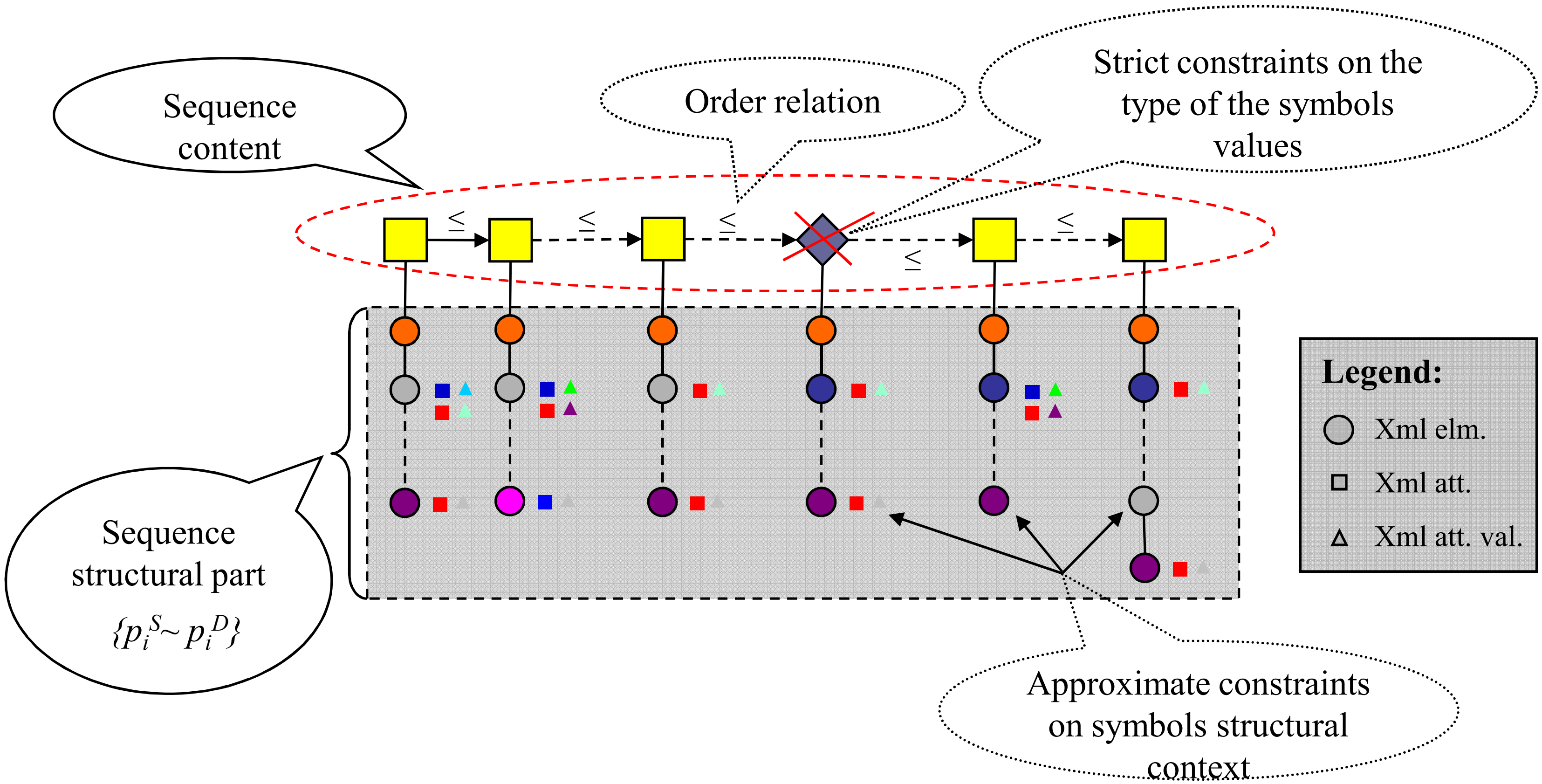
Fig1: Swiss-Prot DB: sequences & annotations

Fig2: MIDI XML File

Observation:

- XML structure indicates the sequential organization of the data.
- **This information should be analyzed, extracted and used in the IR process.**

Sequential Data Model



Sequence Structural Types: **Node**, **Document**, **Collection**

Sequence Extraction

- Heterogeneous XML documents
- Different kinds and types of sequential data
- Users may have highly diversified, subjective and time evolving interests

Hypothesis:

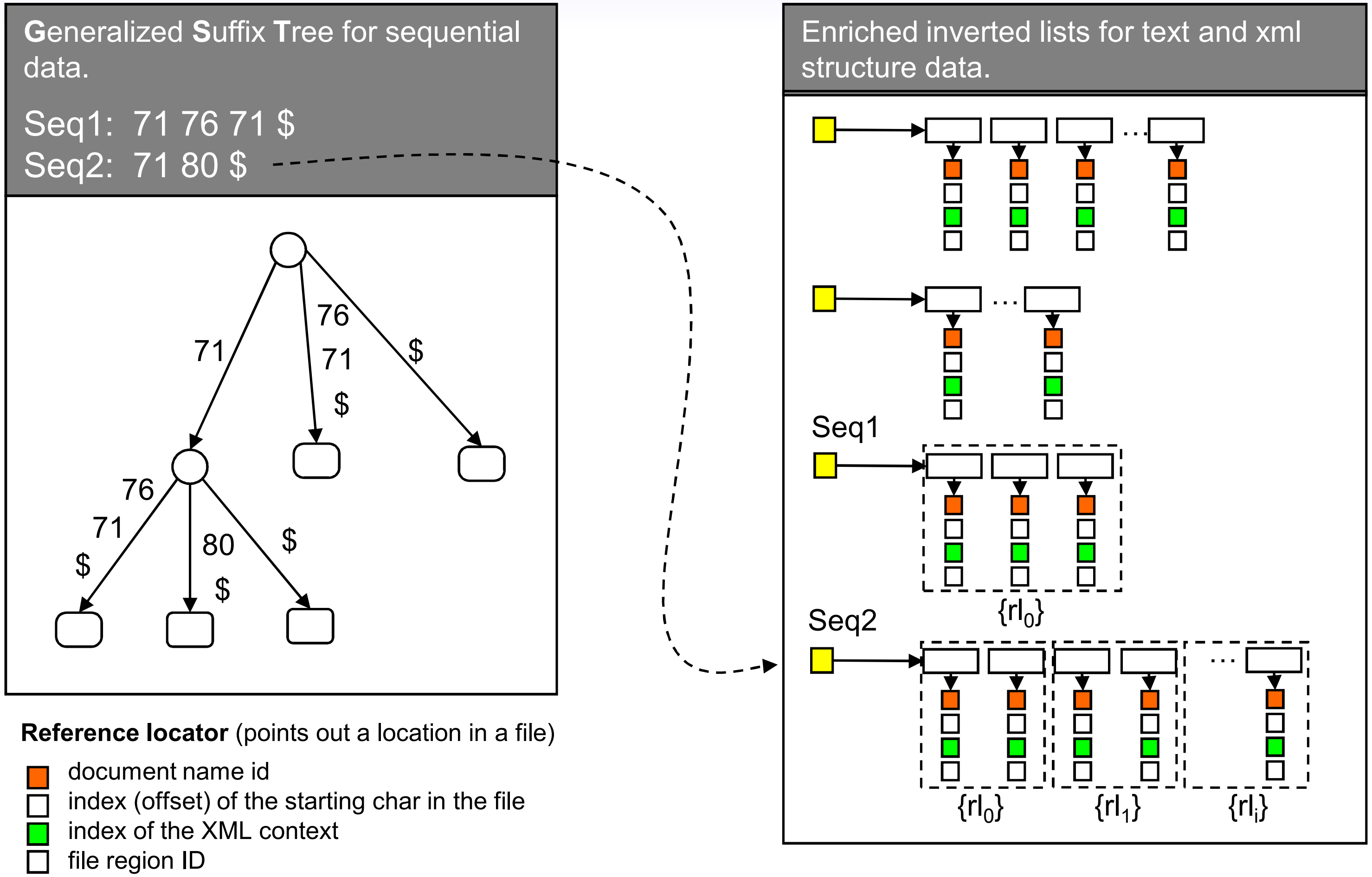
- the users have at least an imprecise, incomplete or fuzzy knowledge of the particular underlying organization of the sequential data in which they are interested in.

⇩

supervised sequence extraction process

```
(makeSeq
    [/MIDIFile/Track/Event/NoteOn(and (return @note NUMBER) (== channel 1))/]
    % return the note attribute values from the first channel of the XML document that have the
    % specified (or similar) structure
    1.0 % threshold for the symbol contextual matching score
    DOCUMENT  % sequence structural type
)
```

Indexing Scheme



- We propose a hybrid index model designed to merge both types of data: semistructured and sequential data.

Sequence Approximate Matching

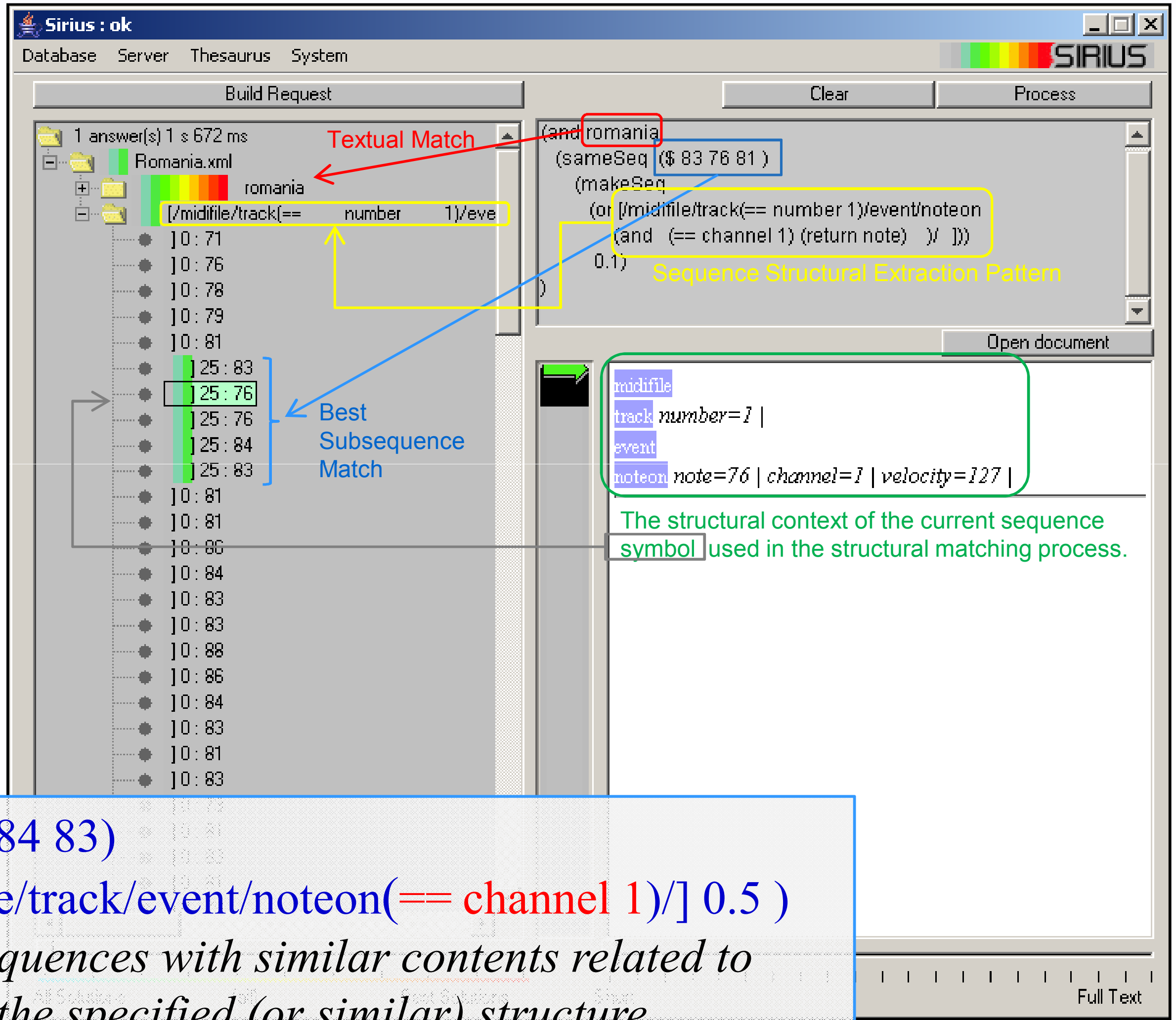
- Sequence δ distance (**dynamic programming)**
  - **Editing Levenshtein distance** (sequences) [Baeza-Yates and Gonnet, 99]
  - **Dynamic Time Warping** (DTW) (time series) [Park et al. 2003]

> retrieve all the subsequences S_i^j similar with a sequential query S_q, having the distance δ less than a specified threshold
> - *the P-against-all problem.* [Gusfield, 1997]

- **Problem complexity reduction** by using
  - a suffix tree as an index structure and
  - a dynamic programming method
- The **best subsequence match score** is aggregated with the **best contextual score** of its symbols in a global sequence score by using a **weighted geometric mean**.

Prototype

- **SIRIUS XML IR** system extension
  - Using the DTW distance
- **Dataset**
  - Small **MIDI XML** file collection enriched with semi-randomly generated meta-structure
- **Requests**
  - **structural requests** semi-randomly generated
  - **sequential requests -** short (<20 symbols) gapped subsequences
  - **complex requests**



(sameSeq ($ 83 76 80 81 84 83)
[/mididb/midifile/track/event/noteon(== channel 1)/] 0.5)

look for (sub) sequences with similar contents related to documents with the specified (or similar) structure

Outline

1. **XML Information Retrieval**
   Structure management, Focused access

2. **XML Multimedia IR**
   Sequential data

3. **XML IR on Specialized Hardware**
   Hardware accelerator

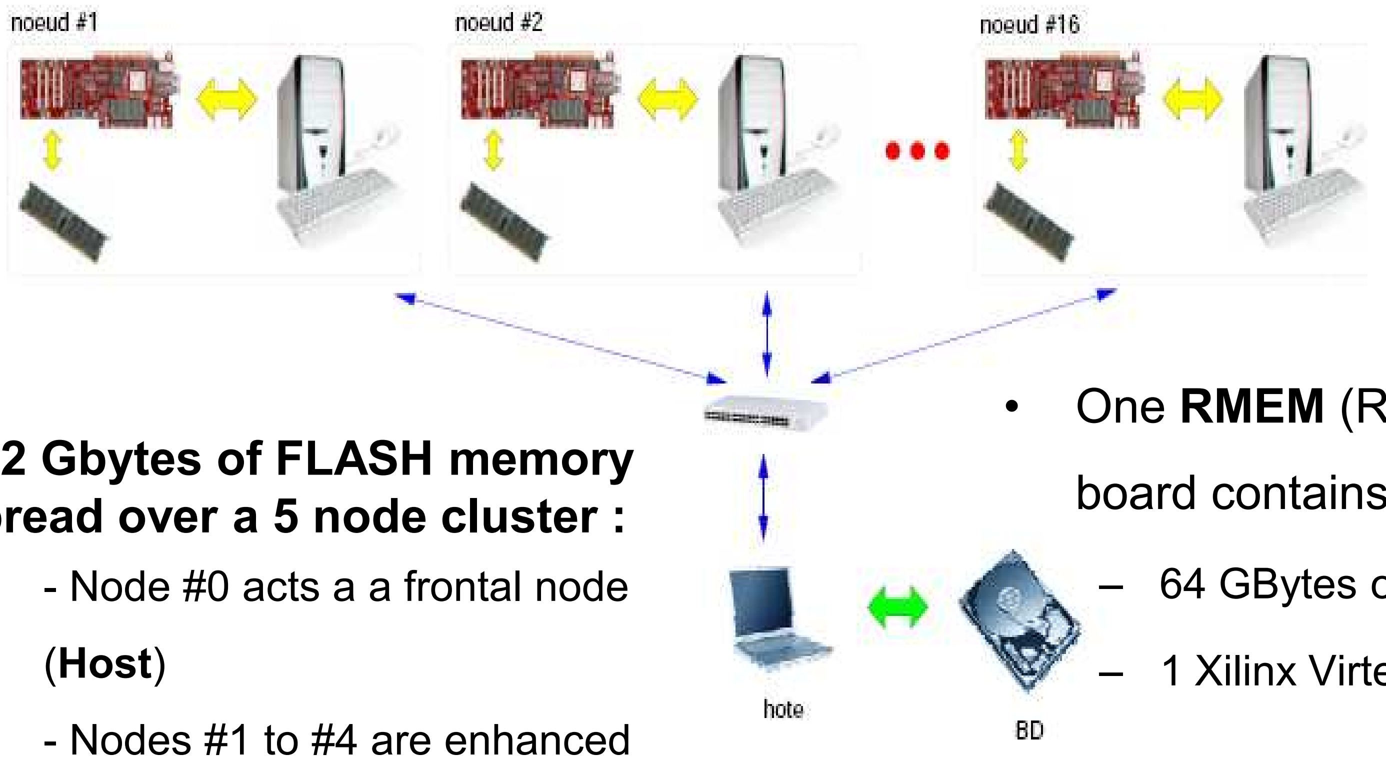
4. **Summary & Future Work**

The ReMIX Project

REconfigurable Memory for massive data IndeXing

- Supported by ACI "Masse de Données"
- **Specialized Hardware** based on two technologies
  - **FLASH memories**: to provide a large data capacity together with a fast access
  - **FPGA devices**: to process and filter accessed data
- **Software**
  - Programming Framework (ReMIX API)
  - Dedicated file system
- **Applications** focus on content-based search
  - genomics, images and **semi-structured text** processing.

The ReMIX System



512 Gbytes of FLASH memory spread over a 5 node cluster :

- Node #0 acts a a frontal node (**Host**)

- Nodes #1 to #4 are enhanced with 2 PCI **RMEM boards**

- One **RMEM** (ReMIX Memory) board contains:
  - 64 GBytes of FLASH memory
  - 1 Xilinx Virtex 2 Pro - XC 2VP30
- The nodes are interconnected through an Ethernet switchbox

Prototype

- **Implementing a subset of the SIRIUS search operators on the ReMIX API 0.93**
  - Input :
    - Index database file
    - List of elementary requests file
  - Output :
    - A file containing the list of documents and scores for the selected XML contexts
- **Index Construction**
  - External program based on **inverted lists** Implemented using a Distributed Hashtable (CURIA, QDBM)
- **Memory Organization**
  - **Term partitioning approach** [Baeza-Yates99]
  - Nodes equally loaded by using a **round-robin strategy**
- **Search Process**
  - Parallel processing of elementary requests on the ReMIX nodes
  - Merging and aggregating operations on the 'host'

Outline

1. **XML Information Retrieval**
   Structure management, Focused access

2. **XML Multimedia IR**
   Sequential data

3. **XML IR on Specialized Hardware**
   Hardware accelerator

4. **Summary & Future Work**

Summary

(C1): **Answer multi-criteria approximate requests**

(C2): **Provide focused access to relevant information**

(C3): **Process large volumes of documents**

Design

- **(C1, C2)** XML search mechanism based on a modified Levenshtein editing distance for XML paths and information fusion heuristics
- **(C1,C2)** Effective and simple strategy for detecting Best Entry Points in XML documents
- **(C1,C2)** A sequence extraction scheme guided by structural patterns for extracting sequential data symbols and contextual information from XML documents with heterogeneous structures
- **(C1)** Hybrid index model for the indexing of textual, structural and sequential data
- **(C1,C2)** A model for representing and searching similar sequences embedded in XML document databases based on two levels of approximation:
  - on their structural context and on their sequential content.
- **(C3)** Contribution to the specification phase of a specialized memory architecture for accelerating content-based search applications

Summary

(C1): **Answer multi-criteria approximate requests**

(C2): **Provide focused access to relevant information**

(C3): **Process large volumes of documents**

Implementation

(C1,C2) Developed a complete XML IR system: SIRIUS
- Indexer, QueryProcessor, GUI, distributed storage repository …

(C1,C2) Dedicated operators for sequence extraction, indexing and similarity search embedded in XML documents

(C2,C3) Prototype tailored for the use of the ReMIX specialized hardware memory architecture
- performs fast approximate structural filtering as support for searching relevant information in XML DB

Evaluation

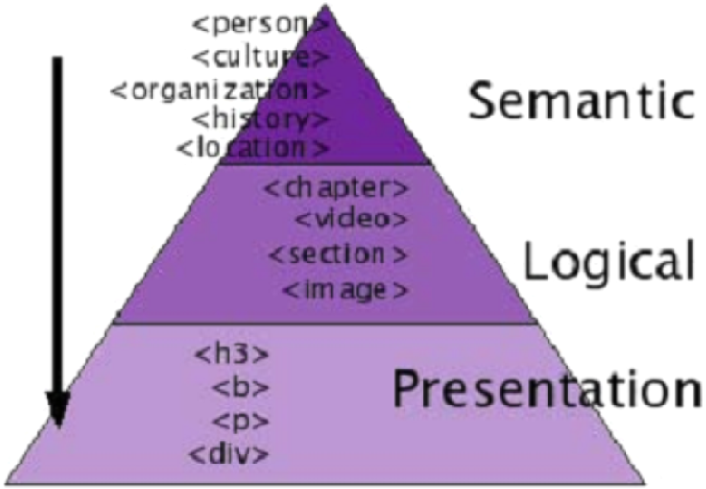
- INEX 2005 & INEX 2006 evaluation campaigns
- Encouraging & good performance results relative to the state of the art XML IR Systems

Future Work @ CWI

- Focus on the interaction between **focused/structured information retrieval** and **named entity annotation** approaches with **semantic Web technology**, **faceted search** and **visualization and interaction techniques** in order to support content representation and discovery, and information seeking and browsing in digital libraries/vast online document repositories.

Future Work @ CWI

- **Context / Working Hypotheses**
  - The documents are poorly annotated
    - mostly presentation and logical tags, less frequent meaningful semantic tags (for a specific user)
  - The users have a vague, imprecise, erroneous or any knowledge at all of the structure of the data
  - The users are able to recognize (and re/use) a useful structure/type/category in relation with their sought information



Types of XML structure
[van Zwol et al., ECIR'07
Chiaramella et al., FERMI'96]

Future Work @ CWI

Some incipient ideas submitted for further refinement…

1. **(Semi-) Structured Datasets with Rich Semantic Annotations**
   - Use named entity recognition and semantic web technologies to **annotate** and **link** the data
   - Use user profiles/domain ontologies to personalize the entity extraction/annotation phase
   - Gate?, Calais?, available annotated datasets(Wikipedia?, news?...)

2. **Apply Adapted Visualization & Interaction Techniques to each Information Type**

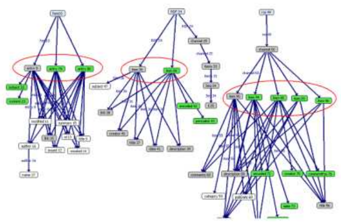
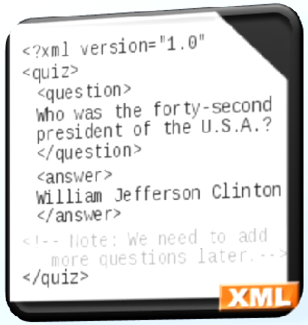


Figure 2: Summary Refinement for item and entry

Future Work @ CWI

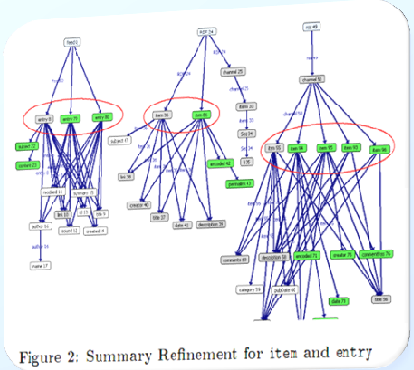
Some incipient ideas submitted for further refinement…

1. **(Semi-) Structured Datasets with Rich Semantic Annotations**
2. **Apply Adapted Visualization & Interaction Techniques to each Information Type**
   - Let the user INTERACT and understand the data and the effects of its queries on that data
   - Expose the structure (Structural summaries?) and the semantic types/categories of the retrieved entities within their context
   - Show the relations between them (i.e. the context of the named entities), refine faceted search?, highlight patterns
   - Adapt the visualizations and the interaction modes to each specific data type (text – snippets with highlighted terms, tag clouds; structure – structural summaries; temporal information - time lines, cycles; locations – maps; persons-, organizations-,…)



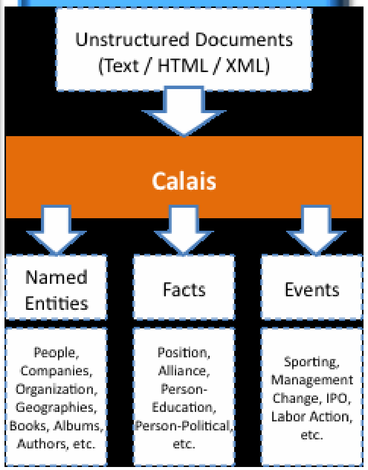
calais
Powered by Thomson Reuters

W3C Semantic Web



Structured Documents

NLP techniques



Semantic Web Technologies

Visualization & Interaction

Future Work @ CWI

- **Applications**
  - Book content analysis, exploration and search tasks
  - (Dynamic(interacting by querying and browsing)/Personalized(users profiles)) Semantic Rich Site Map Navigation
  - Visual Document/Collection Summarization
  - Visual exploration and retrieval of XML document collections
  - Semantic and structural documents/results clustering
  - (Visual) Schema matching with semantic clues / RSS feeds integration (applied on news?)
  - Relevance feedback
  - Recommendation systems
  - …

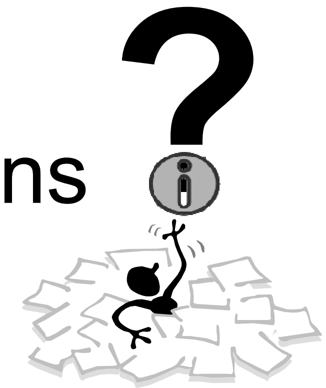
Future Work @ CWI

- **Research Questions** (Among the)
  - How to enrich the documents with valid semantic tags (for the user)?
  - How to support the users in exploring/understanding the (structure/content/organization of the) data ?
    - and how their queries (both in efficiency and relevance) are affected by this structure ?
  - How to support the users in finding the sought information ?
    - and new information related to this (recommendations?, patterns?)
  - How to evaluate the proposed system/interface?

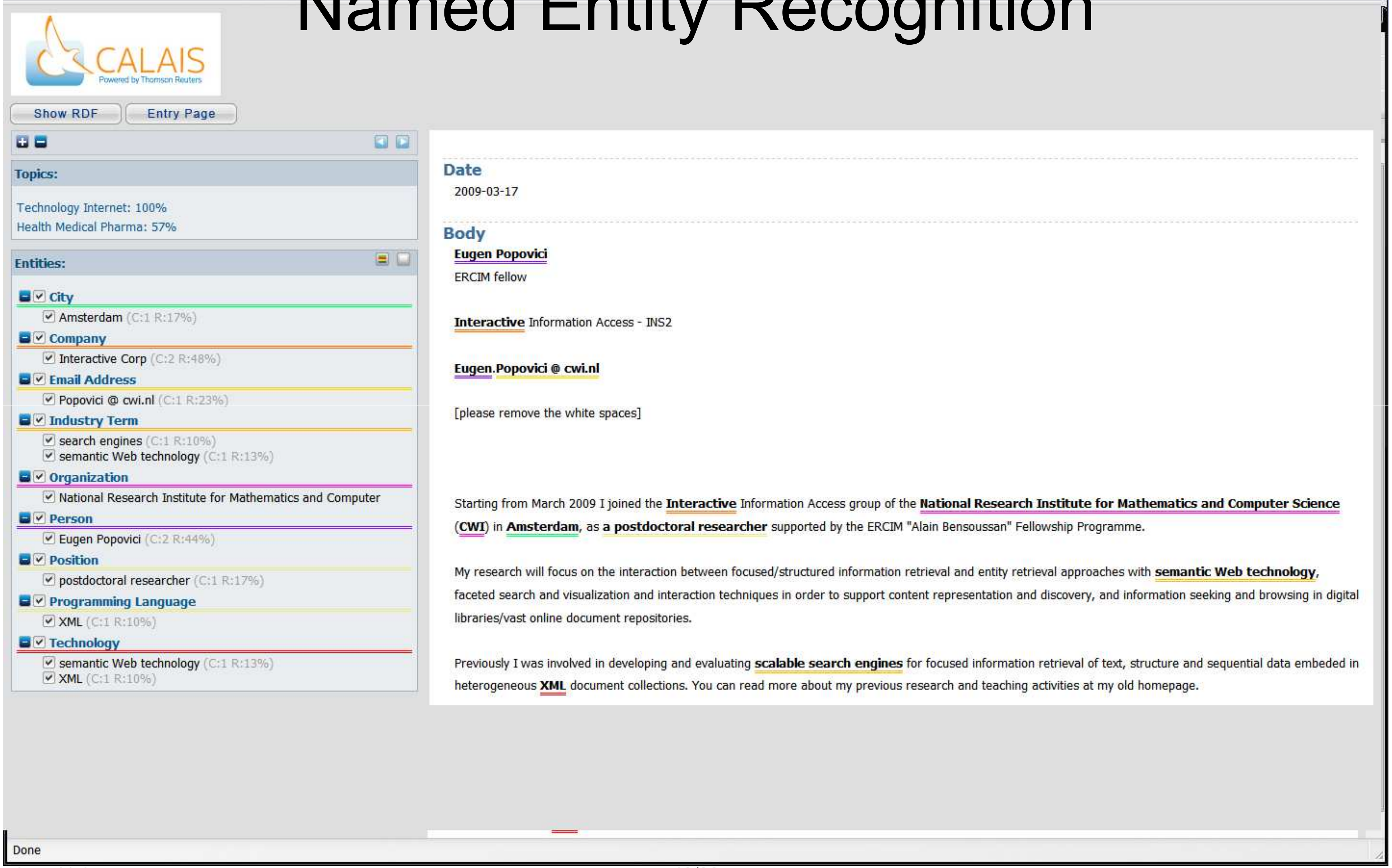
Keywords: XML Documents – Graphical user interfaces – Interactive information retrieval – Schema browsing (and matching) – Answer visualization and exploration – Structural summaries (Data Guides, …) – Focused/Structured IR – Named entities – Faceted search – Semantic Web , …

Dank u wel

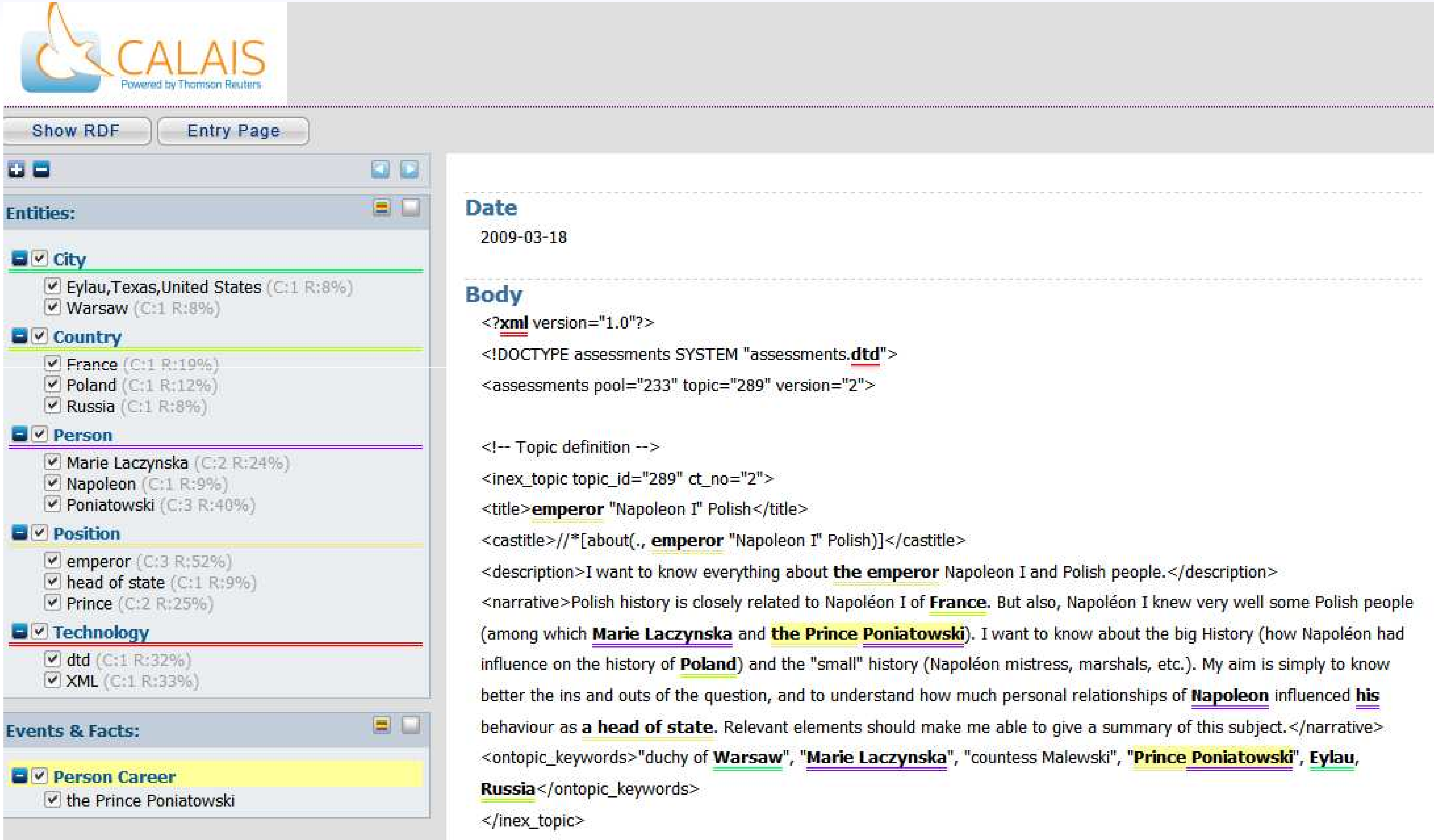
Questions



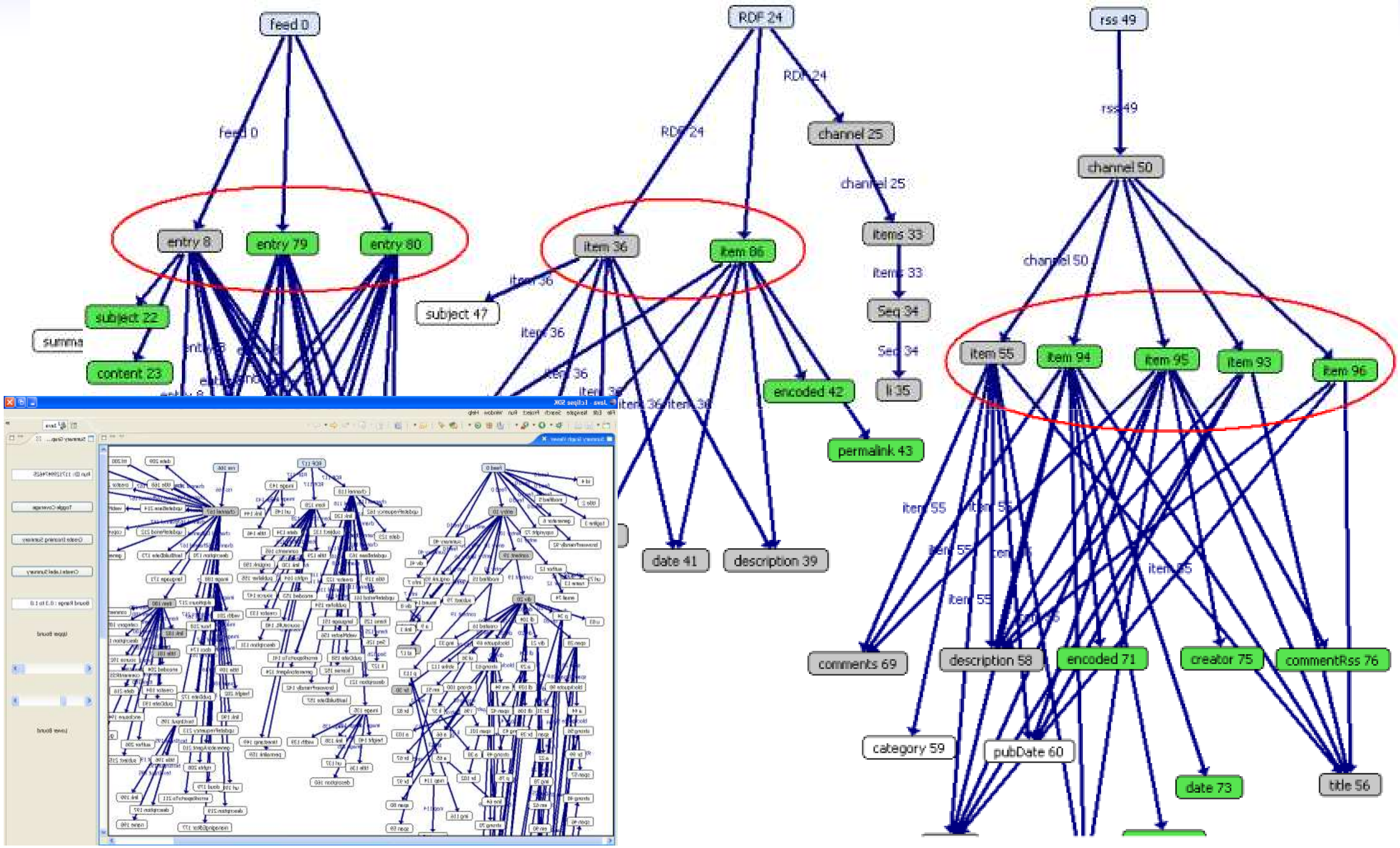
Named Entity Recognition



Named Entity Recognition



Visualizing Structural Patterns



M. S. Ali, Mariano P. Consens, Flavio Rizzolo, Visualizing Structural Patterns in Web Collections WWW'08

Evaluation I

- **Research Questions** (Among the)

  How to evaluate the proposed system/interface?

1. use a collection with existing relevance assessments
   - annotate both the topics and the relevant documents with semantic information about the entities (link the entities with domain ontologies) and try to do something useful with the whole package
2. Users studies on a specific task
   - Site map navigation
   - Content and structure query formulation
   - Summarization
   - Recommendation
   - Clustering
   - Schema matching / News Feeds Integration
   - Search & browse
   - …

Evaluation II

- Use a collection with existing relevance assessments (like the ones provided by the INEX evaluation campaign)
  - Choose a subset of topics that are fitted for named entity recognition and semantic query enrichment (i.e. either the title/the description/or the narrative of the topic makes some reference to a known entity - i.e. we were able to recognize it).
  - Annotate both the topics and the documents with information about the recognized entities. As the number of entities recognized within an article may be large, in a first approach we could restrict the annotations only to entities that were initially recognized within the topic. Link the entities with domain ontologies and try to do something useful with the whole package.
  - Check the named entities (specified within the topics) distribution within the data and within the relevant assessments. Try to find correlations and patterns to be integrated within the retrieval model.
  - Evaluate the system against the assessments, by using only the text, text & structure, text & structure enriched with named entities and text & structure with named entities linked to ontologies. See what works better. If globally the results are bad, go back to the topics and try to establish classes of topics for which the semantic annotation (statistically) improved the results (if any).

Dank u wel

Questions