OLAP on XML Text-Oriented Documents

IRIT (UMR 5505), University Toulouse 1 Capitole, Team SIG/ED (Generalised Information Systems, Data Warehouse)

Why work with documents ?

Analysis data come from

- 20% of numerical data (transactional Databases)
- 80% from documents (not compatible with OLAP)

Warehousing Documents = Document Warehouses ?

OLAP = On-Line Analytical Processing
XML = Extensible Markup Language

Data-Centric XML documents

```
<transactions>
   <transaction id="t0001">
      <customer id="c21">
         <name>Smith</name>
         <address>...</address>
      </customer>
      <products>
        <product>
         <name>LCD TV 52"</name>
         <qty>1</qty>
        </product>
        ...
      </products>
   </transaction>
   <transaction id="t0002">
   ...
   </transaction>
   ...
</transactions>
```

Document-Centric XML documents

```
<is_journal>
   <issue>Volume 34, Issue 4-5</issue>
   <article>
      <title>Preface</title>
      <author> T. B. Pedersen</author>
      <Paragraph>This special section
      contains extended versions of the
      best papers from the ACM Tenth
      International Workshop on Data
      Warehousing and OLAP (DOLAP'07) which
      was held on November 9, 2007, in
      Lisbon, Portugal, as one of workshops
      associated with the ACM Sixteenth
      Conference on Information and
      Knowledge...</Paragraph>
      <Paragraph>...</Paragraph>
      ...
   </article>
   ...
</is_journal>
```

Two types of XML documents

- **Data-Centric** XML Documents
  - Element order **does not** matter
  - Usually highly structured
  - Mainly transactional data
- **Document-Centric** XML Documents
  - Element order **does** matter
  - Usually loosely structured
  - Mainly textual data

4 ways of warehousing XML documents

- **XML Document integration**
  - **Data-centric** only
  - Standard data warehouses
- **XML Data warehousing**
  - **Data-centric** only
  - XML as storage technology
  - Similar to traditional data warehouses
- **XML Document warehousing**
  - **Document-centric**
  - No analysis (**no OLAP**)
  - Information retrieval oriented
  - Analysis limited to "contextualisation"
- **XML Document OLAP…**
  - **Data-Centric** and **Document-Centric**
  - **OLAP analysis**
- **But for this last category**
  - How to analyse/aggregate textual data ?
  - Usage of XML specificities ?

[jIS 2010]

Analysis on Document-Centric XML Document = OLAP Textual Analysis

From Numerical Analysis…

	Institute	Inst1		
	Author	A1	A2	A3
Conference				
DaWaK		2	1	-
ICEIS		1	3	-
CAiSE		-	1	2

Number of publications per author per conference

…Towards Textual Analysis

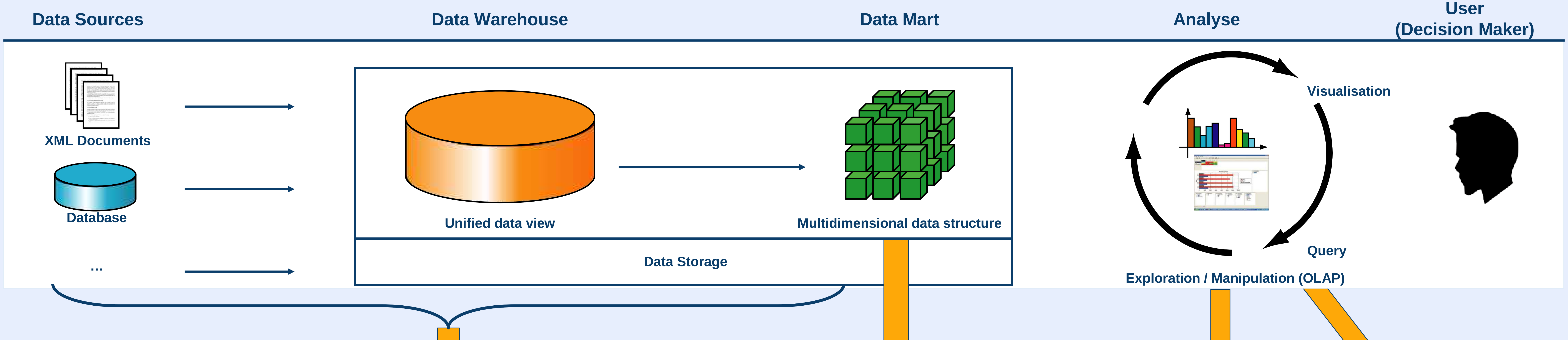
	Institute	Inst1		
	Author	A1	A2	A3
Conference				
DaWaK		XML, Temporal	Data warehouse	-
ICEIS		XML, Temporal DB	XML, Data mining, Constraints	-
CAiSE		-	Data warehouse	Conceptual model, Data mining

Same analysis but with the publication subjects

Some Interesting points

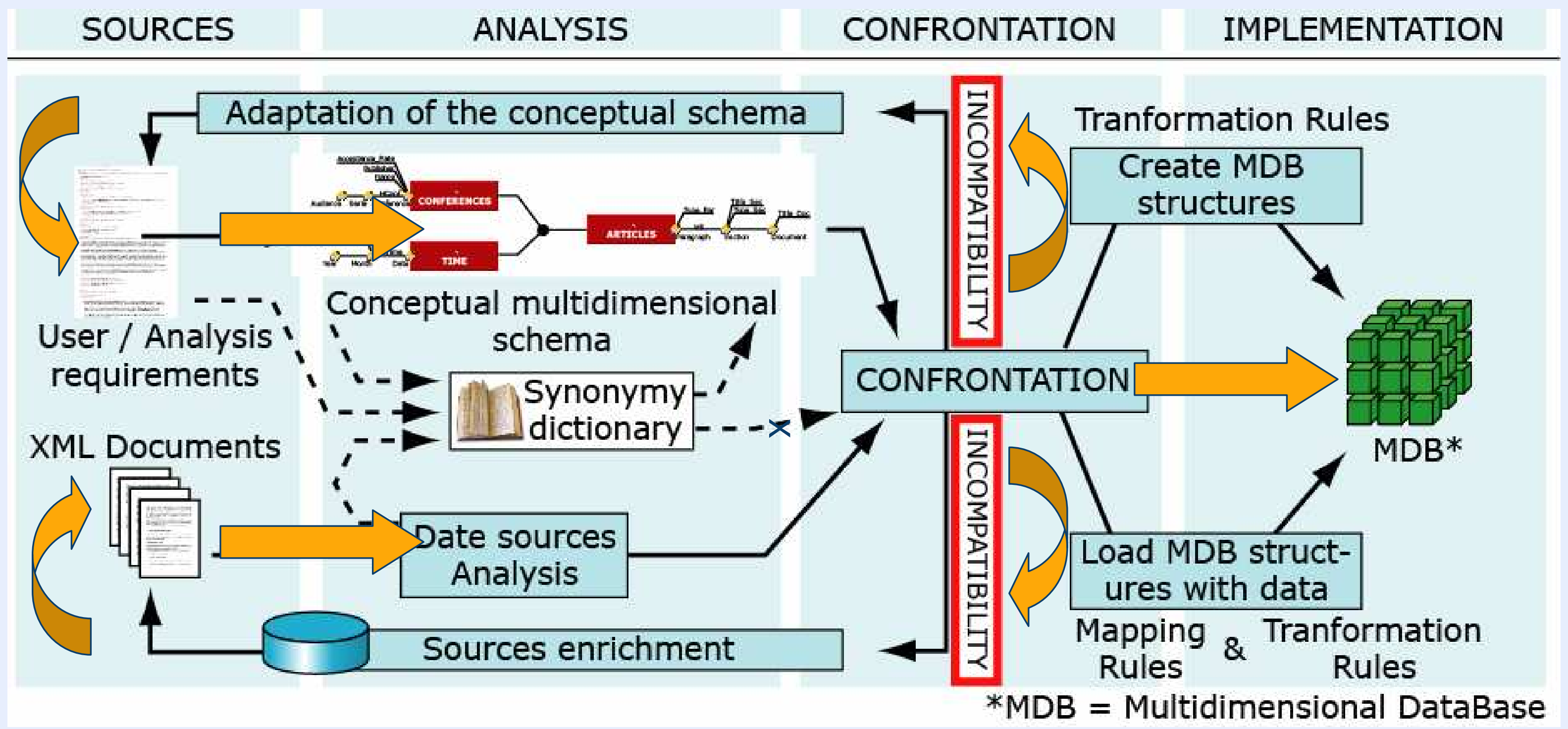
- **OLAP environment**
  - Works well on numerical data
  - Numerous Modelling solutions
  - …
- **XML Documents**
  - Some structure (required by data warehouses)
  - Tools…
- **Textual Data**
  - Information Retrieval Techniques
  - Text Mining Techniques
  - …

The Ideal Environment : OLAP on document-centric XML documents



Contributions

Design Process



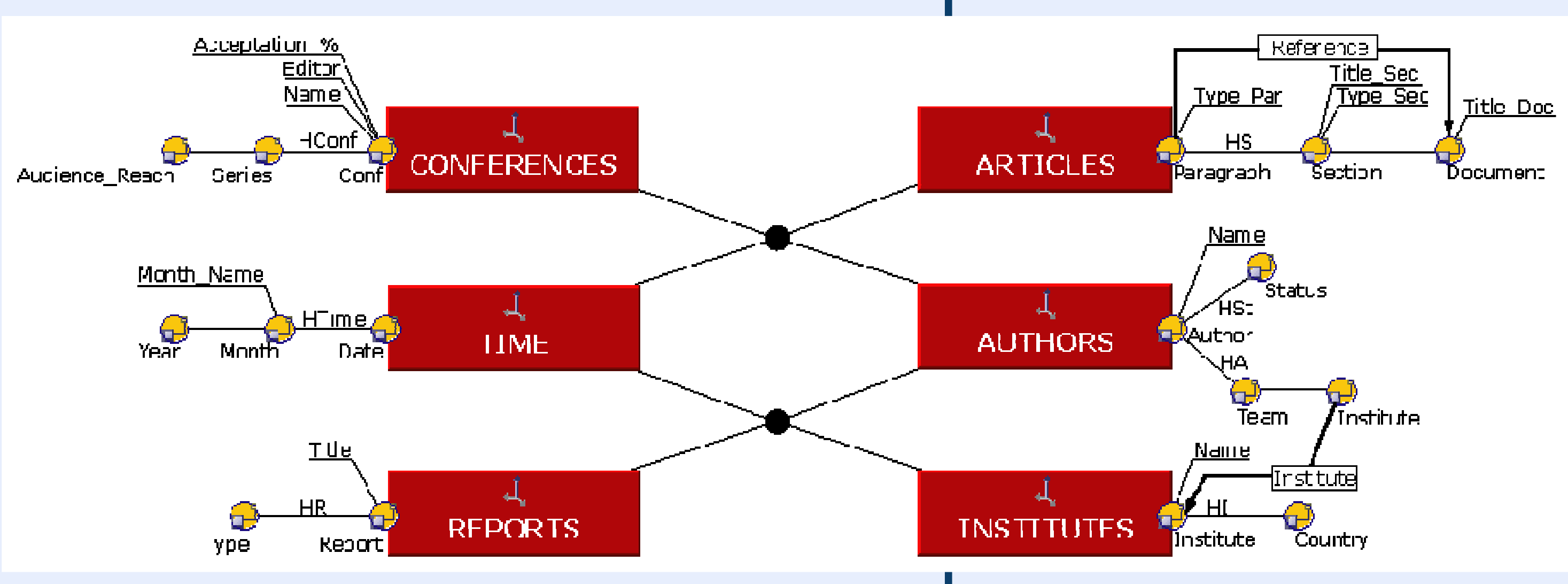
*MDB = Multidimensional DataBase

- Simplified design process (due to the unique dimension concept)
  - Mixed iterative approach
  - Eased design process
  - Semi-Automatic Implementation

[DaWaK 2011]

Conceptual Model GALAXY

Manipulation (OLAP)



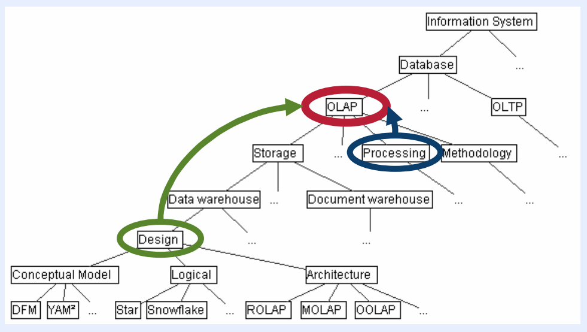
- Specificities
  - Unique **dimension** concept
  - No predefined analysis
  - Design process simplified
  - Closer to document specificities

[ER 2007]

- Adapted manipulation operators
  - Display Analysis
  - Drilling
  - Rotating
  - Restriction (Select)

[ER 2007]

Aggregation Functions



$$tf_{ij}(t) = \frac{n_{ij}(t)}{n_{ij}}$$

$$idf_{ij}(t) = \log \frac{d_{ij} + 1}{d_{ij}(t)}$$

- Adapted textual aggregation
  - Ontology
  - Information Retrieval Term weights

[ICEIS 2007]
[DaWaK 2008]

Implementation / Validation

- **Galaxy in ROLAP environment**
- **Aggregation benchmark (Ontology oriented)**

Future Works

- **Complete Aggregation Environment**
- **On-Line Aggregation**
- **Advanced Visualisation**
- **Textual Data => Complex Data**

Author : **Ronan TOURNIER**

First European Business Intelligence Summer School (eBISS 2011)
July 3 - 8, 2011
Paris, France

with Geneviève PUJOLLE, Franck RAVAT, Olivier TESTE, Gilles ZURFLUH