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

Warehousing Documents = Document Warehouses?
- 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
    - Data-centric
    - No analysis (no OLAP)
    - Information retrieval oriented
    - Analysis limited to “contextualisation”

- XML Document OLAP...
  - Data-Centric and Document-Centric
  - XML analysis
  - But for this last category
    - How to analyse/aggregate textual data?
    - Usage of XML specificities?

Analysis on Document-Centric XML Document = OLAP Textual Analysis

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