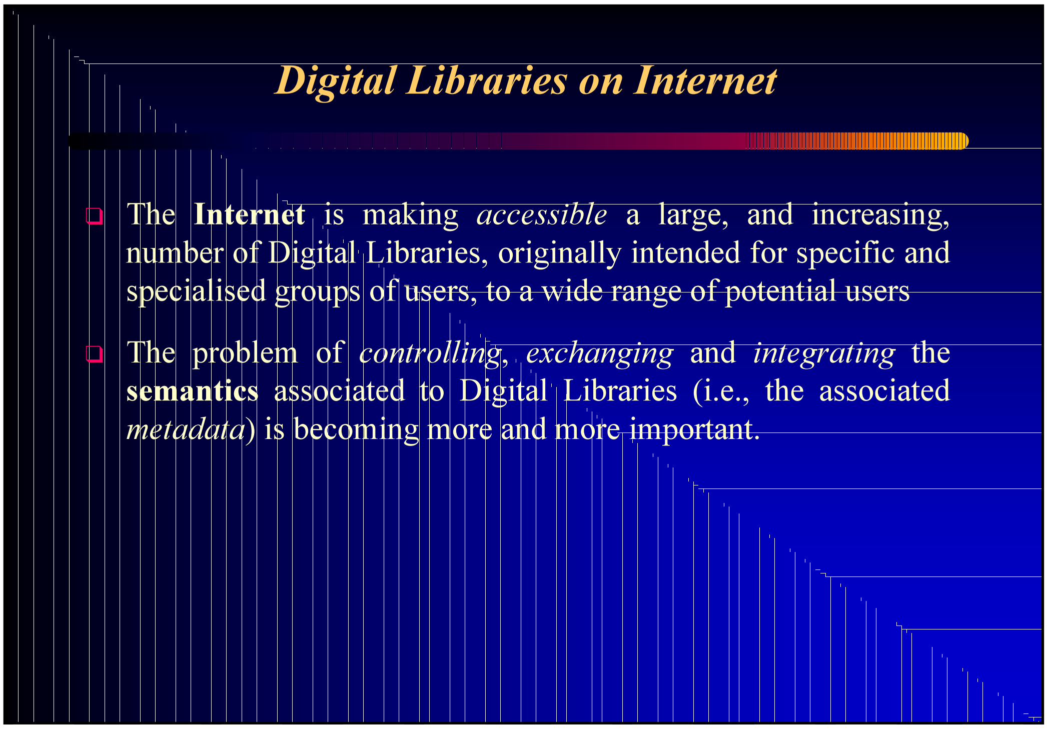
Ontology-based Access to Digital Libraries

Sonia Bergamaschi
University of Modena and Reggio Emilia
Modena Italy

Fausto Rabitti
ISTI - Institute of Information Science and Technology
CNR, Pisa Italy

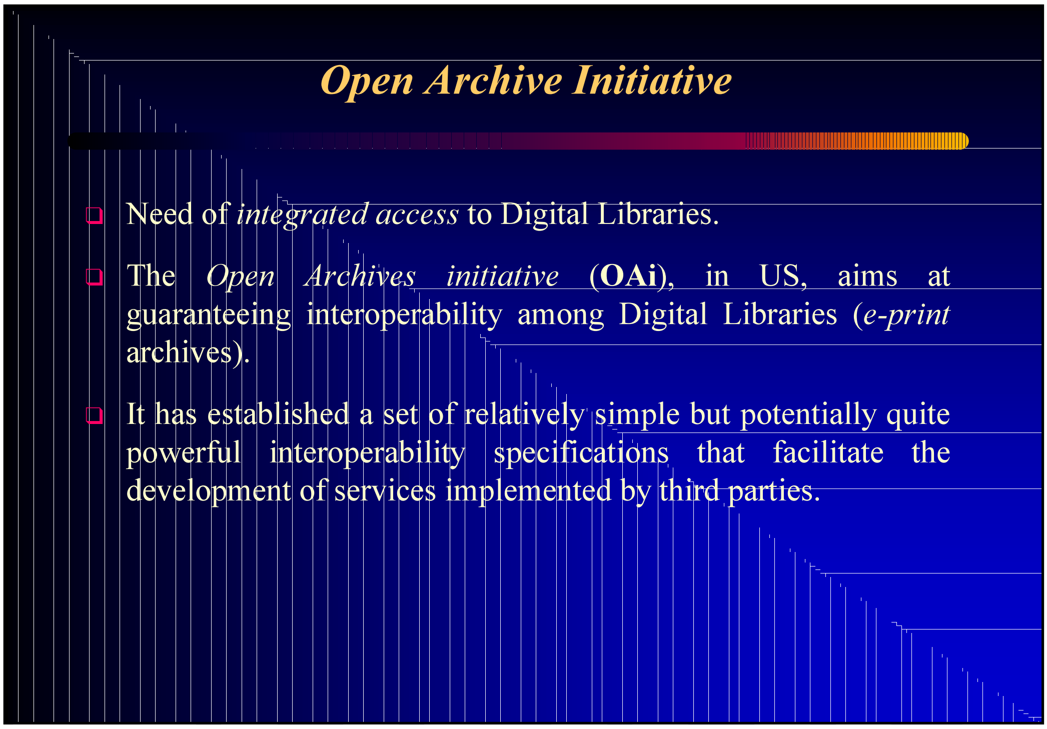
Outline

- Digital Libraries on Internet
- Need of integrated access (Open Archive Initiative)
- Metadata in Digital Libraries
- Impact of XML on Digital Libraries
- Controlling semantics in XML (data and metadata interchange in Digital Libraries)
- Ontology-based approach



Digital Libraries on Internet

- The **Internet** is making *accessible* a large, and increasing, number of Digital Libraries, originally intended for specific and specialised groups of users, to a wide range of potential users
- The problem of *controlling*, *exchanging* and *integrating* the **semantics** associated to Digital Libraries (i.e., the associated *metadata*) is becoming more and more important.



Open Archive Initiative

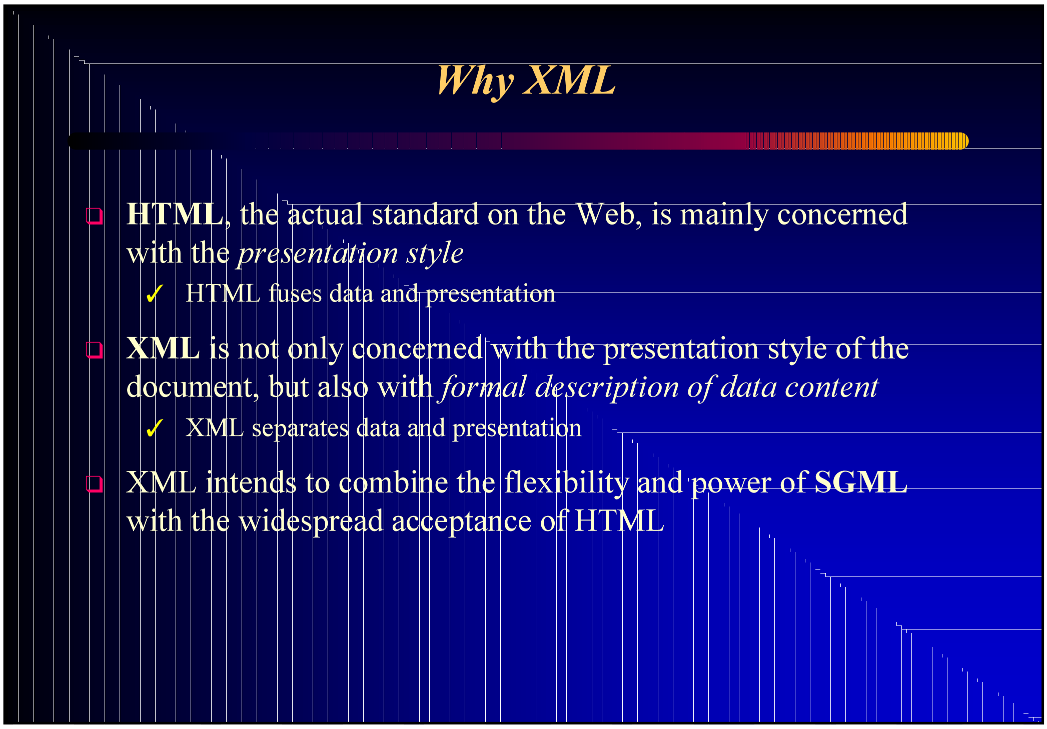
- Need of *integrated access* to Digital Libraries.
- The *Open Archives initiative* (**OAi**), in US, aims at guaranteeing interoperability among Digital Libraries (*e-print* archives).
- It has established a set of relatively simple but potentially quite powerful interoperability specifications that facilitate the development of services implemented by third parties.

Metadata in Digital Libraries

- Metadata in Digital Libraries, for bibliographic data, are usually expressed according to models like Dublin Core or MARC.
- However, there is the need to *generalise* the description of data and metadata made available in a large variety of Digital Libraries.
- The wide acceptance on the Web of XML can be a decisive factor in this direction.

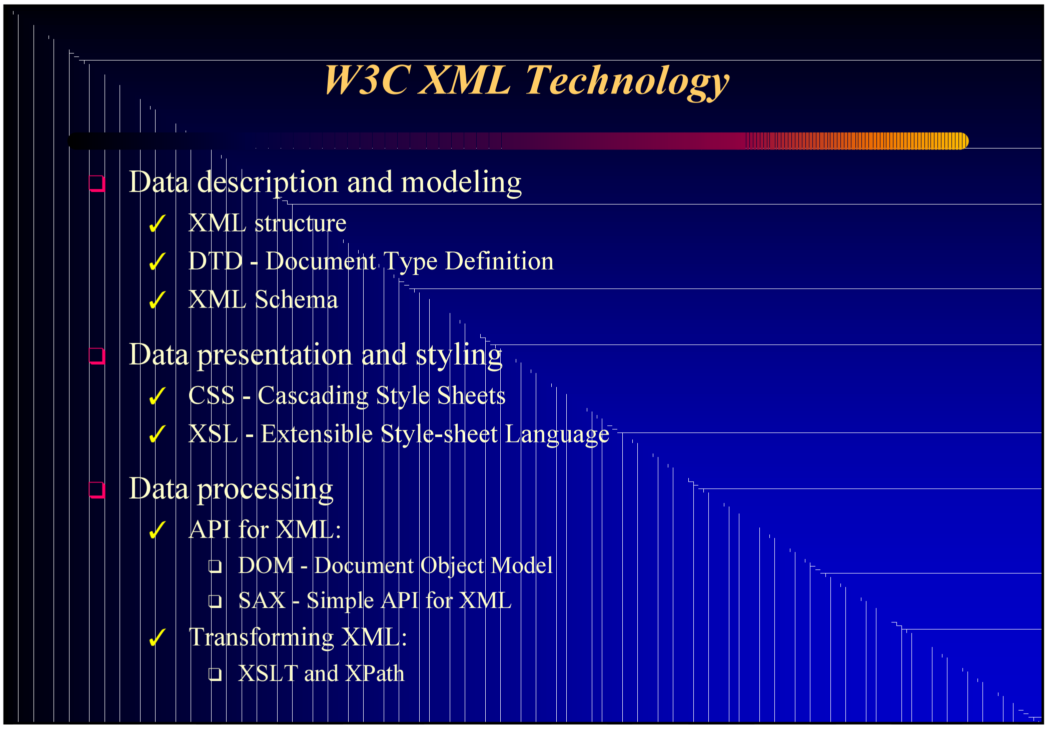
What is XML

- XML: eXtensible Markup Language
  - ✓ XML is a simple, standard way to delimit text data
  - ✓ *the ASCII of the Web:*
    - *use your favorite programming language to create an arbitrary data structure*
    - *share it with anyone using any other language on any other computing platform*
- Proposed by the World Wide Web Consortium (W3C)
- XML is a subset of SGML
  - ✓ SGML - Standard Generalized Markup Language



Why XML

- **HTML**, the actual standard on the Web, is mainly concerned with the *presentation style*
  - ✓ HTML fuses data and presentation
- **XML** is not only concerned with the presentation style of the document, but also with *formal description of data content*
  - ✓ XML separates data and presentation
- XML intends to combine the flexibility and power of **SGML** with the widespread acceptance of HTML



W3C XML Technology

- Data description and modeling
  - ✓ XML structure
  - ✓ DTD - Document Type Definition
  - ✓ XML Schema
- Data presentation and styling
  - ✓ CSS - Cascading Style Sheets
  - ✓ XSL - Extensible Style-sheet Language
- Data processing
  - ✓ API for XML:
    - DOM - Document Object Model
    - SAX - Simple API for XML
  - ✓ Transforming XML:
    - XSLT and XPath

Controlling Semantics in XML

- XML is a powerful and flexible way to convey the *semantics* of data through a *syntax*:
  - ✓ it does not ensures the correctness of the process:
    - ❑ two applications may interoperate via XML and still give different meaning to the same data objects
- XML document tags can be use to describe the *meaning* of the document components. *Controlling* the semantics associated to XML tags will be a decisive task.
- W3C activity on metadata:
  - ✓ **PICS**: Platform for Internet Content Selection
  - ✓ **RDF**: Resource Description Framework

Impact of XML on Digital Libraries

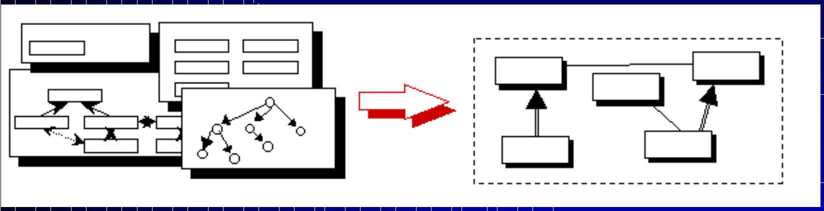
- Controlling the semantics in XML will open new perspective in accessing Digital Libraries, since XML is going to become the new *interoperability standard for distributed Digital Libraries*.
- We foresee a situation where XML will be used in Digital Libraries:
  - ✓ for *exchanging digital documents* (often multimedia) and their multi-modal presentations (via XSL)
  - ✓ for *defining metadata*, using XML DTD or Schema descriptions, with associated RDF (Resource Description Framework) schema descriptions.

Ontology-based approach

- aims to build a Digital Library Ontology representing a *global virtual view* of distributed Digital Libraries
- Mapping rules between local and global views based on a "Common Thesaurus" of terminological relationships able to reconcile different representation of similar concepts.
- The starting point is the MOMIS system

Mediator envirOnment for Multiple Information Sources (MOMIS) Project

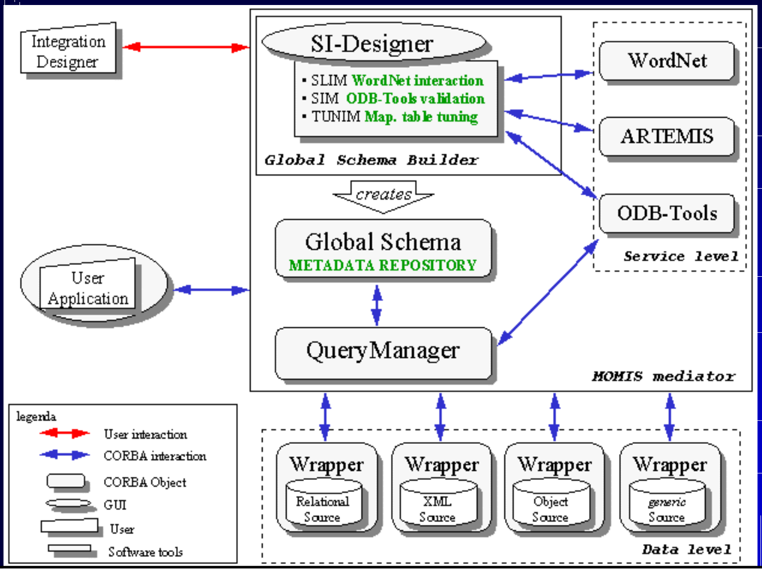
- Information sharing from multiple heterogeneous sources



- **Proposal :**Information Integration to provide a global conceptual schema, allowing a user to pose a query and to receive a single unified answer.

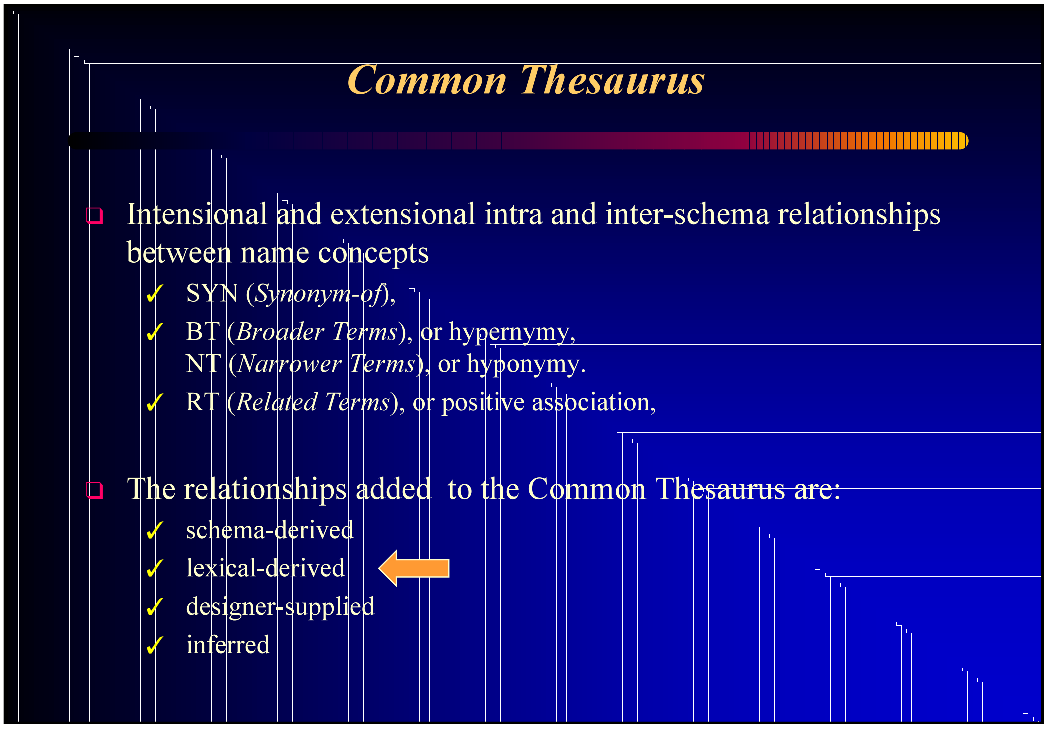
Internet: http://sparc20.dsi.unimo.it

MOMIS Architecture



MOMIS wrapper

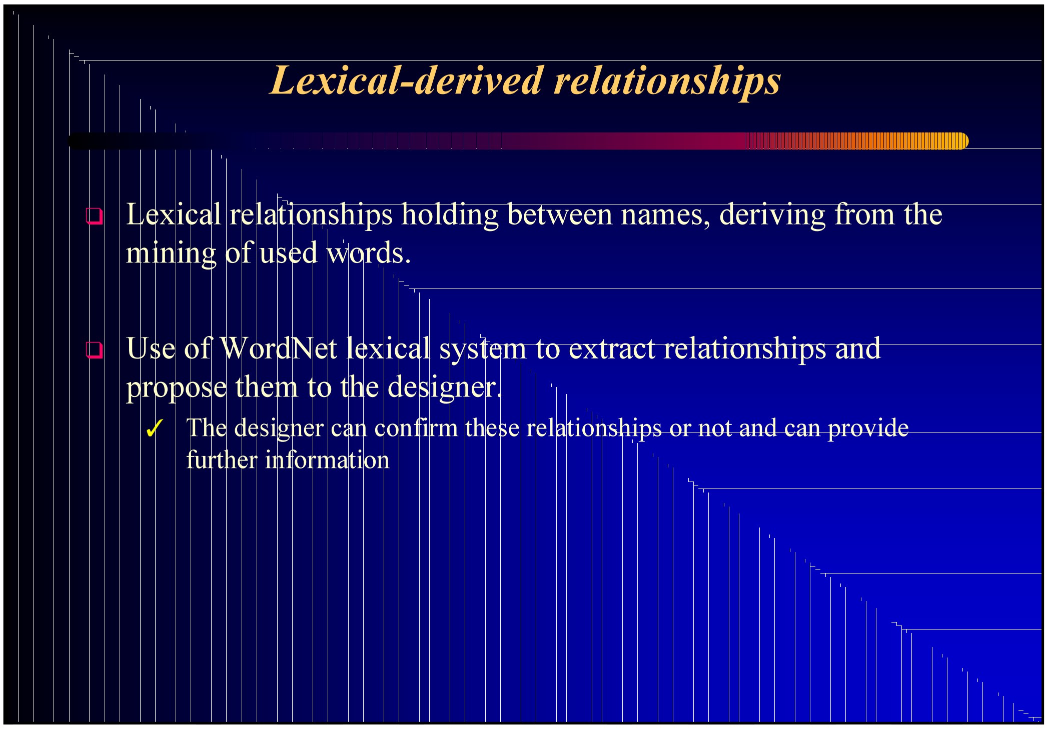
- The wrappers are the access point for the data sources.
- The wrappers present each data source (XML, relational, object, ...) in a common data model (derived from ODMG and I^3/POB proposal)
- An XML wrapper *wraps* data sources that contains *valid* XML data:
  - ✓ Translation phase: from XML-DTD data structures to ODMG data structures
  - ✓ Querying phase: query translation from a ODMG-standard query language to XML query language



Common Thesaurus

- Intensional and extensional intra and inter-schema relationships between name concepts
  - ✓ SYN (*Synonym-of*),
  - ✓ BT (*Broader Terms*), or hypernymy,
    NT (*Narrower Terms*), or hyponymy.
  - ✓ RT (*Related Terms*), or positive association,

- The relationships added to the Common Thesaurus are:
  - ✓ schema-derived
  - ✓ lexical-derived ⬅
  - ✓ designer-supplied
  - ✓ inferred



Lexical-derived relationships

- Lexical relationships holding between names, deriving from the mining of used words.

- Use of WordNet lexical system to extract relationships and propose them to the designer.
  - ✓ The designer can confirm these relationships or not and can provide further information

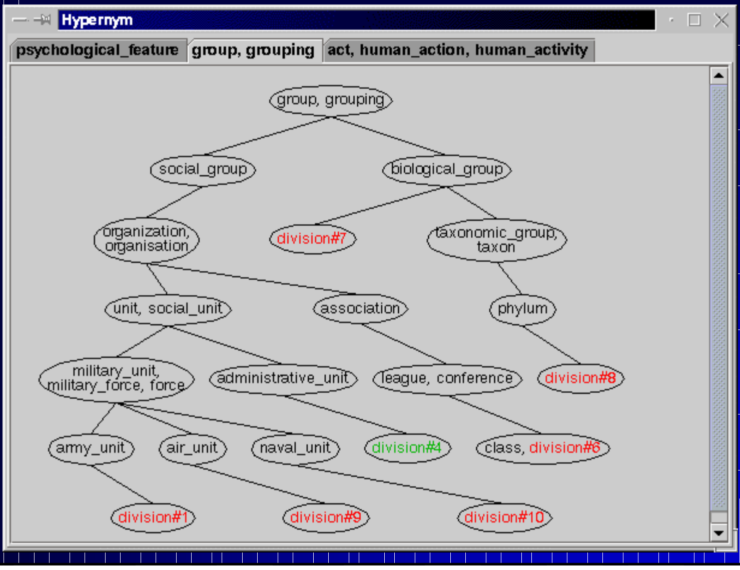
Lexical-derived relationships :an example

NT

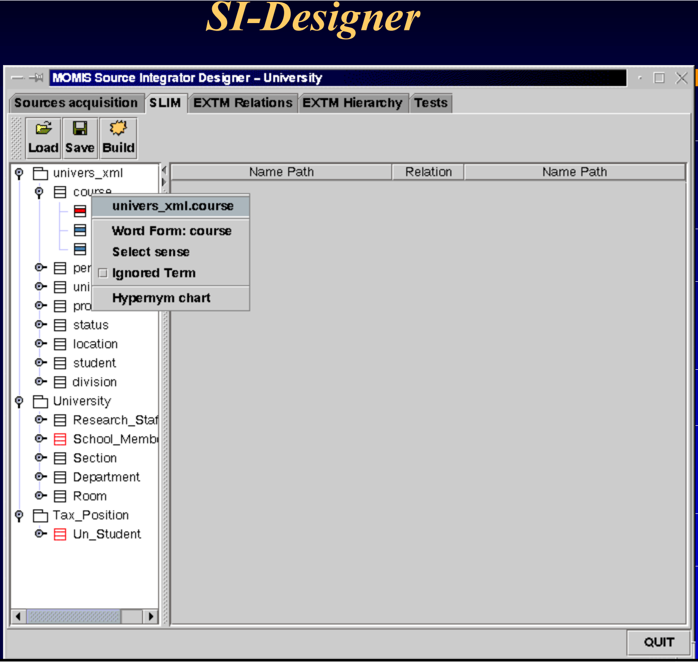
hyponymy

	Base Form	
Meaning (synset)	section	division
department, section -- (a specialized division of a large organization;	section#2	
division -- (an administrative unit in government or business)		division#4

Lexical-derived relationships :an example



SI-Designer



Example of XML Source

```
<!ELEMENT University (Person)*>
<!ELEMENT Person (first_name, last_name, email, Status)>
<!ATTLIST Person Code ID #REQUIRED>
<!ELEMENT Status (Student | Professor)>
<!ELEMENT Student (year, Course*, home_address. rank)>
<!ATTLIST Student StudentId ID #REQUIRED
          tutor CDATA #REQUIRED>
<!ELEMENT Professor(ptitle, Division, rank)>
<!ATTLIST Professor Prof_code ID #REQUIRED
          Office_phone CDATA #IMPLIED>
<!ELEMENT Division (Location, fund, employeenr)>
<!ATTLIST Division description CDATA #REQUIRED
          sector CDATA #REQUIRED>
... . .
```