DiWeb2004 - 3rd International Workshop on Data Integration over the Web

Riga (Latvia) – 8th June 2004

Ontology-based Integration for Sharing Knowledge over the Web

Devis Bianchini and Valeria De Antonellis

University of Brescia (ITALY)

Outline

- Motivations and context
- A semi-automated methodology for the construction of ontologies from XML datasources
- The VISPO experimentation
- Conclusive remarks and future work

Motivations and context

The VISPO (Virtual-district Internet-based Service PlatfOrm) Project:

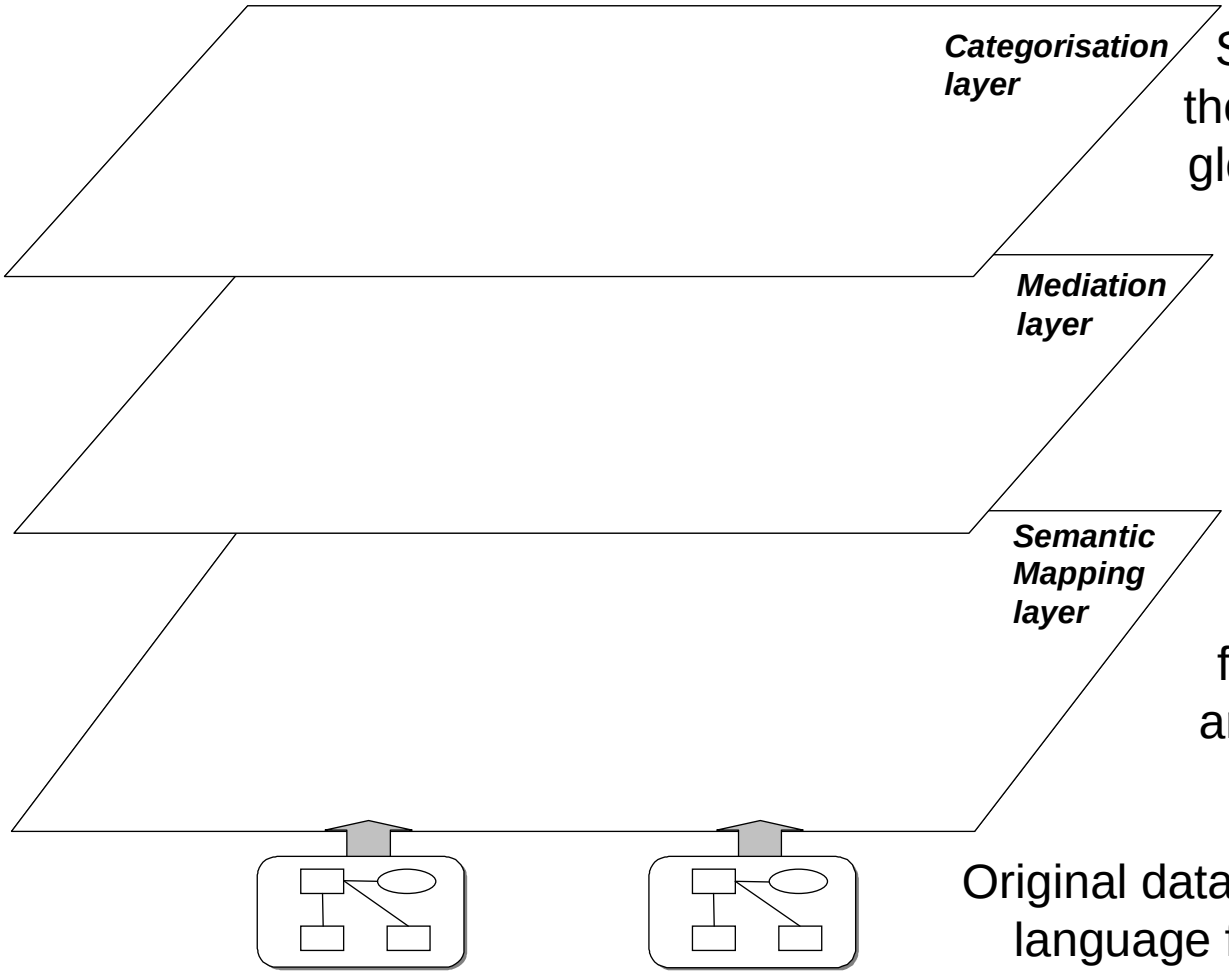
- A consortium of independent member enterprises wish to operate in a cooperative way to exploit business opportunities (*virtual district*) in industrial accessory production market
- Member enterprises cooperate in a Web-based scenario, where XML is the standard adopted for information exchange

Needs

- While XML can resolve format incompatibility in exchanged data, interorganizational cooperation can be obtained if there is a way to mediate among the different conceptual contexts

- Concept ontologies are particularly relevant to improve semantic interoperability (i.e., the difficulty to understand the data)

The three-layer domain ontology



Subject categories, to provide thematic views on the underlying global concepts and datasources

Global concepts, obtained by integrating similar local concepts and organized by means of semantic relationships

Local concepts, extracted from datasources schemas and clustered on the basis of their semantic similarity

Original datasources, expressed using a language for XML schema definition

The methodology for ontology construction

- The approach for three-layer ontology construction is organized into four phases:

  1) data analysis and conceptualisation

  2) integration

  3) synthesis and categorization

  4) implementation

- The involved activities are partially supported by the ARTEMIS software tool environment

Phase 1: Data analysis and conceptualization

Input *= Set of schemas for XML sources to be considered*

Output *= Clusters of XClasses with affinity*

- Each source schema is modelled as a set of *XClasses* (*XFormalism*)
- The level of semantic relationship, called *affinity*, between XClasses is computed by exploiting a thesaurus of weighted terminological relationships (*synonymy*, *hyperonymy*)
- ($\Rightarrow$ ARTEMIS tool environment)
- XClasses are grouped in *clusters* of semantically related concepts ($\Rightarrow$ ARTEMIS tool environment)

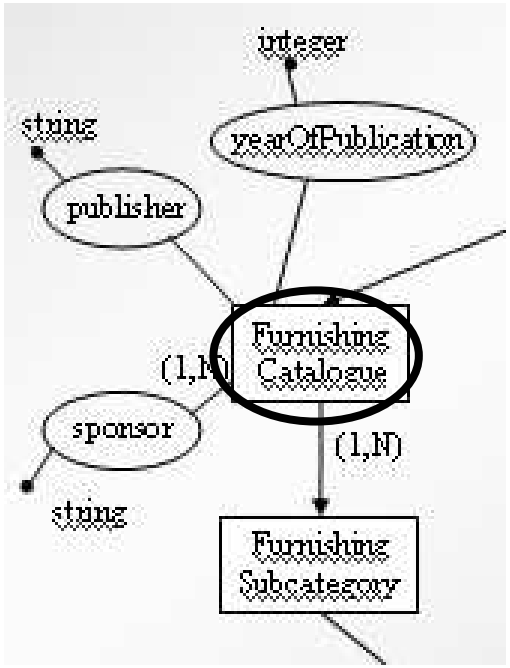
XFormalism

We introduce the XFormalism in order to highlight classes of concepts and their relationships in the sources, with no reference to a particular XML schema formalism (DTD, XML Schema, DSD, …)

A local XClass represents a set of objects with a common structure, described by means of

- a *name*
- a *content model* (empty, text, sequence, choice, mixed)
- a set of *properties* (sub-elements with simple or built-in data types)
- a set of *attributes*
- a set of *references* to other local XClasses
- *cardinality constraints* for properties, attributes and referenced XClasses

A simple example



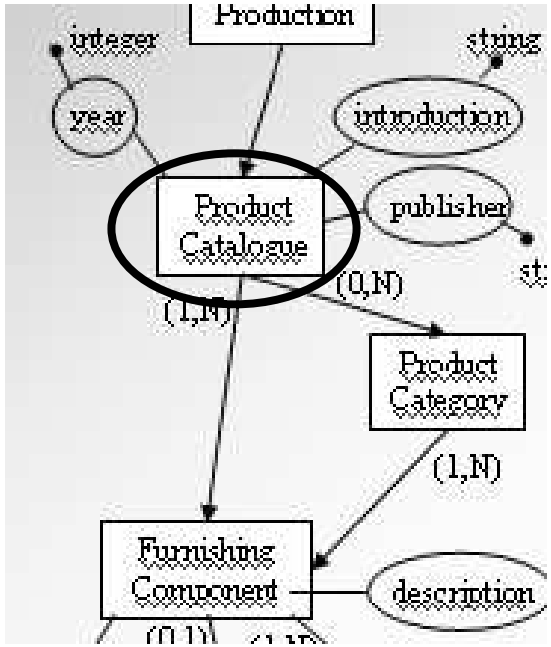
Name:	FurnishingCatalogue
Content Model:	(yearOfPublication, publisher, sponsor,FurnishingSubcategory)
Properties:	{(yearOfPublication,integer,(1,1)),(publisher,string,(1,1)), (sponsor,string,(1,N))}
RefXClasses:	{(FurnishingSubcategory,(1,N))}
Attributes:	{}

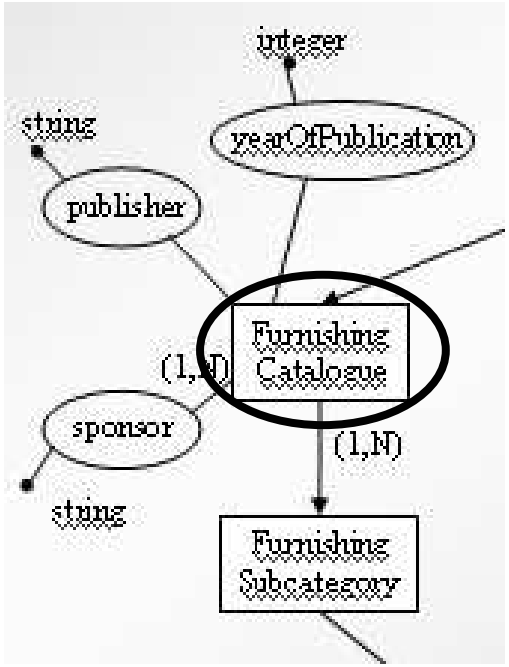
Semantic affinity among local XClasses

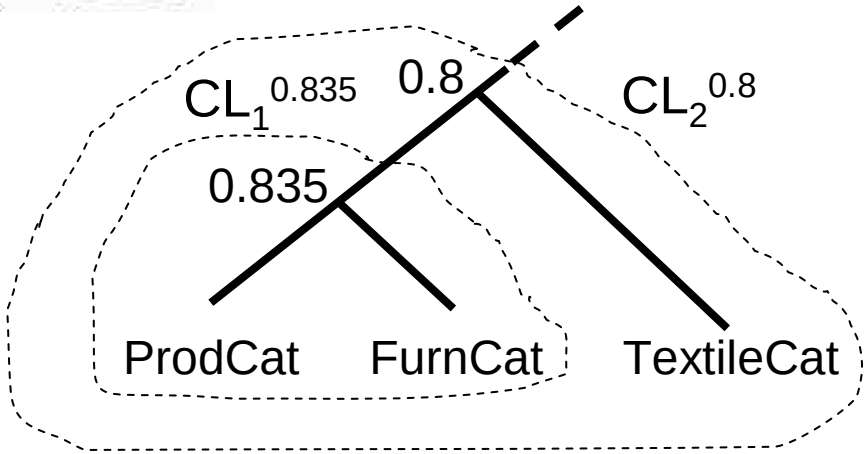
- Local XClasses xc_1 and xc_2 are compared on the basis of:
  - ***Name Affinity*** $NA(xc_1, xc_2) \in [0,1]$ – evaluated by computing the weight of paths of terminological relationships between xc_1, xc_2 in the thesaurus
  - ***Structural Affinity*** $SA(xc_1, xc_2) \in [0,1]$ - evaluated by computing name affinity of their features (properties, attributes, referenced XClasses) and domain compatibility

- A *Global Affinity coefficient* is computed as a weighted sum of Name Affinity and Structural Affinity coefficients

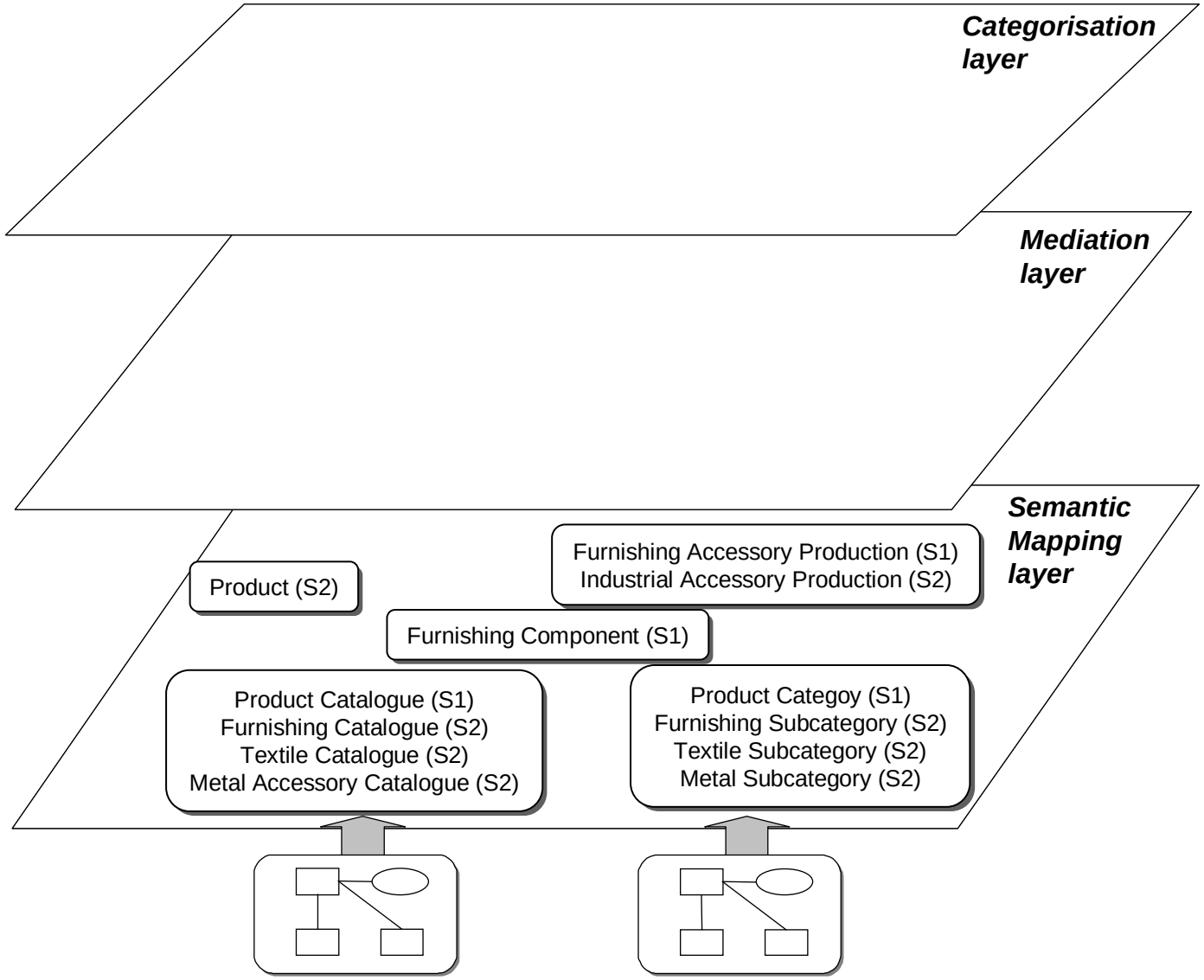
Example


Production
integer
string
year
introduction
Product Catalogue
publisher
(0,N)
(1,N)
Product Category
(1,N)
Furnishing Component
description


integer
string
yearOfPublication
publisher
Furnishing Catalogue
(1,N)
sponsor
(1,N)
string
Furnishing Subcategory


$CL_1^{0.835}$
0.8
$CL_2^{0.8}$
0.835
ProdCat
FurnCat
TextileCat

Three-layer domain ontology



Phase 2: Integration

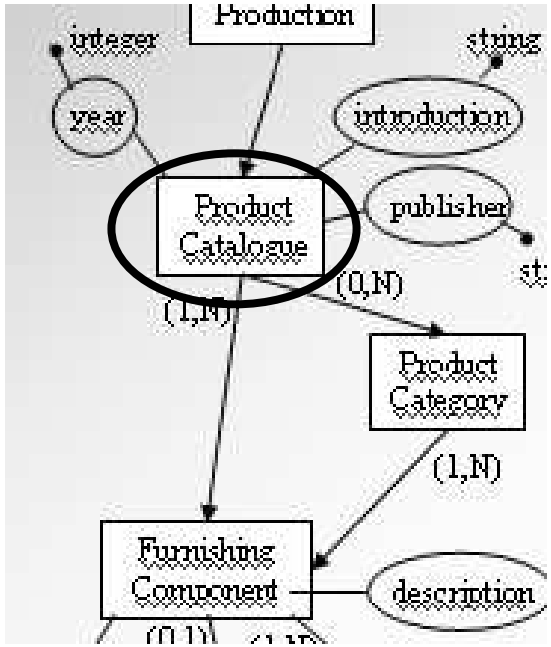
Input = *Clusters of XClasses with affinity*

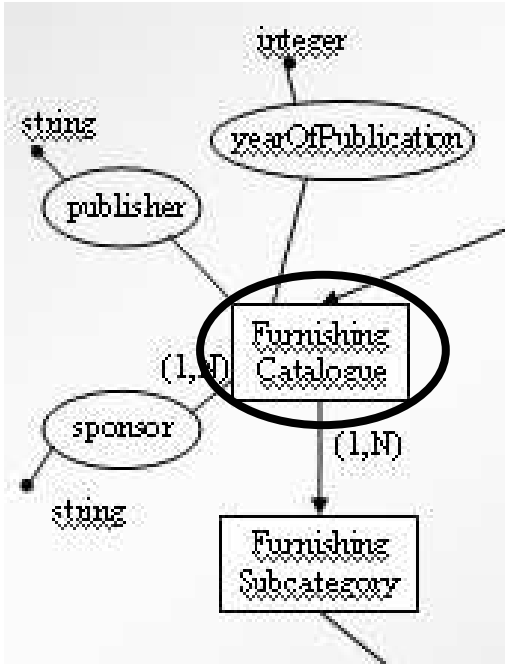
Output = *Global Concepts*

For each cluster of XClasses, its representative global concept is derived by reconciling the properties, links, referenced XClasses and attributes

- ***Name reconciliation:*** the mediated name of two features f_1 and f_2 is name(f_1) or name(f_2) or one of their synonyms or hyperonyms
- ***Type reconciliation:*** the mediated type of two features f_1 and f_2 is type(f_1) or type(f_2), if they have the same type; the type less restrictive, otherwise
- ***Cardinality reconciliation:*** the mediated cardinality of two features and is defined as the less restrictive

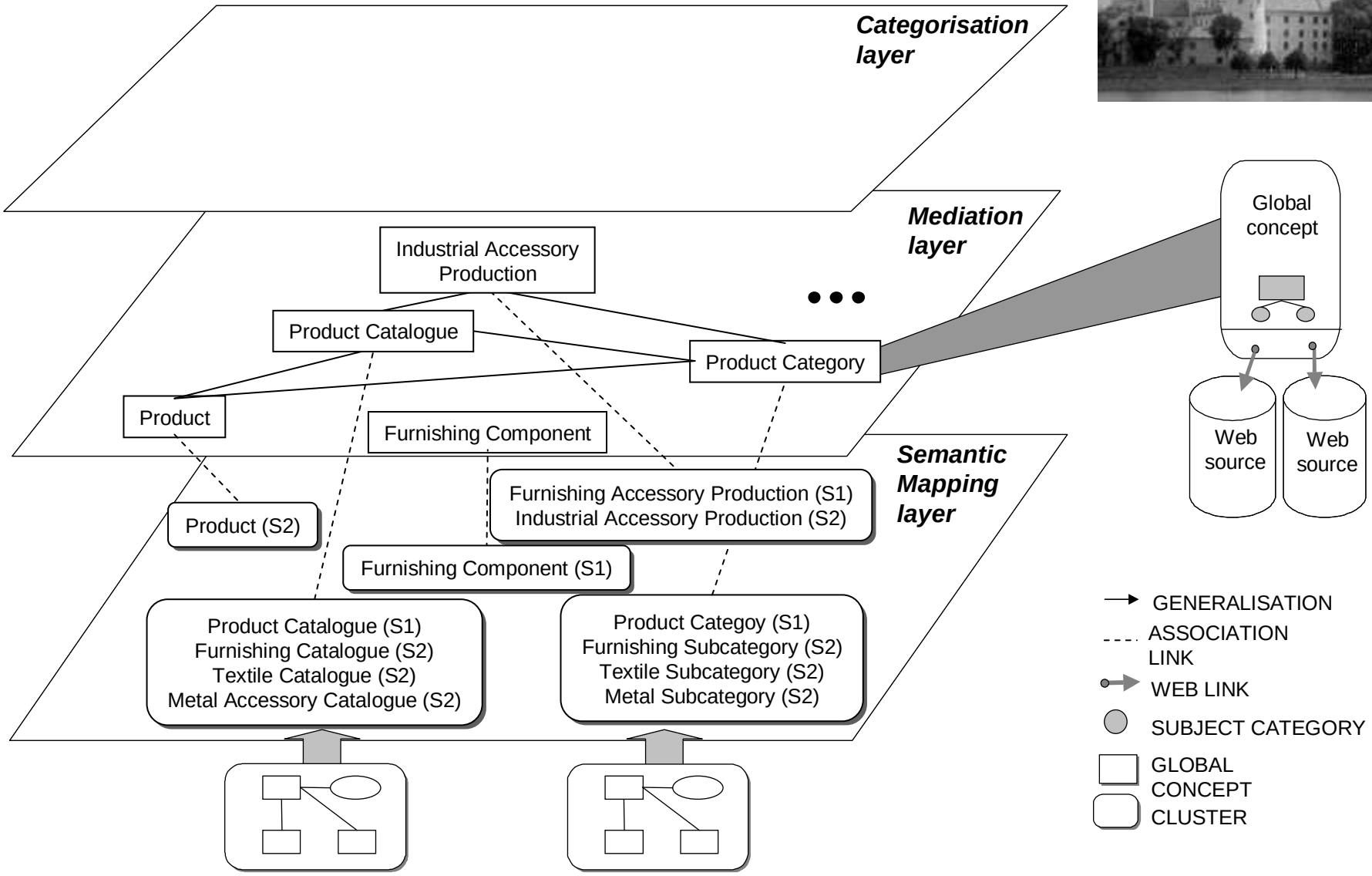
Example





Name:	ProductCatalogue
Content Model:	(introduction, year, publisher, sponsor, ProductCategory, FurnishingComponent)
Properties:	{(introduction,string,(1,1)), (year,integer,(1,1)), (publisher, string,(1,1)), (sponsor,string,(1,N))}
RefX:	{(ProductCategory,(0,N)), (FurnishingComponent,(1,N))}
Attributes:	{ }

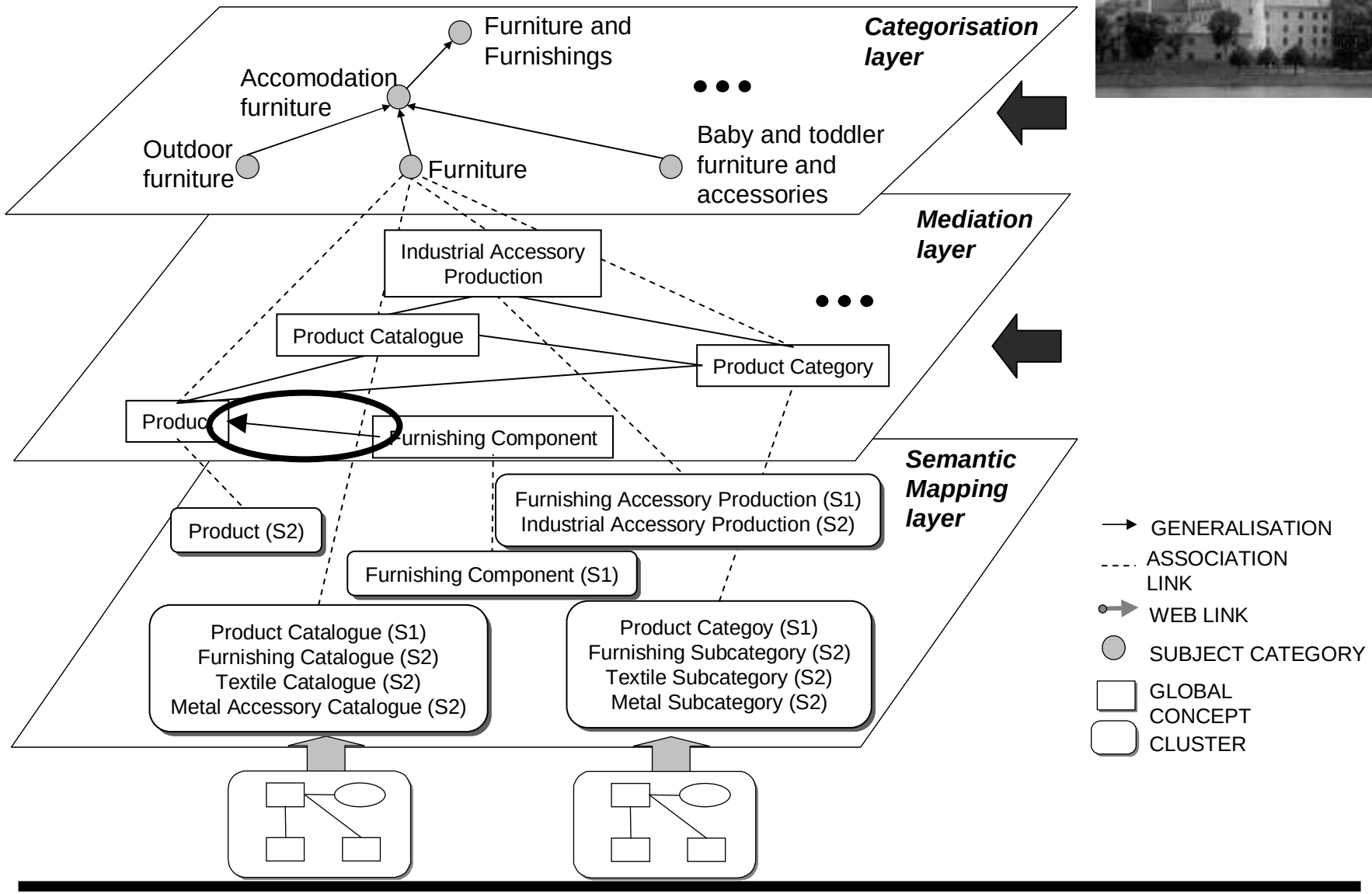
Three-layer domain ontology



Phase 3:
Synthesis and Categorization

- Global concepts are organized with semantic relationships among them (generalization, disjunction, equivalence)
- Global concepts in the mediation layer are associated to one or more subject categories provided in available standard taxonomies with the supervision of a domain expert

Three-layer domain ontology



Phase 4: Implementation

Input *= Informal description of the three layer ontology*

Output *= ODL_I^3 representation of the ontology*

- Ontology concepts and links are formally represented in the ODL_I^3 language
- ODL_I^3 is a subset of the ODMG-93 standard for object databases
- Translation of the ODL_I^3 representation into a Description Logic permits to perform automatically inferences and consistency checks on the ontology
- *Example:*
  - new semantic links can be discovered and added to the ontology;
  - inconsistencies can be discovered and properly treated

The VISPO experimentation

Sources *= three industrial catalogs of professional mechanical tools (~150 local concepts uniformly distributed in the three sources)*

Obtained benefits:

- Starting from a great number of local concepts, a small set of global concepts is obtained, enhancing their consultation and discovery
- Richer descriptions of global concepts with respect to the local ones are created
- Ontological links allow to reach local concepts in each single sources starting from the selected global concepts

Conclusions and future work

- We have presented a methodology for building a three-layer ontology for XML data sources in a specific domain
- The methodology is partially supported by the ARTEMIS tool environment
- The proposed methodology has been succesfully experimented in the VISPO virtual district
- Future work includes:
  - strategies for ontology maintenance
  - design of a inference engine to extract new information from ontological concepts and semantic relationships between them