The *Ecumene* Experience to Data Integration in Cultural Heritage WISs

Alessio Bechini – Univ. of Pisa, Italy
Andrea Tomasi – Univ. of Pisa, Italy
Gianni Ceccarelli – Hyperborea scrl

Outline

- Cultural Heritage Data: Characteristics, Management and Utilization Issues
- Ecumene Project WIS Architecture
- Data Integration and Usage
- The Role of Standards
- Conclusions

Data Organization and Management: User Needs

- Starting points for any kind of analysis:
  - availability of data
  - possibility to retrieve required info
  - Completeness and accuracy of the result set
- Characteristics of data source:
  - Variety of data types
  - Cultural Heritage information spread over local data repositories

Cultural Heritage Data Availability

- In Europe a large amount of digitalized data on Cultural Heritage is even more available
- Such material represents a precious information source for people involved in the humanities
- Management of Cultural Heritage Data has to face with different types and with various source of information
- The data repositories are hold by different institutions, in different countries

Cultural Heritage Data Heterogeneity

Data Category	Description
Artistic & Historical Objects	Works of art (paintings, sculptures, jewels, tissues, etc.)
Historical Buildings	Architectural assets
(Historical) Archives	Archival data
(Historical) Libraries	Books and other written pieces

- There exist various models for data description of works of art, historical buildings, etc.
- The generic item representation relies on the model schema(s) chosen
- Agreement on standard models does not exist for each humanistic discipline
- Neither exist a cross-model describing different object types

Cultural Heritage Data Accessibility

- Cultural Heritage data access is usually accomplished in the specific way used in their own domain;
- Cross-references and cross-access are meaningful but are supported by standard access rules and protocols rather than by a comprehensive data model

Cultural Heritage and WISs

- The design of the data-handling part of a C.H. WIS, must be aimed at providing tools for
  - integration,
  - organization
  - search

  of Cultural Heritage information.
- The WIS services offered must be focused on the collection of a consistent view of the different objects involved in a particular search

Role of a *Semantic Layer*

- One fundamental step: design of the conceptual model for *data-driven* WIS
- This phase is often referred to as the definition of the *semantic layer*, emphasizing the importance of semantic interoperability in systems that gather information from heterogeneous sources

Standards & Interoperability

- How to increase the chances for a data source to interoperate with other databases / applications?
- Important step:
  adoption of standards for metadata representation (such as e.g. Dublin Core)
- Standards are not enough:
  integration cannot be achieved entirely through standardization.

  More flexibility is required.

A Mediated Schema

- A more flexible approach:
  proper design (possibly model-driven) of the semantic layer.
- The definition of a *mediated schema* (or *global schema*) is the central activity for the data integration within the WIS
- A mediated schema is a purely logical schema for the purpose of query issuing, often supported by an actual DB schema

The Ecumene WIS

- Objective of the "Ecumene Project" is the development of a framework to facilitate access to and to improve fruition of assorted data in the cultural/artistic heritage domain.
- We are dealing with a huge amount of data, spanning a very long period of time, relating to any kind of work of art, real estates, as well as many associated historical archives and libraries.
- In the presented case study, local data repositories mainly contains information about goods belonging to the dioceses of the Italian Catholic Church

Against Odds

- Main obstacles to the integration process:
  - Heterogeneity of the involved information systems that are in charge of providing raw data;
  - Assortment of structures for the involved data (e.g. archives, information on architectural assets, descriptions of works of art, etc.).

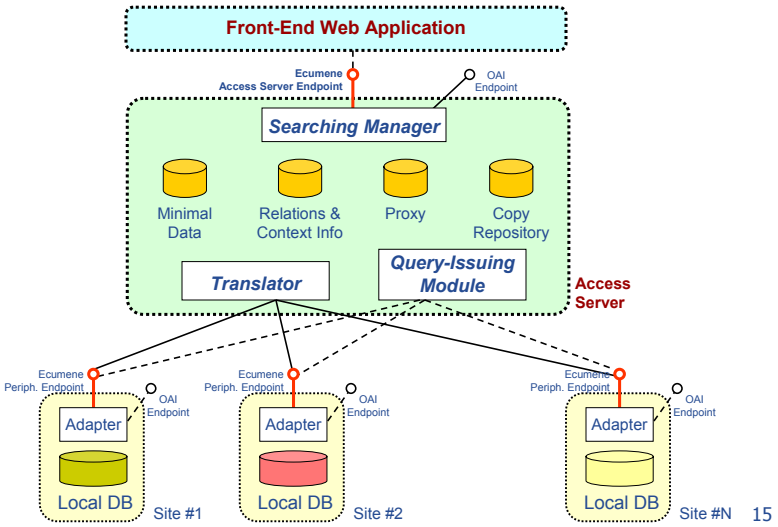
Choosing Weapons

- In the Ecumene project these problems have been overcome *from the technological standpoint* applying the following strategies:
  - To leverage the web infrastructure for the implementation of the communication facilities
  - To make use of XML in dealing with a wide variety of data sets, *and of XML-related technologies* to manage metadata representations

Roles and Layers

- The separation of roles is one of the main concerns in WIS design
- Ecumene follows this approach as well. The system is structured in different functional blocks:
  - A number of assorted local sites, containing sources of raw data;
  - An *access server*, i.e. a block dedicated to the data management, able to provide data access and export search services
  - A *front-end web application*, i.e. the interface to interact with the core WIS

Overall Architecture



System Endpoints

- Each block is an independent logic unit
- *Endpoints* are exposed by blocks to accept incoming service requests.
- Each *endpoint* is associated with a precise *schema* (or set of schemas) for the exchanged data.
- The Access Server presents a main endpoint for data delivering with the richest information content. It has been designed primarily for communication with the front-end web application.

Handling Data

- The Ecumene WIS functionality goes far beyond accessing heterogeneous distributed data: it should help the end user to fully understand the set of collected information.
- The goal is to deliver trustable, un-ambiguous data, and to relate different pieces of information within a unique picture (as typically provided by authority list support).
- The solution adopted in the data model is to define for each data type a minimal set of information and the related *context*

Data Context

- CONTEXT is the set of information describing the main entities related with an object, and the relations existing between them;
- The common entities identified are: persons (individuals and families), organizations, types of organizations, locality, administrative areas;
- A relation can express the role of an entity vs an object, or it can simple represent a reference between the two

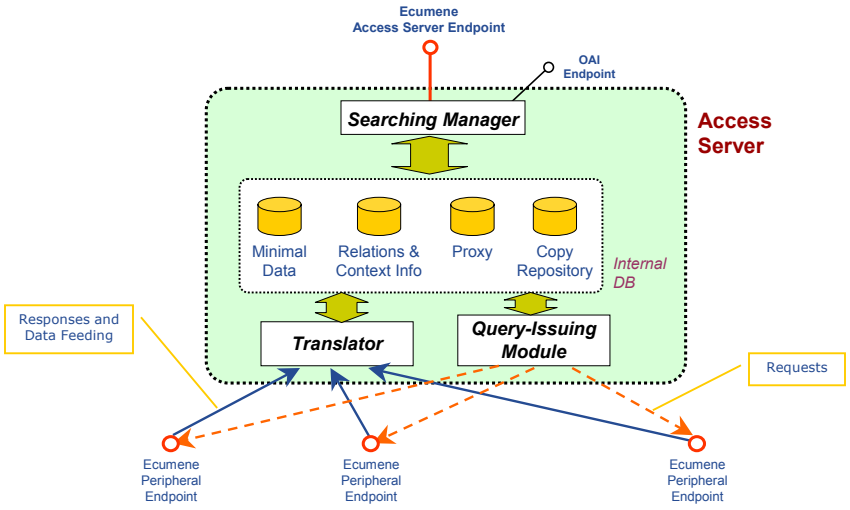
Need for an Internal DB

- Requirements for *context and relations data* drive to plan for new capabilities to the system
- The overall WIS must thus rely on an *internal database* containing the information of the mediated schema (semantic layer)
- The internal DB can contain also the set of minimal data and complements used to represent the various data types

Internal DB functionality

- The *internal database* is logically placed within the access server block, and it can offer *searching facilities*, participating in data-searching tasks and caching data
- The internal DB should be fed also in a asynchronous fashion, and not necessarily as a side-effect of the query process

Access Server Architecture



Access Server & Data Searching

- The *searching manager* provides an interface for info request.
- Upon a query posed *over the mediated schema*, it is in charge of answering making use of data
  - stored within the internal DB
  - stored within peripheral sources.
- A specific problem addressed at this level is re-formulation of the incoming queries into queries that refers to the schema of the local sources as well.

Searching Manager: Responsibilities

- The searching manager is responsible
  - for providing and handling the Ecumene Access Server Endpoint
  - for providing and handling the OAI endpoint (discussed later)
  - for any kind of authentication procedures and security issues in the communication with different categories of users.
- The interaction with peripheral data sources is delegated to:
  - A query-issuing module (for requests)
  - A translator (to gather response and for data feeding)

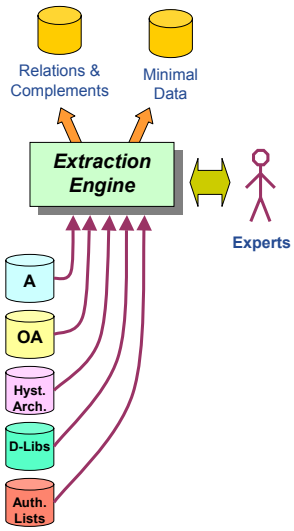
Possible Collected Info

- The outcome from the translator is the following set of information:
  - All the available raw data, organized according to the mediated schema, about specific objects belonging to the cultural heritage domain;
  - *Minimal data/metadata* to be used for searching purposes;
  - Context information (partially);
  - Relations with other data.

Role of Metadata

- In other approaches like Dublin Core and Z39.80, metadata basically represent a tool to retrieve actual data
- In Ecumene, metadata handling represents a crucial step towards the *integrated fruition* of heterogeneous data within a common description framework
- Metadata are stored in the internal DB within the Access Server
- The internal DB contains also *relations* among data/metadata, in order to further support the integrated fruition

Metadata Extraction

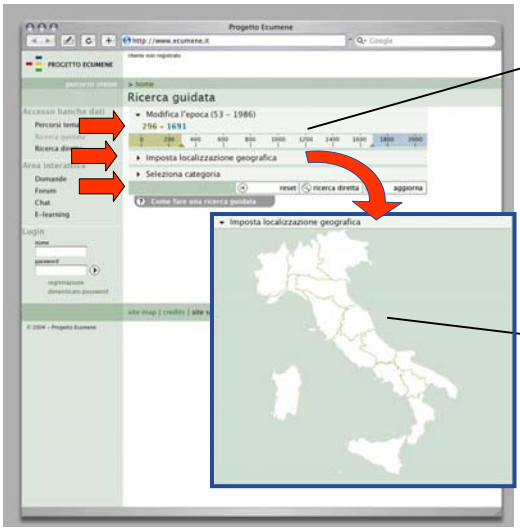


- Metadata, as well as relations and minimal data, have to be extracted from different data categories from local DBs.
- This operation often calls for direct contributions of experts, to validate context info and to setup relations not inferable from raw data
- This "glue" info represents the very contribution of Ecumene to an effective data fruition

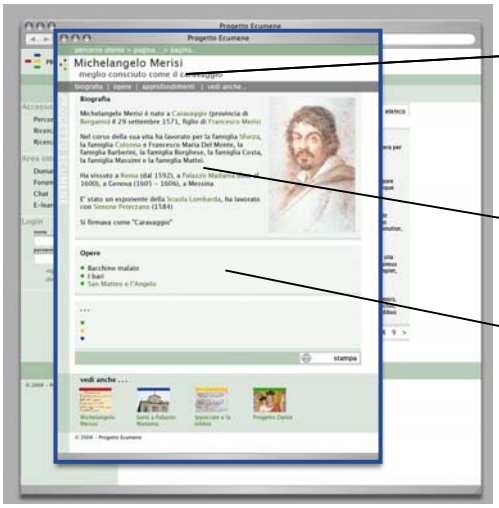
Involved Standards

- EAD (Encoded Archival description)
- The EAD schema is taken into account in designing the mediated schema
- This step is allowed by the information richness within EAD
- EAC (for complements) has been referred to for schemas of context data
- OAI (Open Archive Initiative)
- OAI, together with the related protocol PMH, is taken into account as means to:
  - export data towards clients not belonging to Ecumene
  - Import data from other institutional data sources in related areas
- This goal is met by providing OAI-compliant endpoins

Searching in Practice



Data about People



Conclusions

- Integration and organization of Cultural Heritage data
- The Ecumene project
- An open WIS architecture: towards Cultural Heritage Data sharing over the web
- Compliance with existing standards and richness of information
- A modular approach to scalability in data integration

Tributes

- I.D.S. (Informatica Distribuita e Software) Messina, Italy

- Hyperborea scrl, Pisa, Italy

- UNITELM spa, Padova, Italy

- Conferenza Episcopale Italiana