ATHENA
Access to cultural heritage networks across Europe

ECP-2007-DILI-517005

ATHENA

Implementation plan and access to content of museums through Europeana

Deliverable number	*D7.5*
Dissemination level	*Public*
Delivery date	*30 April 2011*
Status	*Final*
Authors	*Vassilis Tzouvaras, Giorgos Tolias, Giannis Kalantidis, Giorgos Goudelis, Anna Christaki, Arne Stabenau, Fotis Xenikoudakis*

e*Content*plus

This project is funded under the eContentplus programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

1 OJ L 79, 24.3.2005, p. 1.

Table of Contents

1. Executive summary

The implementation plan employed and executed by WP5 & WP7 addresses the selection and delivery of metadata from a wide range of user communities. It involves the definition and adoption of a harvesting metadata schema and the deployment and support of a technical infrastructure to define and use semantic mappings to a spectrum of commonly used data models. Work package 7 provided the ATHENA ingestion system and offered training and support to content providers in order to understand the work flow processes necessary for the delivery of content.

The development of both the content selection policy and the adoption of the common metadata schema often involved a critical review, revision or enhancement of the available metadata schemas and technical standards employed in the cultural domain. This report introduces the main concepts of metadata framework and identifies the most important and prominent metadata standards, vocabularies and representation languages in the cultural domain. Finally, it illustrates an overview of the procedure and tool that was deployed within ATHENA, in order to establish interoperability between provider's metadata and the ATHENA repository. Related material and a more elaborate description of the ingestion platform and its implementation can be found in deliverables D7.1 and D7.4.

Present deliverable complements, D5.5 'Implementation plan for taking the content into Europeana' in documenting related material and highlighting the workflow processes that were followed in order to

- Understand the required steps for the delivery of content
- Maintain and execute an implementation plan in line with the requirements of Europeana
- Implement the plan to ensure museums content gathered during the ATHENA project is searchable and retrievable in a cross sector search within Europeana

2. Introduction

This deliverable reports on the results of Task 7.5 (Creation, management and execution of implementation plan) of WP7 concerning the technical standards and metadata models of interest for the design of the ATHENA harvesting strategy, the technical infrastructure that supports the ingestion work flow and, the resources to train and support the project's users. Its main objective is to provide the information and technical specifications for the alignment of provider's metadata with ATHENA and, for the interoperability of the later with the Europeana repository and relevant standardized and widely used data models for digital cultural heritage. This deliverable introduces the available cultural metadata standards, focusing on their implementation and discussing the prevalent representation languages. Based on the results obtained from the WP3 questionnaire completed by all content providers and the following requirement analysis, LIDO was selected as a suitable metadata schema for the ATHENA needs.

The implementation plan supports the following critical activities:

- migrating from providers' legacy schemas (whether standard or local) to LIDO,
- harvesting or aggregating metadata records that were created using shared community standard or different metadata standards and,
- semantic alignment of the LIDO schema with existing models, especially Europeana Semantic Elements and Europeana Data Model.

The rest of this deliverable is structured as follows:

Understanding Metadata and Terminology: An overview of the basic concepts behind metadata and its use a reader should grasp to have a good understanding of the ingestion process.

Technical Standards: A report on established metadata modelling approaches and their application in digital cultural heritage and, on well-known, machine-understandable representation languages used to serialize the aforementioned schemas.

Metadata Modelling in Europeana: An overview of the evolution of data models for Europeana.

ATHENA Ingestion Workflow: An overview of the procedure, reference metadata model and tool that was be deployed within ATHENA, in order to establish interoperability between provider's metadata and the ATHENA repository.

3. Understanding Metadata and Terminology

3.1 Knowledge representation and metadata

Knowledge Representation is a two sided concept. Knowledge on cultural heritage objects is represented in metadata schemas (mainly in the semantic description of a cultural heritage object, not in the technical or administrative part of a metadata schema). [Synonym: metadata model]. Knowledge on cultural heritage object is also represented in 'controlled vocabularies' or 'knowledge organization systems' of all kinds, therewith controlling the content of several metadata elements or attributes of a metadata schema. [Synonym: authority files].

Metadata; many definitions have been provided for the term metadata, e.g. "a cloud of collateral information around a data object" as defined by Clifford Lynch (director of the Coalition for Networked Information). Metadata (Greek: meta- + Latin: data "information") are defined literally as "data about data" or "information about information", but the term is normally understood to mean structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource. A resource may be anything that has identity and a resource may be digital or non-digital. Operations might include, for example, disclosure and discovery, resource management (including right management) and the long-term preservation of resource. For a single resource different metadata may be required to support these different functions. A metadata record is a file of information, compiled (automatically and/or manually) in the format of the metadata schema concerned, which captures the basic characteristics of a data or information resource (e.g. a cultural heritage object). In other words, metadata refers information that describes information sources or objects, e.g. a Dublin Core record or a record from the catalogue of an archive.

The term metadata is used differently in different communities. Some use it to refer to machine understandable information, while others use it only for records that describe electronic resources. In the library environment, metadata is commonly used for any formal scheme of resource description, applying to any type of object, digital or non-digital. Traditional library cataloguing is a form of metadata; MARC 21 and the rule sets used with it, such as AACR2, are metadata standards. Other metadata schemes have been developed to describe various types of textual and non-textual objects including published books, electronic documents, archival finding aids, art objects, educational and training materials, and scientific datasets.

Metadata is sometimes classified according to the functions it is intended to support. In practice, individual metadata schemas often support multiple functions and overlap the categories:

Descriptive metadata is mainly information to identify and describe the object or information source and what it expresses. These metadata include the author/title cataloguing as well as the subject indexing. In other words, the descriptive metadata include the subgroup of the objective elements that formally describe the object (e.g. identification number, title, creation date, creator name, the language of the object, physical media). And the subgroup of semantic elements (also called analytical metadata) that contain information on the subject of the object to enhance access to the resources contents (e.g. subject keywords, classification codes and abstract). Note that descriptive metadata can be of a technical character, think of for instance 'compression Schema' (this is the algorithm used to compress the audiovisual essence), the number of pages (book), black and white / colour (photograph, film) or specific information on the storage medium or carrier.

Structural metadata describes the logical or physical relationships between the parts of a compound object. For example a physical book consists of sequences of pages to forma a chapter.

Technical metadata describe the technological characteristics of the related object (e.g. data that must be available to be able to use out the material, file locations, authentication and security information, characteristics needed for computer programming and database management)

Administrative metadata provides information for managing and administering the objects concerned (e.g. content provider name, acquisition information, copyrights, location information, record language and number). There are several subsets of administrative metadata; two that sometimes are listed as separate metadata types are:

- *Rights management* metadata, which deals with intellectual property rights and
- *Preservation metadata*, which contains information needed to archive and preserve a resource (as it was published in 1988 by Working Group on Preservation Issued of Metadata constituted by the Research Libraries Group -RLG)

3.2 Metadata Framework

A metadata framework can be viewed as having five key components:

- A *schema* (the categories of information you choose to record)
- *Vocabulary* (specific ‘words' or ‘values' you enter into those categories)
- *Conceptual model* - the underlying model that describes how all the information and concepts inherent in a resource are related to one another
- *Content standard* - practical standards that describe how specific information (e.g. vocabularies) should be entered within metadata schema categories (e.g. Cataloguing Cultural Objects)
- *Encoding* - which is concerned with the way the metadata is presented (e.g. XML)

Based on the above structure of a “metadata framework”, in the rest of this section we attempt to provide some definitions and descriptions of the basic components of a metadata framework along with the description of other key terms related to this framework.

Metadata schema refers to the format and structure of metadata that is often dictated in a set of rules, called metadata schema. It can be defined as:

- A full, logically organised structure of relations between defined (groups) of metadata and the information objects they describe.
- A set of rules for encoding information that supports specific communities of users. A metadata schema consists of several metadata elements. For some elements the input is free (e.g. Title), for other elements the input is guided by syntactical rules or guidelines or even restricted by controlled vocabularies of all kinds (e.g. thesaurus for subject keywords or closed term list for object type).

Metadata element is an item, or an editorial part of metadata. A semantic metadata element is an element from the descriptive metadata that describes the cultural heritage object. A metadata element name is given to a data element in, for example, a data dictionary or metadata schema or registry. In a formal data dictionary, there is often a requirement that no two data elements may have the same name, to allow the data element name to become an identifier, though some data dictionaries may provide ways to qualify the name in some way, for example by the application system or other context in which it occurs. A data element definition is a human readable phrase or

sentence associated with a data element within a data dictionary that describes the meaning or semantics of a data element.

Controlled Vocabulary; A limited set of terms that must be used to index | represent | tag the subject matter | content of documents | objects (indexing tools in use to describe a cultural heritage object). Examples: Alphabetic lists of “approved” words or phrases, thesauri, subject heading systems, classification schemes, ontologies, taxonomies. These examples illustrate that controlled vocabularies are largely applied for subject keywords or generic concept identification. However, controlled vocabularies or lists of preferred terms are also applied for other metadata elements, e.g. person names like author or creator, names of historical people and corporate bodies on the cultural heritage object or as its subject of the cultural heritage object, geographic places (actual location of the cultural heritage object / place of creation / place where the cultural heritage object was found / place as subject of the cultural heritage object) and organisation names. See also: Authority files in this section.

Classification Schemes, taxonomies and Categorization Schemes; these terms are often used interchangeably. Although there may be subtle differences from example to example, in general these types of knowledge representation provide ways to separate entities into buckets or relatively broad topic levels. Some examples provide a hierarchical arrangement of numeric or alphabetic notation to represent broad topics. These types of knowledge representation may not follow the strict rules for hierarchy required in the ANSI NISO Thesaurus Standard (Z39.19) (NISO), and they lack the explicit relationships presented in a thesaurus. Examples of classification schemes include the Library of Congress Classification Schedules (an open, expandable system), the Dewey Decimal Classification (a closed system of 10 numeric sections with decimal extensions), and the Universal Decimal Classification (based on Dewey but extended to include facets). Subject categories are often used to group thesaurus terms in broad topic sets, outside the hierarchical scheme of the thesaurus. Taxonomies are increasingly being used in object oriented design and knowledge management systems to indicate any grouping of objects based on a particular characteristic. "Taxonomy" may also refer to a scheme that presents subject elements in a hierarchical arrangement based on some characteristic.

Thesauri are knowledge organization systems based on concepts, and they show relationships between terms. Relationships commonly expressed in a thesaurus include hierarchy, equivalence, and associative (or related). These relationships are generally represented by the notation BT (broader term), NT (narrower term), SY (synonym), and RT (associative or related). There are standards for the development of monolingual thesauri (NISO, 1998; ISO, 1986) and multi-lingual thesauri (ISO, 1985). It should be noted that the definition of a thesaurus in these standards is often at variance with schemes that are actually called thesauri. There are many thesauri that do not follow all the rules of the standard, but are still generally thought of as thesauri. Many thesauri are very large (more than 50,000 terms). Most were developed for a specific discipline, or to support a specific product or family of products.

Subject headings; this scheme provides a set of controlled terms to represent the subjects of items in a collection. Subject heading lists can be extensive, covering a broad range of subjects. However, the subject heading lists structure is generally very shallow, with a limited hierarchical structure. In use, subject headings tend to be pre-coordinated, with rules for how subject headings can be joined to provide more specific concepts. Examples include the Medical Subject Headings (MeSH) and the Library of Congress Subject Headings (LCSH).

Authority files are lists of terms that are used to control the variant names for an entity or the domain value for a particular field. Examples include names for countries, individuals, and organizations. Non-preferred terms may be linked to the preferred versions. This type of knowledge organization generally does not include a deep organization or complex structure. The presentation may be alphabetical or organized by a shallow classification scheme. There may be some limited hierarchy applied in order to allow for simple navigation, particularly when the authority file is being accessed manually or is extremely large. Specific examples of authority files include the Library of Congress Name Authority File and the Getty Geographic Authority File.

Semantic Network; with the advent of natural language processing, there have been significant developments in the area of semantic networks. These knowledge organization systems structure concepts and terms not as hierarchies but as a network or a Web. Concepts are thought of as nodes with various relationships branching out from them. The relationships generally go beyond the standard BT, NT and RT. They may include specific whole-part relationships, cause-effect, parent-child, etc. One of the most noted semantic network is Princeton's WordNet, which is now used in a variety of search engines.

Ontology is a data model that represents the existing knowledge within a domain and is used to reason about the objects in that domain and the relations between them. Ontologies are used as a form of knowledge representation about the world or some part of it. Ontologies (as defined in www.wikipedia.org) generally describe:

- Individuals (the basic or "ground level" objects); Classes (sets, collections, or types of objects).
- Attributes (properties, features, characteristics, or parameters that objects can have and share).
- Relations (ways that objects can be related to one another).

Therefore thesauri and classification schemes can be regarded as ontologies with a relatively little number of relationships.
Ontologies can represent complex relationships between objects, and include the rules and axioms missing from semantic networks. Ontologies that describe knowledge in a specific area are often connected with systems for data mining and knowledge management.
Upper Ontology (top-level ontology, or foundation ontology); an ontology that describes very general concepts, applicable across all domains. The aim is to have a large number of ontologies accessible under this upper ontology.

Mark-up ontology languages; these languages use a mark-up scheme to encode knowledge, most commonly XML (SHOE, XOL, DAML+OIL, OIL, RDF, RDF Schema, OWL)

The **Semantic Web** provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming. The Semantic Web intent is to enhance the usability and usefulness of the Web and its interconnected resources. A Semantic Web-compatible mark-up guarantees a rich use (mainly in retrieval functionality) of the metadata on cultural heritage objects in combination with several ontologies related to the cultural heritage domain. A domain ontology (or domain-specific ontology) models a specific domain, or part of the world. An ontology on arts can be used to say, for instance that “Picasso” is a “Painter”, and that a “Painter” is an “Artist”. The combination of such ontologies together with indexes automatically provides the end user with several extra ways to navigation through the collection. E.g. this combination can present all cultural heritage objects

from museums in Spain, without the need for the content providing partners to manually add extra metadata to the descriptions of their objects.

An **XML schema** is a description of a type of XML document, typically expressed in terms of constraints on the structure and content of documents of that type, above and beyond the basic syntax constraints imposed by XML itself. An XML schema provides a view of the document type at a relatively high level of abstraction. There are languages developed specifically to express XML schemas. The Document Type Definition (DTD) language, which is native to the XML specification, is a schema language.

A **Data Model** is a model that describes in an abstract way how data are represented in a business organization, an information system or a database management system. This term is ambiguously defined to mean:

- how data generally are organized, e.g. as described in Database management system. This is sometimes also called "database model" or,
- how data of a specific business function are organized logically (e.g. the data model of some business).

While simple data models consisting of few tables or objects can be created "manually", large applications need a more systematic approach. Within the relational database modelling community, the entity-relationship model method is used to establish a domain-specific data model. In computer science, an entity-relationship model (ERM) is a model providing a high-level description of a conceptual data model. Data modelling provides a graphical notation for representing such data models in the form of entity-relationship diagrams (ERD). A conceptual schema, or high-level data model or conceptual data model, is a map of concepts and their relationships, for example, a conceptual schema for a karate studio would include abstractions such as student, belt, grading and tournament." A data model, especially the concepts or entities and relationships of the model, dictate the metadata elements that are needed in the metadata schema that goes along with the data model.

Metadata Crosswalks. The interoperability and exchange of metadata is further facilitated by metadata crosswalks. A crosswalk is a mapping of the elements, semantics, and syntax from one metadata schema to those of another. A crosswalk allows metadata created by one community to be used by another group that employs a different metadata standard. The degree to which these crosswalks are successful at the individual record level depends on the similarity of the two schemes, the granularity of the elements in the target scheme compared to that of the source, and the compatibility of the content rules used to fill the elements of each scheme. Crosswalks are important for virtual collections where resources are drawn from a variety of sources and are expected to act as a whole, perhaps with a single search engine applied. While these crosswalks are key, they are also labour intensive to develop and maintain. The mapping of schemes with fewer elements (less granularity) to those with more elements (more granularity) is problematic.

4. Technical Standards

As explained earlier, metadata are data used to describe other data structured in formats easily understood by machines. One of the most familiar ways to organize metadata is through ontologies. Metadata standards are ontologies that define the vocabulary that describes the concepts and the relations among them in the specified domain of interest. Metadata schema refers to the format and structure of metadata that is often dictated in a set of rules. Many different metadata schemas are being developed in a variety of user environments and disciplines.

It should be noted here that “schemas" is used in a broad sense, to describe a set of categories (i.e. "elements" or "units") of information used to describe resource. Metadata schemas can be differentiated in many different ways, for example:

- Their size and scope (e.g. comprehensive or 'core'; emphasis on description, administration, preservation; concern with single items or collections or both)
- Things they describe (e.g. art images, audio, video, objects, books, places)
- Communities they serve (e.g. libraries, museums, educators)

Furthermore distinctions between schemas, conceptual models, content standards, and encoding standards are often not fixed or discreet. Several metadata schemas describe their underlying conceptual models, provide guidance on what data might to be entered within their categories, or indicate how the metadata should be encoded. Dublin Core, for example, provides all of these.
Since the ATHENA project deals with cultural heritage content, WP3 results have highlighted some of the most important metadata standards and schemas used within the cultural heritage domain and more specifically, descriptive data structure standards for different kinds of community resource descriptions.

In this section we will refer to some of the technical standards that are used for the harvesting and remediation of metadata, specifically:

- Markup languages and schemas for encoding metadata in machine-readable syntaxes.
- Ontologies for semantic mediation between data standards.
- Protocols for distributed search and metadata harvesting, for example, the Z39.50 family of information retrieval protocols (Z39.50,48 SRU/SRW49), SOAP,50 and OAI-PMH.51

4.1 Protocols for distributed search and metadata harvesting

OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting
The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) provides an application-independent interoperability framework based on metadata harvesting. There are two classes of participants in the OAI-PMH framework: Data Providers administer systems that support the OAI-PMH as a means of exposing metadata; and Service Providers use metadata harvested via the OAI-PMH as a basis for building value-added services.

To allow various repository configurations, the OAI-PMH distinguishes between three distinct entities related to the metadata made accessible by the OAI-PMH.

- **resource** - A resource is the object or "stuff" that metadata is "about". The nature of a resource, whether it is physical or digital, or whether it is stored in the repository or is a constituent of another database, is outside the scope of the OAI-PMH.
- **item** - An item is a constituent of a repository from which metadata about a resource can be disseminated. An item is conceptually a container that stores or dynamically generates

metadata about a single resource in multiple formats, each of which can be harvested as records via the OAI-PMH. Each item has an identifier that is unique within the scope of the repository of which it is a constituent. That metadata may be disseminated on-the-fly from the associated resource, cross-walked from some canonical form, actually stored in the repository, etc.

- **record** - A record is metadata in a specific metadata format. A record is returned as an XML-encoded byte stream in response to a protocol request to disseminate a specific metadata format from a constituent item.

The XML-encoding of records is organized into the following parts:

- **Header** contains the unique identifier of the item and properties necessary for selective harvesting. The header consists of the following parts:
  - the *unique identifier* -- the unique identifier of an item in a repository;
  - the *datestamp* -- the date of creation, modification or deletion of the record for the purpose of selective harvesting.
  - zero or more *setSpec* elements -- the set membership of the item for the purpose of selective harvesting.
  - an optional *status* attribute with a value of deleted indicates the withdrawal of availability of the specified metadata format for the item, dependent on the repository support for deletions.
- **Metadata**: a single manifestation of the metadata from an item. The OAI-PMH supports items with multiple manifestations (formats) of metadata. At a minimum, repositories must be able to return records with metadata expressed in the Dublin Core format, without any qualification. Optionally, a repository may also disseminate other formats of metadata. The specific metadata format of the record to be disseminated is specified by means of an argument -- the metadataPrefix -- in the GetRecord or ListRecords request that produces the record. The ListMetadataFormats request returns the list of all metadata formats available from a repository, or for a specific item (which can be specified as an argument to the ListMetadataFormats request).
- **About**: an optional and repeatable container to hold data about the metadata part of the record. The contents of an about container must conform to an XML Schema. Individual implementation communities may create XML Schema that defines specific uses for the contents of about containers. Two common uses of about containers are:
  - *rights statements*: some repositories may find it desirable to attach terms of use to the metadata they make available through the OAI-PMH. No specific set of XML tags for rights expression is defined by OAI-PMH, but the about container is provided to allow for encapsulating community-defined rights tags.
  - *provenance statements*: one suggested use of the about container is to indicate the provenance of a metadata record, e.g. whether it has been harvested itself and if so from which repository, and when. An XML Schema for such a provenance container, as well as some supporting information is available from the accompanying Implementation Guidelines document.

4.2 Ontologies for semantic mediation between data standards

SKOS - Simple Knowledge Organisation System

The Simple Knowledge Organization System (SKOS) is an RDF vocabulary for representing semi-formal knowledge organization systems (KOSs), such as thesauri, taxonomies, classification schemes and subject heading lists. Because SKOS is based on the Resource Description Framework

(RDF) these representations are machine-readable and can be exchanged between software applications and published on the World Wide Web.
SKOS has been designed to provide a low-cost migration path for porting existing organization systems to the Semantic Web. SKOS also provides a lightweight, intuitive conceptual modeling language for developing and sharing new KOSs. It can be used on its own, or in combination with more-formal languages such as the Web Ontology Language (OWL). SKOS can also be seen as a bridging technology, providing the missing link between the rigorous logical formalism of ontology languages such as OWL and the chaotic, informal and weakly-structured world of Web-based collaboration tools, as exemplified by social tagging applications.

The SKOS Core Vocabulary is a set of RDF properties and RDFS classes that can be used to express the content and structure of a concept scheme as an RDF graph. SKOS Core provides a model for expressing the basic structure and content of concept schemes. A 'concept scheme' is defined here as: a set of concepts, optionally including statements about semantic relationships between those concepts. Thesauri, classification schemes, subject heading lists, taxonomies, 'folksonomies', and other types of controlled vocabulary are all examples of concept schemes. Concept schemes are also embedded in glossaries and terminologies.

Data model. The SKOS data model is formally defined as an OWL Full ontology. SKOS data are expressed as RDF triples, and may be encoded using any concrete RDF syntax (such as RDF/XML or Turtle). The SKOS data model views a knowledge organization system as a concept scheme comprising a set of concepts. These SKOS concept schemes and SKOS concepts are identified by URIs, enabling anyone to refer to them unambiguously from any context, and making them a part of the World Wide Web.

Concepts. The fundamental element of the SKOS vocabulary is the concept. Concepts are the units of thought - ideas, meanings, or (categories of) objects and events - which underlie many knowledge organization systems. As such, concepts exist in the mind as abstract entities which are independent of the terms used to label them. The basic features of SKOS concepts are:

- SKOS concepts can be **labelled** with any number of lexical (UNICODE) strings, in any given natural language, such as English or Japanese (written here in hiragana). One of these labels in any given language can be indicated as the preferred label for that language, and the others as alternative labels. Labels may also be "hidden", which is useful where a knowledge organization system is being queried via a text index. SKOS concepts can be assigned one or more notations, which are lexical codes used to uniquely identify the concept within the scope of a given concept scheme. While URIs are the preferred means of identifying SKOS concepts within computer systems, notations provide a bridge to other systems of identification already in use such as classification codes used in library catalogues.
- SKOS concepts can be **documented** with notes of various types. The SKOS data model provides a basic set of documentation properties, supporting scope notes, definitions and editorial notes, among others. This set is not meant to be exhaustive, but rather to provide a framework that can be extended by third parties to provide support for more specific types of note.
- SKOS concepts can be **linked** to other SKOS concepts via semantic relation properties. The SKOS data model provides support for hierarchical and associative links between SKOS concepts. Again, as with any part of the SKOS data model, these can be extended by third parties to provide support for more specific needs.
- SKOS concepts can be **grouped** into collections, which can be labeled and/or ordered. This feature of the SKOS data model is intended to provide support for node labels within thesauri,

and for situations where the ordering of a set of concepts is meaningful or provides some useful information.

- SKOS concepts can be **mapped** to other SKOS concepts in different concept schemes. The SKOS data model provides support for four basic types of mapping link: hierarchical, associative, close equivalent and exact equivalent.

CIDOC – Conceptual Reference Model (CRM)

The CIDOC Conceptual Reference Model (CRM) is a formal ontology that provides definitions and a structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. The purpose of CIDOC CRM is to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. It contributes to the specification of a common ground for domain experts in conceptual modelling. Therefore, it promotes an extensible semantic framework where information deriving from sources such as libraries and archives, may be integrated.

History. The CRM was developed by different teams of experts such as archaeologists, art historians, and computer scientists following the standards of International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). The first schema was analyzed in 1996 under the auspices of the ICOM-CIDOC Documentation Standards Working Group. Since 2000, development of the CRM has been officially delegated by ICOM-CIDOC to the CIDOC CRM Special Interest Group, which collaborates with the ISO working group ISO/TC46/SC4/WG9 to bring the CRM to the form and status of an International Standard. The present model has been accepted as ISO 21127 since September 2006. It contains 80 classes and 132 properties, representing the semantics of various schemata.

Outline. The aim of CIDOC CRM was to gather scientific documentation of cultural heritage collections with a view to enabling wide area information exchange and integration of heterogeneous sources. This means that the information presented should be sufficiently descriptive and precise as it is intended not only for casual browsing but also for usage from the field experts, museums and professionals. The term cultural heritage collections includes all types of material displayed by museums, relevant collections, sites, monuments, that are related to natural history, archaeology, ethnography, historic monuments as well as collections of fine and applied arts. The CIDOC CRM is intended to cover contextual information, the historical, geographical and theoretical background in which items are placed, which reveals much of their significance. Information exchange is achieved through a more abstract perspective, clear from any specific local context. Integration between different sources determines the level of detail in CIDOC CRM. It aims to leverage contemporary technology while it enables communication with other legacy systems.

The above description of CIDOC CRM reveals the intended scope that ontology aims to cover. The practical scope of CIDOM CRM may be defined as the current coverage of the ontology. It refers to documents and sources that have been used in its elaboration.

The initial practical scope of the CIDOC CRM was defined by the International Guidelines for Museum Object Information: The CIDOC Information Categories, published in June 1995 (the Guidelines). This document, edited by a joint team of the CIDOC Data and Terminology and the Data Model Working Groups, resulted from the consolidation of two parallel initiatives: the Information Categories for Art and Archaeology Collections, 1992 and the CIDOC Relational Data Model 1995, both of which had been in gestation since 1980. The Guidelines thus represent the fruit of many years of collective effort and reflection concerning museum information and constituted an

obvious starting point for the development of the CIDOC CRM. The first published version of the CIDOC CRM, Melbourne 1998, covers all the Guidelines, with the exception of elements that fall outside the intended scope of the CIDOC CRM.

Construction Details. The basic principle underlying CRM is the explicit modelling of events. It allows for metadata representation such as creation, use, publication content summarization. Event oriented modelling permits the connection of facts into coherent representations of history. The language provided by CRM permits integration at the schema level. In fact, terminology is separated from schema. That is, classes of the ontology serve to define relationships, the ontology is then used as a schema and the classes that do not refer to range or domain restrictions for some relationship are treated as data. Furthermore, it important to note that any information
CRM conforms to some central ideas. Firstly, any ambiguity of the relationship between entities and their identifiers form a part of the historical reality to be described by the ontology and is not considered as a problem to be resolved. Entities representing the object itself are therefore, separated from those that serve for its identification.
Another idea, to which CRM conforms, is that documentation is considered as a part of the historical reality and may be described together with the documented content itself. Types and classification system form themselves part of the reality. In addition to this, the documented past can be formulated as series of events. Items, places and time form different entities are linked through events creating the impression of historical evolution. Classes that do not refer explicitly to time or space and have temporal boundaries are approximated by outer or inner bounds.
Finally, immaterial objects may be present in events through the respective physical information carriers. Although the carries may be destroyed, the immaterial objects cannot be lost.
The contents of CRM can be presented as distinct units that are linked to each other through relationships that form an IsA hierarchy. Functions supported by the relationships are identification of items by their names, classification of items, decomposition of physical and immaterial entities, temporal entities, place, times and people entities. In addition, functions refer to participation of persistent items in temporal entities, location of temporal entities in space and participation of time and physical things in space and reference of information object to real world items.
CIDOC CRM supports a wide range of classes and relationships that are considered as generic. Furthermore, the fact that terminology is separated from schema favours stability and therefore a considerable chance of agreement on common semantics for schema-level semantics rather than terminology.

FRBR – Functional Requirements for Bibliographic Records

FRBR is a conceptual model for describing information resources within a library context. It describes particular entities (e.g. Item or Person) and their relationships (e.g. Item is owned by Person). Like the CRM, FRBR is not a metadata schema, but a model that can be used to analyse existing schemas or influence the development of new schemas or content standards. It is currently being drawn on in the development of the RDA content standard. FRBR is an international model, published in 1998 by a working group of the International Federation of Library Associations (IFLA). A working group was established in 2002 to review and further develop the standard. One of its tasks is to look at how FRBR and the CRM can be related.
From 1992-1995 the IFLA Study Group on Functional Requirements for Bibliographic Records (FRBR) developed an entity-relationship model as a generalized view of the bibliographic universe, intended to be independent of any cataloguing code or implementation. The FRBR report1 itself includes a description of the conceptual model (the entities, relationships, and attributes or metadata as we'd call them today), a proposed national level bibliographic record for all types of materials,

and user tasks associated with the bibliographic resources described in catalogs, bibliographies, and other bibliographic tools.

Terminology. FRBR offers us a fresh perspective on the structure and relationships of bibliographic and authority records, and also a more precise vocabulary to help future cataloguing rule makers and system designers in meeting user needs. Before FRBR our cataloguing rules tended to be very unclear about using the words "work," "edition," or "item."2 Even in everyday language, we tend to say a "book" when we may actually mean several things. For example, when we say "book" to describe a physical object that has paper pages and a binding and can sometimes be used to prop open a door or hold up a table leg, FRBR calls this an "item."
When we say "book" we also may mean a "publication" as when we go to a bookstore to purchase a book. We may know its ISBN but the particular copy does not matter as long as it's in good condition and not missing pages. FRBR calls this a "manifestation."
When we say "book" as in 'who translated that book,' we may have a particular text in mind and a specific language. FRBR calls this an "expression." When we say "book" as in 'who wrote that book,' we could mean a higher level of abstraction, the conceptual content that underlies all of the linguistic versions, the story being told in the book, the ideas in a person's head for the book. FRBR calls this a "work."

Entities. The JSC is examining AACR2 to update the terminology to be clearer when we mean **work**, **expression**, **manifestation**, and **item**, following these FRBR "**Group 1**" entities.
FRBR's "**Group 2**" entities are **person** and **corporate body** that are related to "Group 1" entities through specific relationships. These relationships reflect the role of the person or corporate body with respect to the work, expression, manifestation, or item. FRBR's model shows us how important such role information is for performing user tasks and for assisting a user to navigate through the bibliographic universe. (Note: This universe may be limited to our local catalogue or may be the realm of global resources available through the Web.) The value of this 'role' information becomes very apparent in light of FRBR. We need to regain the lost link of relator terms and codes in our bibliographic records. It is time to re-examine a change in cataloguing practice that abandoned use of "relator" terms and codes to cut cataloguing costs. In hindsight we can see that decision was unfortunate for future users of our records and should be reversed to allow greater flexibility in manipulating bibliographic data and offering better information to users as they navigate our catalogues.

FRBR "**Group 3**" entities are the subjects of works. These can be **concepts, objects, events, places**, and any of the "Group 1" or "Group 2" entities. For example, you can have a work about another work or a work about a person or corporate body.

Bibliographic Relationships. A lot of attention has been given to the inherent relationships among the entities in the Group 1 hierarchy of work, expression, manifestation, and item. Additionally, there are many other rich content relationships that enable collocation of related items and navigation through the sometimes complex network of the bibliographic universe. Content relationships can be viewed as a continuum from works/expressions/manifestations/ items. Moving left to right along this continuum we start with some original work and related works and expressions and manifestations that can be considered "equivalent," that is, they share the same intellectual or artistic content as realized through the same mode of expression. Next we come to works/expressions/manifestations that are related through a "derivative" relationship. These comprise a range of new expressions, such as translations, different performances, slight modifications and editions that move along the continuum across a magic line where they become a new work yet still related to some original work. To the far right on this continuum we find

'descriptive' relationships that involve new works describing some original work. FRBR reminds us of the importance of these relationships and keeps us focused on those of most importance to meeting user tasks.

Whole/part and part to part relationships are also in FRBR. When we provide bibliographic control for electronic digital resources, we find these whole/part and part to part relationships especially relevant. For example, a Web site may be viewed as the "whole" and the components as its "parts," or we may view the whole digitized resource and its components as the parts that will need to be tracked through technical metadata for storing and displaying that digital information. The part to part relationships include 'sequential' and 'accompanying' or 'companion' relationships. Companion relationships can be either dependent or independent, which will influence how many bibliographic records we would make for the related works and their manifestations. In fact the number of records we make is a decision made up front by the cataloguer based on local policies reflecting local user needs. We may choose to catalogue at various levels: the collection of works (FRBR calls this an aggregation), an individual work, or a component of a work. At the collection level we may include a description of all the parts and should provide access to each component. At the component level we should provide a link to relate to the larger "whole." FRBR reminds us that these relationships are important factors for fulfilling user tasks regardless of what we choose to view as the "whole."

User Tasks. The FRBR user tasks are **find, identify, select, and obtain**.
'Find' involves meeting a user's search criteria through an attribute or a relationship of an entity. This can be seen to combine both the traditional "find" and "collocate" objectives of a catalogue.
'Identify' enables a user to confirm they have found what they looked for, distinguishing among similar resources.
'Select' involves meeting a user's requirements with respect to content, physical format, etc. or to reject an entity that doesn't meet the user's needs.
'Obtain' enables a user to acquire an entity through purchase, loan, etc., or electronic remote access.

FRBR oo

The FRBRoo is a formal ontology intended to capture and represent the underlying semantics of bibliographic information and to facilitate the integration, mediation, and interchange of bibliographic and museum information. The FRBR model was originally designed as an entity-relationship model by a study group appointed by the International Federation of Library Associations and Institutions (IFLA) during the period 1991-1997, and was published in 1998. Quite independently, the CIDOC CRM model was being developed from 1996 under the auspices of the ICOM-CIDOC (International Council for Museums – International Committee on Documentation) Documentation Standards Working Group. The idea that both the library and museum communities might benefit from harmonising the two models was first expressed in 2000 and grew up in the following years. Eventually, it led to the formation, in 2003, of the International Working Group on FRBR/CIDOC CRM Harmonisation, that brings together representatives from both communities with the common goals of: a) Expressing the IFLA FRBR model with the concepts, tools, mechanisms, and notation conventions provided by the CIDOC CRM, and: b) Aligning (possibly even merging) the two object-oriented models with the aim to contribute to the solution of the problem of semantic interoperability between the documentation structures used for library and museum information, such that:

- all equivalent information can be retrieved under the same notions and
- all directly and indirectly related information can be retrieved regardless of its distribution over individual data sources;

- knowledge encoded for a specific application can be repurposed for other studies;
- recall and precision in systems employed by both communities is improved;
- both communities can learn from each other's concepts for their mutual progress;

4.3 Representation Languages and Schemas

XML - Extensible Markup Language
Extensible Markup Language (XML) is a set of rules for encoding documents in machine-readable form. It is defined in the XML 1.0 Specification produced by the W3C, and several other related specifications, all gratis open standards. XML's design goals emphasize simplicity, generality, and usability over the Internet. It is a textual data format with strong support via Unicode for the languages of the world. Although the design of XML focuses on documents, it is widely used for the representation of arbitrary data structures, for example in web services. Many application programming interfaces (APIs) have been developed that software developers use to process XML data, and several schema systems exist to aid in the definition of XML-based languages.

RDF - Resource Description Framework
The Resource Description Framework is a general-purpose language for representing information in the web. RDF's main elements are resources, properties and property values. A resource represents an object in our ontology which is connected through a property to some value which is either a literal or another resource. More than one resource can be interconnected and create a graph.

RDFS - Resource Description Framework Schema
RDFS (RDF Schema) is an extension of RDF that is more expressible by allowing classes, as well as class and property subsumption. It provides mechanisms for describing groups of related resources and the relationships between these resources as well as other characteristic of resources, such as domain and range.

OWL - Ontology Web Language
OWL is a Web Ontology Language. OWL builds on RDF and RDFS and adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties and characteristics of properties (e.g. symmetry), and enumerated classes. It is also designed for use by applications that need to process the content of information instead of just presenting information to humans providing greater machine interpretability of Web content than that supported by RDF, and RDF Schema. OWL has three increasingly-expressive sublanguages: OWL Lite, OWL DL, and OWL Full.

OWL Lite - Ontology Web Language Lite
OWL Lite supports those users primarily needing a classification hierarchy and simple constraints. For example, while it supports cardinality constraints, it only permits cardinality values of 0 or 1. It should be simpler to provide tool support for OWL Lite than its more expressive relatives, and OWL Lite provides a quick migration path for thesauri and other taxonomies. OWL Lite has a lower formal complexity than OWL DL.

OWL DL - Ontology Web Language Description Logics
OWL DL supports those users who want the maximum expressiveness while retaining computational completeness (all conclusions are guaranteed to be computable) and decidability (all computations will finish in finite time). OWL DL includes all OWL language constructs, but they

can be used only under certain restrictions (for example, while a class may be a subclass of many classes, a class cannot be an instance of another class). OWL DL is so named due to its correspondence with description logics.

OWL Full - Ontology Web Language Full

OWL Full is meant for users who want maximum expressiveness and the syntactic freedom of RDF with no computational guarantees. For example, in OWL Full a class can be treated simultaneously as a collection of individuals and as an individual in its own right. OWL Full allows an ontology to augment the meaning of the pre-defined (RDF or OWL) vocabulary. It is unlikely that any reasoning software will be able to support complete reasoning for every feature of OWL Full.

OWL semantics are based on the formalism of Description Logics. OWL Lite and OWL DL are basically very expressive description logics almost equivalent to the SHIF(D+) and SHOIN(D+) Description Logics. Description Logics (DLs) is the most recent name for a family of Knowledge Representation formalisms that represent the knowledge of an application domain by first defining the relevant concepts of the domain (its terminology), and then using this concepts to specify properties (called roles) of objects and individuals occurring in the domain (the world description). Typically we distinguish between atomic (or primitive) concepts, and complex concepts defined by using DL constructors. Different DL languages vary in the set of constructors provided. A DL Knowledge base comprises of two components, the TBox and the ABox. The TBox introduces the terminology, i.e. contains a set of concept descriptions and represents the general schema modeling the domain of interest. The ABox is a partial instantiation of this schema consisting of a set of assertions either relating individuals to classes, or individuals to each other.
One of the most attractive features of DLs is reasoning. Reasoning allows one to infer implicitly represented knowledge from the knowledge that is explicitly contained in the knowledge base. Thus, we distinguish between TBox and ABox reasoning. Many of the applications only require reasoning in the TBox but in a demanding environment ABox reasoning is also essential. Reasoning tasks in a TBox are: satisfiability (consistency), that checks if a knowledge base is meaningful; subsumption, that checks whether all the individuals of a concept are subsumed (also belong) to another concept; equivalence, that checks whether two concepts denote the same set of instances; and disjointness, that checks whether the sets of instances of two concepts are disjoint. On the other hand reasoning tasks in ABox reasoning are: instance checking, that verifies whether an individual belongs to a given concept; consistency of the knowledge base, which checks whether the knowledge base is meaningful; and realization, that finds the most specific concept an individual object is an instance of.

5. Metadata Modelling in Europeana

5.1 ESE - Europeana Semantic Elements Specification

Europeana provides integrated access to digital objects from the cultural heritage organisations of all the nations of the European Union. It encompasses material from museums, libraries, archives and audio-visual archives with the aim of making Europe's multicultural and multilingual riches discoverable together in a common on-line environment. To do this Europeana harvests and indexes the descriptive metadata associated with the digital objects. As there is no one universal metadata standard applied across the participating domains, a set of metadata elements has been developed that will allow a common set of information to be supplied to support the functionality desired by the user and needed for the operation of the underlying system. The Europeana Semantic Elements V3.3 (ESE) is an updated version of the metadata set used in the Europeana prototype in November 2008. It has been amended to include additional elements for the Rhine release of the portal in July 2010. It is a Dublin Core-based application profile providing a generic set of terms that can be applied to heterogeneous materials thereby providing a baseline to allow contributors to take advantage of their existing rich descriptions.

To provide metadata in the ESE format, it is necessary for contributors to map elements from their own metadata format to ESE. In addition to the mapping it is necessary for a normalisation process to be carried out on some values to enable machine readability. In the initial implementation of the Europeana prototype much of the mapping and normalisation was carried out centrally in the Europeana Office. This work is increasingly being passed to data providers or aggregators. An XML Schema has also been produced as a further tool to assist providers in ensuring compliance with ESE. ESE v3.3 is a sub-set of the metadata initially defined in the Europeana Metadata Requirements described in the EDLnet deliverable D2.5 "Europeana Outline Functional Specification".

The ESE v3.3 XML Schema (http://www.europeana.eu/schemas/ese/ESE-V3.3.xsd) is the XML representation of the Europeana Semantic Elements (ESE) specifications v3.3 (http://version1.europeana.eu/web/guest/technical-requirements/). This schema can be used to validate XML instances of Data Sets to be submitted to Europeana. The ESE v3.3 XML Schema extends the DC XML Schema with the addition of elements belonging to the Europeana namespace. The Europeana Semantic Elements (the ESE), consist of the 15 original Dublin Core (DC) metadata elements, a subset of the DC terms and a set of thirteen elements which were created to meet Europeana's needs. The ingestion process currently ignores the xml:lang attribute although it is present in data from some providers. It is anticipated that functionality will soon be in place to take advantage of these attributes in the display of metadata values, in particular where they are provided in one or more languages. Providers are encouraged to include them in all appropriate metadata elements.

5.2 EDM – Europeana Data Model

The Europeana Data Model (EDM) is a new proposal, still under development, aimed at being an integration medium for collecting, connecting and enriching the descriptions provided by Europeana content provider. The purpose of the open structure of EDM is to enable the linking of data, placing it in the vanguard of semantic web developments.

Outline of EDM. The initial development of Europeana was based on Europeana Semantic Elements (ESE) data model which is evolved into EDM. Particularly, ESE was developed in order

to constitute the lowest common denominator of the different data standards used for each one of the heritage sectors. Whereas, EDM reverses this reductive approach and attempts to transcend the respective information perspectives of the sectors that are represented in Europeana.
In addition, EDM has upgraded ESE with respect to its content. In terms of a digitized book, the individual chapters, illustrations and index can be understood both individually and collectively. The same holds for an archival finding aid or fonds with respect to the constituent letters, deeds, manuscripts or other items. Finally, in contrary to ESE, EDM supports the preservation of original data while still allowing interoperability.

The strength of EDM lies on the fact that its development is not based on a specific standard but rather adopts an open, cross-domain Semantic Web based framework. It can accommodate several rich standards like LIDO for museums, EAD for archives or METS for digital libraries.
Apart from its ability to support standards of high richness, it also enables data enrichment from a range of third party sources. In this way, a particular digital object from a specific provider can be enriched by metadata from another provider and at the same time by additional data held from a third party. EDM enables this interoperability while clearly providing the provenance of all the data linking to the digital object.
One of the crucial purposes of EDM is to answer the basic queries “Who?”, “What?”, “When?” and “Where?” for every digital object and to make connections between the networks that will animate Europeana’s content.

Construction Principles. EDM complies with the modeling principles that underpin the approach of the Semantic Web. Therefore, there is no fixed schema that dictates a particular way to represent the data. Instead, the common model of EDM functions as an anchor to which various finer-grained models can be attached. In this way, they become partly interoperable at the semantic level, while the data retain their original expressivity and richness.
One of the main features of EDM is that via the digital representations submitted to Europeana it enables the representation and accessing of the provided objects. It is also able to ingest the descriptive metadata supplied by various providers and at the same time to represent new information added by Europeana. In addition to this, not only it accommodates various description paradigms of the ingested objects, but also enables further enrichment of the objects by connecting the to semantically enriched resources. At the same time, it still allows for different levels of granularity in the descriptions by taking advantage of special features of semantic mapping.
The requirements and principles that EDM follows according to Europeana are:

- Distinct the provided object (book, painting, sculpture), which is the focus of the users’ interest, from its digital representations which are the elements manipulated by information systems like Europeana
- Distinct the provided object from the metadata record describing the object
- Allow for multiple records for the same object, even if they contain contradictory statements with each other
- Support objects that are composed of other objects
- Standard metadata format that can be specialized
- Standard vocabulary format that can be specialized
- Should be based on existing standards

Conceptually, four are the main concepts used in EDM and these are: ore:Aggregation, ore:Proxy, ore: EuropeanaAggregation and ens:WebResource. Following the Object Reuse and Exchange (ORE) model, EDM considers that the provided object, along with its digital representations contributed by any provider, form an aggregation that is represented as the ore: Aggregation class. Each instance of ore: Aggregation relates through the property ore:aggregates to one resource that

represents the provided object and through the property ens:hasView to one or more resources (ens: WebResource) that are digital representations of the object. Each provider contributes a different set of digital representations and a new aggregation connected to the web resources.
Inspired again by ORE model, EDM leverages the proxy mechanism to enable the representation of different views on the same resource. Each provider contributes a separate metadata record using the ore: Proxy resource, in order to represent the description of the provided object as seen from the perspective of the specific provider. A proxy is related to the resource using the ore: proxyFor property and to the provider's aggregation through the ore: proxyIn property.
Finally, Europeana creates it own aggregation, the ens: EuropeanaAggreagation, and proxy in order to be able to add new information to the original object description and representation while keeping a clear distinction from the contributed information.

6. ATHENA Ingestion Workflow

6.1 Overview

The ATHENA ingestion workflow is established to support the consortium needs through close cooperation with all relevant work packages. It was implemented on the metadata interoperability platform of NTUA that was customised and extended to support the project's requirements. Primary focus is to support the aggregation of arbitrary provider organisation data models, adopting the LIDO schema as the reference metadata model, and establishing semantic, machine understandable crosswalks for providers' datasets. The basic steps in the process that leads to semantic interoperability and allows for the ingestion and aggregation of all cultural heritage content within ATHENA and the subsequent publishing of the semantically interoperable metadata, especially for harvesting by the Europeana portal, are as follows:

- registration and access rights for users and their respective organisations, also supporting aggregators
- import and parsing of organisations' metadata records, supporting any proprietary or standardised schema
- analysis (statistical, structural and semantic) of user input in order to provide a detailed overview and to assist the user in subsequent steps with previewing and guiding capabilities
- cross-walk editing and transformation of user metadata records to a reference, well defined schema that will allow for bidirectional interoperability with all standardised outside sources

The adoption of the LIDO Metadata Schema, coupled with the loosely defined providers' input schemas, lead to an aggregation and mapping workflow that allows for an elaborate, visually guided ingestion of metadata in the repository. The fundamental principles include the disassociation of input metadata from existing metadata standards in order to avoid ambiguity over interpretation and, the ability to create and manage transformations that will apply to the actual metadata records, which subsequently (re-)define the input schema in a semantic, machine understandable way, based on its mapping to LIDO.

The ATHENA ingestion tool provides a user friendly environment that allows for the extraction and presentation of all relevant and statistical information concerning input metadata together with an intuitive mapping service for the LIDO schema, and provides all the functionality and documentation required for the providers to define their crosswalks. Transformations are editable and reusable and can be applied incrementally to user input while providing, throughout all steps, best practice examples, previews and visual indications to illustrate and guide user actions. One of the key capabilities lies in the ability to semantically enhance user metadata through conditional mapping of input elements and use of value transformation functions (e.g. string manipulation) that allow for the addition and enrichment of records even with metadata that are not present in the input.

6.2 Metadata Ingestion

In the Cultural Content Metadata Space, the largest technological challenge is to ensure syntactic and semantic interoperability across the different types of metadata that exist in the Cultural Heritage sector. The technical standards enabling interoperability form an important dimension of this work. In order to achieve semantic interoperability we need a common automatic interpretation of the meaning of the exchanged information, i.e. the ability to automatically process the information in a machine-understandable manner. The first step of achieving a certain level of common understanding is a representation language that exchanges the formal semantics of the

information. Then, systems that understand these semantics can process the information and provide web services like searching, retrieval etc.
The following figure illustrates the proposed workflow for ingesting metadata in ATHENA.

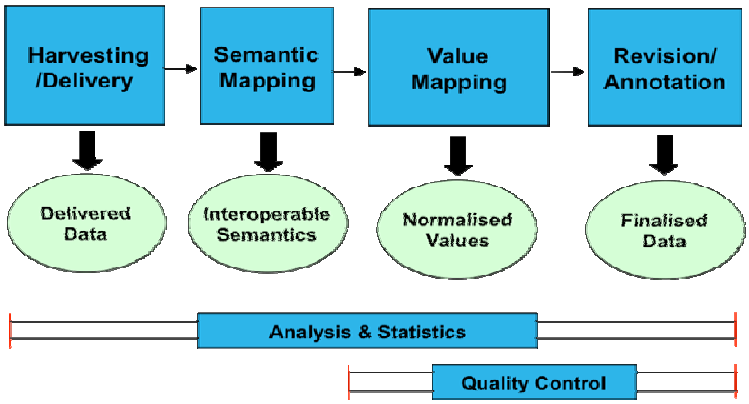

Figure 6.1 Ingestion Workflow

The workflow consists of four phases. Each phase is responsible for specific services all needed to ensure the quality of the ingestion process.
Harvesting/delivery is responsible for collecting the metadata It is an interface for different methods of data delivery including, OAI-PMH, HTTP upload/download FTP upload/download
Semantic Mapping provides the service for assigning semantics to the harvested metadata. It assists providers to manually map their metadata elements to a reference rich schema. Providers that have metadata in supported known formats might be able to omit this step (use stored transformations from selected schemas to the reference schema based on existing crosswalks).
Value Mapping aligns existing element and attribute values with the reference model. In particular:

- It enables providers to resolve data issues, e.g. map own terminology list to selected terminology lists
- It automatically normalizes data e.g. dates, geographical locations, nationality/language, name writing convention to selected vocabulary standards.

Revision/Annotation enables the addition of data that is not in the original metadata (e.g. empty fields, fields that take values from controlled vocabularies).
Analysis & Statistics service provides detailed analysis and statistics of metadata contributed by a provider. (i.e. number of items imported, total values per field etc).
Quality Control automatically checks and reports on Content Provider's data (i.e. missing values, malformed data). Error reports and warnings are produced to facilitate editing the semantic mappings, value mappings and/or edit items until the Provider's data successfully passes the Quality control checks.

6.3 The ATHENA Harvesting Schema

LIDO - Lightweight Information Describing Objects
Lightweight Information Describing Objects (LIDO) was developed by CIDOC Working Group Harvesting and Integration with the purpose of contributing content to cultural heritage repositories. LIDO satisfies the need for a convenient common instrument for providing core data from different collections, data structures or software systems. The necessity for a common schema emerged as it was both time consuming and costly to integrate information from different resources in the same portal, considering that each resource has potentially a different metadata format. LIDO was developed to overcome this situation.

Outline. LIDO is an XML harvesting schema intended for delivering metadata, for use in a variety of online services, from an organization's on line collections database to portals of aggregated resources, as well as exposing, sharing and connecting data on the web. It is capable of supporting the full range of descriptive information about museum objects. Particularly, it supports all kinds of objects, such as art, cultural, technology and natural science and can be used by multilingual portals. It is not intended to be used for proper cataloguing or to support loan and acquisition activities.
The architecture of LIDO is based on a nested set of "wrapper" and "set" elements which structure records in culturally significant ways. The development of its design was inspired by CIDOC CRM resulting into a consistent event-centric schema. Event-centric approaches consider that descriptions of objects should focus on describing the various events in which objects have participated. For instance, the creation, collection and use of an object are defined as events that are related to entities such as dates, places and actors.
The strength of LIDO lies not only on its ability to support extensive range of information, but also on its flexibility. LIDO defines seven groups of information of which only four are mandatory allowing for as large a variety of completeness of information as possible. This enables the organizations to choose which data they wish to provide to a portal and publish online. The mandatory fields are related to the definition of the type of the object described, its title and its record.
The structural elements of LIDO contain "data elements" which hold the information that is being harvested and is delivered to the user of the service environment. It allows an organization to support not only optimized searching and retrieval processes, but also the online presentation of the information and the demonstration of the sources of the data to the user of the portal. To succeed this, it allows the organization to provide indexing and display information and at the same time, supports the recording of information related to the sources of the data within a controlled terminology.

Construction Principles. The construction principles of LIDO are the following:

- To provide a specification and related XML schema that describes cultural materials in a meaningful and comprehensive manner
- To allow the contribution of data and images related to described objects to union catalogues
- A record should provide all the necessary information for display and retrieval of a described object
- Individual data providers should be able to define the level of richness of the contributed metadata records
- Links from contributed metadata back to records in their "home" context should be provided
- It should supply optimized metadata for retrieval and display, with clear distinction between display and indexing elements
- To provide references to controlled environment

Conceptually the information in a LIDO record is organized in seven areas, of which four have descriptive and three administrative characters. The descriptive metadata of an object record hold information about its type, identification, the events that has participated in and the relations to other resources. The administrative metadata hold information about the rights, the record and any digital resource being supplied to the service environment.

History of Lido. LIDO is result of collaborative work of the CDWA Lite, museumdat, SPECTRUM and CIDOC CRM communities. The schema is a combination of the CDWA Lite and museumdat schemas and has been aligned with the SPECTRUM collections management standard.

It is CIDOC CRM compliant and can be used to submit information about all kinds of cultural heritage objects.
CDWA Lite is an XML schema provided for encoding core records for works of art and material culture based on the data elements and guidelines in Categories for the Description of Works of Art (CDWA) and following the data content standard Cataloguing Cultural Objects (CCO) provided by the J. Paul Getty Trust and ARTStor. More details about CDWA will follow in next chapters. museumdat is an XML schema, developed by the Documentation Committee of the German Museums Association, which builds on CDWA Lite but overcomes its focus on art by reconfiguring CDWA Lite elements that takes into account the event-oriented multi-disciplinary approach of the CIDOC Conceptual Reference Model.
CIDOC CRM suggests definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. More details about this schema will follow in next chapters. SPECTRUM is an XML schema based on the UK and international standard for collections management with the same name from the Collections Trust. It suggests a format for exchanging object records between different collection management systems.
LIDO was implemented by the collective efforts and support from: the CDWA Lite Advisory Committee, the Documentation Committee of the German Museums Association, the CDWA Lite – museumdat WorkingGroup, the CIDOC CRM Special Interest Group and the Athena Project.

6.4 Mapping Procedure

For the needs of the ATHENA ingestion service, an import is not required to include the schema used. This simplifies the actual work for the user and at the same time the set of schema components that have to be mapped is reduced to only those that are used, thus reducing redundancy. The Schema Generator module produces the required simplified version of the schema that corresponds to a specific import by the user. When a user triggers the invocation of the mapping tool for a specific import, this module is also invoked. The next step in the workflow is to parse the data for a specific import and generate a tree like structure using HTML elements that represents the schema used. This tree like structure is then transmitted to the mapping Interface in order to create an interactive tree that represents a snapshot of the XML schema that the user is going to use as input for the mapping process.
The Mapping Interface is responsible for creating and presenting an intuitive and visual appealing environment for the user to define mappings, without sacrificing any of the functionality needed to properly achieve the task of schema mapping.
In order to offer a more user friendly environment to perform the task of schema mapping, the tool can be configured to provide to the user groups of high level elements that constitute separate semantic entities. These top level sets of elements are presented on the right side of the mapping Interface as can be seen in Figure 6.2. On the left side of the mapping tool User Interface a tree structure is always present that represents the schema produced by the Schema Generation module for a specific import. The user is able to interact with this tree, expand or collapse the elements of the tree and retrieve brief statistics for each element and its values. An example of the info provided for each element can be found in Figure 6.3

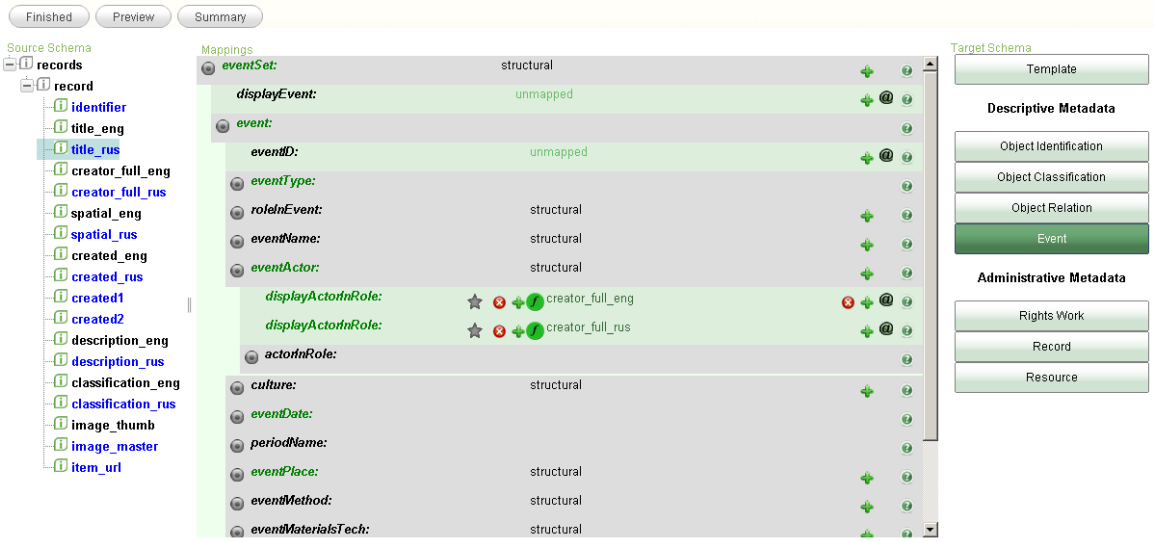

Figure 6.2 Screenshot of the mapping tool.

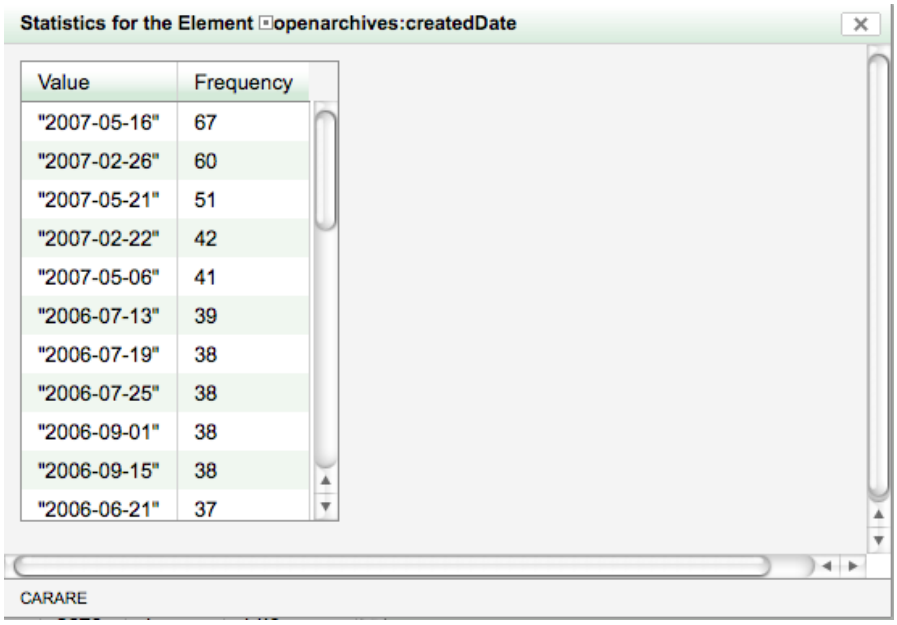

Figure 6.3 Statistics for an input element

When a user wants to create or edit a mapping, he initially has to select one of the top level element groups that are presented on the right side of the mapping interface. Clicking the corresponding button, the set of the sub-elements that are part of that group are presented to him in the middle part of the screen. This part of the user interface has a tree structure of embedded boxes that represents the internal structure of the complex element. The user is able to interact with this structure by clicking to collapse and expand it, similar to what he is able to do the with the tree representation of the input schema. Every embedded box represents an element and the user is able to request and view any information about it that is part of the XML schema.

When a user wants to perform an actual mapping between the input and the target schema, he has to drag and drop any element he wishes from the tree structure on the left part of the user interface to one of the boxes in the middle. When a successful mapping occurs, the user gets notified for the event and he is able to view the mappings in the middle part of the screen. Using the delete button the user is able to delete and correct any mappings he has made so far and repeat the procedure.

The user interface of the mapping tool is completely schema aware regarding the target schema. That means that many operations might be restricted based on constraints that appear in the target XML schema. For example, if an element can be repeated the user is able by using a button that appears on the visual representation of that element to add another one and make a new mapping.

6.5 User Manual

A complete User Manual explaining the functionalities and usage of the ATHENA mapping tool is part of deliverables D7.1 and D7.4, while a full documentation of the MINT services can be found online at:
http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki/User_manual

7. Conclusions

Present report documents the technical requirements and tools for the execution of the ATHENA implementation plan. The web service offered is used to ingest metadata from a diverse group of cultural heritage content providers, homogenise and align them with an established metadata schema standard that guarantees semantic interoperability and, publish them in the Europeana Semantic Elements (ESE) schema for harvesting and presentation in the Europeana portal. It is based on a platform developed by the leader of the WP, which is customised and updated with all current state of the art technologies and the most recent developments in the Europeana family of projects. Functionality includes a user and organisation management scheme that supports appropriate user roles and access rights for simple organisations up to the formation of complex thematic or national aggregators, import of arbitrary metadata schemata used by providers and serialized in xml, a statistical module for input data sources, a visual mapping module that functions as an xslt editor from a data source to a reference metadata schema that enables semantic interoperability, transformation of imported data sources, publishing and exposing aggregated metadata in standard metadata schemata, focusing on Europeana enabled content.

The tool was presented and tested by a wide variety of ATHENA users and collaborators that attended respective WP7 workshops. It was extended to incorporate additional functionalities based on user feedback and recent developments in the field of digital cultural heritage and the web of data. Ingestion process was initiated in regional workshops that were organised within ATHENA in order to register and train providers. Ingestion was supported continuously throughout the duration of the project and will remain available until at least one year after its end.