EUROPEAN COLLECTED LIBRARY OF ARTISTIC PERFORMANCE

www.ECLAP.eu

Grant Agreement No 250481

DE4.1
Metadata descriptors Identification and Definition

Version: 10.1
Date: 15/02/2011

Project Title: ECLAP Project Number: ICT-PSP-250481 Deliverable Number: DE4.1 Accessibility: public (PU) Work-Package contributing to the Deliverable: (WP4, WP4.1) Nature of the Deliverable: report Status: Final

Contractual Date of Delivery: 31/01/ 2011 Approve for quality control by: Paolo Nesi Finally approved by coordinator: Paolo Nesi Actual Date of Delivery: 15/02/2011

Document responsible: Natasa Sofou Email address: natasa@image.ntua.gr Affiliation acronym: NTUA
Authors: • Natasa Sofou (NTUA) • Vassilis Tzouvaras (NTUA) • Nasos Drosopoulos (NTUA)

- Arne Stabenau (NTUA)
- Ivan Bruno, Paolo Nesi (DSI)

Revision History:

Revision	Date	Author	Organization	Description
0.2	**10/01/2011**	**Vassilis Tzouvaras**	**NTUA**	
0.3	**15/01/2011**	**Natasa Sofou**	**NTUA**	
0.4	**20/01/2011**	**Natasa Sofou**	**NTUA**	
0.5	**29/01/2011**	**Nasos Drosopoulos**	**NTUA**	
0.6	**30/01/2011**	**Natasa Sofou**	**NTUA**	
0.7	**08/02/2012**	**Nassos Drosopoulos**	**NTUA**	
0.8	**10/02/2011**	**Natasa Sofou**	**NTUA**	
0.9	**14/02/2011**	**Natasa Sofou**	**NTUA**	
10.0	**15/02/2011**	**Paolo Nesi**	**DSI**	**Polishing and closing**

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Catalogue:

Title	**Metadata descriptors Identification and Definition**
Identifier.de	DE4.1
Identifier.ISBN	
Creators	Natasa Sofou (NTUA), Vassilis Tzouvaras (NTUA), Nasos Drosopoulos (NTUA), Arne Stabenau (NTUA), Ivan Bruno (DSI).
Subject	Metadata descriptors, metadata ingestion process
Description	Report describing the metadata descriptors and ingest process
Keywords	Performing art, metadata, ingestion process, metadata mapping
Publisher	ECLAP
Date	15/02/2011
Format	Document
Type	PDF or DOC
Language	EN

Citation Guidelines

Author(s) name Surname, Deliverable number, Deliverable title, ECLAP Project, DD/MM/YY, URL: univocally determined on http://bpnet.eclap.eu

ECLAP Copyright Notice

Please note that:

- You can become affiliated with ECLAP. This will give you access to a great amount of knowledge, information related to ECLAP services, content and tools. If you are interested please contact ECLAP coordinator Paolo Nesi at info@eclap.eu. Once affiliated with ECLAP you will have the possibility of using the ECLAP for your organisation.
- You can contribute to the improvement of ECLAP by sending your contribution to ECLAP coordinator Paolo Nesi at info@eclap.eu
- You can attend ECLAP meetings that are open to public, for additional information see www.eclap.eu or contact ECLAP coordinator Paolo Nesi at info@eclap.eu

Table of Contents

1 Executive Summary and Report Scope

WP4 involves the selection and delivery of content and metadata for a wide range of user communities as well as the definition of the harvesting metadata schema and its semantic mappings to a spectrum of commonly used standards. It will involve the development of a realisable content selection strategy that appeals to the broadest range of user communities, that most effectively supports and integrates with the Europeana initiative, but that also stays in line with content providers' holdings. The work package includes the mapping of proprietary metadata sets to the common metadata schema. The work package will also include the uploading of content and metadata. The development of both the content selection policy and the common metadata schema necessarily involves a review, revision or enhancement of the available metadata schemas in the cultural domain.

Main WP4 objectives include:

- To collect and create online access to theatrical content catalogues with their metadata and descriptors;
- To collect and create online access to performing art content at item level listed in these open access catalogues across Europe.
- To support cataloguing, metadata and programme content with additional contextual information for a range of users to integrate with the Europeana initiative.
- To define metadata and descriptors coming from performing art institutions and suitable for posting on Europeana
- To define interoperability map among several different models for metadata and descriptors for performing art content with respect to the semantic meaning of Europeana classification model.
- To collect and automatically produce rich media content for further enrichment and augmentation integrating descriptors, annotations and digital resources with features and intelligence for their perusal.

The present deliverable is the first deliverable of WP4.

This report includes the identified metadata and their corresponding semantics in relationship with those of Europeana, taking into account all kind of annotations coming from the several partners' providers. Specifically it introduces the main concepts of metadata framework and identifies the most important and prominent metadata standards, vocabularies and representation languages in the cultural domain. It summarizes the main results (with respect to metadata) obtained by the ECLAP survey questionnaire completed by ECLAP content providers. It provides identification and definition of the metadata schema that will be used to harvest and homogenize ECLAP annotations as well as its relation with the Europeana Data Model. In addition, it illustrates an overview of the procedure and tool that will be deployed within ECLAP, in order to establish interoperability between provider's metadata and the ECLAP repository.

2 Introduction

This deliverable reports on the results of Task 1 (Metadata/descriptors Identification and Definition) of WP4 (Content Provision and Augmentation) concerning the identification of metadata standards of interest and the definition of the ECLAP harvesting metadata schema and work flow. Its main objective is to provide the

ECLAP ingestion system and portal with the information and technical specifications for the alignment of provider's metadata with ECLAP and, for the interoperability of the later with the Europeana repository and relevant standardized and widely used data models for digital cultural heritage. This report introduces the available cultural metadata standards as well as the prevalent representation languages. It discusses identified metadata and their corresponding semantics in relationship with those of Europeana, taking into account all kinds of annotations coming from different content providers that participate and contribute to ECLAP project. Based on the results obtained from a questionnaire completed by all content providers and the following requirement analysis, a suitable metadata schema was developed so as to adapt to ECLAP needs.

The process of metadata identification and definition in ECLAP supports subsequent critical activities:

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

The results of this task are also based upon the efforts and contributions of the members of the ECLAP working group??, the ECLAP DE211-User Requirement and Use Cases.

The rest of this deliverable is structured as follows:

- *Understanding Metadata and Terminology:* An overview of the basic concepts behind metadata and its use that a reader should grasp to have a good understanding of the deliverable content.
- *Metadata Standards and Representation Languages:* Two sections reporting on established metadata models and their application in digital cultural heritage and, on well-known, machine-understandable representation languages used to serialize the aforementioned schemas.
- *ECLAP Harvesting Schema:* Identification and definition of the metadata schema that will be used to harvest and homogenize ECLAP annotations. Its relation with the ECLAP portal metadata and the Europeana Data Model.
- *ECLAP Ingestion Workflow:* An overview of the procedure and tool that will be deployed within ECLAP, in order to establish interoperability between provider's metadata and the ECLAP repository. The ECLAP harvesting schema implementation is discussed and specifically its association with the ECLAP system and its alignment with relevant data models and standards that will allow for efficient and semantic aggregation of providers' digital resources.

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 cataloging 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:

1. **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, abstract). Note, that the descriptive metadata and especially the semantic elements are the scope of DE4.1. Note also: 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.
2. **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.
3. **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)
4. **Administrative** metadata provides information for managing and administering the objects concerned (e.g. content provider name, acquisition information, copyrights, location information, language of record, record 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:

1. A *schema* (the categories of information you choose to record)
2. *Vocabulary* (specific ‘words' or ‘values' you enter into those categories)
3. *Conceptual model* - the underlying model that describes how all the information and concepts inherent in a resource are related to one another
4. *Content standard* - practical standards that describe how specific information (e.g. vocabularies) should be entered within metadata schema categories (e.g. Cataloguing Cultural Objects)
5. *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. [1]
- A set of rules for encoding information that supports specific communities of users **Errore. L'origine riferimento non è stata trovata.**. 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. For the definitions of the several types of controlled vocabularies the following sources is used: [3], [4].

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.

An **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.

Markup ontology languages; these languages use a markup 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 markup 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 labor intensive to develop and maintain. The mapping of schemes with fewer elements (less granularity) to those with more elements (more granularity) is problematic.

4 Standards Landscape

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 E-Clap project deals with cultural heritage content, in this section we present some of the most important metadata standards and schemas used within the cultural heritage domain. It should be noted that this document does not attempt to categorise the schemas and standards available. Instead it provides a list of all the necessary standards and technology components to facilitate intracommunity knowledge sharing most related to the ECLAP project:

- *Descriptive data structure standards* for different kinds of community resource descriptions.
- *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 Descriptive Data Structure Standards

4.1.1 MPEG Multimedia Metadata

The ISO/IEC Moving Picture Experts Group (MPEG) has developed a suite of standards for coded representation of digital audio and video. Two of the standards address metadata: MPEG-7, Multimedia Content Description Interface (ISO/IEC 15938), and MPEG-21, Multimedia Framework (ISO/IEC 21000).

MPEG-7 [11] defines the metadata elements, structure, and relationships that are used to describe audiovisual objects including still pictures, graphics, 3D models, music, audio, speech, video, or multimedia collections. It is a multipart standard that addresses:

- *Description Tools* including Descriptors that define the syntax and the semantics of each metadata element and Description Schemes that specify the structure and semantics of the relationships between the elements.
- A *Description Definition Language* to define the syntax of the Description Tools, allow the creation of new Description Schemes, and allow the extension and modification of existing Description Schemes.
- *System tools*, to support storage and transmission, synchronization of descriptions with content, and management and protection of intellectual property.

Descriptors for visual and audio are defined separately using a hierarchy of elements and subelements. For visual objects there are descriptors for Basic Structure, Color, Texture, Shape, Motion, Localization, and Face Recognition. Audio descriptors are divided into two categories: low-level descriptors that are common to audio objects across most applications, and high-level descriptors that are specific to particular applications of audio. The cross-application low-level descriptors cover Structures and Features (temporal and spectral). The domain-specific high-level descriptors include such elements as Musical Instrument Timbre, Melody Description, and Spoken Content Description. The Description Schemes are based on XML, and can be expressed in textual form suitable for editing, searching, filtering, and human readability; or in a binary form for storage, transmission, and streaming delivery. Since the full description of a multimedia

object can be quite complex, the standard provides for a Summary Description Scheme geared to browsing and navigation.

The standard envisions that search engines could use MPEG-7 metadata descriptions to identify audiovisual objects in entirely new ways, such as digitizing a musical phrase played on a keyboard and then retrieving a list of musical pieces that contain the sequence of notes; drawing some lines on an electronic drawing tablet and retrieving images with similar graphics; or using a voice excerpt to retrieve related speech files, photographs, video clips, and biographical information of the speaker. These retrieval mechanisms are outside the scope of MPEG-7, but the standards developers wanted to accommodate these futuristic capabilities and have included many interoperability requirements beyond the typical metadata elements.

MPEG-21 [12] was developed to address the need for an overarching framework to ensure interoperability of digital multimedia objects. The multi-part standard is not yet fully completed but is intended to include the following:

- *Part 1*: Vision, Technologies and Strategy provides the overview of the complete vision and plan for the framework. It was issued as an ISO technical report (ISO/ IEC TR 21000:1-2001) and is available as a free download from ISO's publicly available standards website. A second edition of the vision document is underway to address comments and suggestions received from other organizations following the initial publication.
- *Part 2*: Digital Item Declaration, issued in 2003, describes a model for defining Digital Items. It includes a description of the syntax and semantics of each of the Digital Item Declaration elements and a corresponding XML schema.
- *Part 3*: Digital Item Identification, also issued in 2003, describes how to uniquely identify Digital Items and how to link Digital Items with related information such as descriptive metadata. • Part 4: Intellectual Property Management and Protection is still in development. It is intended to define the framework for ensuring interoperability of intellectual property management tools, including authentication, and accommodates the Rights information defined in the following two parts.
- *Part 5*: Rights Expression Language, issued in 2004, is a machine-readable language that can declare rights and permissions.
- *Part 6*: Rights Data Dictionary is still in development. It will define a standard set of terms to be used with the Rights Expression Language. It is also expected to include specifications for mapping and transforming rights metadata terminology. The Rights Data Dictionary and Expression Language are being viewed as models for the handling of intellectual property metadata for applications beyond audiovisual.
- *Part 7:* Digital Item Adaptation, also in development, is intended to standardize networking and interoperability description tools. Included in this part will be User Characteristic description tools that specify user preferences. There are some seven additional parts identified and in various stages of development that deal with technical interoperability issues of less specific relevance to metadata. All of the published parts are available from ISO as ISO/IEC 21000-[part#].

4.1.2 Dublin Core

Dublin Core [13] is a standard for cross-domain information resource description. Its name "Dublin" is due to its origin at a 1995 invitational workshop in Dublin, Ohio, while "core" because its elements are broad and generic, usable for describing a wide range of resources. It provides a simple and standardised set of

conventions for describing things online in a machine understandable way making them easier to find. Dublin Core is a metadata standard used mainly to describe content of multimedia essence, such as video, sound, image, text and composite media. Early Dublin Core workshops popularized the idea of "core metadata" for simple and generic resource descriptions. Dublin Core achieved wide dissemination as part of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [31] and has been ratified as IETF RFC 5013 [14], ANSI/NISO Standard Z39.85-2007 [15], and ISO Standard 15836:2009 [16].

Starting in 2000, the Dublin Core community focused on "application profiles", the idea that metadata records would use Dublin Core together with other specialized vocabularies to meet particular implementation requirements. During that time, the World Wide Web Consortium's [17] work on a generic data model for metadata, the Resource Description Framework (RDF) [18], was maturing. As part of an extended set of DCMI Metadata Terms, Dublin Core became one of most popular vocabularies for use with RDF, more recently in the context of the Linked Data movement [19].

The consolidation of RDF motivated an effort to translate the mixed-vocabulary metadata style of the Dublin Core community into an RDF-compatible DCMI Abstract Model (2005). The DCMI Abstract Model was designed to bridge the modern paradigm of unbounded, linked data graphs with the more familiar paradigm of validatable metadata records like those used in OAI-PMH. A draft Description Set Profile specification defines a language for expressing constraints in a generic, application-independent way. The Singapore Framework for Dublin Core Application Profiles defines a set of descriptive components useful for documenting an application profile for maximum reusability.

The Dublin Core metadata standard includes two levels: Simple and Qualified. Simple Dublin Core, also known as Dublin Core Metadata Element set, is a vocabulary of fifteen properties for use in resource description. Qualified Dublin Core is an ongoing process to develop exemplary terms extending or refining the Dublin Core Metadata Element set. There exist two broad classes of qualifiers. These are Element Refinement and Encoding Scheme.

- *Element Refinement*: makes the meaning of an element narrower or more specific. A refined element shares the meaning of the unqualified element but with a more restricted scope. A client that does not understand a specific element refinement term should be able to ignore the qualifier and treat the metadata value as if it were an unqualified element.
- *Encoding Scheme*: identifies schemes that aid in the interpretation of an element value with the help of controlled vocabularies and formal notations or parsing rules. If an encoding scheme is not understood by a client or agent the value may still be useful to a human reader.

The Dublin Core metadata standard can be encoded in many syntax formats such as XML and RDF/XML. When considering an appropriate syntax, it is important to note that Dublin Core concepts and semantics are designed to be syntax independent and are equally applicable in a variety of contexts, as long as the metadata is in a form suitable for interpretation both by search engines and by human beings. In order to be able to proceed we present the following terminology concerning XML, RDF/XML and Dublin Core:

- *Resource*: anything that has identity. Familiar examples include an electronic document, an image, a service and a collection of other resources. Not all resources are network "retrievable"; e.g., human beings, corporations, and bound books in a library can also be considered resources.
- *Property*: a specific aspect, characteristic, attribute, or relation used to describe a resource.

- *Record:* some structured metadata about a resource, comprising one or more properties and their associated values.

Dublin Core in XML. One of the many syntax forms Dublin Core can be encoded in is XML. XML provides a very simple framework for the encoding of Dublin Core Elements. In the following we provide some simple rules regarding the XML syntax. These as well as recommendations, guidelines and examples for XML implementations can be found in [20].

A Dublin Core record is made up of one or more properties and their associated values. Each property is an attribute of the resource being described and can be either one of the simple elements or one of the qualified elements and may be repeated. Moreover each value is a literal string and may be associated to an encoding scheme, which always has a name. Finally each literal string value may have an associated language (e.g. en-GB).

Dublin Core in Ontology Definition Languages. Although XML provides a very simple framework for encoding Dublin Core elements its abstract syntax does not provide semantics to the content described. Taking into account that cultural heritage are mostly indexed on the basis of divergent metadata standards, that hampers the combination and opening up of the cultural content a more semantic language needs to be supported by Dublin Core. This is why Dublin Core can be encoded in Ontology Definition Languages [21]. RDF and OWL provide semantics making interoperability easier among knowledge bases. Moreover, through some tools provides reasoning such as checking the consistency and validity of metadata as well as extracting implicit from explicit knowledge.

Unlike the XML syntax, in the RDF/XML case there is a different syntax regarding unqualified and qualified elements. Representing unqualified Dublin Core elements requires a bit more extra information. Every assertion is made about a fixed resource. Each resource is identified between a pair of rdf: Description tags.

4.1.3 CDWA - Categories for the Description of Works of Art

Categories for the Description of Works of Art (CDWA) is a framework for describing and accessing information of cultural heritage resources [23]. It provides access to art databases for describing and collecting information about works of art, architecture, other material culture, groups and collections of works, and related images. This framework provides 532 categories in which to describe works of art. Some of these are bound to represent the minimum information necessary to ensure a sufficient description for work identification. These categories are considered to be the core. The CDWA includes discussions, basic guidelines for cataloging, and examples.

History. DWA was developed by the US-based Art Information Task Force (AITF), with funding from the J. Paul Getty Trust, National Endowment for the Humanities (NEH), and the College Art Association (CAA). The purpose was to permit a dialog between cultural communities in order to develop guidelines for describing and presenting works of art, architecture, groups of objects, and visual and textual surrogates. The current version of CDWA (2.0) was published in 2000 and revised in 2006. CDWA is now maintained by the Getty Research Institute, who has developed CWDA Lite. CDWA Lite is an XML based schema that can be utilized in databases that using the Open Archives Initiative (OAI) harvesting protocol. It is intended to describe core records for works of art and material culture based on CDWA and CCO.

Outline of CDWA. The goal of CDWA is to provide guidelines in order to reach an agreement on the information that should accompany a surrogate work of art. CDWA contributes to data integrity and longevity while information systems evolve and data migrates to new systems. The common set of rules provided, form a common ground for curators, registrars, researchers, information managers, and systems

vendors. These rules ensure compatible and accessible information. Existing information systems as well as new ones may both be developed upon CDWA framework. In addition, end- users are provided with system independent, reliable and consistent information. CDWA aims to matching data deriving from different systems that conform to different standards. Therefore, it provides the necessary compatibility level for matching with other schemes such as MARC, Dublin Core, EAD, METS and DACS.

Construction Principles. CDWA recommends a relational data structure. Separate files or authorities are maintained for works and objects, related textual or visual materials, locations/place, persons/corporate bodies, generic concepts, and subjects.

Information on authority of persons, places, generic concepts, and subjects may be important for retrieval of the work, but it is recorded in separate authority files rather than in records about the work itself. These files are then linked to all relevant work files while they are recorded only once and easily updated. Authorities described in CDWA should be hierarchical; given that authority entities often require multiple broader contexts, a polyhierarchical structure is recommended.

Categories and subcategories of art information may differ depending on the end- user. Even the core categories which are meant to represent the minimum information necessary to describe and identify a work may be adapted to the needs of the information system to be served. As a result to this end- users are provided with helpful browsing options specified according to their needs. An example of the user friendly orientation is the classification category which provides broader and less detailed description of objects.

CDWA assumes information for display to be in a format and with syntax that is easily read and understood by users. Such free-texts or concatenated displays may contain all the nuances of language necessary to relay the uncertainty and ambiguity that are common in art information.

As far as retrieval is concerned, CDWA assumes that key elements should be formatted to allow for indexing. CDWA recommends that indexing should be performed by knowledgeable catalogers who reliably interpret the meanings of their indexing terms, in contrast to automated methods that perform parsing of a text.

In CDWA, display fields are often described as free-text fields (which may alternatively be concatenated from controlled fields, if necessary); indexing fields are intended to be controlled fields. CDWA advises the use of controlled vocabularies; CDWA describes when categories should be controlled by a simple controlled list (e.g., Classification), an authority (e.g., Creator), or by consistent formatting of certain information (e.g., Earliest and Latest Dates) to ensure efficient end-user retrieval.

4.1.4 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 [36]. 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 [37] 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.

4.1.5 SPECTRUM

SPECTRUM documentation standard [22] (Standard ProcEdures for CollecTions Recording Used in Museums) was created by the mda (previously the Museum Documentation Association) in the UK. SPECTRUM is a guide to "good practice for museum documentation, established in partnership with the museum community. It contains procedures for documenting objects and the processes they undergo, as well as identifying and describing the information which needs to be recorded to support the procedures"1. It includes information on the minimum UK standard for museum documentation. A simplified version, called SPECTRUM Essentials, is available for smaller museums. SPECTRUM is a well-respected standard internationally, and is increasingly used as the basis for international interchange of museum data. An XML DTD has been produced for SPECTRUM which serves as a system-neutral interchange format for museum data that is based on SPECTRUM. The SPECTRUM XML DTD is intended to be a universal interchange format for museum collections management systems that are based on SPECTRUM or that can map to SPECTRUM. The SPECTRUM XML DTD "will allow different collections management systems to exchange complete museum records that are compliant with the SPECTRUM standard. It will also provide a means for testing software compliance and archiving records"2 in a system-neutral form. The mda has now completed a project to map the information requirements of several major archival description standards to SPECTRUM's units of information.

4.1.6 IMS – Instructional Management Systems

The IMS Learning Resource Meta-data Information Model [32] identifies a subset of IEEE LOM meta-data elements to be used to describe learning materials in various types of learning systems. It provides open technical specifications for interoperable learning technology and standards for delivering learning products and services. IMS collaborates with the organizations that develop and maintain profiles. Profiles are collections of one or more specifications/standards and extensions that an adopter community, such as a Governmental Department, or region has selected as a requirement for procurement.

At a 1997 meeting of the EDUCOM consortium (now EDUCAUSE), a group of higher education institutions and their vendor partners established an effort to develop open, maket-based standards for online learning, including specifications for learning content meta-data. This resulted in the formation of the Instructional Management Systems (IMS) Cooperative. The IMS Learning Resource Meta-data Information Model, a metadata standard, is one of the first products of this group.

The LOM and IMS schemas are based upon a hierarchical model that groups elements into categories of information. The top level of the hierarchy is the root element, which is actually at the record level. A root can have sub-elements that are either "branches" or "leaves". A branch is a sub-element that has further sub-elements. A leaf is a sub-element that has no further sub-elements.

The IMS Best Practice Core specifies 9 top "branches" which each contain several branches and leaves. These categories are: general, lifecycle, metametadata, technical, educational, rights, relation, annotation, and classification. Several of the LOM elements are similar to Dublin Core elements.

4.1.7 AMICO - Art Museum Image Consortium

The AMICO metadata vocabulary [24] is mainly used in the collection of art museum images. The AMICO metadata vocabulary using the DC and CDWA vocabularies provides a framework for the specification of images and multimedia files.

4.1.8 MARC21- Machine-Readable Cataloguing

MARC is a family of metadata standards for representing library resources. Although chiefly used by libraries to describe bibliographic material (books or periodicals), it is also sometimes used to describe non-book material (e.g. images) or archival collections. MARC is a very extensive and formalised standard, with hundreds of potential categories and a rigid way of encoding its data. In the past, individual countries developed their own versions of MARC (e.g. UKMARC), but many are now converging to the current version: MARC 21 (published in 1999 and maintained by the US Library of Congress) [25], [26]. MODS provides a sub-set of MARC encoded as XML, but there are also efforts underway to provide an XML encoding of the larger MARC standard (MARCXML). Many libraries have relied on the Anglo-American Cataloguing Rules for guidance on how to enter data within MARC, but this will be replaced by the new RAD content standard.

4.1.9 MODS – Metadata Object Description Schema

The Library of Congress' Network Development and MARC Standards Office, with interested experts, developed the Metadata Object Description Schema (MODS) [27] in 2002 for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications. As an XML schema it is intended to be able to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records. It includes a subset of MARC fields and uses language-based tags rather than numeric ones, in some cases regrouping elements from the MARC 21 bibliographic format. As of June 2009 this schema is in its third version (version 3.3). MODS is expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained by the MODS Editorial Committee with support from the Network Development and MARC Standards Office of the Library of Congress.

4.1.10 METS - Metadata Encoding and Transmission Standard

METS [28] is a standard for encoding metadata within an XML format. Although it contains descriptive and administrative elements of its own, a key function of the METS standard is to structure or "package" other metadata or data for exchange or delivery. METS can embed or link to other XML-based metadata (e.g. MODS, MIX, PREMIS or TEI, see below). Any number or type of digital files can be described and linked together by a METS record, enabling it to represent very complex digital resources (e.g. a whole digitised book, with bibliographic data, images and transcribed text). METS grew out of work in the mid 1990s on the Making of America II (MOA2) digitisation programme sponsored by the US Digital Library Federation. It is

now maintained by the US Library of Congress. METS 1.1 was released in 2001; the current version is 1.5, released in 2005.

4.1.11 EAD – Encoded Archival Description

The EAD Document Type Definition (DTD) [29] is a data structure standard for encoding archival finding aids. It defines the structural elements and designates the content of descriptive guides to archival and manuscript holdings following the syntax of the Standard Generalized Markup Language (SGML – ISO 8879). The Society of American Archivists (SAA) is the owner of the intellectual component of the EAD, and the SAA EAD Working Group (EAD Roundtable) is overseeing its development. The MARC Standard Office of the Library of Congress (LC) acts as the maintenance agency and provides access to the online documentation.

EAD enables standardized exchange of descriptive data contained in specific types of archival finding aids known either as archival inventories or manuscript registers. It provides tools for a detailed, multilevel description, structured display, navigation, and searching. It plays a similar role for archival materials to that played by a MARC record for library holdings. While collection-level MARC records contain bibliographic information about archival and manuscript holdings, full text finding aids provide detailed descriptions of collections essential for understanding their content and research value.

4.1.12 VRA - Visual Resources Association Core

The Visual Resources Association is a multi-disciplinary organization dedicated to furthering research and education in the field of image management within the educational, cultural heritage and commercial environments. The Association is committed to providing leadership in the visual resources field, developing and advocating standards, and offering educational tools and opportunities for the benefit of the community at large. Since the 1980s, VRA has worked on creating standards to describe images. Since every visual resources collection seemed to use different and variant standards, the Association has worked towards creating a usable and common standard. Somewhat based on the Dublin Core model, the Core has grown from a list of elements to describe art and architectural images to a data standard (with an XML schema to promote the sharing of records) for describing images of cultural heritage.

Specifically, the VRA Core [30] is a data standard for the cultural heritage community. It consists of a metadata element set (units of information such as title, location, date, etc.), as well as an initial blueprint for how those elements can be hierarchically structured. The element set provides a categorical organization for the description of works of visual culture as well as the images that document them. Establishing an official encoding of the data elements into a data format (such as XML) is a logical next step in the development of efficient systems for cataloging, retrieval, and record sharing. To this end, the VRA Data Standards Committee has developed an XML Schema for the VRA Core 4.0 metadata element set to be used primarily for record sharing and exchange purposes.

4.1.13 IPTC - International Press Telecommunications Council

The IPTC's older standard for metadata, the Information Interchange Model (IIM), defined a large set of metadata properties. In the early nineties a subset of this IIM was adopted as the well known "IPTC Headers" supported in JPEG, TIFF and PSD files by Adobe Photoshop. The IPTC Core [35] is the latest revision of the International Press Telecommunications Council IIM schema, designed to use the Extensible Metadata Platform (XMP) developed by Adobe in 2001. XMP was first introduced in Adobe Photoshop 7.0 and Adobe Acrobat 5.0. With XMP, legacy IPTC IIM metadata, new and customized metadata and metadata from other

standards (such as EXIF camera data) can be recorded using a common data format based on XML. This information can be stored inside JPEG, TIFF, and PSD files as well as other file types including PDF. Since its introduction a number of other software vendors have added XMP support in their products.

The new IPTC Core XMP schema uses properties from the older IIM standard in addition to new metadata properties, and it specifies XMP to store metadata in files. The IPTC Core custom panels provide a view of metadata that's labeled and organized so it is easier for photographers to use the IPTC Core schema. The IPTC Core panels concentrate all key sets of properties needed by photographers into four sections or panels.

The first focuses on the **Contact Information**, the second on basic information about the **Content** of the photograph, the third with more abstract **Image information** relating to the photograph. The final section, **Status information**, relates to image management, workflow and copyright.

4.1.14 IMS Global Learning Consortium

IMS Global Learning Consortium (usually known as ITIMS or IMS GLC) [32] is a global, non-profit, member organization that strives to enable the growth and impact of learning technology in the education and corporate learning sectors worldwide. Their main activity is to develop interoperability standards and adoption practice standards for distributed learning.

IMS Global Learning Consortium, Inc. (IMS) is developing and promoting open specifications for facilitating online distributed learning activities such as locating and using educational content, tracking learner progress, reporting learner performance, and exchanging student records between administrative systems. IMS has two key goals: (1) Defining the technical specifications for interoperability of applications and services in distributed learning, and (2) supporting the incorporation of the IMS specifications into products and services worldwide. IMS endeavors to promote the widespread adoption of specifications that will allow distributed learning environments and content from multiple authors to work together (in technical parlance, 'interoperate').

The scope for IMS specifications and standards cover most of the data elements used in "distributed and collaborative learning." IMS specifications promote the adoption of learning and educational technology and allow selection of best of breed products that can be easily integrated with other such products. These include a wide variety of technologies that support or enhance the learning experience, such as web-based course management system, learning management systems, virtual learning environments, instructional management systems, student administrative systems, ePortfolios, assessment systems, adaptive tutoring systems, collaborative learning tools, web 2.0 social learning tools, learning object repositories, and so forth. These include technologies and products that support learning situations that involve support for collaborative learning involving learners and instructors. The learners may be in a traditional educational environment (i.e., a school classroom in a university), in a corporate or government training setting, or at home.

4.2 Protocols for distributed search and metadata harvesting

4.2.1 OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [31] 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.

4.3 Ontologies for semantic mediation between data standards

4.3.1 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 **labeled** 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 catalogs.

- 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.

4.3.2 CIDOC – Conceptual Reference Model (CRM)

The CIDOC Conceptual Reference Model (CRM) [34] 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 modeling. 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 archeologists, 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 modeling of events. It allows for metadata representation such as creation, use, publication content summarization. Event oriented modeling 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.

4.3.3 FRBR – Functional Requirements for Bibliographic Records

FRBR [39] 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 cataloging 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 anguage. 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 catalog 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 cataloging 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 catalogs.

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 catalog 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. So what are these FRBR user tasks? Briefly, they 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 catalog.
- **'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.

4.3.4 FRBR oo

The FRBRoo [40] 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;

4.4 Representation Languages and Schemas

4.4.1 XML - Extensible Markup Language

Extensible Markup Language (XML) [6] 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.[6] 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.

4.4.2 RDF - Resource Description Framework

The Resource Description Framework [7] 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.

4.4.3 RDFS - Resource Description Framework Schema

RDFS (RDF Schema) [8] 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.

4.4.4 OWL - Ontology Web Language

OWL is a Web Ontology Language [9]. 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.

4.4.5 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.

4.4.6 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.

4.4.7 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) [10] 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 Modeling in Europeana Standars

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 [41]. 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 [42] 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 Results obtained from ECLAP survey

The aim of the ECLAP survey is to collect information about 'individual' collections that a content provider is prepared to submit to ECLAP. The collected information will be used by technical partners to develop the IPR Wizard and the metadata and content collection system.

The ECLAP survey was divided in two main sections. The first one was to collect general information related to content partners. The second section's aim was to collect additional information related to collections provided by Content Partners to ECLAP from a technical point of view. Each content partner had to fill in a technical survey form (in the form of Excel file) providing information for each content set.

The ECLAP General survey had 2 sections:

- Content Provider Information
- Metadata and terminology

Each partner has to answer the same number of question in general questionnaire part. The technical questionnaire can be found in Appendix A.

In the technical questionnaire each partner had to provide technical information regarding content type, number of items, IPR, metadata etc. A table providing a summarized view of results with emphasis to metadata is available in section 6.3.

Along with the completed questionnaire each partner had to provide a sample file of metadata in xml. These files would provide a more complete view of how each content provider organizes the available metadata and serve as test data for ECLAP mapping tool.

6.1 Information Schemes (Metadata)

6.1.1 Metadata Types

The table below shows how many provider organisations use which information schemes (metadata). Note that where a scheme does not appear in answers to the survey it is not included in the table. Additionally, where there are multiple collections from the same organisation, using the same schemes, they only contribute once to the table. Organizations with more than one type (e.g. library, museum, archive, etc) appear with more than one entry in the table.

Metadata Standard	**Type of provider**				
	Archive	**Library**	**Museum**	**University**	**Other**
Dublin Core	2	1		2	
MARC	1			1	
EAD	1				
FRBR	1				
CDWA	1		1	1	
No Standard	2	1	1	3	1

Metadata Format	**Type of provider**				
	Archive	**Library**	**Museum**	**University**	**Other**
XML	3	1	1	5	
Other	2	1	1	1	1

Metadata upload method	**Type of provider**				
	Archive	**Library**	**Museum**	**University**	**Other**
FTP		1		1	1
HTTP	2	1	1	3	1
SHTP	2		1	2	1
OAI-PMH	2				1

6.1.2 Metadata Conclusions

Based on the results and the analysis of the provided questionnaires and samples the most commonly used standards among content providers are Dublin Core, MARC, EAD, CDWA and a simplified customization of FRBR. The majority of content providers do not use any standard for the metadata. The first conclusion reached so far is that Dublin Core is a popular metadata scheme**.** This scheme has been fashionable over the last few years for public access to cultural material. The challenge for a Dublin Core-based system is whether it can support the rich nature of performing arts data as exemplified by the cultural domain standards. Secondly, a significant number of providers use no standard for their metadata schemes, on in-house customized schemes. This is difficulty for automated ingestion of cultural data by ECLAP as well as Europeana.

6.2 Metadata Terminology

6.2.1 Overview

Regarding metadata terminology the following table summarizes the number of partners that use standards for different areas of terminology.

Standard area	Number of partners using standards
Geographic names	5
Date formats	11
Time Periods	3
Subjects	9
Person and organization authorities	4

Regarding standards for metadata terminology, some of them are published standards whereas other are developed by the provider, as illustrated in the following table:

Standard area	Provider developed	Published standards
Geographic names	4 (28,57%)	1 (7,14%)
Time periods	2 (14,28%)	1 (7,14%)
Subjects	8 (57,14%)	1 (7,14%)
Person and organization authorities	3 (21,42%)	1 (7,14%)

6.2.2 Date Format Standards

Regarding date format standards, 11 out of 14, partners, that is 78,5%, answered that they used a standard for date format of their items. Among the date format standards they use there is almost no overlap. The date format standards used by content providers can be found in the following table.

Standard	Number of partners using standard
DD-MM-YYYY	2
YYYY/MM/DD	1
DD/MM/YY	1
DD/MM/YYYY	1
YY-MM-DD	1
YYYY-MM-DD	2
text format	2
YYYY (year of creation)	1
dd-mm-jjjj. For video time: tt:tt:tt (hours, minutes, seconds)	1
jjjj-mm-dd tt:tt:tt for xml exports	1
year of production + day of arrival of the screener at the festival (dd/mm/yy)	1

6.3 Summarized results of technical questionnaire

A table of summarized results for each partner with respect to metadata obtained by the technical questionnaire can be found in the following table.

	CONTENT TYPE	METADATA UPLOAD METHOD	METADATA STANDARD	METADATA FORMAT	METADATA LANGUAGE	METADATA SAMPLE
UNIROMA	video	HTTP	no standard	xml	Italian	✓
CTFR	image	HTTP	Dublin Core	xml	Italian	
		SFTP	EAD			
	text	OAI-PMH	MARC			
B&G	video	OAI-PMH	based on FRBR but customized	xml	Dutch	✓
	image					
ITB	image	FTP	Dublin Core	xml	Catalan	✓
UvA	video	HTTP	Dublin Core	xml	Dutch or English	✓
ESMAE	video	HTTP	MARC	xml	Portuguese	✓
	audio					
	text		no standard			
	image					
UCLM						
FIFF	video	no existing metadata				
	audio					
	text					
	image					
OSZMI	video	HTTP	no standard	other formats	Hungarian	✓
	audio					
	text					
	image					
Bellone	video	FTP	no standard	other formats	French	
	audio					

	text image					
UCAM	image video	SFTP	CDWA	xml	Dutch	
MUZEUM	missing					
IKP	missing					
UG	Video Audio Text	FTP, SFTP	Dublin Core	xml	English	✓
	Image Animation Html		no standard	other formats		

7 ECLAP Ingestion Workflow

The basic workflow in ECLAP for massive content and metadata ingestion will be made of the following steps:

- content partners will provide metadata using the ECLAP Metadata Ingestion Service (EMIS) portal, metadata should be provided as XML, it will be directly uploaded as a file or harvested them via a OAI-PMH access. In alternative ECLAP content partners can also provide single items and metadata via direct upload on ECLAP Social Service Portal.
- each content partners will map their own metadata XML structure to the ECLAP metadata XML format , this will be done using the EMIS portal to define a XSLT that will be used in the mapping phase; Both original metadata sets as provided by Content Partners and the mapped result, together with the XSLT map will be also posted/provided to DSI portal or FTP.
- in case of a OAI-PMH access the EMIS will crawl the content partners archives and it acquires the original metadata;
- when the original metadata is acquired it is mapped to the ECLAP metadata format and stored and it will be available to the ECLAP Social Service Portal (ESSP) via OAI-PMH access;
- the ESSP regularly crawls the EMIS to acquire metadata represented using the ECLAP metadata format and in the original metadata format;
- for each metadata record acquired the ESSP will download the content files associated with it, ESSP will use the content urls defined in the ECLAP metadata (content may be also acquired using Hard Disks, FTP, DVD/CD Data ROM in case content size is too high for a rapid download);
- each piece of content item acquired will be formatted, adapted and a complete media content will be produced in the ESSP portal with the associated metadata;
- each produced content will be available only internally in the ESSP portal for:
  - metadata enrichment,
  - metadata translation
  - metadata edit, review and validation
  - IPR definition
  - Content publication on process;
  - …

these activities are going to be orchestrated using an internal workflow management tool integrated in the ESSP;

- when the content and the associated metadata will be ready they will be available (on the basis of the IPR defined) for access to the identified users according to the IPR rules and restriction imposed by the content owner.
- metadata will be published on Europeana, so that Europeana users will be capable to reach the ESSP portal to access at the content item according to the IPR rules.

More details about ECLAP ingestion Workflow can be found in *DE3.1 –Infrastructure: ingestion and processing content and metadata*. In the next section we focus on the metadata ingestion and mapping procedure followed within ECLAP project. http://bpnet.eclap.eu/drupal/?q=en-US/home&axoid=urn:axmedis:00000:obj:a345a84f-6fdf-4f84-a412-88094ce363e2&section=search_base

7.1 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 E-clap.

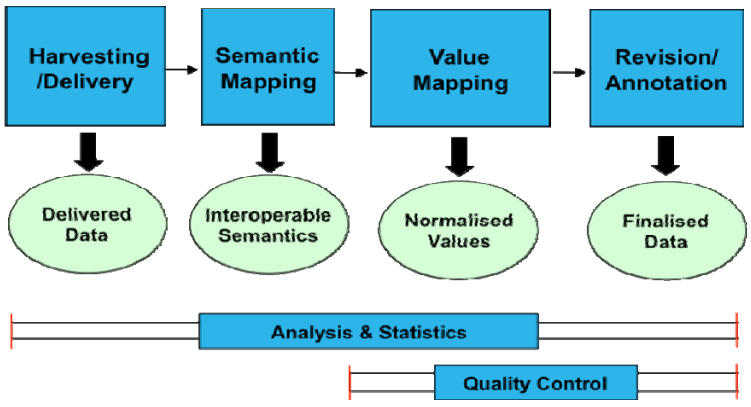


Figure 7.1 *Ingestion Workflow*

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 will be an interface for different methods of data delivery including, OAI-PMH, HTTP upload/download FTP upload/download

Semantic Mapping will provide the service for assigning semantics to the harvested metadata. It will assist to manually map Providers fields 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 will take existing attribute values and produce different/edited values. In particular:

1. It will enable providers to resolve data issues, e.g. map own terminology list to selected terminology lists
2. It will then automatically normalize data e.g. dates, geographical locations, nationality/language, name writing convention to selected vocabulary standards.

Revision/Annotation will enable 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 will provide detailed analysis and statistics of metadata contributed by a provider. (i.e. number of items imported, total values per field etc).

Quality Control will automatically check and report on Content Provider's data (i.e. missing values, malformed data). Error reports and warnings will be produced to facilitate editing the semantic mappings, value mappings and/or edit items until the Provider's data successfully passes the Quality control checks.

7.2 ECLAP Harvesting Schema

The WP4 working group evaluated the relevant metadata schemas for the heritage domain to identify the requirements for the schema to be uses to mediate between content provider's native data and the data models implemented by ECLAP portal and Europeana. The group identified four metadata schemas as being of particular relevance to the ECLAP project; Dublin Core, CIDOC initiatives and especially LIDO as a CRM compliant harvesting schema, FRBR and, the Europeana Data Model (EDM) itself. Those schemas were analysed with respect to mappings and alignment between the local proprietary data models and the identified schemas. Following this analysis, the advantages and disadvantages of the different schemas were highlighted, looking in particular at the modelling of concepts such as 'audiovisual resource', 'agent', 'event', 'time span' and 'location'.

In recommending the implementation of a metadata schema for the ECLAP ingestion procedure, the task group took into consideration the strengths of the CIDOC CRM model with respect to modelling events related to the cultural heritage object, the approach of EDM towards spatial and temporal information and the features of the LIDO schema with respect to harvesting and aggregating metadata for varied information resources. The ECLAP ingestion metadata schema will be implemented as a harvesting XML schema intended for delivering metadata to the ECLAP service environment about an organisation's online collections and digital objects. It has been established by the project to ensure interoperability between the native metadata held by heritage organisations and the standards and schemas used in ECLAP and Europeana.

Therefore the metadata schema for the ECLAP ingestion procedure will consist of three main complex types, each one for the ingestion of different kinds of annotations. The first of them will address the *Aggregated Cultural Heritage Object* (as the term is defined in the Europeana Data Model) and will be used for the collection of metadata about the physical object itself. Such information will be the physical object's type, the identifier for this object by the provider etc. The target metadata for ingestion of the next complex type will concern web resources related to the psychical thing. Examples of the web resources that would be ingested here include the provider's web representation of the object's metadata together with its digital resources (thumbnails, video etc). The third category of complex type, constructed using ore:Proxy, is implemented for the ingestion of the main bulk of descriptive and administrative metadata; examples in this case include the data provider, rights, provenance etc. To facilitate the use of the ingestion schema by the

providers and, to highlight the alignments with existing domain standards, this category will be further classified. Hence, sub-categories will be made for the ingestion of metadata according to the data models of Europeana, Dublin Core and the Dublin Core Terms and to specifically tackle partonomic and other relations between the object and other information and non-information resources. Most notably, a subcategory will be constructed for the modelling of historical and other events that represent important points of reference in the object's lifetime before and after its acquisition. The Event structure, as it is introduced in various data models (CIDOC, FRBR, Europeana), facilitates the registration of important events that the object was present at (such as creation, find, use, etc) together with participating actors (people or organizations) at a certain time period and location.

The ECLAP harvesting XML schema may be employed for delivering metadata for use in a variety of online services such as the ECLAP portal and Europeana, as well as for exposing, sharing and connecting data on the web. It intends to represent the full range of descriptive information about cultural heritage objects, providing support for multilingual environments. On the other hand, it is not developed as a full data exchange format or as a complete cataloguing model for a collection management system. Individual data providers can evaluate and decide on how light or right they want their contributed metadata records to be depending on their home systems and the ECLAP portal's requirements. The schema allows for the delivery of both data and resources relating to objects and includes appropriate fields for the inclusion of links back to records in their home portal. The schema will allow the provision of references to controlled vocabularies and authority files for terms and entities and will attempt to provide optimised metadata for both retrieval and display by distinguishing between respective elements.

7.3 Mapping Procedure

For the needs of the ECLAP 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* 7.1 Ingestion Workflow. 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 7.2*

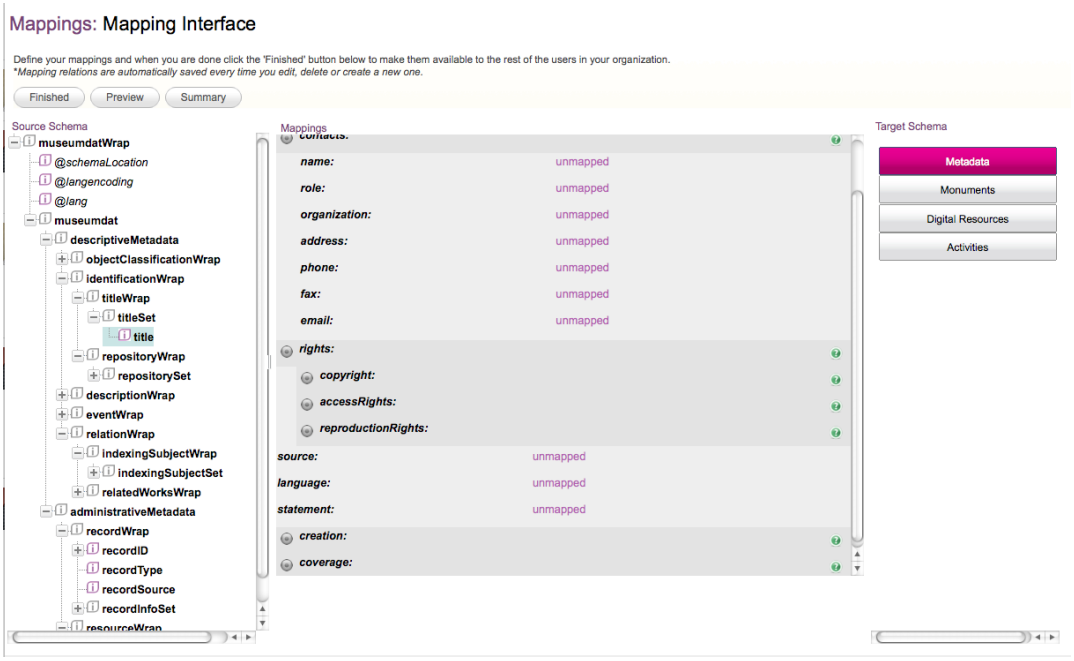


Figure 7.2 *Screenshot of the mapping tool.*

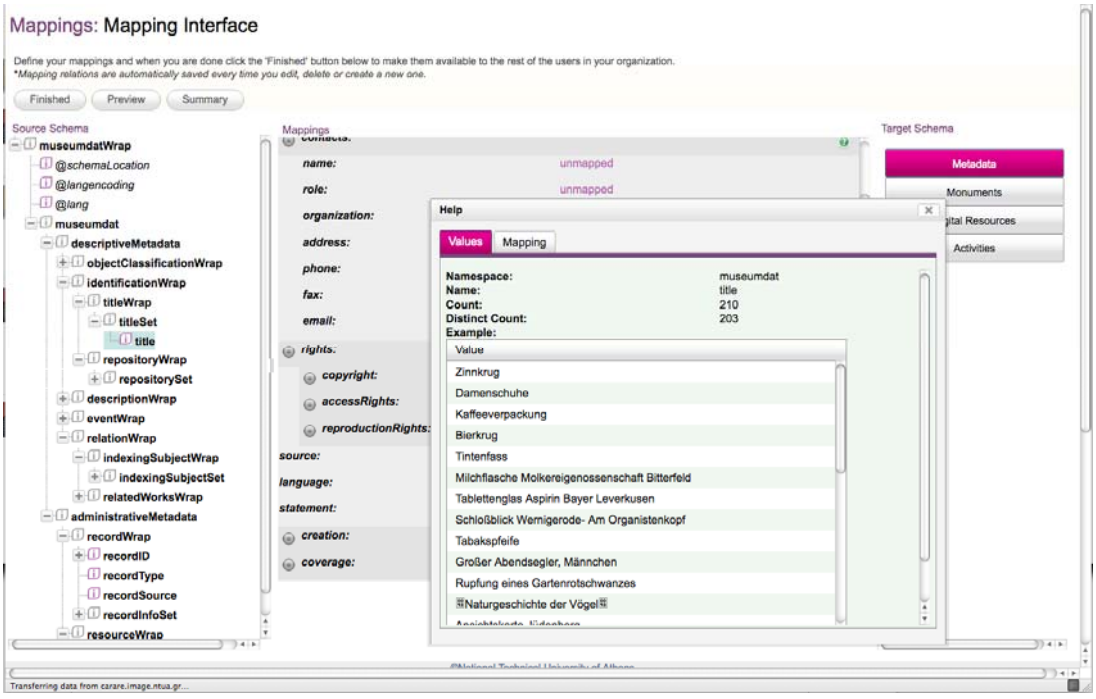


Figure 7.3 Statistics for an input element

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.

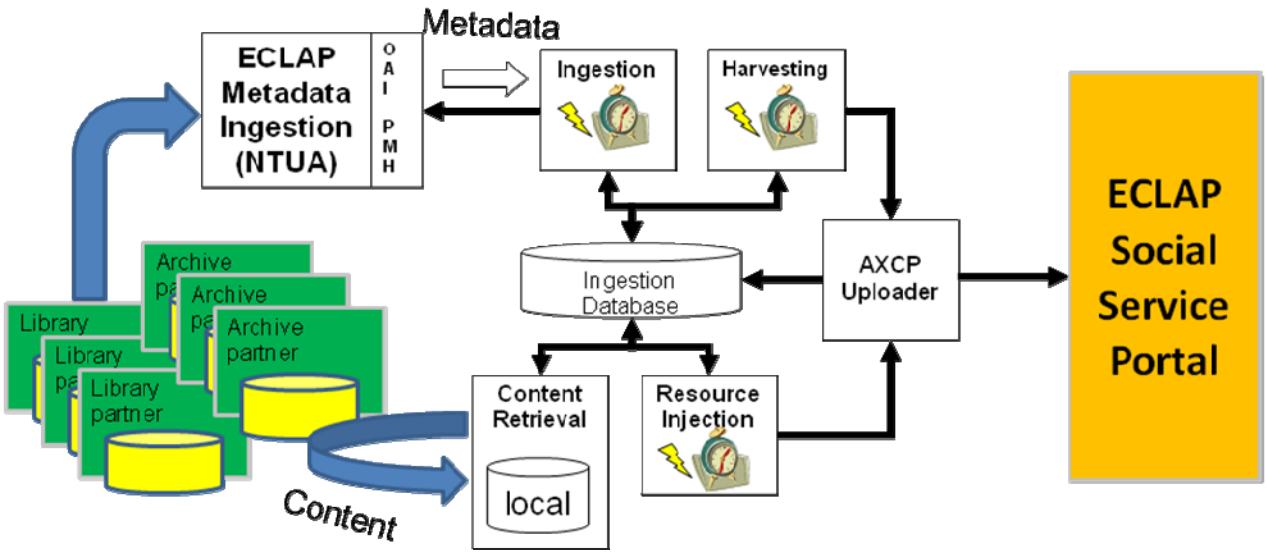
The user interface of the mapping tool is completely schema aware regarding the target schema. That means that many operations might be restricted based on constraints that appear in the target XML schema. For example, if an element can be repeated the user is able by using a button that appears on the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

7.4 Content & metadata ingestion

Content and Metadata Ingestion will be the process that will allow importing metadata and content coming from ECLAP partners and Digital Archives. The process will allow ingesting both massively and singularly metadata and digital resources.

It could be happen that some digital resource since too big to get via Internet has to be provided by using physical device, so that we will have the need to start producing ECLAP content just with metadata and then when the digital resource will be available injecting it off-line in the corresponding ECLAP content..

This section describes how the ECLAP portal will work with ECLAP Metadata Ingestion service by NTUA. The picture below describes the processes that will be involved in content ingestion:



The activities will be divided as following:

- Content and Metadata Retrieval: retrieving the metadata from the EMIS service
    - **Ingestion** will be a periodic process that will monitor and get metadata of new content or a new version of metadata for just retrieved contents.

- **Harvesting** will be a periodic process that will monitor for new retrieved metadata and will call the content production and posting on ECLAP portal by the AXCPUploder process.

- **Content Retrieval**: retrieving and organize contents coming from the partners via internet or via physical devices
- **Resource Injection**: it will be a periodic process that will monitor for digital contents availability and will call the content production to update the existing ECLAP contents just created only with metadata in order to inject/fill with the digital resource by the AXCPUploder process.

All processes will share and update a Database of metadata (Ingestion Database) where all metadata and other information will be stored and used to manage the whole life cycle of contents ingestion.

For more detailed description of the aforementioned process please refer to DE3.1: http://bpnet.eclap.eu/drupal/?q=en-US/home&axoid=urn:axmedis:00000:obj:a345a84f-6fdf-4f84-a412-88094ce363e2&section=search_base

8 User Manual

A User Manual explaining the functionalities and usage of ECLAP mapping tool is available online in HTML and can be found at: http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki/User_manual. In http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki/ you can also find some screencasts of the mapping tool. An updated pdf version of the user manual will be provided in future and posted on the ECLAP portal.

9 Bibliography

[1] Metadata in the audiovisual production environment: an introduction / Annemieke de Jong. – Hilversum: Nederlands Instituut voor Beeld en Geluid, 2003

[2] Multimatch project deliverable on metadata - D2.1 First Analysis of Metadata in the Cultural Heritage Domain

[3] Taxonomy of Knowledge Organization Sources/Systems (1). - Draft June 7, 2000 (revised July 31, 2000)

[4] http://nkos.slis.kent.edu/KOS_taxonomy.htm Last viewed 2006-09-14.

[5] Description of Work, eclap

[6] Extensible Markup Language (XML), http://www.w3.org/XML/

[7] Resource Description Framework (RDF) http://www.w3.org/RDF/ .

[8] RDF Vocabulary Description Language 1.0: RDF Schema http://www.w3.org/TR/rdf-schema/ .

[9] Web Ontology Language (OWL) http://www.w3.org/2004/OWL/ .

[10] Borgida, M. Lenzerini, and R. Rossati. The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press.

[11] José M. Martínez. MPEG-7 Overview, http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

[12] Jan Bormans, Keith Hill. MPEG-21 Overview, http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm .

[13] Dublin Core Metadata Element Set, http://dublincore.org/documents/dcmi-terms/ .

[14] http://www.ietf.org/rfc/rfc5013.txt

[15] http://www.niso.org/kst/reports/standards?step=2&gid=&project_key=9b7bffcd2daeca6198b4ee5a848f9beec2f600e5

[16] http://www.iso.org/iso/search.htm?qt=15836&searchSubmit=Search&sort=rel&type=simple&published=on

[17] http://www.w3.org/

[18] Resource Description Framework, http://www.w3.org/TR/rdf-primer/

[19] Linked Data Movement, http://linkeddata.org/

[20] Guidelines for implementing Dublin Core in XML, http://dublincore.org/documents/dc-xml-guidelines/index.shtml

[21] Expressing Qualified Dublin Core in RDF / XML, http://dublincore.org/documents/2002/05/15/dcq-rdf-xml/

[22] SPECTRUM Metadata Standard, http://www.mda.org.uk/spectrum.htm.

[23] Categories for the Description of Works of Art (CDWA) Metadata Standard, http://www.getty.edu/research/conducting_research/standards/cdwa/ .

[24] Art Museum Image Consortium (AMICO) Metadata Standard, http://www.amico.org/AMICOlibrary/dataspec.html .

[25] http://www.loc.gov/marc/

[26] MAchine-Readable Cataloguing 21 (MARC21) Metadata Standard, http://www.bl.uk/services/bibliographic/marc21move.html .

[27] MODS Metadata Object Description Schema http://www.loc.gov/standards/mods/

[28] METS http://www.loc.gov/standards/mets/

[29] EAD official website http://www.loc.gov/ead/

[30] Visual Resources Association (VRA) Core, http://www.vraweb.org/vracore3.htm .

[31] Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm#Introduction.

[32] IMS Metadata Standard, http://www.imsglobal.org/metadata/index.html .

[33] Alistair Miles, Dan Brickley. SKOS Core Vocabulary Specification, http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102/.

[34] CIDOC-Conceptual Reference Model (CRM), http://cidoc.ics.forth.gr/ .

[35] IPTC Core, http://www.iptc.org/IPTC4XMP/ .

[36] LIDO v0.9 Specification Document: http://www.lido-schema.org/schema/v0.9/lido-v0.9-specification.pdf

[37] Introduction to Lido (Regina Stein): http://www.athenaeurope.org/getFile.php?id=559

[38] http://www.english-heritage.org.uk/

[39] FRBR http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

[40] FRBRoo http://www.cidoc-crm.org/frbr_inro.html

[41] EDM Data Model Primer (http://version1.europeana.eu/c/document_library/get_file?uuid=718a3828-6468-4e94-a9e7-7945c55eec65&groupId=10605)

[42] Europeana Data Model presentation (for v5.2) (http://version1.europeana.eu/c/document_library/get_file?uuid=76eff9ae-5a70-409c-87b0-baf46ede7bd9&groupId=10602)

[43] Lois Mai Chan, and Marcia Lei Zeng. Metadata Interoperability and Standardization – A Study of Methodology Part II, http://www.dlib.org/dlib/june06/zeng/06zeng.html

[44] NISO (National Information Standards Organization). (2004). *Understanding metadata*. Bethesda, MD: NISO Press. http://www.niso.org/standards/resources/UnderstandingMetadata.pdf .

[45] CC:DA (ALCTS/CCS/Committee on Cataloging: Description and Access). (2000). Task Force on Metadata: Final report, June 16, 2000. http://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html .

[46] A. Taylor. The Organization of Information. 2nd ed. Westport, CN: Libraries Unlimited, 2004.

[47] P. Johnston, Metadata and interoperability in a complex world. Ariadne, 37. http://www.ariadne.ac.uk/issue37/dc-2003-rpt/ .

[48] Maria del Mar Roldan-Garcia and Jose F. Aldana-Montes: A Survey on Disk Oriented Querying and Reasoning on the Semantic Web, ICDEW '06: Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), 2006, 58.

[49] Broekstra, J., Kampman, A., Harmelen, F. (2002). "Sesame: A Generis Architecture for Storing and Querying RDF and RDF Schema". 1st Internacional Semantic Web Conference (ISWC2002).

[50] RDQL - A Query Language for RDF W3C Member Submission 9 January 2004. http://www.w3.org/Submission/2004/SUBM-RDQL .

[51] The SeRQL query language, http://www.openrdf.org/doc/sesame/users/ch06.html#d0e1056 .

[52] B. McBride. Jena: Implementing the RDF Model and Syntax Specification. Steffen Staab et al (eds.): Proceedings of the second international workshop on Semantic Web. SemWeb2001.

[53] K. Wilkinson, C. Sayers, and H. Kuno "Efficient RDF Storage and Retrieval in Jena2", Int. Conf. on Semantic Web and Databases, 2003.

[54] KAON. The Karlsruhe Ontology and Semantic Web Framework. Developer's Guide for KAON 1.2.7. January 2004. http://km.aifb.unikarlsruhe.de/kaon2/Members/rvo/KAON-Dev-Guide.pdf.

[55] Wood, D., Gearon, P., Adams, T. Kowari: A Platform for Semantic Web Storage and Analysis. XTech Conference 2005.

[56] Tucana Technologies, iTQL Commands, http://kowari.org/271.htm .

[57] Pan, Z. and Heflin, J. DLDB: Extending Relational Databases to Support Semantic Web Queries. In Workshop on Practical and Scaleable Semantic Web Systems, ISWC 2003.

[58] Finin, Labrou and Mayfield, A brief introduction to the knowledge interchange format, http://www.cs.umbc.edu/kse/kif/kif101.shtml .

[59] Horrocks, I. The FaCT system. International Conference Tableaux'98, number 1397 in Lecture Notes in Artificial Intelligence. Springer-Verlag, 1998.

[60] Grosof, B.N., Horrocks, I., Volz, R., and Decker, S. Description Logic Programms: Combining Logic Programms with Description Logic. In Proceedings of the 12th International World Wide Web Conference. 2003.

[61] Weithoener, T., Liebig, T., Specht, G. Storing and Querying Ontologies in Logic Databases. The first International Workshop on Semantic Web and Databases. VLDB 2003.

[62] Raghu Ramakrishnan, Divesh Srivastava, S. Sudarshan, and Praveen Seshadri. The CORAL Deductive System. VLDB Journal: Very Large Data Bases, 3(2):161-210, 1994.

[63] Horrocks, I., Li, L., Turi, D., Bechhofer, S. The Instance Store: Description Logic Reasoning with Large Numbers of Individuals. 2004.

[64] ECLAP DE3.1 infrastructure: ingestion and processing content and metadata http://bpnet.eclap.eu/drupal/?q=en-US/home&axoid=urn:axmedis:00000:obj:a345a84f-6fdf-4f84-a412-88094ce363e2&section=search_base

10 Glossary

AMICO	Art Museum Image Consortium
CDWA	Categories for the Description of Works of Art
CRM	Conceptual Reference Model
DC	Dublin Core
EAD	Encoded Archival Description
EDM	Europeana Data Model
ESE	Europeana Semantic Elements Specification
FRBR	Functional Requirements for Bibliographic Records
IMS	Instructional Management Systems
IPTC	International Press Telecommunications Council
LIDO	Lightweight Information Describing Objects
MARC21	Machine-Readable Cataloguing
METS	Metadata Encoding and Transmission Standard
MODS	Metadata Object Description Schema
MPEG	Moving Pictures Expert Group
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OWL	Ontology Web Language
OWL DL	Ontology Web Language Description Logics
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema

SKOS	Simple Knowledge Organisation System
SPECTRUM	Standard ProcEdures for CollecTions Recording Used in Museums
VRA	Visual Resources Association
XML	Extensible Markup Language

Appendix A – ECLAP General Questionnaire & Results

ECLAP General Questionnaire

ECLAP Survey - First Part

1. Content provider information

Please give the following information about the organisation providing the content:

1.1. Name

1.2. Organization

1.3. Contact person

1.4. What sort of organisation are you? (multiple answers possible, select from the drop-down)

Organisation type…

Organisation type…
Organisation type…
Organisation type…
If you selected 'Other' please state your organisation type

Contact details of the person in charge of IPR issues

1.5. Name

1.6. Role

1.7. Email

1.8. Telephone

ECLAP Survey

2. Metadata and terminology

Geographic Name Terminology

2.1 - Do you use a standard set of terms for geographic names? (YES/NO)

2.2 - If 'Yes' is the source for the terms? (Select the correct answer)

Developed by the provider

2.3 - Include the contact details of the person who we can ask for more information.

2.4 - For published terminology, please give a reference for each source used, if available. If the source is available on-line please give its URL (web address). For example: TGN

Date Format and Time Period Terminology

2.5 - Is a standard date format (or formats) used to describe the digital objects in this collection? (YES/NO)

2.6 - Which date format do you use?

2.7 - Do you use a standard set of terms for time periods? (YES/NO)

2.8 - If 'Yes' is the source for the terms? (Select the correct answer)

Developed by the provider

2.9 - Include the contact details of the person who we can ask for more information.

2.10 - For published terminology, please give a reference for each source used, if available. If the source is available on-line please give its URL (web address)

Subject Terminology

2.11 - Do you use a standard set of terms for subjects associated with the content (for example: object name, subject, and iconography)? (YES/NO)

2.12 - If 'Yes' is the source for the terms? (Select the correct answer)

Developed by the provider

2.13 - Include the contact details of the person who we can ask for more information.

available on-line please give its URL (web address). For example: AAT, Iconclass, SWD/RSWK, RAMEAU, LCSH, DDC, UDC

Person and Organisation Terminology

2.15 - Do you use standard authorities for persons and organisations? (YES/NO)

2.16 - If 'Yes' is the source for the terms? (Select the correct answer)

2.17 - Include the contact details of the person who we can ask for more information.

2.18 - For published terminology, please give a reference for each source used, if available. If the source is available on-line please give its URL (web address). For example: LCNA, ULAN

ECLAP General Questionnaire Results

Organisation	**Type of organization**	**Geographic name terminology**		
		standard set of terms for geographic names	**Source of terms**	**Reference source for published terminology**
UNIROMA	Research centre, University	No		
CTFR	Archive, L.T.D	No		
B&G	Archive	Yes in-house thesaurus, the GTAA, which has a	developed by provider	http://ems01.mpi.nl:8080/GTAABrowser/

		Geographical 'axis,		
ITB	Library, Archive	No		
UvA	University	No		
ESMAE	School Polytechnic	Yes	Published standard	UNESCO THESAURUS / SIPORBAS
UCLM	University	No		
FIFF	Association	Yes -for films related data (a list of countries)	Developed by provider	
OSZMI	Museum, Library, Archive	No	Developed by provider	
BELLONE	Archive, other	Yes, Own thesaurus	Developed by provider	
UCAM	Museum, Archive, Research Center	No		
MUZEUM	Artistic and cultural production, publishing and dissemination	No	Developed by provider	To be determined
IKP	University	No	Developed by provider	
UG	University	Yes	Developed by the provider	

Organisation	**Date format and time period terminology**			
	standard date format	**standard set of terms for time periods**	**Source of terms**	**Reference source for published terminology**
UNIROMA	Year of creation	No		
CTFR	dd/mm/yyyy but not always	No		

B&G	dd-mm-jjjj for extranet search results jjjj-mm-dd tt:tt:tt for xml exports	No		Not applicable
ITB	Text format?			
UvA	dd-mm-jjjj. For video time codes we use: tt:tt:tt (hours, minutes, seconds)	No		
ESMAE	YYYY/MM/DD	Yes	Published standard	UNESCO THESAURUS / SIPORBASE
UCLM	Text format	No		
FIFF	year of production + day of arrival of the screener at the festival (dd/mm/yy)	No		
OSZMI	YY-MM-DD	No	Developed by provider	
BELLONE	None	Yes	Developed by provider	
UCAM	Yes, yyyy-mm-dd	No		
MUZEUM	No, Mixed	No	Developed by provider	
IKP	No		Developed by provider	
UG	Yes for most collections but not all yyyy-mm-dd	Yes	Developed by provider	

Organisation	**Subject terminology**		
	standard set of terms for subjects associated with the content (e.g.	**Source of Terms**	**Reference source for published terminology (e.g. AAT, Iconclass, SWD/RSWK, RAMEAU, LCSH,**

	object name, subject, and iconography)		**DDC, UDC)**
UNIROMA	Yes	Developed by provider	
CTFR	Yes	Developed by provider	
B&G	Yes, in-house thesaurus, the GTAA, which has Subject and Genre 'axes	Developed by provider	http://ems01.mpi.nl:8080/GTAABrowser/
ITB	No		
UvA	Yes	Developed by provider (The thesaurus of the Theatre Institute Netherlands)	
ESMAE	Yes	Published standard	GETTY / MUSAURUS / UNESCO
UCLM	No		
FIFF	Yes - for films related data	Developed by provider	
OSZMI	No	Developed by provider	
BELLONE	Yes	Developed by provider	
UCAM	Yes	Developed by provider	
MUZEUM	No	Developed by provider	
IKP	No	Developed by provider	
UG	Yes	Developed by provider, LCSH	http://www.ahds.ac.uk/metadata/ahds-controlled-subject-terms.htm ; http://id.loc.gov/authorities/

Organisation	**Person and organisation terminology**		
	Standard authorities for persons and organisations	**Source of terms**	**Reference source for published terminology**
UNIROMA	No		
CTFR	No	Developed by provider	
B&G	Yes, in-house thesaurus, the GTAA,	Developed by provider	

	which has Subject and Genre 'axes		
ITB	No		
UvA	No		
ESMAE	Yes	Published standard	SIPORBASE / REGRAS PORTUGUESAS DE CATALOGAÇÃO
UCLM	No		
FIFF	No		
OSZMI	No	Developed by provider	
BELLONE	Yes	Developed by provider	
UCAM	No		
MUZEUM	No	Developed by provider	
IKP	No	Developed by provider	
UG	Yes	Developed by provider	