



Project no. 033104

MultiMatch

Technology-enhanced Learning and Access to Cultural Heritage

Instrument: Specific Targeted Research Project

FP6-2005-IST-5

D2.2.1 Semantic Web Encoding

MultiMatch Knowledge Representation and Interoperability
with Cultural Heritage Domain Standards

Start Date of Project: 01 May 2006

Duration: 30 Months

Netherlands Institute for Sound and Vision

Final Version

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)

Document Information

Deliverable number: D2.2.1
Deliverable title: Semantic Web Encoding
Due date of deliverable: February 2007
Actual date of deliverable: March 2007
Author(s): Neil Ireson, Johan Oomen, Hanneke Smulders.
With contributions by Giuseppe Amato, Rafael Carrasco, Martha Lawson, Sam Minelli and Carol Peters.
Participant(s): ISTI-CNR, Netherlands Institute for Sound and Vision, University of Sheffield, University of Amsterdam
Workpackage: WP2
Workpackage title: Content Selection and Preparation
Workpackage leader: Netherlands Institute for Sound and Vision
Est. person months: 7.9
Dissemination Level: PU
Version: Final Version
Keywords [interoperability, semantic web, metadata, metadata schemas, controlled vocabularies, knowledge representation, cultural heritage]

Abstract

This deliverable identifies the dimensions of the content the cultural heritage institutes have provided to test MultiMatch prototype 1. Three main outcomes are reported. The common interoperability schema in combination with the user requirements resulted in a MultiMatch data model and functional metadata model which are at the same time a refinement and an extended version of Dublin Core (DCMI Metadata Terms).

A second result of this task is the definition of the semantic web encoding (including common metadata schemas and ontology mapping providing common understanding on the file formats to be made accessible by the current and future Cultural Heritage institutes).

A third result is the mapping of the MultiMatch metadata schema to Dublin Core Metadata Element Set (deflating the schema for exchange purposes) and to the reference model CIDOC-CRM (which is believed to become increasingly important for knowledge representation and interoperability in the Cultural Heritage domain). This deliverable will be revised at month 25, when the experimental results of the first prototype have been processed with the current versions of the data model and metadata schema and feedback is available.

Table of Contents

Abstract	2
Table of Contents	3
Executive Summary.....	5
1 Introduction	7
1.1 Metadata Working Group.....	8
1.2 Outline of the Document	9
1.3 Terminology	9
2 MultiMatch Metadata Development.....	12
2.1 “External” Issues: Interoperability in the CH Domain	13
2.2 “Top-down” Issues: Meeting the specification of the user requirements	14
2.3 “Bottom-up” Issues: Representing the concepts which are present in the data.....	15
2.4 Conclusion.....	15
3 The Content: Data with accompanying Metadata	17
3.1 Themes of the Content.....	17
3.2 Metadata encoding.....	19
3.3 Distribution and overview	19
3.4 Alinari content for the first prototype.....	20
3.4.1 Selection of the Content	20
3.4.2 Metadata schema and controlled vocabularies in use.....	20
3.4.3 Metadata for first prototype	21
3.5 Biblioteca Virtual Miguel de Cervantes content for the first prototype	22
3.5.1 Selection of Texts and Web pages.....	22
3.5.2 Metadata schema and controlled vocabularies in use.....	22
3.5.3 Metadata for first prototype	23
3.6 Netherlands Institute for Sound and Vision content for the first prototype.....	25
3.6.1 Selection of the Content	25
3.6.2 Metadata schema and controlled vocabularies in use.....	25
3.6.3 Metadata for first prototype.....	27
3.7 Web data.....	29
CH-related subset of Wikipedia	29
Cultural heritage white list crawl.....	29
3.8 UvA Audio Corpus.....	30
3.9 The European Library.....	31
3.10 The common metadata from the provided content	32
4 Generic Metadata Schemas	33
4.1 Dublin Core (and Refinements).....	33
4.2 MPEG-7.....	33
4.3 Functional Requirements for Bibliographic Records	34
4.4 CIDOC CRM.....	35
4.5 Generic metadata schema: final remarks	36
5 MultiMatch Metadata Schema.....	36
5.1 Focus of knowledge representation in MultiMatch	36
5.2 MultiMatch Data Model	37
D2.2.1 Semantic Web Encoding	

Overview of entities in the MultiMatch data model.....	38
5.3 Introduction to the MultiMatch metadata schema.....	40
5.3.1 The formats of the MultiMatch metadata schema.....	40
5.3.2 Explanation of the metadata format tables.....	41
5.4 Issues in Application of the MultiMatch Metadata Schema.....	42
5.4.1 Closed Lists / Controlled Vocabularies.....	42
5.4.2 Relationships in the MultiMatch data model.....	42
5.4.3 Alternative values and Multilingual related metadata issues.....	43
5.4.4 Processing (free text) ambiguous date information.....	44
5.5 Mapping of first prototype metadata to the MultiMatch Metadata Schema.....	44
5.5.1 Mapping of Alinari first prototype metadata to the MultiMatch Metadata Schema.....	44
5.5.2 Mapping of Cervantes first prototype metadata to the MultiMatch Metadata Schema.....	47
5.5.3 Mapping of Sound and Vision first prototype metadata to the MultiMatch Creation format.....	48
6 The MultiMatch ontology and interoperability.....	49
6.1 Interoperability in MultiMatch.....	49
6.2 The MultiMatch Ontology.....	50
6.3 Standardised values and Controlled Vocabularies.....	52
6.4 AAT (Arts and Architecture Thesaurus).....	53
6.5 ULAN (Getty Unified List of Artist Names).....	54
6.6 TGN (Thesaurus of Geographic Names).....	55
6.7 Interoperability: further work.....	56
7. Concluding remarks.....	57
References.....	58
Acronyms.....	59
Annex 1: Artist list used for Content selection.....	60
Annex 2: Sample record Alinari.....	62
Annex 3: Biblioteca Cervantes xsd schema.....	63
Annex 4: Sample record Sound and Vision ('element' range).....	64
Annex 5: Draft version of MultiMatch OWL Ontology.....	66
Annex 6: Relationship between MultiMatch Creation and Dublin Core / CIDOC CRM.....	74
Annex 7: Relationship between MultiMatch Actor and CIDOC CRM Actor.....	75
Annex 8: MultiMatch metadata schema.....	76
Annex 9: MultiMatch Metadata xsd schema.....	93
Annex 10: Example illustrating the application of the MultiMatch metadata schema.....	101
Annex 11: TEL sample record.....	112

Executive Summary

The main objective of Deliverable 2.2 is to describe the knowledge representation framework to be adopted in MultiMatch. The deliverable is released in two versions: a preliminary version at month 10 (D2.2.1) and a final version at month 25 (D2.2.2). This preliminary version reports the knowledge representation for the first system prototype and, as such, focuses mainly on the construction of a “common metadata schema” to describe the content of prototype 1. We expect to revise and further extend this schema as work progresses and as a result of the feedback received from the experience gained with the first system prototype. We also consider the problem of interoperability with other collections and present an ontological representation of the MultiMatch data.

Content for the first prototype

A preliminary goal was to provide a descriptive overview of the metadata schemas and semantic resources (i.e. thesauri, controlled vocabularies) widely used within the organisations belonging to the specific sub-domains. This was reported in deliverable D2.1: First Analysis of Metadata in the Cultural Heritage Domain. However, the starting point for the current work was to study the content and associated metadata actually to be used in the first prototype. This consisted in the first place of the material selected by the cultural heritage institutions members of the MultiMatch consortium in different languages and different media, plus data extracted from Wikipedia, crawled from domain-specific sites on the Web and provided by The European library, as shown in the following table:

Origin	Alinari	Sound and Vision	Biblioteca Virtual Miguel de Cervantes	Wikipedia	White list crawl	UvA Audio Corpus	The European Library
Quantity	5.000 stills (jpg)	900+ video's (mpeg-1)	9.000 texts	65.000 Wikipedia articles	40.000 pages	20 hours audio	1.6 million records
Metadata dimensions	Proprietary Dublin Core	Proprietary Dublin Core	Proprietary Dublin Core	UvA web content format	UvA web content format	UvA web content format	OAI-DC
Content location	Alinari repository	Hard disk, metadata on BSCW	Project BSCW	OCLC SFTP	OCLC SFTP	OCLC SFTP	OCLC SFTP

The MultiMatch metadata schema

The next step was to attempt to build a common metadata schema to which the proprietary schema of our content providers could be mapped. Three main factors influenced this work:

- The need to meet the specification of the user requirements. In terms of the influence on the metadata, the user requirements give an indication as to the concepts that are required by the user and the relative importance of those concepts. From the expert users survey in D.1.2 we can conclude that, on average, experts tend to classify searches for information about creators (authors) and creations (works of art and masterpieces) as their most common search tasks. Therefore, in MultiMatch we have initially decided to focus two types of specialised searches on creators and creations, although specialised searches focused on other relevant categories will also be considered.
- The need to represent the concepts that are present in the data. By examining the data it is possible to determine the issues that arise when representing concepts relating to cultural heritage objects. Therefore the metadata should consider the ability to adequately represent the concepts that are extractable from the data itself, so that concepts (including possibly unforeseen concepts) can be suitably represented.
- The need for interoperability: mapping from content provider legacy metadata and to current “standard” metadata schemas. Interoperability is concerned with the capability of different

information systems to communicate. In the research conducted by MultiMatch, it became clear that next to the Dublin Core element set, also the entity-relationship model FRBR and CIDOC conceptual reference model can prove valuable solutions to provide the level of interoperability required.

In addition, some MultiMatch specific aspects impacted on the building of the metadata schema, namely multilinguality, the automatic information extraction procedures and multimodal search facilities.

In the cultural heritage domain, a vast amount of research is conducted in the field of metadata and metadata interoperability. In the first place, the MultiMatch metadata schema is described in the context of the four standards most relevant for interoperability (see D2.1, section 5.2.2 and 5.3.2):

- Dublin Core: because it is in use throughout the entire cultural heritage domain.
- MPEG-7: because it can handle multimedia in a way appropriate for MultiMatch.
- FRBR: because it provides a data model with relationships and a hierarchy that are probably useful for MultiMatch.
- CIDOC CRM: because it provides a reference model for the cultural heritage domain.

For each of these standards, its applicability for the MultiMatch project is discussed in the deliverable.

However, in the end we have decided to build the MultiMatch metadata schema upon DCMI Metadata Terms. Overall the focus of the schema is on representing the kind of data of interest to the user (D1.3) in combination with the focus of the project on Creators and Creations. The data model and the metadata schema are both constructed in such a way, and with an eye on interoperability, so that they can easily be extended when necessary. The MultiMatch schema is intended to be automatically populated where possible, otherwise the elements will remain empty. Prototype 1 will test this first proposal for the MM metadata schema. Annex 10 provides an Example illustrating the application of the MultiMatch metadata schema.

The MultiMatch ontology and interoperability

In order for other systems to be able to use the (meta)data published by MultiMatch it is not only necessary to publish that data but provide it in a standard machine-readable form (syntax) and provide semantic mark-up so that the meta(data) can be correctly interpreted. The techniques developed to enable the Semantic Web are being employed for this purpose. This involves publishing the metadata described above as an ontology, in an XML – RDF – RDFS – OWL form. The OWL description gives the basic class structure of the metadata entities and the properties associated with those classes, described in the tables above. It is not simply a flat, record-like data model; the metadata relationship attempts to capture some of the fundamentally important structure of the Cultural Heritage domain. The OWL model gives a representation of the current class structure.

Testing and revising the initial schema

The MultiMatch metadata schema needs to be evaluated in the first prototype. In month 25 this schema will be updated in D2.2.2, according to:

- the experiences with populating the schema from the content provided for the first prototype: how feasible did it prove to be, to extract metadata, concepts and relationships;
- the user evaluation: does the metadata schema satisfy the user requirements for search and presentation of results, and does it enable a satisfactory interoperability level between the different collections indexed for the first prototype.

1 Introduction

According to the title specified in the project technical annex (DoW), the declared objective of this document should be to provide “Semantic Web Encoding”. In practice this title was found to be misleading. Although a semantic web encoding is one of the aims of MultiMatch, the key area of work covered by Task 2.2 and reported in this deliverable is the construction of a “common metadata schema”. For this reason, to better represent our objectives, the title is extended with “MultiMatch Knowledge Representation and Interoperability with Cultural Heritage Domain Standards”.

The DoW also states that the MultiMatch common metadata schema should be derived primarily from an analysis of the “ontologies”, i.e. controlled vocabularies and metadata, provided by the Cultural Heritage (CH) institutions within the project, where “common” is intended as in use among the MultiMatch CH Institutes. However during the initial study, it became apparent that this scope was too limited to derive an efficacious metadata schema. Firstly, although the major data of the MultiMatch Content Providers (CPs) are of different media types, domains and languages (i.e. Dutch video, Spanish literature and Italian photographs), they do not cover the whole CH domain. It was shown in D2.1 that each of the Cultural Heritage Sectors tends to use a metadata schema specifically tailored towards the needs of the users within that sector, and this was certainly apparent from the analysis of the CPs. Secondly, an analysis of the metadata used by a given institution does not give the whole picture as it may be more concerned with the indexing and archiving of data rather than the information search requirements.

Therefore this document has extended the originally intended scope of this activity as defined in the DoW and, in addition to analysing the CPs’ metadata, also pays considerable attention to the results of User Requirements and Functional Specification documents produced by Work Package 1, in order to ensure the outcomes of this deliverable satisfy these requirements. In addition, to overcome the proprietorial nature of the CPs’ ontologies, which acts as a hindrance to interoperability, a greater emphasis was placed on the generic metadata schema and models used in the CH domain. This follows the user-centred design strategy adopted by the project, and extends the utility of the document, beyond solely providing a common metadata schema that allows for interoperability between the CPs for the first prototype. Instead the aim is to provide a more complete picture of the metadata that is required to meet the wider needs of the MultiMatch system.

Given the current state of work in MultiMatch there is a realisation that those requirements outside the first prototype, e.g. interoperability beyond the CPs and encoding of automatically extracted semantic annotations, whilst considered in this document will be finalised in later deliverables, when the results of the first prototype and the initial work in WP 4/5 can be evaluated.

The wide scope of influences considered in the metadata developmental process is due to the fact that metadata can be considered to be at the heart of the MultiMatch project; acting as the “glue” between the user (interface) and content (data). Many of the questions the users might ask the system can only be supported if the system can represent and retrieve the required information from the metadata. In addition, the MultiMatch system does not work in isolation with a predefined set of users but provides information search facilities in a broader domain. Therefore, when developing the appropriate metadata schema, there are a number of requirements that must be taken into consideration. These can be divided into three basic areas:

- **“External”** - from the legacy metadata and current standards.
The CPs are providing metadata that will be mapped onto the MultiMatch metadata schema. To a large extent this metadata gives an indication of those concepts that are important to the CP; this aspect has thus been taken into consideration during the development of the metadata. On a higher level, metadata is the key to interacting with the wider digital cultural heritage domain. MultiMatch is very much aware of the impact the project can have in working towards the i2010 Digital Libraries initiative. In effect, the scope of this research goes beyond the direct requirements expressed within the development of the MultiMatch system as it will not be an isolated system, but rather an open system that enables multifarious users to explore and interact with online internet-accessible cultural heritage content, across media types and language boundaries.

- **“Top-down”** - from the user requirements.
To ensure that user needs are met, the metadata has been derived from the functional specifications for Prototype 1, which in turn were derived from the user requirements.
- **“Bottom-up”** - from the content (data).
Meeting these needs also necessitated that the metadata can satisfactorily represent those concepts that are expressed within (and extracted from) the data.

Thus, the document initially focuses on what are referred to as “external” influences, these are more in line with the originally envisaged scope of the deliverable, i.e. analysis of the CPs metadata. In addition the generic metadata schemas and reference models relevant to MultiMatch are described and their relationship to the development of the MultiMatch schema is considered. The influence of the User Requirements on the metadata is implicit in Deliverables 1.2 and 1.3. However the effects of considering both the generic metadata schema and the user requirements are explicitly presented in the metadata element tables, which summarise the reasoning behind the development of the schema. In terms of the “bottom-up” issues, i.e. how the need to represent particular CH concepts affects the metadata, these are in part implicitly considered by the CP and generic metadata schemas, and in particular by the domain reference models. Issues of particular relevance to MultiMatch, e.g. multilinguality and automatically extracted metadata, are highlighted to show how they must be considered and will be more fully examined in D2.2.2. The impact of these three factors is discussed further in Chapter 2.

This document outlines how the development of the MultiMatch metadata schema has been considered in the context of Semantic Interoperability to ensure that the aims of the project are met. Although examples of semantic web encodings of the current metadata schema are provided, the final encoding is, in part, reliant on the results of the automatic semantic annotation by WP4 for the second prototype. The focus here is on the interoperability required by the first prototype, i.e. with the MultiMatch CH Institutes.

1.1 Metadata Working Group

The development of the MultiMatch search engine can be divided into four areas:

Data Collection

- crawl the Internet to identify websites with CH information, locating relevant texts, images, audio and videos
- likewise identify relevant material via an in-depth crawling of selected CH institutions, accepting and processing any semantic web encoding of the information retrieved

Data Analysis

- automatically classify the results, in a semantic-web compliant fashion, based on document content, metadata, context, and on the occurrence of relevant CH concepts
- automatically extract relevant information that will then be used to create cross-links between related material, such as biographies, exhibitions of work, critical analyses, etc.

Indexing

- organise and further analyse the material crawled to serve focused queries generated from user-formulated information needs

Search and Retrieval

- interact with the user to obtain a more specific definition of information requirements
- organise and display search results in an integrated, user-friendly manner, allowing users to access and exploit the information retrieved regardless of language barriers

As metadata issues play a role in each of these areas, a special working group was initiated with representatives from different work-packages and perspectives to study the issues involved in creating the MultiMatch metadata schema. The starting point was ‘D2.1 First Analysis of Metadata in the Cultural Heritage Domain’, which listed forty knowledge representation standards and documented their usage. To coordinate the group activities, with respect to the work presented in this document, numerous telephone

conferences were held between December and February and the group gathered for meetings in Amsterdam (8-9 February 2007) and Madrid (28-29 March 2007).

As it became increasingly evident that for this work to be effective there was a need to focus beyond the MultiMatch CPs in order to identify a knowledge representation framework that would be suitable to handle heterogeneous cultural heritage content and that would not preclude interoperability, MultiMatch organised a joint Delos-MultiMatch workshop (15 February 2007) in order to bring together a set of experts (both researchers and practitioners) working on the representation of cultural heritage digital objects. The workshop gave us the opportunity to investigate the current state-of-the-art, and to discuss and compare ideas with experts in the field and with others who have already faced or are currently addressing the same issues.¹

1.2 Outline of the Document

The structure of this document is as follows:

- Chapter 1 (this chapter) gives a broad overview of the issues, methodology and terminology and results discussed in the rest of the document.
- Chapter 2 describes the issues addressed when developing the MultiMatch metadata schema.
- Chapter 3 presents the Content Providers meta(data) and controlled vocabularies, and the content selection process for the first prototype.
- Chapter 4 describes the generic metadata schemas, reference models that are used in the CH domain and their relationship to the MultiMatch metadata schema.
- Chapter 5 describes in detail the MultiMatch data model plus metadata schema.
- Chapter 6 outlines the approach adopted towards achieving the required level of interoperability with other collections including developing an ontological representation of the MultiMatch metadata schema, which is a vital component of the MultiMatch system.
- Chapter 7 concludes the document and indicates the next steps WP2 will take towards the second prototype.

1.3 Terminology

In order to have a clear understanding of the document the following table provides a definition of some of the key specific terminology used in the MultiMatch project.

creations	<p>In the context of MultiMatch, a Creation describes a physical cultural heritage object, i.e. it is:</p> <ul style="list-style-type: none"> • a unique man-made object; • the result of a distinct intellectual or creative creation process which can be a Text, Still Image, Video, Audio or Physical Object; • an Endeavour in terms of the Functional Requirements for Bibliographic Records (FRBR) reference model, in that it represents a “Work” in its abstract term and also the actual “Manifestation” of that work. <p>Note, that a digital still image, e.g. a digital photograph, can be a Creation. A thumbnail of that original work of art would then be the Digital representation of that photograph.</p>
-----------	--

¹ See http://www.delos.info/index.php?option=com_content&task=view&id=533&Itemid=278 for the online proceedings and presentations of the workshop.

Cultural Heritage object	Creation as an entity in the MultiMatch data model.
Data model	"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. A conceptual schema, or high-level data model or conceptual data model, is a map of concepts and their relationships." ²
Digital representations	<p>A digital representation is a visual surrogate or reproduction of a Creation.</p> <p>In many cases the digital representation will be an image (think of scans or digital photographs of paintings, books, sculptures, videos etc.). Also, a thumbnail of an image of a Creation is a digital representation.</p> <p>However, for Audio and Video Creations (the original work of art can be analogue or digital) there will also be digital representations of several kinds. If possible, every kind of digital representation, regardless what type of creation it refers to, should be described in one format.</p> <p>For example, an analogue artwork can be a movie or film (type of Creation = Video). This Creation is likely to have more than two digital representations:</p> <ul style="list-style-type: none"> • Images (the keyframes that summarize the movie) • a Browse Copy in MPEG-1 • a high quality digital representation in MPEG-2 or MPEG-4 or something else.
Metadata	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).
Metadata element	A 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.
Metadata schema	"Full, logically organised structure of relations between defined (groups) of metadata and the information objects they describe." [De Jong 2003] "a set of rules for encoding information that supports specific communities of users." [Baca 1989]
Ontology	<p>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 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).³ Therefore thesauri and classification schemes can be regarded as ontologies with a relatively small number of relationships.</p> <p>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.</p>
Semantic Web	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.

² www.wikipedia.org

³ Definition taken from: www.wikipedia.org

Works of art	<p>Within MultiMatch the following five types of original works of art, represented in digital form, can be described, although other types of Cultural Heritage objects are not excluded:</p> <ul style="list-style-type: none">• Text, e.g. books, poems etc.• Sound, e.g. radio programmes, radio plays, other recorded speech or sounds, a music playback file format, an audio compact disc.• Still Image, e.g. paintings, photographs, drawings, maps.• Video, e.g. TV programmes and films.• Physical Object, e.g. sculpture, vase, cooking pot, any inanimate, three-dimensional object or substance that is the result of a distinct intellectual or creative creation process.
--------------	--

2 MultiMatch Metadata Development

One of the first activities of the MultiMatch project was to provide an overview of current practice regarding knowledge representation in the cultural heritage domain in Deliverable 2.1: First Analysis of Metadata in the Cultural Heritage Domain. As metadata standards facilitate interoperability and exchange of information between systems and organisations, this overview provided the basis for the approach towards interoperability that will be adopted within MultiMatch.

The research for Deliverable D2.1 showed that each of the Cultural Heritage Sectors tends to use a metadata schema specifically tailored towards the needs of the users within that sector. Whilst there are crosswalks available between some of these schemas, the primary method to achieve interoperability between the multifarious schemas is to map to and from the Dublin Core. However, whilst the Dublin Core offers a flexible representation, it is less expressive than the specific schemas so there is a degree of information loss during this mapping process. To an extent this can be alleviated by using the standard refinements of the Dublin Core, such as the Qualified Dublin Core or the Visual Resources Association Core Categories (VRA Core 4.0)⁴. An alternative is to use more expressive standards like MPEG-7/21⁵, or to map to reference models such as FRBR⁶ and CIDOC⁷ that offer more expressiveness. Biblioteca Virtual Miguel de Cervantes, one of the MultiMatch partners, can export their metadata to a CIDOC-CRM compatible format. Sound and Vision has applied the reference model FRBR into its metadata schema iMMix. However these reference models are not currently widely adopted methods to provide interoperability. Chapter 4 provides a more in-depth discussion of these issues.

The next step, after D2.1, was to choose from or build upon these standards and design the appropriate knowledge representation approach for the project. As stated in the DoW, the intention was not to invent a new "standard". In practice this involves the consideration of the three general areas of influence upon the metadata development in order to ensure that the metadata is expressive enough, and has involved close consultation with the other work packages:

- To interoperate with content provider metadata and other standards in the domain, it had to be considered work being done in WPs 4 and 5, given that this involves applying generic semantic interoperability technologies but tailoring them to the Cultural Heritage Domain.
- To meet the requirements expressed by the final users of the system involves WP1 issues, primarily expressed in the User Requirements and Functional Specification documents (D 1.2 and 1.3).
- To express the concepts which are present and to be extracted from the data, involves work being done in WPs 4 and 5, through examination of the data in light of the automatic classification on information extraction discussed in the State of the Art (D1.1), which will be applied within the project.

These three influences are referred to as "external", "top-down" and "bottom-up", respectively, and are described in more detail in the sections below.

⁴ <http://www.vraweb.org/projects/vracore4/index.html>

⁵ MPEG-7 is formally called *Multimedia Content Description Interface*. Thus, it is *not* a standard which deals with the actual encoding of moving pictures and audio. [MPEG-7: Overview of MPEG-7 Description Tools](#)

The MPEG-21 standard, from the Moving Picture Experts Group, aims at defining an open framework for multimedia applications. ISO 21000. Specifically, MPEG-21 defines a "Rights Expression Language" standard as means of sharing digital rights/permissions/restrictions for digital content from content creator to content consumer. <http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm>

⁶ Functional Requirements for Bibliographic Records. <http://www.ifla.org/VII/s13/frbr/frbr.htm>

⁷ The Committee on Documentation of the International Council of Museums (ICOM-CIDOC) focuses on the documentation requirements and standards of museums, archives, and similar organizations. CIDOC has defined a Conceptual Reference Model (also known as the CIDOC CRM).

2.1 “External” Issues: Interoperability in the CH Domain

The “external” influences on the development of the MultiMatch metadata refer to the other metadata schemas that have to be taken into consideration. In broader terms the issues relate to Semantic Interoperability, i.e. the capability of different information systems to communicate information consistent with the intended meaning of the encoded information (as intended by the creators or maintainers of the information system). This communication may take various forms such as the transfer, exchange, transformation, mediation, migration or integration of information. It involves processing of the shared information so that it is consistent with the intended meaning

Initially, for the first prototype, this mainly involves the CP metadata resources. The three partners of MultiMatch, that provide their catalogues, apply the following standards for knowledge representation:

- Alinari
 - a proprietary metadata schema based upon Dublin Core
 - the schema is mapped to Dublin Core in XML for exchange purposes
 - a proprietary thesaurus for subject indexing in Italian and English (a selection is also translated into French, German, Polish and Spanish)
 - together with several proprietary authority lists in English, French, German, Italian and Spanish.
- Biblioteca Virtual Miguel de Cervantes
 - the MARC21 metadata schema
 - the schema is mapped to Dublin Core in XML, IEEE LOM⁸, MARC⁹-XML and CIDOC-CRM for exchange purposes
 - the UDC classification schema in Spanish for subject indexing
 - a controlled vocabulary with subject headings in Spanish based on the one in use by the University Library of Alicante.
- Netherlands Institute for Sound and Vision
 - a proprietary metadata model that makes use of the FRBR concepts
 - the schema is mapped to Dublin Core in XML and IEEE LOM for exchange purposes
 - a thesaurus for subject indexing in Dutch (the de-facto standard for audiovisual archives in the Netherlands)
 - several proprietary authority lists in Dutch several proprietary authority lists in Dutch.

It is clear that the CP metadata schemas are tailored to the focus of the respective collections:

Cultural Heritage institute	Specific domain
Alinari	Photography
Cervantes	Spanish Literature
Sound and Vision	Dutch Broadcasting

Chapter 3 gives a thorough analysis of the CP metadata, and associated control vocabularies to determine how they will be used (mapping into) MultiMatch.

⁸ Institute of Electrical and Electronics Engineers, Learning Objects Metadata (Standard for describing learning objects)

⁹ MARC is an acronym for MACHine-Readable Cataloging. It defines a bibliographic data format and provides the protocol by which computers exchange, use, and interpret bibliographic information.

However, MultiMatch is intended to be used in a wider CH domain than that represented by the CPs. The influence of the generic CH domain has been considered primarily by examining the generic interoperability resources used in the domain, especially the reference models (FRBR and CIDOC) which have been specifically designed to capture the information within the domain. The influence of these “external” resources on MultiMatch is discussed in Chapter 4.

2.2 “Top-down” Issues: Meeting the specification of the user requirements

The MultiMatch project is following a user-centred design strategy and so a good deal of the initial effort was directed at interviewing the end-users of the system to determine their requirements (see Del 1.2). These user requirements have been translated into a functional specification that is used to guide the development of MultiMatch so that the final system will meet the needs (and expectations) of the end-users (Del 1.3). Obviously there is not a unique concept of an end-user, and different users, having different levels of expertise and information search needs, will present different requirements that should be considered in the system development.

In terms of the influence on the metadata, the user requirements give an indication as to the concepts that are required by the user and the relative importance of those concepts.

Many of the findings of the user requirements are, as one would expect, intuitive, such as the fact that text and images are the media types of primary importance, and that creators and their works are the main concepts for search, with typical queries involving: proper names, places, and titles. There were also less definite requirements, such as clearer semantics in faceted browsing and clustering so that the categorisation (clustering) of the search results “make sense”. This strongly points to the need to use standard (and well defined) vocabularies and subject taxonomies.

In MultiMatch all search facilities will be translingual, i.e. the user can formulate queries in a given language and can retrieve results in any of the languages supported by MultiMatch (i.e. primarily English, Spanish, Italian and Dutch). Users expressed a strong desire to have control over the process, i.e. selection of the appropriate translation in the case of a word with multiple meanings and the option of not translating certain expressions. The notion of a controllable and transparent search process was thus seen as a major requirement on the system, a key reason for this is the ability to determine the “authoritativeness” of any information that results from a user search.

From the expert users survey it can be concluded that, on average, experts tend to classify searches for information about creators (authors) and creations (works of art and masterpieces) as their most common search tasks. Therefore, the MultiMatch partners have initially decided to focus two types of specialised searches on creators and creations, although specialised searches focused on other relevant categories will also be considered.

For the top-down approach to the development of the MultiMatch metadata schema the user requirements, as reported in D1.3, have been used as a starting point. In other words, the descriptions of the requirements have been translated into required metadata elements for search and presentation purposes. The outcome of the top-down approach is fully reported in Chapter 5, via the Schema tables that have a separate column referring to D1.3.

2.3 “Bottom-up” Issues: Representing the concepts which are present in the data

Whilst it is crucial to provide a representation to fulfil the user requirements, these will not fully specify the requirements of all users: only a selection of users have been interviewed (expert users), needs are dynamically changing and users’ concepts of what they want are very much dependent on their experience of what is available. In addition, by examining the data, it is possible to determine the issues that arise when representing the concepts relating to cultural heritage objects.

Therefore the metadata must be able to adequately represent the concepts that are extractable from the data itself (including concepts possibly unforeseen in the user requirements process). In the MultiMatch project this data includes digital representations of Cultural Heritage objects (images, texts, etc.) and data describing those objects, which is presented in a human rather than machine-readable format.

However when examining the data it is seen that values can be ambiguous (an individual can be known by several names), have varying degrees of precision (a location can be a country, province or town) or may be contested (the attribution of a creation may be in doubt).

The ambiguity, imprecision and uncertainty of the data is accentuated by its multilingual nature and fundamentally by the use of automatic techniques to extract information from it. Cultural heritage organisations employ professional cataloguers to annotate the objects in their collections, using a metadata model and cataloguing rules formalise the way objects are described in the catalogue and keywords are usually assigned to an object from a given thesauri or controlled vocabulary. However, within MultiMatch this annotation process will use automatic techniques. Thus, where a concept is seen to have conflicting values, this could be due to a genuine difference of opinion in the Cultural Heritage domain, an error in the data, or an error in the extraction process.

One of the key features which was highlighted by the user requirements analysis is the need to represent the authoritativeness of the information presented to the user. It is therefore important for the MultiMatch representation to be able to express the ambiguous, imprecise and uncertain nature of the information to the user. It is also potentially useful/necessary to provide an “audit-trail” to the source(s) and process(es) which have been used to acquire the information.

In addition to extracting metadata from textual data, MultiMatch will also extract metadata from other media types; audio, still images and video (although to some extent audio (transcripts) can be seen as noisy text and video is seen as (keyframe) images and audio (text). The metadata which is extracted from images represents the underlying, low-level features of an image; however, what the user requires is the linking of this to a meaningful (semantic) representation of the image. Thus the metadata format must cope with low-level and high-level concepts and the links between these representations.

2.4 Conclusion

All three influences, “external”, “top-down” and “bottom-up”, are taken into consideration in the MultiMatch data model and metadata schema presented in Chapter 5. However it should be noted that whilst this tri-partite division is used to more clearly differentiate the different influences, they are obviously heavily interrelated. For example, the user requirements express the need to retrieve information about “Creators”, all of the metadata schemas, used both by the MultiMatch CP and more widely in the CH domain, have some form of representation of a creator, and within the data the “Creators” are represented in various formats. Thus the MultiMatch schema must also contain the concept of a creator and the representation of that concept within the schema needs to consider all the three issues.

These concepts have been taken into consideration, to an extent, by the metadata schemas and reference models used in the CH domain and examined in Chapters 3 and 4. However, many of these issues are not fully considered in this deliverable. As a result of the work of WPs 4 and 5, the relevant concepts will be extracted from the data and indexed, and this will result in a clearer understanding of the requirements placed on the metadata by this process. Thus, aspects which relate to the data, such as multilinguality and the information extraction process, are very specific to the MultiMatch system; these will be further examined in D2.2.2.

3 The Content: Data with accompanying Metadata

3.1 Themes of the Content

The consortium partners selected a well-balanced corpus of test data for training and testing of the first prototype.

In order to structure the selection of content from the three cultural heritage institutions (Alinari, Sound and Vision and Biblioteca Virtual Miguel de Cervantes (BVMC)), the consortium partners compiled a list of top-ranking artists in their collections. This list covers authors in the domains of:

- fine arts (primarily European painters)
- literature (with a focus on Spanish literature)

The cumulated list is included as Annex 1. This list has been compiled in order to ensure that comparisons between objects in the various collections can already be made in the first prototype, when the system is tested on a small corpus.

In addition to the content from the three cultural heritage institutions, the University of Amsterdam has provided access to many thousands of Wikipedia articles and Websites related to European cultural heritage websites. Finally, The European Library Office in The Hague has kindly offered MultiMatch a test corpus of 1.6 million metadata records from the European Library portal.

Figure 1 provides an overview of the content provided for the first prototype.

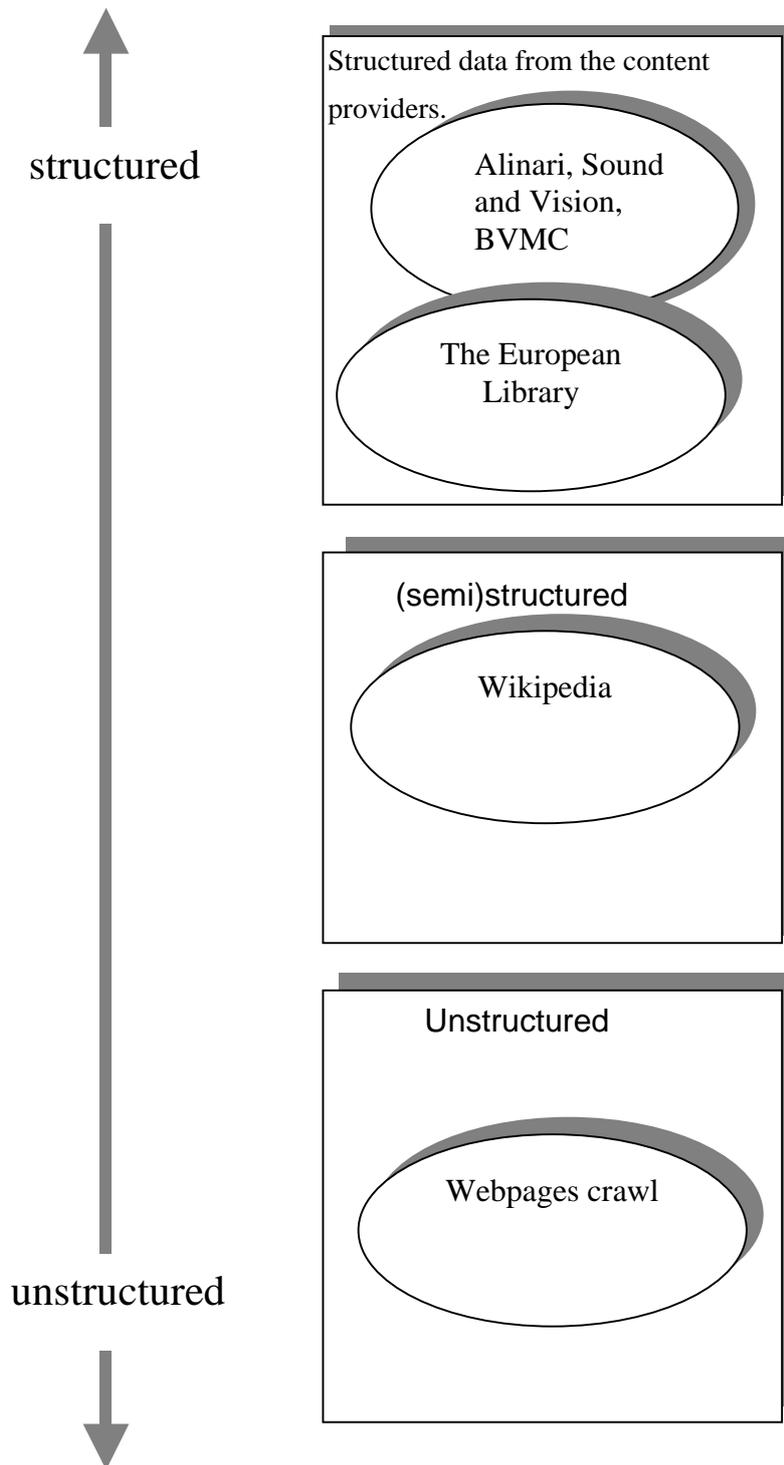


Figure 1: Content for the first prototype: structured and unstructured sources.

3.2 Metadata encoding

The three cultural heritage institutions offer their metadata in XML format. For the first prototype, Dublin Core is used as export format to which the mapping to the MultiMatch schema is performed. For the second prototype, it is expected that the richer 'native' catalogue entries will be used.

The Web data is delivered in a proprietary format defined by the University of Amsterdam. The deliverable D4.1 "Text/Image/Speech/Video Indexing components and documentation for 1st prototype"¹⁰ will describe the UvA Web data distribution format. It will show the content of the files available for a dump along with some naming conventions, the structure of bins, and details on the media files and how to map between them and referring documents.

The test collection of The European Library is delivered in XML format; see Annex 11 for a sample record.

3.3 Distribution and overview

For reasons of convenience, the content (including both the digital objects and accompanying metadata in XML) is stored on different areas:

- SFTP server: hosted by consortium partner OCLC PICA for large files. Every consortium member has their own directory on this server.
- BSCW workspace: hosted by consortium partner CNR as the document store of the project.
- Alinari repository: location of the images provided by consortium partner Alinari
- Hard disks: to host the video files from Sound and Vision

The following table summarizes the content available in the first prototype; the following paragraphs provide a more detailed description.

Origin	Alinari	Sound and Vision	Biblioteca Virtual Miguel de Cervantes	Wikipedia	White list crawl	UvA Audio Corpus	The European Library
Quantity	5.000 stills (jpg)	900+ video's (mpeg-1)	9.000 texts	65.000 Wikipedia articles	40.000 pages	20 hours audio	1.6 million records
Metadata dimensions	Proprietary Dublin Core	Proprietary Dublin Core	Proprietary Dublin Core	UvA web content format	UvA web content format	UvA web content format	OAI-DC
Content location	Alinari repository ¹¹	Hard disk, metadata on BSCW ¹²	BSCW ¹³	SFTP ¹⁴	SFTP ¹⁵	SFTP ¹⁶	SFTP ¹⁷

¹⁰ This deliverable due at 12 months is only available in a preliminary version and has not yet been completed.

¹¹ <http://project.alinari.it/drivemultimatch/htdocs/index.php?>

¹² <http://dlib.sns.it/bscw/bscw.cgi/0/50493>

¹³ <http://dlib.sns.it/bscw/bscw.cgi/0/51103>

¹⁴ homer.multimatch.hostedbyfdi.net/

¹⁵ homer.multimatch.hostedbyfdi.net/

¹⁶ homer.multimatch.hostedbyfdi.net/

¹⁷ homer.multimatch.hostedbyfdi.net/

3.4 Alinari content for the first prototype

3.4.1. Selection of the Content

The images have been selected by Alinari professionals and image curators. The selection of images initially addresses the MultiMatch user group needs (education and cultural tourism). The Alinari team selected images from the main historical periods and artistic styles; this selection was also influenced by the languages available. Alinari provides MultiMatch with a collection of images with metadata in the following four languages: English, French, Italian and Spanish.

The most important art domains are covered: architecture, painting, sculpture, plus also ceramics and mosaics. Themes covered range from history to fine arts, landscapes and events. The images represent European culture taking into account various aspects in order to satisfy the multicultural perspective required by MultiMatch.

The group of experts has included a sampling of a series of important events (historical, territorial and cultural) that go transversally through the European history. The great wars, natural catastrophes, literary awards, festivals and episodes in the world of entertainment, milestones in history and in the history of science.

A significant effort was made to represent local history and culture in the different countries with special attention to those of particular interest to the project by including important artists from the past: musicians, poets, artists (Van Gogh, Giotto, Michelangelo, Velazquez, Gauguin, Picasso, Matisse, Durer, Vermeer, Goya, Granach, Gaudì, Rembrandt), the most important European royal dynasties (i.e. the Hapsburgs, the Savoy, English, Dutch and Spanish royal families, etc. up to east European countries). Thus the collection includes faces, deeds, depictions, portraits and self-portraits of personalities in the fields of culture, history, religion, mythology, science, literature, athletics, and entertainment.

3.4.2 Metadata schema and controlled vocabularies in use

Description of the **metadata schema** in use at Alinari

Name	Alinari internal schema based on the ICCD and University of Florence guidelines;
Type	<ol style="list-style-type: none">1. a proprietary metadata schema based on a national standard, namely: ICCD (Istituto Centrale del Catalogo)2. a proprietary metadata schema: Dublin Core with additional qualifiers3. an international standard, namely: Open Archive Initiative
Management	Alinari
Short description	<p>The http://business.alinari.it website presents an initial selection of around 200.000 pictures including historical nineteenth and twentieth century vintage prints in sepia and black and white, and colour photographs from 1920 to the present, chosen by teams of experts in the fields of traditional and electronic publishing, audiovisual, television and cinema communication, history of photography.</p> <p>The schema is only in use by Alinari.</p>
Number of elements	Internal schema: 30+ elements
URL(s) documentation	www.alinari.com (documentation is not open to the general public)
XML encoding available	Yes

Export formats offered to MultiMatch	Dublin Core Metadata Element Set (DCMES), see Annex 2 for an example record in XML.
--------------------------------------	---

Description of the **controlled vocabularies** in use at Alinari

Name	1- The thesaurus is built as a hierarchical tree with 61 primary classes and 8000 key words 2- Authority lists with structured categories and controlled vocabulary for person names, period names, geographic locations 3-MPEG visual descriptors (only for R&D) in CBR system
Type	Not applicable
Management	Alinari
Short description	GEOGRAPHICAL: functional in the search for sites. For Italy the thesaurus allows searching by region, province and municipality. For other countries, the search can be by nation. ICONOGRAPHIC: a dictionary of selected terms allows a search via subject, ordered in 61 iconographic classes from Agriculture to Zoology. PERIODS AND STYLES regards searches involving pictures of art and allows for a search by art style and/or historical period. TYPE OF WORK OF ART offers a subdivision of subjects of artistic and historical-artistic importance and therefore allows searches by type of object and/or architectural complex.
Number of elements	200,000
Available in language	200,000 in EN, IT; 50,000 in EN, IT, PL, GE, SP, FR
XML encoding available	Yes

3.4.3 Metadata for first prototype

For the first prototype the proprietary metadata schema of Alinari is converted/deflated into the Dublin Core Metadata Element Set as follows:

DCMES	Values from Alinari XML files
Identifier	<imagecode>
Title	<titolo>
Creator	<artista>
Date	<data_opera>
Description	<specifica_luogo_scatto>
Coverage	<luogo_scatto>
Subject	<periodo_stile>, <tipo_opera>, <eventi>, <personaggi> and <keywords>
Relation	<url_img_thumb> and <url_img_med>

3.5 Biblioteca Virtual Miguel de Cervantes content for the first prototype

This section describes the content for the first prototype, provided by Biblioteca Virtual Miguel de Cervantes. The www.cervantesvirtual.com site presents a subset of 9.000 works, including text, facsimile and journals, and 1.600 web pages, all extracted from BVMC catalogue.

3.5.1 Selection of Texts and Web pages

The whole set of 9.000 texts contains a wide variety of themes and authors, mainly in text or facsimile formats, but also journals. Themes covered are mostly literary works, which cover about 50% of content provided. There is also historical/geographical content, about 10% of the content. As well as texts about fine arts, about 5%, social sciences, about 5%, and other themes like philosophy, psychology, religion, mathematics, etc, which cover the rest of the content package provided.

The main language of all these texts (about 90%) is Spanish, but there are also some works with total or partial content in other languages, such as Catalan (5%), Portuguese (1%), Latin (1%), English (0,5%), French (0,5%) and others (German, Galician, Greek and Italian). Some of the most important authors included in this content package are Lope de Vega, Emilia Pardo Bazán, Mariano José de Larra, Miguel de Cervantes and Ricardo Guillén. The "identifier" of each work leads to the URL where the table of contents (i.e., the links to the downloadable content) of the creation can be found.

Furthermore, a set of 1.600 web pages has been provided in order to have heterogeneous material for web crawling. It contains the whole set of web pages with available metadata in the www.cervantesvirtual.com web site at the moment. Themes included in this set are, amongst others, contemporary Spanish poetry, Spanish cinema, contemporary Chilean poetry, literature for children and young people, Spanish theatre, photography, etc.

The main language for (almost) all web pages is Spanish (about 98%), but there are also some pages in Catalan or Portuguese. The "identifier" of each web page in the metadata content leads to the main page of the content that can be crawled to extract every piece of information needed.

3.5.2 Metadata schema and controlled vocabularies in use

Description of the **metadata schema** in use at the Biblioteca Virtual Miguel de Cervantes.

Name	MARC21
Acronym	MARC
Status / version	Binary
Type	International standard metadata format for bibliographic information
Management	Library of Congress
Short description	It provides the mechanism by which computers exchange, use, and interpret bibliographic information, and its data elements make up the foundation of most library catalogues today.
Number of elements	About 105 main fields (combinations and depth of sub-fields are unlimited)
URL(s) documentation	http://www.loc.gov/marc/
Export formats offered to MultiMatch	LOM, CIDOC-CRM, MARC-XML and Dublin Core, see an example in Annex 3.

The following **controlled vocabularies** are in use at the Biblioteca Virtual Miguel de Cervantes:

- the UDC classification schema in Spanish for subject indexing;
- a controlled vocabulary with subject headings in Spanish based on the one in use by the University Library of Alicante.

3.5.3 Metadata for first prototype

A preliminary analysis was performed in order to select the type of contents and metadata in the Biblioteca Virtual Miguel de Cervantes to be used for the MultiMatch testing purposes. As a consequence, the first prototype of the content package includes:

1) Dublin Core descriptions of 9000 texts in the Miguel de Cervantes digital library, including all texts, journals and facsimile hosted by the BVMC web (mostly Spanish and South American literature).

Each entry also includes the URL where the table of contents (i.e., the links to the downloadable content) of the creation can be found.

2) Dublin Core descriptions of 1600 web pages with available metadata in the www.cervantesvirtual.com web site. This subset will be later expanded within the MultiMatch project.

No fine distinction is made for this prototype between different types of content (i.e. biography, review, creation) or between the actual creator and secondary creators (such as editors or translators).

This information will be included in a second version.

The information is delivered as XML files. Software filters converting the original metadata (MARC) to other formats (LOM, CIDOC-CRM, MARC-XML) are also available.

Pending tasks (April 2007):

1. Metadata encoding will be revised in future package versions to conform to the UTF-8 standard.
2. Meta-information about the subject will be added using Universal Decimal Classification.
3. Meta-information about secondary creators and content type will be added.

For the first prototype the provided metadata are in a format that is a subset of Dublin Core. Each record or XML document contains the following metadata elements:

Metadata element	Definition for Text works	Definition for Web pages
tag value	"text"	"Collection"
dc:type	"Text Data"	"Text Data"
dc:language	main language (ISO639-2)	main language (ISO639-2)
dc:creator	creator of source work. Not the name of the creator of the web edition of the Table of Contents (= Digital representation of the original creation). Including birth and death dates. Attention: This metadata element can be repeated if there are different authors for different editions (in MARC differentiated as primary author and secondary author).	creator of page content
dc:title	title of source creation	page title
dc:publisher	"Biblioteca Virtual Miguel de Cervantes"	"Biblioteca Virtual Miguel de Cervantes"
dc:date	web page date	web page date
dc:description	additional description	additional description
dc:subject	This element is empty for the first	This element is empty for the first

	prototype. This element will contain the UDC codes for the second prototype. E.g. 800 (for literary works)	prototype. This element will contain the UDC codes for the second prototype. E.g. 800 (for literary works)
dc:identifier	URL of resource	URL of resource
dc:coverage	description of content with UDC category in Spanish. E.g. Literatura. This element is also empty for the first prototype.	description of content with UDC category in Spanish. E.g. Literatura. This element is also empty for the first prototype.

An example of the provided text records in RDF.

The creation "La vida es sueño" was written by Pedro Calderón de la Barca and first published in 1936. However, the catalogue description is initially made for the 1997 edition: the text adapted and annotated by Evangelina Rodríguez Cuadros. The source is modified further before being published on-line (for instance, typos are corrected, web format differs from source one), then leading to a new edition, the web edition.

The Cervantes metadata quotes previous creators, so both Calderón de la Barca (primary author in MARC) and E. Rodríguez (secondary author in MARC) but excludes our Institution as creator. The XML for this example follows (note that two creators are given).

```
<rdf:Description>
<dc:type>Text Data</dc:type>
<dc:language>spa</dc:language>
<dc:creator>Calderón de la Barca, Pedro 1600-1681</dc:creator>
<dc:title>La vida es sueño Pedro Calderón de la Barca ; edición de Evangelina Rodríguez Cuadros</dc:title>
-
<dc:description>Edición digital a partir de la edición de Evangelina Rodríguez Cuadros, Madrid, Espasa-
Calpe, 1997, 18ª ed. </dc:description>
<dc:subject>Teatro español -- Siglo 17º</dc:subject>
<dc:creator>Rodríguez Cuadros, Evangelina</dc:creator>
-
<dc:identifier>http://www.cervantesvirtual.com/servlet/SirveObras/02448397211915617422202/index.htm<
/dc:identifier>
</rdf:Description>
```

3.6 Netherlands Institute for Sound and Vision content for the first prototype

This section describes the content for the first prototype, provided by Netherlands Institute for Sound and Vision.

3.6.1 Selection of the Content

For the first prototype the following content is selected:

Video	<p>Sound and Vision has released two corpora of video material in January and February 2006. These are provided in MPEG-1 format distributed on hard-disks:</p> <p>899 video items of television programs from various non-fictional genres. Including documentaries, current affairs programmes, popular science etc. . This is part of collection Sound and Vision selected for the TRECVID 2007 (http://www-nlpir.nist.gov/projects/trecvid/) competition. The material mainly covers the 1999-2005 period, but also includes material from the mid-eighties. The XML provided follows the .Dublin Core standard.</p> <p>40 items linked to the top-40 list of artists MultiMatch has compiled. This extra selection was made to meet the first prototype requirements. The XML provided follows the .Dublin Core standard.</p>
Text	<p>The complete catalogue in XML format, following the iMMix schema, described below. This covers a few hundred thousand records describing the Sound and Vision collection; archive material from the year 1898.</p>

3.6.2 Metadata schema and controlled vocabularies in use

Description of the **metadata schema** in use at Sound and Vision:

Name	iMMix schema, based on International Federation of Library Associations and Institutions (IFLA)
Type	<p>Proprietary standard.</p> <p>References for the form and content of the relevant metadata are Dublin Core, SMPTE and P_META. The IFLA model FRBR has been the most important reference for modelling the metadata. XML (and AXF) are used for the exchange format.</p>
Management	local
Short description	<p>The metadata model defines the way the metadata should be structured. It is roughly divided in four stages: concept, actual realisation, physical embodiment, and carrier. Those four stages represent different layers in the model.</p> <ul style="list-style-type: none"> • Work: the name of the intellectual and artistic concept or idea, which is the foundation of one or more realisations. • Realisation: a realisation is an elaboration of a concept: a specific single or multiple productions. Every realisation has a clear structure and form of content compared to other realisations of the same work. In the case of several productions the realisation contains all data that are valid for the underlying expressions. • Series: a series is a group of expressions, usually decided by the makers or producers. A series has a beginning and an end. • Expression: an expression is a concrete result of a realisation or series. It is the actual production. • Selection: a selection is a part of an expression. For example a news item. • Publication: a publication of an expression; i.e. a television broadcast. • Position: the position of the publication on a carrier. • Carrier: description of a carrier.

	This schema is only in use by Sound and Vision.
Number of elements	85 (will be extended)
Extra information on application	In addition to the metadata model which defines the way the metadata are to be structured, two other means have been developed: Formats: these define in which way the fields and metadata are to be presented to the documenter. Intentions: these define which metadata should be available to fulfil the specific information needs of a specific target group. The idea is that a broadcast professional prefers objective annotations whereas, for a visitor of the future Beeld en Geluid Media experience, a more catchy description is needed. This principle also facilitates the desire to add domain specific and other additional data to a description.
URL(s) documentation	http://www.prestospace.org/project/deliverables/D15-1_Analysis_AV_documentation_models.pdf
URL guidelines for application	http://www.prestospace.org/project/deliverables/D15-1_Analysis_AV_documentation_models.pdf
XML encoding available	Yes, see Annex 4 for an example record of the XML that is provided.

The following **controlled vocabularies** are in use at Sound and Vision:

- the GTAA thesaurus in Dutch for subject indexing: each facet is connected to a different metadata element (e.g. the subject facet to the subject keyword element, the genre facet to the genre element and the person names facet to the person-as-subject element and to the person-as-speaker element);
- several proprietary authority lists in Dutch. E.g. an open list of geographic names in Dutch for the Location-element and a closed list of publication types.

Description of the **GTAA thesaurus** in use at Sound and Vision

Name	Common Thesaurus for Audiovisual Archives
Acronym	GTAA
Status / version	Not applicable
Type	Hierarchical thesaurus, proprietary
Management	Sound and Vision
Short description	The GTAA thesaurus is the controlled vocabulary used at The Netherlands Institute for Sound and Vision. GTAA stands for the Common Thesaurus for Audiovisual Archives; it is the result of the collaborative work of different institutions concerned with audiovisual documents indexing, including the Filmmuseum Amsterdam.
Number of elements	It contains 159.831 preferred terms, 1.900 non-preferred terms, and 88 categories.
Available in language	Dutch
XML encoding available	Yes, also in RDF
Extra information on application	The GTAA is a general thesaurus with multiple facets: subjects, genres, persons, makers, names and locations. Only the subject facet, which contains the keywords, is structured. The terms in the subject facet are related to others via the related term, broader term and narrower term relations. The types of information of interest for MultiMatch (keywords, persons, locations, names, makers and genre) are very closely related to the different facets of the GTAA.

Applied by the following organizations e.g.	Sound and Vision, Filmmuseum Amsterdam
URL(s) documentation	http://ems01.mpi.nl/CHOICE/ (the browser is also accessible online)
URL guidelines for application	http://www.cs.vu.nl/~guus/papers/Assem06b.pdf#search=%22GTAA%20malaise%22

3.6.3 Metadata for first prototype

For the first prototype the rich metadata format iMMix is converted, in other words deflated, to DCMES. In Annex 4 a sample record illustrates the format in which the metadata that accompany the selected digital objects are provided. One of the types of information that is thus lost for the first prototype, is the type of Creator and the type of Subject.

An example in XML:

```
<asset>
<assets_id>17911</assets_id>
<title>NOORDERLICHT (expr_id:22404)</title>
<creator>VPRO</creator>
<creator>Blok, Tessel</creator>
<creator>Etten, Hansje van</creator>
<creator>Hoff, Niels van &apos;t</creator>
<creator>Somer, Madeleine</creator>
<creator>Spiegel, Karin</creator>
<creator>Vink, Anja</creator>
<creator>Wassink, Jos</creator>
<subject>woonwijken</subject>
<subject>achterbuurten</subject>
<subject>scholen</subject>
<subject>opvoeding</subject>
<description>BG_34877-out.wmv</description>
<description>Teleblik</description>
<description>expressie_id:#22404#</description>
<description>Afwisselend aandacht voor de Marshall Middle School in Houston, Texas, VS en een
middelbare school in de Amsterdamse Bijlmermeer. Op beide scholen worden methoden toegepast waarin de
leerlingen worden geleerd gedisciplineerd en ordelijk te werken en sociaal met elkaar om te gaan. De
Marshall Middle School heeft de methode van prof. H. Jerome reiberg, pedagoog, inmiddels met succes
toegepast. De school in de Bijlmermeer staat nog in de beginfase en wordt begeleid door de Amerikaanse
socioloog Bowen Paulle. Aan de orde komen: de aanpak en zienswijze van Freiberg en Paulle; de problemen
die sommige groepen jongeren met zich meebrengen; bedreigingen door leerlingen richting de leraren; de
behoefte van leerlingen aan orde, discipline en structuur; de regels op de Marshall Middle School; het
strenge en effectieve strafsysteem als onderlegger van Freibergs methode.
INTERVIEWS hierover met:
- prof. H. Jerome Freiberg, pedagoog aan de Universiteit van Houston;
- Arend Pouwels, adj. dir. middelbare school Bijlmermeer;
- Bowen Paulle, socioloog aan de Universiteit van Amsterdam;
- leerlingen van de Marshall Middle School;
- Juan Gonzales, dir. Marshall Middle School.
SHOTS: achterstandswijk in Houston; ext. Marshall Middle School in Houston; int. school in de
Bijlmermeer met groepjes jongeren oa tijdens de lessen; int. Marshall Middle School in Houston oa tijdens
de lessen en bewaking op het schoolterrein.</description>
<description_abstract>Wetenschappelijk magazine met uiteenlopende onderwerpen.
```

Wekelijks programma met reportages over wetenschappelijke onderwerpen. In deze aflevering aandacht voor opvoedkundige methoden die er voor moeten zorgen dat de chaotische situaties op sommige scholen in mn achterstandswijken weer op orde worden gebracht.</description_abstract>

<publisher>VPRO</publisher>

<date_created>31-10-2005</date_created>

<date_issued>18-04-2002</date_issued>

<type>Video</type>

<language>NL</language>

<coverage_spatial>Amsterdam</coverage_spatial>

<coverage_spatial>Bijlmermeer</coverage_spatial>

<coverage_spatial>Houston</coverage_spatial>

<coverage_spatial>Nederland</coverage_spatial>

<coverage_spatial>VS</coverage_spatial>

<rights>Voor informatie over het gebruik van archiefmateriaal kunt u contact opnemen met de Klantenservice van het Nederlands Instituut voor Beeld en Geluid:

klantenservice@beeldengeluid.nl .</rights>

<rights>For information on the use of archive material, please contact the Customer Service department of the Netherlands Institute for Sound and Vision: klantenservice@beeldengeluid.nl .</rights>

<collection>Beeld en Geluid in Academia - ACADEMIA TOTAAL</collection>

<collection>Beeld en Geluid in Academia - Teleblik</collection>

<mediafile>

<filename>/mmc/17911/BG_34877-out.wmv</filename>

<server>windowsmedia</server>

<format>WindowsMedia</format>

<duration>0</duration>

<bitrate>0</bitrate>

<filesize>0</filesize>

<system>N/A</system>

<language>EN</language>

<protection type="combined">

<realm/>

<realm/>

</protection>

</mediafile>

</asset>

3.7 Web data

CH-related subset of Wikipedia

The University of Amsterdam has crawled 65,000 Wikipedia articles related to the cultural heritage domain. The material is made available in the web content format designed by The University of Amsterdam.

Wikipedia is a multilingual, Web-based, free content encyclopedia project. Wikipedia is written collaboratively by volunteers; its articles can be edited by anyone with access to the Web site. Wikipedia uses wiki software to allow multiple authors to edit the encyclopedia simultaneously and to archive their changes in a database. Relevant articles may be found by following hyperlinks from one article to another, in the same way as cross-references in traditional works.¹⁸ With its multi-million pages and extensive coverage, Wikipedia is a rich source of information for MultiMatch. The subset for the first prototype includes pages in: English, Italian, Dutch and Spanish.

The Wikipedia articles are considered as Web pages in the MultiMatch metadata model. The metadata structure accompanying each article contains the following bottom-up elements:

Title; Originating URL; Last modified date; Fetch date; Top-facet (either Museum-Institution or Museum-Artist); Subject (the Wikipedia classes specifying artist and museums categories); Description; Language and Other URLs that the page links to. These elements are mapped to the MultiMatch metadata schema (see Annex 8.2.)

Cultural heritage white list crawl

The white list crawl data covers 40,000 total pages. The selection was gathered from different starting points: see below. Web pages cover four languages (English, Italian, Dutch and Spanish) and are encoded in the format provided by the University of Amsterdam team.

Starting point: UK	
1.	http://en.wikipedia.org/wiki/Museums_in_England
2.	http://www.pp.ij4u.or.jp/~murai/frame.gb.text.html
3.	http://en.wikipedia.org/wiki/Museums_in_Scotland
4.	http://en.wikipedia.org/wiki/Museums_in_Wales
5.	http://en.wikipedia.org/wiki/Museums_in_Northern_Ireland
6.	http://en.wikipedia.org/wiki/Museums_in_the_Republic_of_Ireland
7.	http://www.artnut.com/europe.html#Anchor-35882
Starting point Spain	
1.	http://en.wikipedia.org/wiki/List_of_museums#Spain
2.	http://dmoz.org/World/Espa%c3%b1ol/Regional/Europa/Espa%c3%b1a/Artes/Museos_y_centros_culturales/
3.	http://www.pp.ij4u.or.jp/~murai/frame.es.text.html
4.	http://www.artnut.com/europe.html#Anchor-380
5.	http://vlmp.museophile.org/spain.html
Starting Point Italy	
1.	http://www.musei-it.net/vlmp/ (all regions)
2.	http://www.pp.ij4u.or.jp/~murai/frame.it.text.html
Starting point: Netherlands	
1.	http://dir.yahoo.com/Regional/Countries/Netherlands/Arts_and_Humanities/Museums__Galleries__and_Centers/
2.	http://dir.yahoo.com/Regional/Countries/Netherlands/Provinces/Noord_Brabant/Cities/Breda/Entertainment_and_Arts/
3.	http://dir.yahoo.com/Regional/Countries/Netherlands/Arts_and_Humanities/Humanities/History/Museums_and_Memorials/

¹⁸ <http://en.wikipedia.org/wiki/Wikipedia>

4.	http://dir.yahoo.com/Regional/Countries/Netherlands/Society_and_Culture/Museums_and_Exhibits/
5.	http://dir.yahoo.com/Society_and_Culture/Museums_and_Exhibits/By_Region/Countries/Netherlands/Complete_List
6.	http://www.artnut.com/europe.html
7.	http://mediatheek.thinkquest.nl/~klb040/kunst/pagina_bekijken.php?pagina_id=18
8.	http://www.museumserver.nl/nederland/alfabet/front_nl.htm

From the crawled web pages of museum websites and the four subsets of Wikipedia the following three types of metadata are automatically extracted:

- administrative metadata, like the mime type or fetch and modification dates (not for websites, only for web pages)
- semantic metadata and
- links found in the document (to other documents or to multimedia files).

Element	Description	Range/Optional
<typesource>	The identifier for the original dataset source used to generate that dump.	(wikipedia whitelisterawl focusedcrawl)
<language>	The language the document is written in ⁴ .	(en es it nl)
<targetdomain>	The target domain (see footnote 2).	(en es it nl)
<top-facet>	The top-level topic category	(Museum-Institution Museum-Artist)

² the meaning is slightly different depending on the source: for crawled documents, it matches the country associated with the white list containing the domain for the page. As for Wikipedia articles, it corresponds to the language of the Wikipedia dump.

⁴ This corresponds to the language of the Wikipedia dump or to the language as identified by Nutch (character n gram-based) language identifier plugin.

3.8 UvA Audio Corpus

A small audio corpus for first MM prototype, includes:

- about 10 Spoken Wikipedia Articles (English and Dutch)
- selections from 3 authors recorded from BVMC (Spanish)
- podcasts, both professional and consumer generated.

Total size is approximately 20 hours. The UvA Audio Corpus is located on the sftp server and is in the following languages: (from more to less) Dutch, English, Spanish and Italian.

The output of the speech recognizer (transcripts of speech files) will be provided in an XML format with some accompanying metadata (e.g. id, language, start time, end time). The speech transcript generation will be described in D4.1 Text/Image/Speech/Video Indexing components and documentation for the first prototype. This deliverable is due Month 12.

3.9 The European Library

The European Library is a service on the World Wide Web that offers access to the resources of the 47 national libraries of Europe. The resources, both digital and non-digital, include books, magazines, journals, audio recordings and other material. The European Library portal offers free searching, and delivers metadata records as well as digital objects: some free, others at cost. The global management of The European Library is based on a consortium of 23 subscribing national libraries, all in charge of maintaining and developing the portal services.

MultiMatch has access to a test collection of bibliographic data provided by The European Library in several European languages. The metadata format is Dublin Core: see Annex 11 for a sample record in XML.

3.10 The common metadata from the provided content

Analysing the mapping of the catalogue descriptions that are provided by Alinari, Cervantes and Sound and Vision as XML documents to the MultiMatch metadata schema, the conclusion is that the following metadata elements are common to all three organisations:

Creation.Title¹⁹

Creation.Related Actor

Creation.Related Actor-Type (= Creation)

Creation.Related Actor-Date (creation date)

Creation.Subject

Creation.Description²⁰

The above is the result of the first simple mapping of metadata elements. There are several extra elements that can be populated automatically. E.g. a default value for the Archiving location of each creation. Mapping will probably be more complex in the future.

However, the influence of the bottom-up approach (see section 2.4) cannot in itself define the MultiMatch metadata schema. The influences of the top-down approach (the user requirements, see section 2.3) and the external influences (interoperability with existing standards, see section 2.2) will also define how the MultiMatch metadata schema will evolve. One of the aims of MultiMatch is to provide a knowledge representation for Creators and Creations by which the various relationships that are implicit in the data can be made explicit. See section 5 for further details, especially section 5.5. The values of the content providers' metadata will directly populate the MultiMatch metadata schema.

¹⁹ However, many of the photographs that are described in the Alinari content do not have a title of their own. Actually, dc:<titolo> contains what can be called a description-title, describing what the photograph is depicting. For example: 'The small port of Portofino, with moored sailboats. In foreground are some row boats.' See also section 5.5.1.

²⁰ In different ways, all three sets of XML files contain values for the MultiMatch Creation.Description element. See also section 5.5.

4 Generic Metadata Schemas

In the following section the four schemas most relevant for interoperability (see D2.1, section 5.2.2 and 5.3.2) are described:

- Dublin Core: because it is in use throughout the entire cultural heritage domain.
- MPEG-7: because it can handle multimedia in a way appropriate for MultiMatch.
- FRBR: because it provides a data model with relationships and a hierarchy that are probably useful for MultiMatch.
- CIDOC CRM: because it provides a reference model for the cultural heritage domain.

4.1 Dublin Core (and Refinements)

There is a variety of standard metadata schemas that have been developed and applied to Cultural Heritage data. The main ones were reviewed in D2.1. Of these, Dublin Core²¹ (DC) is the most widely used standard, and whatever metadata schema is implemented the relevant parts must be possible to map it to the DC.

Dublin Core is used to represent information on specific cultural heritage items or objects. That is, in the vocabulary of FRBR: “Dublin Core deals with manifestations of a particular work or conceptual item”. The original focus of the standard was the representation of documents and this is its main area of applicability. Within MultiMatch there are thus a number of limitations that should be made explicit when mapping to Dublin Core:

- there is a need to describe more than just the resources itself; it is equally important to have a structured description of the content of the resource (i.e. what is depicted within an image).
- there are concepts which are not readily covered by Dublin Core such as Actors and Multimedia objects
- Dublin Core has a single (or limited for DCMI Metadata Terms) representation of relations between resources.
- Dublin Core is primarily concerned with the Creation of the resource whilst for Cultural Heritage Objects the modifications which occur to the resource during its existence are also of importance.

There are a number of standard refinements of DC available (e.g. DCMI Metadata Terms and VRA Core) and mapping to these extensions increases the information which can be expressed.

DCMI Metadata Terms²² including the refinements will form the basis on which the MultiMatch metadata schema is built.

4.2 MPEG-7

The area of metadata for low-level multimedia features is an evolving one, although the MPEG-7/21 descriptors seem to offer a powerful (and increasingly standard) way of representing and communicating such information. [Cox 2006] The MPEG-7 metadata standard provides a set of elements for describing audiovisual material, including its semantic content. It can be used to describe what (e.g. people, objects) exists in a multimedia content, their relationships (e.g. friends, family members), when they appear in the content and where they appear on the video image (or in the real world).

²¹ DC Metadata Element Set: <http://dublincore.org/documents/dces/>

²² <http://dublincore.org/documents/dcmi-terms/>

There are two main limitations in the adoption of MPEG-7 for interoperability in MultiMatch.

- Firstly it is designed to represent audio-visual material rather than Cultural Heritage objects, for example it is not designed to describe physical objects or their properties. In an attempt to overcome this issue it is interesting to note that there has been research into the harmonisation of MPEG-7 and CIDOC CRM [Hunter 2001-2002, Tsinaraki 2007].
- The second and probably more severe restriction is that there is currently little use of MPEG-7 as a standard in the Cultural Heritage domain, in part due to the first limitation.

Where there is a need for interoperability of audio-visual material MPEG-7 offers a very expressive standard as it has content descriptors which provide coverage from low-level to semantic content. As more cultural collections are digitised it is possible that this standard will become more important.

4.3 Functional Requirements for Bibliographic Records

The reference model FRBR is comprised of three groups of entities:

- Group 1 entities represent the products of intellectual or artistic endeavour and are the foundation of the FRBR model:
 - *Work* is a “distinct intellectual or artistic creation.”
 - *Expression* is “the specific intellectual or artistic form that a work takes each time it is realized.”
 - *Manifestation* is “the physical embodiment of an expression of a work. As an entity, manifestation represents all the physical objects that bear the same characteristics, in respect to both intellectual content and physical form.”
 - *Item* is “a single exemplar of a manifestation. The entity defined as item is a concrete entity.” [IFLA 2006]
- Group 2 entities are person and corporate body, responsible for the custodianship of Group 1’s intellectual or artistic endeavour.
- Group 3 entities are subjects of Group 1 or Group 2’s intellectual endeavour, and include concepts, objects, events, places.

One of the key features of FRBR is the distinction between Work, Expression, Manifestation and Item when representing an artistic endeavour. However the categorisations of a given endeavour into one of these entities is a somewhat subjective judgement, and the work on FRBR continues to attempt to define the boundaries between these types. Within MultiMatch the FRBR distinction is not related to the key objectives nor has it been seen to be an issue for users during the examination of their requirements. In addition the State of the Art in techniques for automatic information extraction and classification are not sufficient to disambiguate FRBR entities, especially as this can be problematic even for professional archivists.

Although within the project the MultiMatch metadata interaction with FRBR will not directly be considered, the metadata schema does not exclude this interaction. This would primarily involve the dissection of Creation into the four FRBR entities, determining which properties apply to which entities. For example, the Creation of the Mona Lisa can be divided into the Work (the Mona Lisa by Leonardo Da Vinci), Expression (the painting of the Mona Lisa), Manifestation (the painting of the Mona Lisa) and the actual Item (the painting of the Mona Lisa hanging in the Louvre). There is an ongoing activity on harmonisation between the FRBR and CIDOC models²³, which may ease the integration with FRBR in the future.

²³ The 8th FRBR - CIDOC CRM Harmonization meeting took place, with representatives of the [CIDOC CRM SIG](#) and the IFLA FRBR Review Group on October 25-27, 2006.

4.4 CIDOC CRM

The CIDOC Conceptual Reference Model²⁴ is a core ontology describing the semantics of schema and data structure elements used for museum object documentation. The central idea of the CIDOC CRM is that the notion of historical context can be abstracted as things, people and ideas meeting in space–time. [Doerr 2003] Another reason why CIDOC focuses on events is that it allows modelling of the complex creation of objects in cultural heritage, since there are many objects which are not the result of a single creation. For example casts of sculptures, such as *The Thinker* by Rodin, where the original sculpture was made by Rodin, but then cast by someone else years after the original sculpture had been made - and possibly even after Rodin's death.²⁵

The stated aim for CIDOC CRM is to aid in both the construction of metadata schema and the interpretation of schema for interoperability. Thus, for the MultiMatch project to look further into the CIDOC CRM reference model (CIDOC) was a logical next step after reporting in D2.1 the metadata standards in use in the cultural heritage domain, as it provides a reference model specifically for the cultural heritage domain. The outcome of this initial mapping, together with the experiences in other projects that worked with CIDOC, especially the eCHASE project [Sinclair 2005] and the Museum24 project [Szász, 2006] made it clear that:

- it was not easy/straightforward to derive a metadata schema from this reference model because the complexity of the scenarios that CIDOC attempts to represent make the model very detailed. Yet at the same time there are limitations (such as Collections have to be made up of Physical Objects) which impose constraints on using the model within MultiMatch.
- it is rather complex to translate the reference model, which captures the semantics of complex scenarios within the cultural heritage domain, into presentable metadata elements in such a way that the users can understand.
- whilst CIDOC CRM is widely promoted in the research community there is little evidence of it being used for interoperability in actual applications.

Thus CIDOC CRM did not (as was hoped) provide a clear representation of the CH domain on which to base the metadata schema. The group responsible for CIDOC CRM has also released a simplified implementation of the use of CIDOC as the CRM Core schema. The schema reduces CIDOC to four fundamental principles:

- participation in events
- part whole relation
- reference
- classification.

However whilst this does give an indication as to how CIDOC CRM can be applied, CRM Core in itself does not provide sufficient expressiveness to capture certain information which is expressible in Dublin Core (e.g. Format and Rights information). As the application of CIDOC CRM in a focused search engine is not in itself a topic of scientific inquiry within the MultiMatch project, it was decided to apply CIDOC to the extent needed for MultiMatch. This means that the MultiMatch metadata will be interpreted in terms of CIDOC with the metadata elements that were specified during the construction of the MultiMatch metadata schema being mapped to CIDOC CRM where applicable which will then offer a medium for semantic interoperability (see 6.2 The MultiMatch Ontology).

²⁴ http://cidoc.ics.forth.gr/official_release_cidoc.html

²⁵ http://cidoc.ics.forth.gr/crm_core/core_examples/balzac.html provides an illustration how CIDOC models this example. See at: http://cidoc.ics.forth.gr/working_editions_cidoc.html#crm_core for more examples of CRM Core

4.5 Generic metadata schema: final remarks

In terms of semantic interoperability the intention in MultiMatch is to allow for flexibility and coverage (providing information to the widest possible user base) and expressiveness (as informative as possible). Currently mapping to and from the Dublin Core will provide the widest coverage; however, this will mean that much of the richness of information which is within the MultiMatch metadata will be lost. It is thus necessary to provide the metadata in more expressive representations (MPEG-7, VRA, FRBR, CIDOC), allowing the receiver to utilise such information, obviously ensuring that the semantics behind the metadata schema are clearly stated and adhered to during the process of populating the metadata.

5 MultiMatch Metadata Schema

This section explains in detail how the MultiMatch data model plus metadata schema were conceived and describes the model (section 5.2) itself. Section 5.3 explains how the MultiMatch metadata schema is presented in a spreadsheet and in Annex 8. The schema includes the mapping to DCMES, DCMI Metadata Terms and CIDOC CRM and the three DC-formats of Alinari, Cervantes and Sound and Vision. The last three mappings are presented in section 5.5. Section 5.4 introduces several issues that need special attention in the application of the MultiMatch metadata schema.

5.1 Focus of knowledge representation in MultiMatch

The goal of T2.2 was to choose a standard or a combination of standards that could be applied to fulfil the needs for knowledge representation in MultiMatch satisfactory. However none of the partners is using a standard that can come up with all expectations. Considering all the influences as discussed in section 2 - 4, the MultiMatch metadata schema is finally built upon DCMI Metadata Terms. Where possible the metadata element names directly match the DCMI element names. The main difference is in the introduction of more entities than just Creation as well as in the qualification of DC Relation which is fully expected to be refined within the DC standard.

The present version of the MultiMatch metadata schema is the result of the metadata development described in this section. Overall the focus of the schema is on representing the kind of data of interest to the user (D1.3) in combination with the focus of the project on Creators and Creations. This implies that the schema concentrates on search and presentation elements without explicitly incorporating all user interface functionality that will be presented on-the-fly. For the sake of this project the metadata schema is moderately complete, as it was not our goal to build a metadata schema that would completely cover the cultural heritage domain. However, the data model and the metadata schema are both constructed in such a way, and with such an eye on interoperability, that they can easily be extended when necessary. For example, in view of the future users of MultiMatch, Exhibitions are included as an Event type. Other relevant Events that can be associated to a Creation are Creation (the creation process), Publication, Discovery, Editing, Translating and Restoration. The reference model CIDOC CRM provides classes and relation types for such events. When necessary the MultiMatch metadata schema can be extended for these purposes.

Finally, the MultiMatch schema is intended to be populated with values automatically where possible, otherwise the elements stay empty. Prototype 1 is testing the possibilities. D2.2.2 will present the definite MM schema.

5.2 MultiMatch Data Model

In the MultiMatch project the concepts present in the data includes direct representations of Cultural Heritage objects (images, texts, etc.) and data describing those objects. The MultiMatch metadata schema should be of a hierarchical structure. Only a structural organization of the description of the cultural heritage objects makes it possible for the sub-parts of the content to be annotated with much more detail. The hierarchical metadata model makes it possible for sub-parts to inherit information from other levels, so it is not necessary to repeat the same information for every object or (sub) item of that object.

Furthermore there is a need to describe more than just the resources themselves. Therefore there is a need for several entities (each with its own metadata format) and a data model to point out the relations (e.g. Is A, Is part of, Is depicted in, Is related to) between them. This goes further then the Dublin Core format for describing cultural heritage objects on their own.

In a typical scenario of the cultural heritage domain the following is described:

- Creators, in other words artists or authors: record their birth place and birth date, who they worked with, with which country they are mostly associated, who was their teacher, who were their pupils, etc. The Union List of Artist Names (ULAN) will form a solid basis for the Creator records in MultiMatch. ULAN will also provide a controlled vocabulary for Creator person names.
- Other Actors concerning Creations: Publishers and Contributors of all kinds, Discoverers etc. Note that all Actors can be persons or organisations.
- Creations: the actual physical works of art. Original works of art can be a photograph, a documentary film, a monument, a painting, a sculpture, a vase etcetera..
- Digital representations: multimedia objects must be digitally captured to be represented in MultiMatch. For example, a painting or a historic building is digitally photographed; a documentary is digitally represented by a selection of keyframes or by a low-resolution browsing copy (MPEG-1). An actual image can then be used for publication on a museum website.
- Collections of Creations or Digital representations, like Catalogues or Websites.

An example to illustrate the need for a data model is the fact that within MultiMatch photographs will be indexed that are a creation or an artwork in themselves, next to the many photographs (or still images or digital representations) of original physical works of art (e.g. paintings and historical buildings). So it should be possible to distinguish between descriptions of original creations and descriptions of their digital representations in whole or part.

The focus of MultiMatch on creations and creators, or on works of art and artists, obviously does not diminish the need for a data model, as the above example illustrates. There will still be complex relationships, e.g. a Creator can be depicted in a Creation of his own hand or in a Creation by another Creator. There can also be multiple representations of the same real life physical object. And sometimes there will be data within MultiMatch about the textual and image representations as well as about the original creation (in the case of digital AV documents). Within the MultiMatch data model the identified entities can be related as needed. And the sets of metadata that belong to the identified entities can also point to the digital essence (or content files), which will be stored separately.

Overview of entities in the MultiMatch data model

The MultiMatch data model that is proposed in this Deliverable, consists of six entities and in total 13 sub-entities.

The acknowledged entities and their basic relations are presented in Figure 2. The other relations between these entities that the MultiMatch metadata schema aims to present explicitly are listed in section 5.4.2. See also Annex 8.

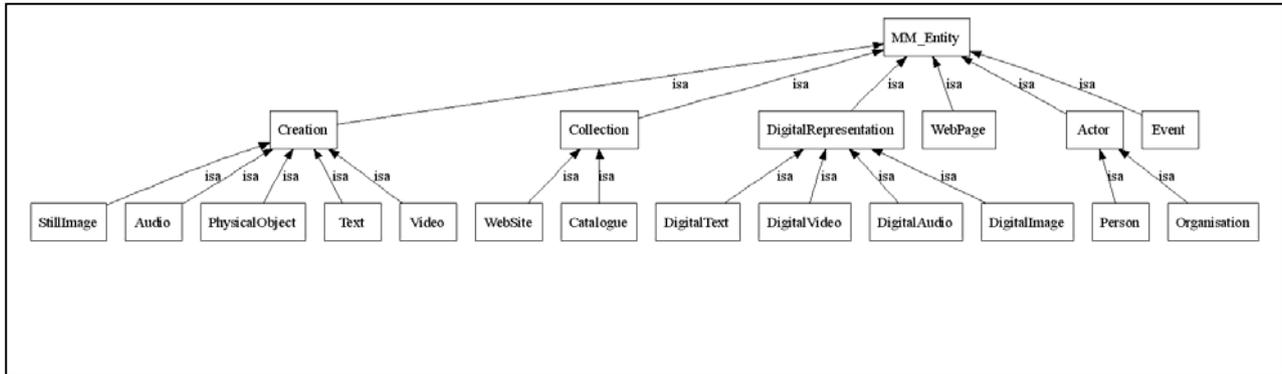


Figure 2: The basic graphical representation of the MultiMatch data model.

For the MultiMatch data model the following entities and sub-entities (See also section 1.3 Terminology) are identified:

Entity	Subentity	Definition
Collection		A Collection is an aggregate of Creations or Representations
	Catalogue	A Catalogue is a set of logically ordered descriptions of documents or other creations, including a reference to the physical location where the cultural heritage object is actually archived. In this context the archiving location is the location of the MultiMatch content provider. See also Annex 8.1.
	Website	A Website, in this context, is the website of a Cultural Heritage institute, for example a museum, or it is a website that presents Cultural Heritage content. A Website could also refer to a News service on cultural heritage. For the first prototype the Website records will concern 200 museum sites per language plus a website description per Wikipedia version. See also Annex 8.1.
Web page		The Web page format describes a page from a crawled or indexed Website. For example a Wikipedia article or a page of a museum web site. A Web page is part of a Website. See also Annex 8.2.
Actor	Person Organisation	The Actor format describes Persons or Organisations primarily responsible for creating the content of the resource (the Creator, Artist, Author) or responsible for making the resource available (Publisher) or responsible for making contributions to the content of the resource (Contributor) or otherwise related to the Creation (e.g. Discoverer, Conservator, Archivist etc.). See also Annex 8.3.
Creation	Text Still Image Video Audio Physical Object	In the context of MultiMatch, the Creation metadata format describes a unique man-made object which can be a Text, Still Image, Video, Audio or Physical Object. See also Annex 8.4.
Digital Representation	DigitalText DigitalImage DigitalVideo DigitalAudio	The Digital representation metadata format describes the visual surrogate or reproduction of a Creation. In many cases the digital representation will be an image. However, for Audio and Video Creations there will also be digital representations of several kinds. See also Annex 8.5.
Event		The Event format is a hook for several kinds of entities from the cultural heritage domain with strong geospatial and temporal aspects, like historical events (CIDOC). See also Annex 8.6.
	Exhibition	The only type of Event currently envisaged as a potential requirement in MultiMatch is an Exhibition of Creations. Exhibitions could be seen as a specific type of collection, but the entity has strong geospatial and temporal aspects, like Event.

5.3 Introduction to the MultiMatch metadata schema

The MultiMatch metadata schema provides metadata formats for the identified entities of the MultiMatch data model. Sometimes the format had to differ per subentity to be able to describe specific properties that do not apply for all subentities of an entity. The full presentation of each format can be found in Annex 8, annex 9: shows the MultiMatch Metadata xsd schema.

5.3.1 The formats of the MultiMatch metadata schema

The MultiMatch metadata schema contains metadata elements describing the following (sub)entities²⁶:

Entity of the MultiMatch data model	Own format	Subentity of the MultiMatch data model	Own format
Collection	No		
		Catalogue	Yes
		Website	Yes
Web page	Yes		
Actor	Yes		
		Person	Yes, has some extra metadata elements compared to Organisation. ²⁷
		Organisation	Yes, has less metadata elements than the Actor.Person format.
Creation	Yes		
		Text	Yes 1)
		Still Image	No
		Video	Yes
		Audio	Yes
		Physical Object	No
Digital Representation	Yes		
Event	Yes		
		Exhibition	No

²⁶ These metadata formats are worked out in detail in an Excel spreadsheet that is stored in the MultiMatch document store (<http://dlib.sns.it/bscw/bscw.cgi/0/54346>).

²⁷ For the identified types of Actors and Creations the formats are largely the same. The first column in the tables of Annex 8 shows the entity or subentities the property relates to.

5.3.2 Explanation of the metadata format tables

The tables that present the MultiMatch metadata schema in Annex 8 can contain the following columns:

Column name	This column presents
Entity	The name of the (sub)entity to which the metadata element relates.
Explanation	Short explanation of the semantics of the metadata element concerned. Indicating if the element contains free text or has a controlled vocabulary. When applicable, default values are listed in this column. As the first prototype will experiment with the populating of these elements, there are no indications which metadata elements are mandatory. The listed quotes are citations of D1.3.
MultiMatch metadata	The name of the MultiMatch metadata element. When possible the name of the Dublin Core standards is used.
D1.3 Section reference	Section and page number of D1.3, where the functional specification can be found that was the reason for the metadata element described to be part of the MultiMatch metadata schema. Often a copy of the text concerned is listed in the Explanation column.
Dublin Core	The results of the initial mapping of the MultiMatch metadata schema to the Dublin Core Metadata Element Set (DCMES).
DCMI Metadata Terms	The results of the initial mapping of the MultiMatch metadata schema to DCMI Metadata Terms.
CIDOC	The results of the initial mapping of the MultiMatch metadata schema to CIDOC-CRM.
Alinari	The results of the initial mapping of the MultiMatch metadata schema to the DC format in which catalogue descriptions of Alinari are provided.
Cervantes	The results of the initial mapping of the MultiMatch metadata schema to the DC format in which catalogue descriptions of Cervantes are provided.
Sound and Vision	The results of the initial mapping of the MultiMatch metadata schema to the DC format in which catalogue descriptions of Sound and Vision are provided.

5.4 Issues in Application of the MultiMatch Metadata Schema

5.4.1 Closed Lists / Controlled Vocabularies

In the metadata formats presented in the previous subsections, there are several references to controlled vocabularies or closed lists to be applied for specific metadata elements. The use of such vocabularies provides consistency and allows for interoperability, these issues will be discussed in the next chapter.

In this first version of the metadata schema there are a number of elements which do not have standard controlled vocabularies defined, however the values are expected to be selected from a closed list. Initially, for the first prototype, the majority of the non-free text metadata elements will be populated by the controlled vocabularies that are used by the content providers. These controlled vocabularies must be studied to decide how they can be used for semantic annotation and translanguag searching in the MultiMatch context. However the lists will be refined during the further analysis of the content for the classification and semantic annotation processes in order to ensure that they reflect the indexed material and provide the user with the necessary search functionality, including browsing and clustering functionality.

5.4.2 Relationships in the MultiMatch data model

One of the key features of the metadata is that it is not merely a flat data model but attempts to represent explicitly the important relationships in the Cultural Heritage domain. In the first prototype the focus will be on providing information on those relations which are key to fulfilling the main user requirements, namely

- Relation Web page – Creators. For example: Web page A describes Creator B;
- Relation Web page – Creation. For example: Web page A reviews Creation B;
- Relation Actor – Creations. For example: Actor A has created Creation B; Actor A has contributed to Creation C; Actor A has copied Creation D.
- Relation Actor – Actors. For example: Actor A is teacher of Actor B.
- Digital Representation – Creation. For example: Creation A is depicted in Digital Representation B.

The Entity-relationship model (ERD) model below provides the conceptual data model or semantic data model of the MultiMatch schema.²⁸

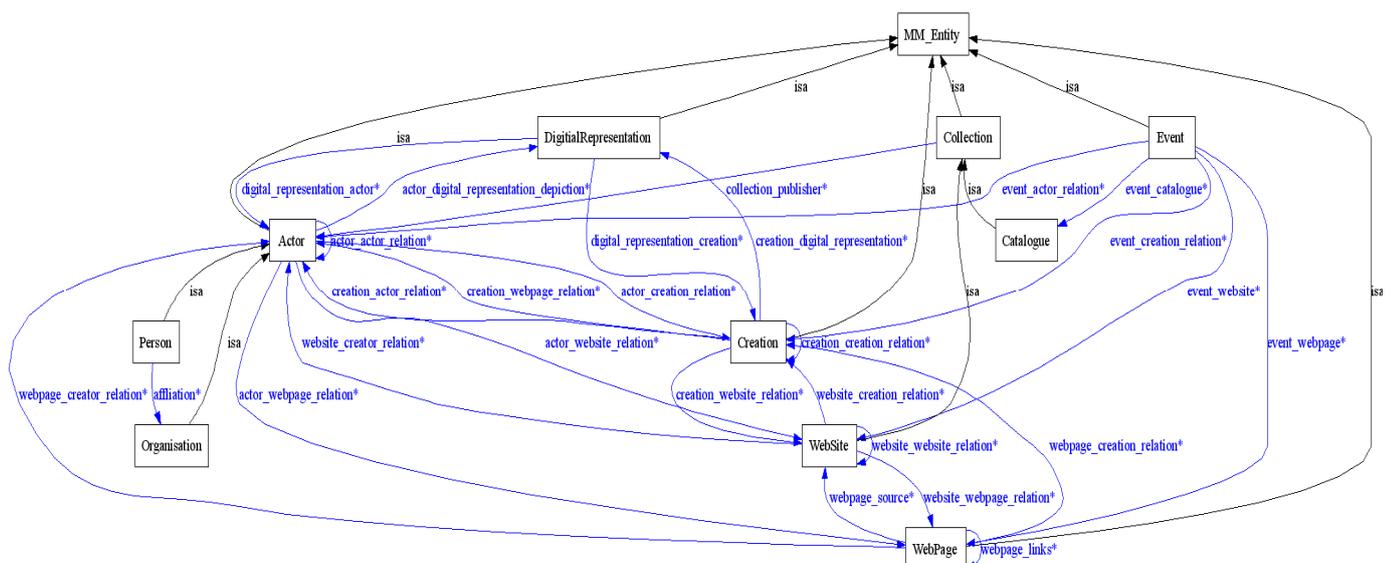


Figure 3. Entity-relationship model

²⁸ A high-resolution version is available at the MultiMatch document store.

It should be noted that these are potentially 1:n or n:n relationships and the relationships are defined by a closed list of relation types (see previous subsection). In addition, in the second prototype it may be decided to describe the relationships with further properties (which are currently not in the MultiMatch metadata schema) if it is deemed possible and useful to do so. Such properties include:

- A score or degree of association: the second prototype might want to distinguish weak from strong relationships. For example if a Video is about seven Creators, but for 70 % about two of those Creators, the question is can two different weights automatically be given to the relations of these two groups of Creators with that Video?
This score could also be applied for the relevancy of a relationship. D1.3, section 1.2, page 8-9, refers to Creators most closely related to the query. However this weight seems to be derived from the query (most closely related to the query), so this would mean that the ranking is calculated on the fly.
- A name for a subject of a relationship: to characterize the relationship. For example, "review" for a relation between Website A and Creation B, or "bibliography" for a relation between Web page A and Creator B. D1.3 suggests that keywords characterising a relationship are presented on the fly in the graphical depiction of the "author's network".
- A temporal coverage property: a relationship can be true during a specific period.
- Truth values: to indicate how confident (certain) the information extraction process is that the relationship actually exists. Note that there may be a trade-off between certainty, for example:

```
<Identifier>Picasso</Identifier>
  <Type>Actor.artist</Type>
  <Birth_Place>Europe</Birth_Place>
  <Birth_Place weight=0.9>Spain</Birth_Place>
  <Birth_Place weight=0.7>Malaga</Birth_Place>
  <Birth_Place weight=0.3>France</Birth_Place>
```

5.4.3 Alternative values and Multilingual related metadata issues

Given the multilingual focus of this project, the use of controlled vocabularies means that the translated of the vocabulary values provides a predefined, accepted translation. However the free-text linguistic elements require multiple language alternatives thus the values must be qualified in some linguistic namespace (i.e. RFC1766), e.g.

```
<Description RFC1766="en">the cat sat on the mat</Description>
<Description RFC1766="nl">de kat zat op de mat</Description>
```

Also some values such as Creation titles and names may have alternative values and alternative translations of same value, thus this information can be easily represented in the metadata.

```
<Title>
  <Preferred><en>Mona Lisa</en>
    <it>La Gioconda</it>
  </Preferred>
  <Alternative><it>Monna Lisa</it></Alternative>
  <Alternative><fr>La Joconde</fr></Alternative>
</Title>
```

5.4.4 Processing (free text) ambiguous date information

Ambiguous dates are common in Cultural Heritage data. Next to exact dates, there will be estimated (creation) dates both in a normalised numerical date format (ISO 8601) and in free text (for example 'late 19th century' or 'the sixties'). How to deal with these ambiguous dates is a representation and indexing research issue. The solution is likely to fall somewhere between the two extremes:

- collapse all dates into a single normalised numeric to ease indexing/retrieval
- express all the nuances of the ambiguous/certain/imprecise dates and work out some method for optimising indexing/retrieval.

Further experiments will provide a proper functionality for when the query for "authors born on 1960".

In the current version of the MultiMatch metadata schema the creation date is listed as one of the possible Creation.Related Actor –Dates. The initial definition of the MultiMatch Date format allows for the definition of approximate dates, e.g. circa 1942, circa 1985-12, 1920-29, 1910-1912, circa mid 1900-1999, as well as exact dates 0634-05-17 AD. Note that this still refers to a (possible uncertain) point in time and not a range.

This property is likely to refer to a set of metadata elements in order to be able to describe:

- start date
- end date
- whether these dates are estimated, and to what extent, or exact
- other relevant type of date if creation date is unknown
- additional annotation to explain the ambiguous date.

To illustrate the above, the cataloguing rules of Sound and Vision write the following on this issue:

"For example, if a programme is recorded somewhere in the month January of 1930,

then Starting date = 1930-01-01 (yyyy-mm-dd) and Precision of date= 'Exact day unknown'

For example, if a programme is recorded somewhere in 1930,

then Starting date = 1930-01-01 (yyyy-mm-dd) and Precision of date= 'Exact date unknown'

For example, if a programme is recorded somewhere in the thirties of the twentieth century,

then Starting date = 1930-01-01 (yyyy-mm-dd) and Precision of date= 'Exact date unknown' and Annotation on recording= 'the thirties'."

5.5 Mapping of first prototype metadata to the MultiMatch Metadata Schema

The Dublin Core metadata formats provided by the three Cultural Heritage institutions (see Annex 2, 3 and 4) are mapped to the MultiMatch Metadata Schema. In some cases there is not a direct 1-1 mapping between the semantics of the source metadata and the target MultiMatch schema. For example, much of the Alinari data is concerned with information about the photograph itself (being in this case a Digital Representation) rather than what is being depicted (the Creation), thus the title of the photograph, whilst encompassing the "Creation" being represented, also includes other information such as the title of the photograph or foreground and background details.

The initial mapping, for the first prototype, presented here will make certain assumptions about the source metadata semantics. All the values will be mapped as textual fields, with the assumption that the information required is contained, implicitly, within the text. Later work in the project will examine the mapping (annotation) of textual values onto predefined, semantically explicit, values.

5.5.1 Mapping of Alinari first prototype metadata to the MultiMatch Metadata Schema

Mapping of Alinari metadata to populate the MultiMatch Digital Representation format:

MultiMatch metadata element of the Digital Representation entity	Values mapped from Alinari XML files
Identifier - Source	<imagecode cod>
Representation source	default "www.alinari.com"
Type	This might be automatically extracted from <titolo>. "Detail of" = Part. If "Detail of" is not present in <titolo>, then = Whole.
Format	default ".jpg"
Rights Holder	<fotografo>
Copyrighted	default "Yes"

Mapping of Alinari metadata to populate the MultiMatch Creation format:

MultiMatch metadata element of the Creation entity	Values mapped from Alinari XML files
Title	<titolo> This can be ambiguous for Alinari; can be both the title of the work photographed and photograph
Source	default "Alinari"
Related Actor	<artista> = creator
Related Actor -Type	default type = "Creation" for the names from <artista>
Related Actor - Date	<data opera>
Subject	<periodo_stile> (controlled list for style period); <tipo_opera> (controlled list for type of work); <eventi> (controlled list for event); <personaggi> (controlled list for proper names); <keywords> (controlled list for other subject keywords).
Description	Actually the description of what the photograph is depicting is put in <titolo>. Alinari content has no title in the common sense (such as the artist gives to a work of art): Alinari has a description-title as example: 'The small port of Portofino, with moored sailboats. In foreground are some row boats.' Also <specifica_luogo_scatto> example: Palazzo Vitelli, Porta Sant'Egidio. This element gives the detailed place, not the town name.
Location	<luogo_scatto>
Archive Location	default "Florence, Italy"
Type	dc:type is always "Still Image". This can be specified with the values from <tipo_opera>
Link to Digital representations	<url_img_thumb> <url_img_med> OR the link to GIFT

5.5.2 Mapping of Cervantes first prototype metadata to the MultiMatch Metadata Schema

Mapping of Cervantes metadata to populate the MultiMatch Digital representation format:

MultiMatch metadata element of the Digital Representation entity	Values mapped from Cervantes XML files
Representation source	URL in dc:identifier
Type	default "Part", as it concerns Table of Content.
Format	default "html"
Rights Holder	default "Cervantes"

This information concerns the Table of Contents of the selected books.

Mapping of Cervantes metadata with tag value "text" to populate the MultiMatch Creation format:

MultiMatch metadata element of the Creation entity	Values mapped from Cervantes XML files
Title	dc:title
Source	default "Cervantes"
Related Actor	Metadata concerning books dc:creator = creator (may be anonymous) (may contain birth-death date: if so, this might be added to the Creator record concerned.)
Related Actor - Type	default type = "Creation"
Related Actor - Date	The creation date, publication date might be derived from dc:description
Subject	dc:subject (UDC Classification ²⁹) not currently in the data.
Description	dc:description
Type	dc:type default value = "Text"
Language	dc:language

Mapping of Cervantes data with tag value "Collection" to populate the MultiMatch Web page format:

MultiMatch metadata element of the Web page entity	Values mapped from Cervantes XML files
Title	dc:title
Identifier	dc:identifier is URL to the content.
Date - modified	dc:date
Source	"www.cervantesvirtual.com"
Subject	dc:subject is UDC code (second prototype)
Description	dc:description
Language	dc:language

²⁹ For the first prototype the following codes of the Universal Decimal Classification (UDC) are in use:

- UDC code starting with 8: mostly literary works);
- UDC code starting with 9: historical/geographical content;
- UDC code starting with 7: fine arts;
- UDC code starting with 3: social sciences;
- several codes for the other themes like philosophy, psychology, religion, mathematics, etc.

5.5.3 Mapping of Sound and Vision first prototype metadata to the MultiMatch Creation format

Mapping of Sound and Vision metadata to populate the MultiMatch Creation format:

MultiMatch metadata element of the Creation entity	Values mapped from Sound and Vision XML files
Identifier - Source	<description> BG_[0-9]+.mpg </description> No iMMix-ID in first prototype.
Title	<title> and <title_alternative>
Source	default "Sound and Vision"
Related Actor	<creator> (be sure to distinguish between persons and organisations for prototype TWO metadata! Creator field is ignored in prototype due to this)
Related Actor - Date	<date_issued> = creation date
Subject	<subject> taken from the GTAA thesaurus. This includes all facets:: Person name; Proper name; Location (as subject) and Genre. In the 1st prototype all these terms are entered in <subject> not possible to separate Locations from Person Names.
Description	<description> <description_abstract>
Location	<coverage_spatial>
Archive Location	default "Hilversum, The Netherlands"
Format	default "MPEG-1"
Type	<type> This is always "Video"
Related Creations	Creation names might be extracted from Summary and Description elements of the iMMix descriptions. See Annex 10. For first prototype this information is in <description>.
Related Creations -Type	See above. Plus the MM ontology could provide this value.
Link to Digital representations	Can be extracted from the URL in File Type.
Copyrighted	default 'No'
Language	<language>

6 The MultiMatch ontology and interoperability

In this section the MultiMatch metadata, presented in the previous sections, is considered in terms of its semantic interoperability in the CH domain. This involves examining the initial work on developing a well-formed ontological representation; how the metadata can be represented in a semantic web encoding and the standardisation of concept values.

6.1 Interoperability in MultiMatch

As was discussed in the introduction to this document, interoperability can take place on a number of levels and for systems such as MultiMatch that wish to provide a useful service in a given domain it is necessary to adequately address the requirements of interoperability. The architectural design of MultiMatch aims at providing components within a service-based architecture with standard specification of interfaces to facilitate reusability and interaction with other systems (see WP3 Deliverables). The previous sections of this deliverable have presented the basis for the information MultiMatch will attempt to capture (extract) from the Cultural Heritage data and present both to the user and also publish to other systems. Cultural heritage has a long history of using thesauri and subject reference systems to provide access to objects in collections. One of the key challenges is currently how to provide access (interoperate) to/from multiple collections; this question forms the focus of the Tirrenia workshop organised by the MultiMatch metadata working group.

With respect to such interoperability there are a number of levels at which different systems can interact; one way of viewing these levels is presented in Figure 4 [Gradmann 2007]. Interoperability in MultiMatch concerns the Syntactic, Pragmatic and Semantic levels shown in the figure. In terms of Syntax, there will be mappings from the Content Provider's metadata and from/to standard schemas such as Dublin Core (and some extensions to DC such as the DCMI Metadata Terms and possibly VRA).

Functional/Pragmatic interoperability is primarily the concern of the architecture discussed in the Deliverables produced by Work Package 3 (see D 3.1 and 3.2). Semantic interoperability also requires syntactic interoperability, thus the MultiMatch Schema must be represented in semantic encoding; i.e. XML³⁰, RDF(S)³¹, OWL³². In addition these encodings must be linked into relevant, available semantic representations, such as the RDF(S) schema provided by the Dublin Core Metadata Initiative³³ and reference models more specific to the CH domain such as CIDOC CRM³⁴. In addition to having clearly defined semantics of metadata elements the values associated with those elements will be populated, where possible, from standard controlled vocabularies.

³⁰ <http://www.w3.org/TR/2006/REC-xml-20060816/>

³¹ <http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/>

³² http://en.wikipedia.org/wiki/Web_Ontology_Language

³³ Dublin Core Metadata Element Set, Version 1.1 (<http://purl.org/dc/elements/1.1/>), DCMI elements and DCMI qualifiers (<http://purl.org/dc/terms/>) and DCMI controlled vocabulary (<http://purl.org/dc/dcmitype/>).

³⁴ CIDOC CRM v4.2 Encoded in RDFS (http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs).

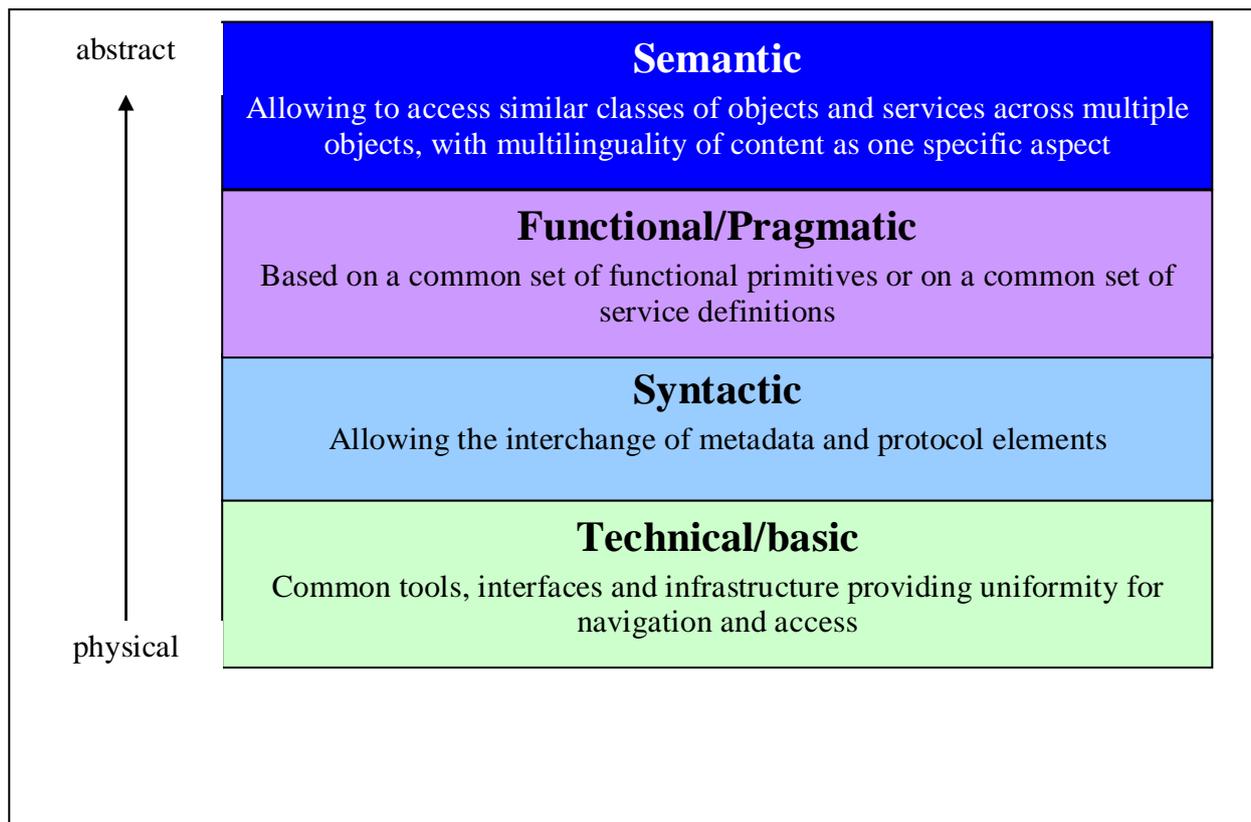


Figure 4: Interoperability Abstraction Levels (see presentation by Stefan Gradmann at Tirrenia Workshop)

6.2 The MultiMatch Ontology

In order for other systems to be able to use the (meta)data published by MultiMatch it is not only necessary to publish that data but provide it in a standard machine-readable form (syntax) and provide semantic markup so that the meta(data) can be correctly interpreted. The techniques developed to enable the Semantic Web are being employed for this purpose. This involves publishing the metadata described above as an ontology, in an XML – RDF – RDFS – OWL form. Annex 5: Draft version of MultiMatch OWL Ontology shows the first draft of the OWL ontology of the MultiMatch metadata.

The OWL description gives the basic class structure of the metadata entities (shown above in Figure 2) and the properties associated with those classes, described in the tables above. The previous chapter emphasised that one of the key features of the metadata schema is that there are a number of inter-relationships between the elements. It is not simply a flat, record-like data model; the metadata relationship attempts to capture some of the fundamentally important structure of the Cultural Heritage domain. The OWL model, shown below in Figure 5, gives a representation of the current class structure. A graphical representation which includes the structure of property values is too complicated to display meaningfully. In addition connections between the MultiMatch schema and other schema are made explicit, i.e. the connections to the Dublin Core and CIDOC CRM classes.

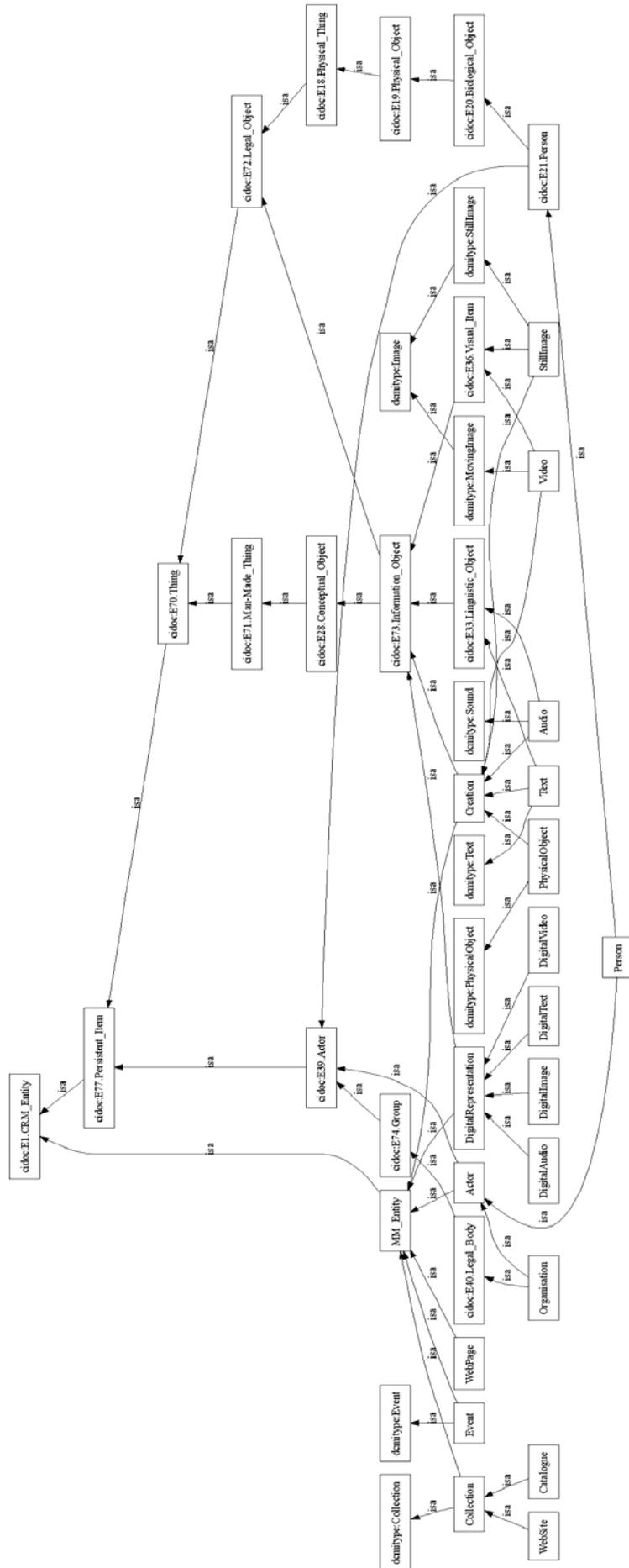


Figure 5: The MultiMatch Ontology
D2.2.1 Semantic Web Encoding

Examples of the relationships can be more clearly seen in the graphical representations in Annex 6: Relationship between MultiMatch Creation and Dublin Core / CIDOC CRM, and Annex 7: Relationship between MultiMatch Actor and CIDOC CRM Actor.

6.3 Standardised values and Controlled Vocabularies

In addition to the ontology providing a clear description of the schema semantics it also must define, as far as possible, the semantics of the values used to populate that schema. This is done by using relevant standards and controlled vocabularies or thesauri whenever they are available.

From the survey conducted for D2.1 it emerged clearly that the uptake of international established controlled vocabularies is quite limited. Local and nationally established/managed vocabularies are therefore predominant. Part of the reason for this is that the existing international controlled vocabularies are still not available in every European language. A clear example of this is shown by the Content Providers in MultiMatch who use a variety of (often proprietary), mainly monolingual, controlled vocabularies.

MultiMatch will require the use of controlled vocabularies to enable:

1. input control and search assistance via closed lists of preferred terms;
2. the browse functionality in the MultiMatch user interface. Dalmau affirms that structured forms of browse and search can be successfully integrated into digital collections to significantly improve the user's discovery experience. [Dalmau 2005]
3. supporting multilingual searching via multilingual term lists or thesauri;
4. the reinforcement of the semantic background information via the associations and background information the controlled vocabularies can provide will be an important part of the knowledge structures that MultiMatch intends to build/construct (see also: section 4.3). Classifications and thesauri can be seen as ontologies with a limited number of relationships between concepts.

From an analysis of the various controlled vocabularies (given their availability, support, coverage etc.) it was decided to adopt the Getty Thesauri³⁵ as the main initial controlled vocabularies. The three Getty Thesauri are:

- Getty Arts and Architecture Thesaurus (AAT) => artist descriptions
- Getty Unified List of Artist Names (ULAN) => creators names and information
- Getty Thesaurus of Geographic Names (TGN) => geospatial information

These thesauri are widely used and developed and supported by a large organisation and are published in a variety of formats, including XML which eases integration. Also the MultimediaN consortium has kindly made its RDF/OWL version of the Getty thesauri available for application in the MultiMatch project. Although there is a cost associated with acquiring the full thesauri it is possible to access the value definitions from the Getty Website³⁵, examines of this, for each thesaurus are shown in the sections below.

There are however a number of issues which will need to be addressed with the use of these (and other) vocabularies. For example, none of the Getty vocabularies cover all the four MultiMatch languages, therefore a translation exercise will be required on the terms used. For example the TGN generally provides place names in the local language (of the place) and English, whilst it would be preferable to have representations in all four of the MultiMatch languages. It may also be necessary to provide extensions to the thesauri to cover values which it does not contain, for example an artist not recorded in ULAN. Where such terms or entities are discovered (by the information extraction process) which are not in the thesauri, a standardised way of representation these novel values will have to be used so that the vocabularies are naturally extended in a way which facilitates interpretation by the user.

³⁵ http://www.getty.edu/research/conducting_research/vocabularies/

6.4 AAT (Arts and Architecture Thesaurus)

The primary motivation for choosing AAT is that it has a very broad coverage of terms to describe artistic works. AAT is a structured vocabulary currently containing around 131,000 terms and other information about concepts. Terms in AAT may be used to describe art, architecture, decorative arts, material culture, and archival materials. Terms for any concept may include the plural form of the term, singular form, natural order, inverted order, spelling variants, various forms of speech, and synonyms that have various etymological roots. Among these terms, one is flagged as the preferred term, or *descriptor*

The scheme proceeds from 34,000 abstract concepts to around 131,000 concrete physical artifacts (things).

Sample Record

The screenshot shows the 'Full Record Display' for the term 'graffiti' in the Art & Architecture Thesaurus Online. The interface includes a search bar, a 'Help' button, and a 'Click the icon to view the hierarchy' instruction. The record ID is 300015613 and the record type is 'concept'. The main text defines 'graffiti' as 'casual scribbles or pictographs on walls, stones, or other surfaces' and provides a note about its use in ancient Greek vase painting. The 'Terms' section lists 'graffiti' as the preferred form and 'graffito' as an alternative. The 'Facet/Hierarchy Code' is VVC. The 'Hierarchical Position' is shown as a tree structure: Objects Facet > Visual and Verbal Communication > Visual Works > <visual works> > <visual works by location or context> > graffiti. The 'Additional Parents' section lists 'Objects Facet', 'Visual and Verbal Communication', 'Information Forms', '<information forms>', '<document genres>', '<document genres by form>', and 'inscriptions'. The 'Sources and Contributors' section lists several sources, including 'Avery Index (1963-)', 'CONARC Subjects: I (CSH (1988-))', 'Clark, Elston and Hart, Understanding Greek Vases (2002) 96', 'Mayer, Dictionary of Art Terms (1969)', 'RIBA, Architectural Keywords (1982)', and 'RILA, Subject headings, unpub. (1975-1990)'. The 'Subject' section lists 'Vandalism' from 'RIBA, Architectural Keywords (1982)'. The 'Notes' section lists 'Clark, Elston and Hart, Understanding Greek Vases (2002) 96' and 'Mayer, Dictionary of Art Terms (1969)'.

Figure 6: Sample record from AAT.

AAT Facets

The following facets are available for facet browsing:

- **Associated Concepts:** This facet contains abstract concepts and phenomena that relate to the study and execution of a wide range of human thought and activity, including architecture and art in all media, as well as related disciplines. Also covered here are theoretical and critical concerns, ideologies, attitudes, and social or cultural movements (e.g., *beauty*, *balance*, *connoisseurship*, *metaphor*, *freedom*, *socialism*).
- **Physical Attributes:** This facet concerns the perceptible or measurable characteristics of materials and artifacts as well as features of materials and artifacts that are not separable as components. Included are characteristics such as size and shape, chemical properties of materials, qualities of texture and hardness, and features such as surface ornament and colour (e.g., *strapwork*, *borders*, *round*, *waterlogged*, *brittleness*).
- **Styles and Periods:** This facet provides commonly accepted terms for stylistic groupings and distinct chronological periods that are relevant to art, architecture, and the decorative arts (e.g., *French*, *Louis XIV*, *Xia*, *Black-figure*, *Abstract Expressionist*).
- **Agents:** The Agents facet contains terms for designations of people, groups of people, and organizations identified by occupation or activity, by physical or mental characteristics, or by social role or condition (e.g., *printmakers*, *landscape architects*, *corporations*, *religious orders*).
- **Activities:** This facet encompasses areas of endeavour, physical and mental actions, discrete occurrences, systematic sequences of actions, methods employed toward a certain end, and processes occurring in materials or objects. Activities may range from branches of learning and professional fields to specific life events, from mentally executed tasks to processes performed on or with materials and objects, from single physical actions to complex games (e.g., *archaeology*, *engineering*, *analyzing*, *contests*, *exhibitions*, *running*, *drawing* (*image-making*), *corrosion*).

- **Materials:** The Materials facet deals with physical substances, whether naturally or synthetically derived. These range from specific materials to types of materials designed by their function, such as colorants, and from raw materials to those that have been formed or processed into products that are used in fabricating structures or objects (e.g., *iron, clay, adhesive, emulsifier, artificial ivory, millwork*).
- **Objects:** The Objects facet is the largest of all the AAT facets. It encompasses those discrete tangible or visible things that are inanimate and produced by human endeavour; that is, that are either fabricated or given form by human activity. These range, in physical form, from built works to images and written documents. They range in purpose from utilitarian to the aesthetic. Also included are landscape features that provide the context for the built environment (e.g., *paintings, amphorae, facades, cathedrals, Brewster chairs, gardens*).

6.5 ULAN (Getty Unified List of Artist Names)

The motivation behind the adoption of ULAN is that it contains well structured and comprehensive information about a large number of artists (the main group of Creators in terms of MultiMatch metadata, see Annex 1). Hopefully as well as directly providing a valuable source of information it will provide data with which to “bootstrap” the information extraction process with respect to creators.

ULAN is a structured vocabulary currently containing around 293,000 names and other information about artists. Names in ULAN may include given names, pseudonyms, variant spellings, names in multiple languages and names that have changed over time (e.g., married names). Among these names, one is flagged as the *preferred name*.

Although it is displayed as a list, ULAN is structured as a thesaurus, compliant with ISO and NISO standards for thesaurus construction; it contains hierarchical, equivalence, and associative relationships. The focus of each ULAN record is an artist. Currently there are around 120,000 artists in the ULAN. In the database, each artist record (also called a *subject* in this manual) is identified by a unique numeric ID. Linked to each artist record are names, related artists, sources for the data, and notes. With every related artist the relationship type is indicated with: associate of; partner of; parent of; child of; teacher was; student was. The temporal coverage of the ULAN ranges from Antiquity to the present and the scope is global.

The Display Biography typically contains the following information for a given artist: nationality, major roles, birth and death dates. Note that biographical information for the same artist is often expressed differently by the various contributing institutions.

Figure 7: Sample Record of ULAN

6.6 TGN (Thesaurus of Geographic Names)

TGN is a structured vocabulary currently containing around 1,106,000 names and other information about places. Names for a place may include names in the vernacular language, English, other languages, historical names, names and in natural order and inverted order. Among these names, one is flagged as the *preferred name*.

TGN is a thesaurus, compliant with ISO and NISO standards for thesaurus construction; it contains hierarchical, equivalence, and associative relationships. Note that TGN is not a GIS (Geographic Information System). While many records in TGN include coordinates³⁶, these coordinates are approximate and are intended for reference only.

The focus of each TGN record is a place. There are around 912,000 places in the TGN. In the database, each place record (also called a *subject*) is identified by a unique numeric ID. Linked to the record for the place are names, the place's *parent* or position in the hierarchy, other relationships, geographic coordinates, notes, sources for the data, and *place types*, which are terms describing the role of the place (e.g., *inhabited place* and *state capital*). The temporal coverage of the TGN ranges from prehistory to the present and the scope is global and the record types include:

- administrative places, defined by administrative boundaries and conditions, including inhabited places, nations, and empires.
- physical places, refers to physical features, defined by their physical characteristics on planet Earth, including mountains, rivers, and oceans.

³⁶ Geographic coordinates indicating the position of the place, expressed in degrees/minutes and decimal fractions of degrees. Latitude (Lat.) is the angular distance north or south of the equator, measured along a meridian. Longitude (Long.) is the angular distance east or west of the Prime Meridian at Greenwich, England. Bounding coordinates and elevation may also be included.

6.7 Interoperability: further work

One of the aims within this project is to enable semantic interoperability with the information in the MultiMatch system. This requires that the information is represented in some form of semantic encoding which allows for an ontological representation of the MultiMatch metadata. This chapter shows the initial work towards this end, with a representation of the metadata in an OWL model. As part of an ontology it is necessary to have standardisation in the way values, for the concepts in the ontology, are represented. This provides an understanding of the semantics behind the values which is essential both in terms of presenting meaningful information to the users and for interoperability. Key to this is the use of standard, widely available controlled vocabularies. However problems can arise if the vocabulary is not extensive enough to suit the requirements of the system to which it is applied.

Further work will examine the most appropriate vocabularies to use in MultiMatch and how these can be used for interoperability, particularly allowing mapping from other (proprietary) vocabularies. Also to consider how these resources can be extended to cover the specific requirements of MultiMatch, such as representation in the four languages, and the incorporation of novel values.

7. Concluding remarks

D2.2.1 is part of the work in task 2.2. The DoW indicates the objectives of this task as follows '...identify the dimensions of the content the cultural heritage institutes will be able to provide to test the prototypes. The result of this task will be a document defining the semantic web encoding (including common metadata schema and ontology mapping providing common understanding on the file formats to be made accessible by the CH institutes).'

This deliverable meets these objectives, taking as reference point the common interoperability schema and the metadata definitions and ontologies already in use at the cultural heritage institutions. The aim was not to introduce a new schema at the different archive sites, but rather work towards a common standard that could be mapped to. The result was the MultiMatch metadata schema described in detail in section 5 and Annex 8. This schema is provided in both OWL and an XSD format.

The following standards for descriptive metadata are intended to be used within MultiMatch:

- for internal use, within the Semantic Web of MultiMatch: the MultiMatch metadata schema. This is an extended version of DCMI Metadata Terms including DC refinements.
- for exchange, import and publication purposes: the Dublin Core standard with the 15 elements will be used. For this purpose the MultiMatch metadata schema is mapped to DCMES.
- for further interoperability within the Cultural Heritage domain: the MultiMatch metadata schema is mapped to the reference model CIDOC-CRM.

The MultiMatch metadata schema needs to be evaluated in the first prototype. In month 25 this schema will be updated in D2.2.2, according to experiences with populating the schema from the content provided for the first prototype, user evaluations and the level of interoperability achieved between the different collections indexed for the first prototype. The current metadata model is flexible and can be easily modified in light of the results.

References

- Baca 1998. Baca, M., ed., Introduction to Metadata, Los Angeles: Getty Research Institute, [Web site] 1998, [2006 April 4], Available: http://www.getty.edu/research/conducting_research/standards/intrometadata/
- Sinclair (2005) eCHASE: Sustainable Exploitation of Electronic Cultural Heritage (Poster). . Sinclair, P., Lewis, P., Martinez, K., Addis, M., Prideaux, D., Fina, D. and Da Bormida, G. In Proceedings of 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, IEE Savoy Place.
- Cox 2006. Descriptive Metadata for Television: An End-to-End Introduction. Mike Cox , Ellen Mulder,Linda Tadic. Focal Press 2006
- Dalmau 2005. Integrating thesaurus relationships into search and browse in an online photograph collection / Michelle Dalmau et al. - Library Hi Tech. - Vol.23 no.3, 2005; p. 425-452.
- Doerr 2003. Towards a Core Ontology for Information Integration. Martin Doerr, Jane Hunter and Carl Lagoze. Journal of Digital Information, Volume 4 Issue 1. 2003
- De Jong 2003. Metadata in the audiovisual production environment : an introduction / Annemieke de Jong. – Hilversum: Nederlands Instituut voor Beeld en Geluid, 2003
- Hunter. 2001. Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In Proc. of the International Semantic Web Working Symposium (SWWS), July 30 - August 1, 2001
- Hunter 2002. Combining the CIDOC CRM and MPEG-7 to Describe Multimedia in Museums, Museums on the Web 2002, Boston, April 2002.
- Gradmann 2007. Position paper on Interoperability and Multilingualism / presentation by Stefan Gradmann of the EC Working Group on Interoperability and Multilingualism DELOS-MultiMatch Workshop on Semantic-Driven Interoperability for Cultural Heritage Digital Objects, held at Tirrenia, Pisa, February 2007 (see http://www.delos.info/index.php?option=com_content&task=view&id=533&Itemid=278)
- IFLA 2006. Functional Requirements for Bibliographic Records - Final Report <http://www.ifla.org/VII/s13/frbr/frbr1.htm#3.2> Latest Revision: 12 October 2006
- Szász (2006) Cultural Heritage on the Semantic Web – the Museum24 project Barnabás Szász, Arto Saraniva, Katalin Bognár1, Michal Unzeitig, Mikael Karjalainen. Digital Semantic Content across Cultures conference Paris, the Louvre May 4-5, 2006
- Tsinarakis 2007. Ontology-Driven Interoperability for MPEG-7. DELOS-MultiMatch Workshop on Semantic-Driven Interoperability for Cultural Heritage Digital Objects, held at Tirrenia, Pisa, February 2007 (see http://www.delos.info/index.php?option=com_content&task=view&id=533&Itemid=278)

Acronyms

AAT	Art and Architecture Thesaurus
CIDOC CRM	CIDOC Conceptual Reference Model\
CP	Content Providers
DCMES	Dublin Core Metadata Element Set
DCMI	Dublin Core Metadata Initiative
FRBR	Functional Requirements for Bibliographic Records
LCSH	Library of Congress Subject Headings
IEEE-LOM	Standard for Learning Object Metadata
MARC	Machine Readable Cataloguing
MPEG-7	Multimedia Content Description Interface
MPEG-21	Moving Picture Experts Group, MPEG-21 standard
OWL	Web Ontology Language
RDF	Resource Description Framework
RDF-S	RDF Schema
TGN	Thesaurus of Geographic Names
UDC	Universal Decimal Classification code
ULAN	Union List of Artists Names
VRA Core	Visual Resources Association Core Categories
W3C	World Wide Web Consortium
XML	Extensible Markup Language

Annex 1: Artist list used for Content selection

Leonardo da Vinci	alinari
Peter Paul Rubens	alinari
Guido Reni	alinari
Marc Chagall	alinari
Pieter Brueghel (the Younger)	alinari
Giotto di Bondone	alinari
Paul Cézanne	alinari
Francisco Y Lucientes Goya	alinari
Raffaello Sanzio (Raphael)	alinari
Andrea Pisano	alinari
Pierre Auguste Renoir	alinari
Michelangelo Buonarroti	alinari
Pablo Picasso	alinari
Vincent van Gogh	alinari
Pierre Bonnard	alinari
Sandro Botticelli	alinari
Umberto Boccioni	alinari
George Grosz	alinari
Albrecht Durer	alinari
Andy Warhol	alinari
Piet Mondriaan	added by Sound and Vision
Francis Picabia	added by Sound and Vision
René Magritte	added by Sound and Vision
El Lissitzky	added by Sound and Vision
Rem Koolhaas	added by Sound and Vision
Gerrit Rietveld	added by Sound and Vision
Harry Mulisch	added by Sound and Vision
Pier Paolo Pasolini	added by Sound and Vision
Julio Cortázar	added by Sound and Vision
Fay Weldon	added by Sound and Vision
Miguel de Cervantes	Cervantes
Gabriel García Márquez	Cervantes
Vargas Llosa	Cervantes
Isabel Allende	Cervantes
Camilo José Cela	Cervantes
Arturo Pérez Reverte	Cervantes
Tirso de Molina	Cervantes
Lope de Vega	Cervantes
Federico García Lorca	Cervantes
Francisco Ayala	Cervantes
Camilo José Cela	Cervantes
Carmen Laforet	Cervantes

Emilia Pardo Bazán	Cervantes
Miguel de Cervantes	Cervantes
Dulce María Loynaz	Cervantes
Margo Glantz	Cervantes
Claude Monet	Additions from Artcyclopedia
Salvador Dali	Additions from Artcyclopedia
Henri Matisse	Additions from Artcyclopedia
Rembrandt	Additions from Artcyclopedia
Gustav Klimt	Additions from Artcyclopedia
Paul Gauguin	Additions from Artcyclopedia
Amedeo Modigliani	Additions from Artcyclopedia
Jackson Pollock	Additions from Artcyclopedia
Titian	Additions from Artcyclopedia
Edgar Degas	Additions from Artcyclopedia
Caravaggio	Additions from Artcyclopedia
Joan Miro	Additions from Artcyclopedia
Frida Kahlo	Additions from Artcyclopedia

Annex 2: Sample record Alinari

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE alinari SYSTEM "alinari.dtd">
<alinari>
<imagecode cod="ACA-F-005353-0000">
<titolo> Detail of the coffered ceiling of Palazzo Vitelli at Porta Sant'Egidio, Perugia </titolo>
<data> 1890 ca </data>
<luogo_scatto> Perugia</luogo_scatto>
<specifica_luogo_scatto> Palazzo Vitelli, Porta Sant'Egidio </specifica_luogo_scatto>
<localita_raffigurata> </localita_raffigurata>
<specifica_localita_raffigurata> </specifica_localita_raffigurata>
<fotografo> Alinari, Fratelli </fotografo>
<oggetto> Glass plate </oggetto>
<tecnica> N Silver salt gelatin </tecnica>
<referenza> Alinari Archives-Alinari Archive, Florence </referenza>
<data_opera> 1528-1556 ca. </data_opera>
<artista> Gherardi, Cristoforo </artista>
<periodo_stile> High Renaissance;;Renaissance;;Renaissance-Baroque styles and periods;;Europe;;First
and Second Millennium A.D. </periodo_stile>
<tipo_opera> Ceiling;;Movable Modern Work </tipo_opera>
<eventi> </eventi>
<personaggi> </personaggi>
<keywords> Wood;;Object;;Applied Arts and Crafts;;Ceiling;;Building Elements,Decoration and
Sections;;Architecture;;Painting;;Work;;Painting </keywords>
<url_img_thumb>http://project.alinari.it/img/thumb/ACA-F-005353-0000.jpg</url_img_thumb>
<url_img_med>http://project.alinari.it/img/med/ACA-F-005353-0000.jpg</url_img_med>
<url_img_high>http://project.alinari.it/img/high/ACA-F-005353-0000.jp2</url_img_high></imagecode>
...
</alinari>
```

Annex 3: Biblioteca Cervantes xsd schema

```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xmlns:x="http://www.w3.org/XML/1998/namespace"
  xmlns="http://purl.org/dc/elements/1.1/"
  targetNamespace="http://purl.org/dc/elements/1.1/"
  elementFormDefault="qualified"
  attributeFormDefault="unqualified">

  <xs:import namespace="http://www.w3.org/XML/1998/namespace"
    schemaLocation="http://www.w3.org/2001/03/xml.xsd">
  </xs:import>

  <xs:element name="title" type="elementType"/>
  <xs:element name="creator" type="elementType"/>
  <xs:element name="subject" type="elementType"/>
  <xs:element name="description" type="elementType"/>
  <xs:element name="publisher" type="elementType"/>
  <xs:element name="date" type="elementType"/>
  <xs:element name="type" type="elementType"/>
  <xs:element name="identifier" type="elementType"/>
  <xs:element name="language" type="elementType"/>
  <xs:element name="coverage" type="elementType"/>

  <xs:group name="elementsGroup">
  <xs:sequence>
    <xs:choice minOccurs="0" maxOccurs="unbounded">
      <xs:element ref="title"/>
      <xs:element ref="creator"/>
      <xs:element ref="subject"/>
      <xs:element ref="description"/>
      <xs:element ref="publisher"/>
      <xs:element ref="date"/>
      <xs:element ref="type"/>
      <xs:element ref="identifier"/>
      <xs:element ref="language"/>
      <xs:element ref="coverage"/>
    </xs:choice>
  </xs:sequence>
</xs:group>

  <xs:complexType name="elementType">
    <xs:simpleContent>
      <xs:extension base="xs:string">
        <xs:attribute ref="x:lang" use="optional"/>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>

</xs:schema>
```

Annex 4: Sample record Sound and Vision ('element' range)

```
</xsd:element>
- <!--
  Define the asset complex type
-->
= <xsd:element name="asset">
  = <xsd:complexType>
    = <xsd:sequence>
      - <!--
        Dublin Core fields
      -->
      <xsd:element name="title" type="xsd:string" />
      <xsd:element name="title_alternative" type="xsd:string"
        minOccurs="0" maxOccurs="unbounded" />
      <xsd:element name="creator" type="xsd:string"
        maxOccurs="unbounded" />
      <xsd:element name="subject" type="xsd:string" minOccurs="0"
        maxOccurs="unbounded" />
      <xsd:element name="description" type="xsd:string"
        maxOccurs="unbounded" />
      <xsd:element name="description_abstract" type="xsd:string"
        minOccurs="0" maxOccurs="unbounded" />
      <xsd:element name="publisher" type="xsd:string"
        minOccurs="0" maxOccurs="unbounded" />
      <xsd:element name="contributor" type="xsd:string"
        minOccurs="0" maxOccurs="unbounded" />
      <xsd:element name="screenshot" type="xsd:string"
        minOccurs="0" />
      <xsd:element name="date_created" type="xsd:date"
        minOccurs="0" />
      <xsd:element name="date_created_time" type="xsd:time"
        minOccurs="0" />
      <xsd:element name="date_available" type="xsd:date"
        minOccurs="0" />
      <xsd:element name="date_available_time" type="xsd:time"
        minOccurs="0" />
      <xsd:element name="date_valid" type="xsd:date"
        minOccurs="0" />
      <xsd:element name="date_valid_time" type="xsd:time"
        minOccurs="0" />
      <xsd:element name="date_issued" type="xsd:string"
        minOccurs="0" maxOccurs="unbounded" />
      <xsd:element name="date_issued_time" type="xsd:time"
        minOccurs="0" />
    = <xsd:element name="type">
      = <xsd:simpleType>
        = <xsd:restriction base="xsd:string">
          <xsd:pattern value="[vV]ideo|[aA]udio" />
        </xsd:restriction>
  </xsd:sequence>
    </xsd:complexType>
  </xsd:element>
</!--
-->
```

```

    </xsd:simpleType>
</xsd:element>
<xsd:element name="source" type="xsd:string" minOccurs="0"
  maxOccurs="unbounded" />
<xsd:element name="language" type="xsd:string"
  minOccurs="0" maxOccurs="unbounded" />
<xsd:element name="relation" type="xsd:string" minOccurs="0"
  maxOccurs="unbounded" />
<xsd:element name="coverage_spatial" type="xsd:string"
  minOccurs="0" maxOccurs="unbounded" />
<xsd:element name="coverage_temporal" type="xsd:string"
  minOccurs="0" maxOccurs="unbounded" />
<xsd:element name="rights" type="xsd:string" minOccurs="0"
  maxOccurs="unbounded" />
<xsd:element name="collection" type="xsd:string"
  minOccurs="0" maxOccurs="unbounded" />
= <xsd:element name="creator_username" type="xsd:string"
  minOccurs="0">
  = <xsd:annotation>
    <xsd:documentation>User who should be the creator
      of that asset. If not specified, the user that
      uploads XML is chosen.</xsd:documentation>
  </xsd:annotation>
</xsd:element>

```

Annex 5: Draft version of MultiMatch OWL Ontology

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:dcmitype="http://purl.org/dc/dcmitype/"
  xmlns="http://multimatch.org/metadata-1.0.owl#"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:protege="http://protege.stanford.edu/plugins/owl/protege#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:cidoc="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xml:base="http://multimatch.org/metadata-0.1.owl">
  <owl:Ontology rdf:about="">
    <owl:imports rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs"/>
    <owl:imports rdf:resource="http://purl.org/dc/elements/1.1"/>
    <owl:imports rdf:resource="http://purl.org/dc/terms"/>
    <owl:imports rdf:resource="http://purl.org/dc/dcmitype"/>
  </owl:Ontology>
  <rdfs:Class rdf:ID="WebSite">
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >Website of a CH institute or with CH content; could also be a News service.
  </rdfs:comment>
    <rdfs:subClassOf>
      <rdfs:Class rdf:ID="Collection"/>
    </rdfs:subClassOf>
  </rdfs:Class>
  <rdfs:Class rdf:ID="Organisation">
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      ></rdfs:comment>
    <rdfs:subClassOf>
      <rdfs:Class rdf:ID="Actor"/>
    </rdfs:subClassOf>
    <rdfs:subClassOf
      rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#E40.Legal_Body"/>
  </rdfs:Class>
  <rdfs:Class rdf:ID="DigitalText">
    <rdfs:subClassOf>
      <rdfs:Class rdf:ID="DigitalRepresentation"/>
    </rdfs:subClassOf>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      ></rdfs:comment>
  </rdfs:Class>
  <rdfs:Class rdf:ID="PhysicalObject">
    <rdfs:subClassOf>
      <rdfs:Class rdf:ID="Creation"/>
    </rdfs:subClassOf>
    <rdfs:subClassOf
      rdf:resource="http://purl.org/dc/dcmitype/PhysicalObject"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      ></rdfs:comment>
  </rdfs:Class>
  <rdfs:Class rdf:ID="Catalogue">
    <rdfs:subClassOf>
      <rdfs:Class rdf:about="#Collection"/>
    </rdfs:subClassOf>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >A Catalogue is a set of logically ordered descriptions of documents or
      other creations, including a reference to the physical location where the
      cultural heritage object is actually archived.
      In this context the archiving location is the location of the MultiMatch content
      provider.</rdfs:comment>
  </rdfs:Class>
</rdf:RDF>
```

```

</rdfs:Class>
<rdfs:Class rdf:ID="DigitalAudio">
  <rdfs:subClassOf>
    <rdfs:Class rdf:about="#DigitalRepresentation"/>
  </rdfs:subClassOf>
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  ></rdfs:comment>
</rdfs:Class>
<rdfs:Class rdf:ID="Person">
  <rdfs:subClassOf>
    <rdfs:Class rdf:about="#Actor"/>
  </rdfs:subClassOf>
  <rdfs:subClassOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#E21.Person"/>
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  ></rdfs:comment>
</rdfs:Class>
<rdfs:Class rdf:ID="WebPage">
  <rdfs:subClassOf>
    <rdfs:Class rdf:ID="MM_Entity"/>
  </rdfs:subClassOf>
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >Part of a crawled or indexed Website. E.g. a Wikipedia
article.</rdfs:comment>
</rdfs:Class>
<rdfs:Class rdf:ID="Video">
  <rdfs:subClassOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#E36.Visual_Item"/>
  <rdfs:subClassOf>
    <rdfs:Class rdf:about="#Creation"/>
  </rdfs:subClassOf>
  <rdfs:subClassOf rdf:resource="http://purl.org/dc/dcmitype/MovingImage"/>
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  ></rdfs:comment>
</rdfs:Class>
<rdfs:Class rdf:about="#Collection">
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >The term "collection" can be applied to any aggregation of physical or
digital items. Those items may be of any type, so examples might include
aggregations of natural objects, created objects, "born-digital" items, digital
surrogates of physical items, and the catalogues of such collections (as
aggregations of metadata records). The criteria for aggregation may vary: e.g.
by location, by type or form of the items, by provenance of the items, by source
or ownership, and so on. Collections may contain any number of items and may
have varying levels of permanence. A "collection-level description" provides a
description of the collection as a unit: the resource described by a collection-
level description is the collection, rather than the individual items within
that collection.
For MultiMatch we identify at least two types of Collection: Catalogue and
Website.</rdfs:comment>
  <rdfs:subClassOf>
    <rdfs:Class rdf:about="#MM_Entity"/>
  </rdfs:subClassOf>
  <rdfs:subClassOf rdf:resource="http://purl.org/dc/dcmitype/Collection"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Audio">
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  ></rdfs:comment>
  <rdfs:subClassOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#E33.Linguistic_Object"/>
  <rdfs:subClassOf>
    <rdfs:Class rdf:about="#Creation"/>

```

```

    </rdfs:subClassOf>
    <rdfs:subClassOf rdf:resource="http://purl.org/dc/dcmitype/Sound"/>
</rdfs:Class>
<rdfs:Class rdf:about="#MM_Entity">
    <rdfs:subClassOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#E1.CRM_Entity"/>
    </rdfs:Class>
    <rdfs:Class rdf:ID="StillImage">
        <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
        ></rdfs:comment>
        <rdfs:subClassOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#E36.Visual_Item"/>
        <rdfs:subClassOf>
            <rdfs:Class rdf:about="#Creation"/>
        </rdfs:subClassOf>
        <rdfs:subClassOf rdf:resource="http://purl.org/dc/dcmitype/StillImage"/>
    </rdfs:Class>
    <rdfs:Class rdf:ID="Text">
        <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
        ></rdfs:comment>
        <rdfs:subClassOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#E33.Linguistic_Object"/>
        <rdfs:subClassOf>
            <rdfs:Class rdf:about="#Creation"/>
        </rdfs:subClassOf>
        <rdfs:subClassOf rdf:resource="http://purl.org/dc/dcmitype/Text"/>
    </rdfs:Class>
    <rdfs:Class rdf:about="#Actor">
        <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
        >Person or Organisation primarily responsible for making the content of the
resource (the Creator, Artist, Author) or responsible for making the resource
available (Publisher) or responsible for making contributions to the content of
the resource (Contributor) or otherwise related to the Creation (e.g.
Discoverer, Conservator, Archiver etc.).</rdfs:comment>
        <rdfs:subClassOf rdf:resource="#MM_Entity"/>
        <rdfs:subClassOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#E39.Actor"/>
        </rdfs:Class>
        <rdfs:Class rdf:about="#DigitalRepresentation">
            <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
            >A digital representation is a visual surrogate or reproduction of a
Creation.</rdfs:comment>
            <rdfs:subClassOf rdf:resource="#MM_Entity"/>
            <rdfs:subClassOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#E73.Information_Object"/>
            </rdfs:Class>
            <rdfs:Class rdf:ID="DigitalVideo">
                <rdfs:subClassOf rdf:resource="#DigitalRepresentation"/>
                <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
                ></rdfs:comment>
            </rdfs:Class>
            <rdfs:Class rdf:ID="DigitalImage">
                <rdfs:subClassOf rdf:resource="#DigitalRepresentation"/>
                <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
                ></rdfs:comment>
            </rdfs:Class>
            <rdfs:Class rdf:ID="Event">
                <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
                >An Event. The only type of Event currently envisaged is an Exhibition of
Creations.</rdfs:comment>
                <rdfs:subClassOf rdf:resource="#MM_Entity"/>

```

```

    <rdfs:subClassOf rdf:resource="http://purl.org/dc/dcmitype/Event"/>
  </rdfs:Class>
  <rdfs:Class rdf:about="#Creation">
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >In the context of MultiMatch, the Creation metadata format describes the
    physical cultural heritage object, i.e. it is
    - a unique man-made object;
    - the result of a distinct intellectual or creative creation process which can
    be a Text, Still Image, Video, Audio or Physical Object;
    - an Endeavour in terms of the FRBR reference model, in that it represents a
    Work in its abstract term and also the actual Manifestation of that
    work.</rdfs:comment>
    <rdfs:subClassOf rdf:resource="#MM_Entity"/>
    <rdfs:subClassOf
    rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#E73.Information_Obj
    ect"/>
  </rdfs:Class>
  <owl:ObjectProperty rdf:ID="MM.actor_name">
    <rdfs:domain rdf:resource="#Actor"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Name of the Actor, either a person name (can be based on ULAN (the
    preferred name) or on the controlled vocabularies of the content providers) or
    an organisation name (in the case of an anonymous artist or when the person name
    is unknown) .

```

Names in ULAN may include given names, pseudonyms, variant spellings, names in multiple languages, and names that have changed over time (e.g., married names).</rdfs:comment>

```

    <rdfs:subPropertyOf
    rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P131F.is_identified
    _by"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="MM.title">
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Title of a Collection, Webpage or Creation.</rdfs:comment>
    <rdfs:subPropertyOf
    rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P102F.has_title"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="MM.description">
    <rdfs:domain rdf:resource="#MM_Entity"/>
    <rdfs:subPropertyOf
    rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P3F.has_note"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >A free text description of the entity.</rdfs:comment>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="MM.related_actor_date">
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >The date associated with the relation between the Actor and the
    Entity.</rdfs:comment>
    <rdfs:subPropertyOf>
      <owl:ObjectProperty rdf:ID="MM.related_actor"/>
    </rdfs:subPropertyOf>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="MM.date_of_birth">
    <rdfs:subPropertyOf
    rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P98B.was_born"/>
    <rdfs:domain rdf:resource="#Person"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Date of birth of the Actor.
    Via ULAN or indexed material.
    This date is part of the Profile presentation.</rdfs:comment>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="MM.affiliated_organisation">

```

```

    <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P107B.is_current_or
_former_member_of"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Name of the organisation that the Actor as a person is affiliated with.
Controlled vocabulary from ULAN ((Related) Corporate bodies) combined with those
from the content providers.
E.g. Magnum for a specific group of photographers or the name of a monastery for
an unknown monk who has written a specific manuscript.</rdfs:comment>
    <rdfs:domain rdf:resource="#Organisation"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.related_webpage">
    <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P67B.is_referred_to
_by"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Webpage that relates to this Entity.</rdfs:comment>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.date_captured">
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >The date when the information used to derive the entity was captured in the
MultiMatch system.</rdfs:comment>
    <rdfs:domain rdf:resource="#MM_Entity"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.actor_depiction">
    <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P62B.is_depicted_by
"/>
    <rdfs:domain rdf:resource="#Actor"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >This element points to content where the Actor can be seen or heard
(depicted).
In case of Video Creations:
Link to the set of representative keyframes where the Actor can be seen.
Link to the Video scenes (timecodes) where the Actor can be seen.
In case of Audio and Image Creations:
Link to the part of the Transcript of speech where the Actor is speaking.
Link to an image of a portrait of the Actor. The portrait is part of the Profile
presentation.</rdfs:comment>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.publisher">
    <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P14F.carried_out_by
"/>
    <rdfs:domain
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#E65.Creation"/>
    <rdfs:range rdf:resource="#Actor"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Actor responsible for publishing a Creation</rdfs:comment>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.actor_identifier">
    <rdfs:domain rdf:resource="#MM_Entity"/>
    <rdfs:subPropertyOf
rdf:resource="http://purl.org/dc/elements/1.1/identifier"/>
    <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P131F.is_identified
_by"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >A unique identifier for an Actor. Preferably from a controlled list of
names (ULAN), however if this does not include the necessary actor a MultiMatch
specific identifier must be used.</rdfs:comment>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.related_website">

```

```

    <rdfs:domain rdf:resource="#MM_Entity"/>
    <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P67B.is_referred_to
_by"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Website that relates to this Entity.</rdfs:comment>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.related_creation_date">
    <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P117F.occurs_during
"/>
    <rdfs:subPropertyOf>
        <owl:ObjectProperty rdf:ID="MM.related_creation"/>
    </rdfs:subPropertyOf>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Date associated with the related creation and this entity, e.g. Creation
Date, Publication Date.</rdfs:comment>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.place_of_death">
    <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P100B.died_in"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >If applicable, the name of the place of death of the Actor.
Can be derived from the Died element from the ULAN concerned or from the indexed
material. Combination with TGN.</rdfs:comment>
    <rdfs:domain rdf:resource="#Person"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.language">
    <rdfs:domain rdf:resource="#MM_Entity"/>
    <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P2F.has_type"/>
    <rdfs:range
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#E56.Language"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.place_of_birth">
    <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P98B.was_born"/>
    <rdfs:domain rdf:resource="#Person"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.related_actor_type">
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >The type of relation between the Actor and the Entity.</rdfs:comment>
    <rdfs:subPropertyOf>
        <owl:ObjectProperty rdf:about="#MM.related_actor"/>
    </rdfs:subPropertyOf>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="Tags">
    <rdfs:domain rdf:resource="#MM_Entity"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >The keywords or tags used to describe the entity.</rdfs:comment>
    <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P3F.has_note"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.actor_main_role">
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >A closed list to indicate the main roles of the Actor.
Minimum, relevant Actors that can be associated to the Dublin Core Roles, i.e.
Creator (primarily responsible for making the content of the resource);
Publisher (responsible for making the resource available); Contributor (a "catch
all" for other roles, that is responsible for making contributions to the
content of the resource).
These categories will be derived from a closed list that consists of at least
two levels. E.g. Creator.Painter.

```

The narrower type level can be derived from the controlled vocabulary form ULAN (Roles element) in combination with the controlled vocabularies on types of actors in the catalogues of the content providers. Also the MARC list can be applied: <http://www.loc.gov/marc/sourcecode/relator/relatorlist.html>

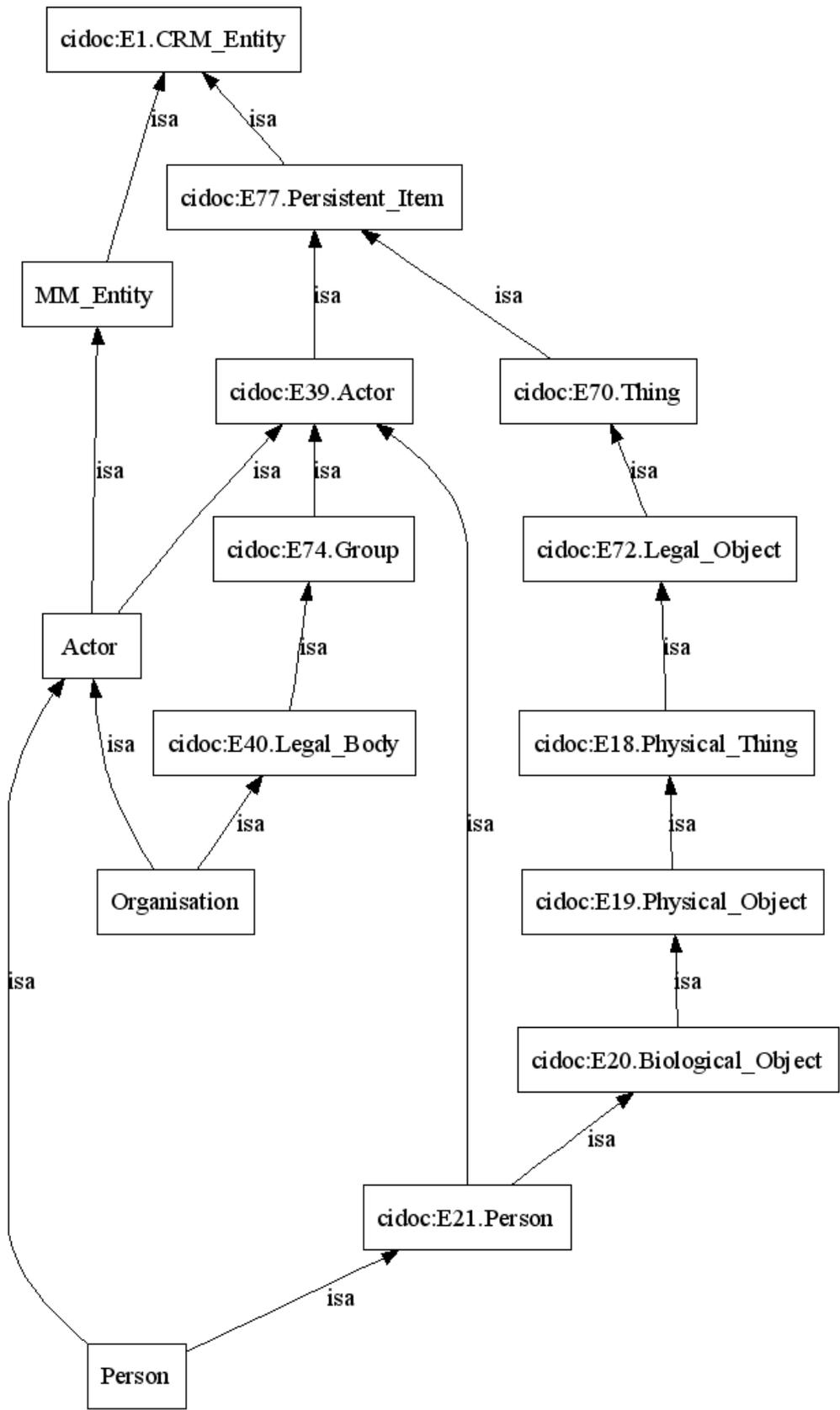
Roles are the major professional roles or activities performed by the artist throughout his or her lifetime (e.g., artist, architect, sculptor). For a corporate body, roles include the major activities or purpose of the firm, institution, or other corporate body (e.g., studio, manufactory, workshop).

```

This information is part of the Profile presentation (Category).</rdfs:comment>
  <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P2F.has_type"/>
  <rdfs:domain rdf:resource="#Actor"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.date_of_death">
  <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P100B.died_in"/>
  <rdfs:domain rdf:resource="#Person"/>
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
>If applicable, the date of death of the Actor.
Derived from ULAN or indexed material.
This date is part of the Profile presentation.</rdfs:comment>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="#MM.related_actor">
  <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P131F.is_identified
_by"/>
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
>An Actor related to an Entity.</rdfs:comment>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="MM.related_creation_type">
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
>The type of relation between the entity and Creation.</rdfs:comment>
  <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P2F.has_type"/>
  <rdfs:subPropertyOf>
    <owl:ObjectProperty rdf:about="#MM.related_creation"/>
  </rdfs:subPropertyOf>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="#MM.related_creation">
  <rdfs:domain rdf:resource="#MM_Entity"/>
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
>A creation related to this entity.</rdfs:comment>
  <rdfs:subPropertyOf
rdf:resource="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P1F.is_identified_b
y"/>
</owl:ObjectProperty>
<rdf:Property rdf:ID="MM.source">
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
>The source of information from which this Actor was derived. For example
the Web site/pages or content providers.</rdfs:comment>
  <rdfs:domain rdf:resource="#MM_Entity"/>
</rdf:Property>
</rdf:RDF>

```


Annex 7: Relationship between MultiMatch Actor and CIDOC CRM Actor



Annex 8: MultiMatch metadata schema

This Annex presents the metadata formats identified per (sub)entity. For the identified types of Actors and Creations the formats are largely the same. The first column of the tables concerned shows the entity or subentities the property relates to.

Annex 8.1: The formats for Catalogue and Website

The Collection entity can be applied to any aggregation of physical or digital items. Those items may be of any type, so examples might include aggregations of natural objects, created objects, "born-digital" items, digital surrogates of physical items, and the catalogues of such collections (as aggregations of metadata records). The criteria for aggregation may vary: e.g. by location, by type or form of the items, by provenance of the items, by source or ownership, and so on. Collections may contain any number of items and may have varying levels of permanence. A "collection-level description" provides a description of the collection as a unit: the resource described by a collection-level description is the collection, rather than the individual items within that collection.

For MultiMatch at least two types of Collection are identified: Catalogue and Website.

A Collection in DCMI Metadata Terms: `dcmitype:collection`. In CIDOC-CRM Collection refers to a collection of physical objects which means that a MM Collection can not directly be a subclass of `E78.Collection` as a Website is a collection of information objects. Thus only subclasses of Collection which refer to physical objects will be subclasses of `E78.Collection`

The Catalogue format:

Explanation	MultiMatch metadata	D1.3 Section reference	Dublin Core	DCMI Metadata Terms	CIDOC
Unique identifier, to distinguish the different catalogues within MultiMatch.	Identifier		Identifier	dc:identifier	P131F.is_identified_by <E41.Appellation>
Name of the content provider responsible for the catalogue.	Publisher		Publisher	dc:publisher	P94B.was_created_by <E65.Creation> P2F.has_type[Publication] P14F.carried_out_by <E39.Actor>
A short description of the catalogue, indicating a.o. the period, the type of creation of the collection of objects described. Free text.	Description		Description	dc:description	P3F.has_note

The Website format:

Explanation	MultiMatch metadata	D1.3 Section reference	Dublin Core	DCMI Metadata Terms	CIDOC
A label generated by automatic classification. High level categorization of the intention or focus of the website. Closed list: cultural heritage, educational or tourism. Possibly with subclasses per type (WP4 development).	Type	1.2	Type	dc:type	P2f.has_type<E55.Type>
Title of Website, as dictated in the HTML Title tag. E.g. <title>Louvre Museum Official Website</title>	Title		Title	dc:title	P102F.has_title <E35.Title>
URI of the home page of the Website	Identifier	1.2	Identifier	dc:identifier	P131F.is_identified_by <E41.Appellation>
The date when the information used to derive the entity was captured in the MultiMatch system.	Date - captured		N/A	N/A	N/A
Name of CH institute or News service or other organisation responsible for the content of the Website or of the RSS (news) feed.	Publisher		Publisher	dc:publisher	P94B.was_created_by <E65.Creation> P2F.has_type[Publication] P14F.carried_out_by <E39.Actor>
The cultural heritage site profile will contain a.o.: "5. The tag cloud of the terms which best describe its web contents." The description of the website is either extracted from of the homepage or the summary of all pages of the website. And it can potentially be stored as keywords (for presenting purposes: tag cloud).	Tags	1.2 page 11 and 2.3	Subject	dc:subject	P3F.has_note
A snippet of text describing the Website	Description	1.2 page 11 and 2.3	Description	dc:description	P3F.has_note
The language the Website is written in. Closed list: Dutch; English; Italian; Spanish. To be automatically extracted.	Language	1.2	Language	dc:language	P2F.has_type<E56.Language>
The cultural heritage site profile will contain a.o.: "2. A list of creators most closely associated with the site and rendered as cultural objects." Note that this may be a long list.	Related Creators	1.2 page 11	Relation	dc:relation	P67F.refers_to<P67.1.has_type<E55.Type>> <E21.Person>
The cultural heritage site profile will contain a.o.: "3. A list of works of art most closely associated with the site and rendered as cultural objects." Note that this may be a long list.	Related Creations	1.2 page 11	Relation	dc:relation	P67F.refers_to<P67.1.has_type<E55.Type>> <E71.Man-Made_Thing>
The cultural heritage site profile will contain a.o.: "4. The network of cultural	Related Websites	1.2 page 11	Relation	dc:relation	P67F.refers_to<P67.1.has_type<E55.Type>>><E73.Information_Object>

heritage sites most closely associated to the site." Note that this may be a long list.					
The cultural heritage site profile will contain a.o.: "1. Web pages associated to the site rendered as title plus snippet and classified."	Related Web pages	1.2 page 11	Relation	dc:relation	P67F.refers_to<P67.1.has_type<E55.Type>> <E73.Information_Object>

Annex 8.2: The format Web page

Explanation	MultiMatch metadata	D1.3 Section reference	Dublin Core	DCMI Metadata Terms	CIDOC	Cervantes
Title of Web page	Title	1.2	Title	dc:title	P102F.has_title<E35.Title>	dc:title
URI of Web page	Identifier	1.2	Identifier	dc:identifier	P131F.is_identified_by<E41.Appellation>	dc:identifier is URL to the content.
Date the Web page was last modified.	Date - modified		Date	dcterms:modified	P124B.was_transformed_by<E81.Transformation> P117F.occurs_during<E2.Temporal_Entity>	dc:date
The date when the content of the page was captured, indexed or crawled for MultiMatch.	Date - captured		N/A	N/A	N/A	
URI of the homepage of the website whereto this page belongs. Via this link, the Publisher of the Website becomes also the Publisher of the linked Web page.	Source		Relation	dc:relation - is part of	P106B.forms_part_of<E73.Information_Object>	www.cervantesvirtual.com
Categories to classify the Web page. These categories will be derived from a closed list that consists of two levels. E.g. reviews.artwork. Minimum list for the first level: biographical data, works of art, source, reviews, news and other.	Type	1.2 page 10 1.3 page 14	Type	dc:type	P2F.has_type<E55.Type>	
Subject keywords to indicate the topic (e.g. art period, creator, creation or any other subject) of the Web page. In the case of Wikipedia articles: this element presents two levels. First the top-facet, namely either Museum-Institution or Museum-Artist. The second level presents the Wikipedia classes, the more specific artist and museums categories. For other web pages: the value of the keyword tag (not in the first prototype).	Subject		Subject	dc:subject	P2F.has_type<E55.Type>	dc:subject is UDC code (second prototype)
"possibly (unless it is done dynamically) a brief keyword-based summary of the content in the four languages of the prototype."	Tags	1.3 page 14	Subject	dc:subject	P3F.has_note	
Presentation element. "Text documents will be summarized using query-biased snippet generation techniques which display the text in its original language or a summary translation at least in English."	Description	2.3	Description	dc:description	P3F.has_note	dc:description
Language the Web page is written in. Closed list: Dutch; English; Italian; Spanish. To be automatically extracted. For Wikipedia adopted from the Wikipedia	Language	1.2	Language	dc:language	P2F.has_type<E56.Language>	dc:language

version.						
Identifier of Creator to which the content is related. Possibly link to ULAN and/or Wikipedia entry. Reference element or browsable relation.	Related Creator		Relation	dc:relation	P67F.refers_to	
Identifier of creations to which the Web page is related.	Related Creations		Relation	dc:relation	P67F.refers_to	
All outgoing links from this page. The links may refer to other pages of the same Website or to other pages of other Websites (cultural heritage and more). This may be all the links from a page or a selected set, i.e. those which are related to CH.	Links		Relation	dc:relation	P67F.refers_to <P67.1.has_type <E55.Type>> <E73.Information_Object>	

Annex 8.3: The format for Actor

Note, that the metadata elements that are in bold, are only applicable for the sub-entity Actor.Person.

Explanation	MultiMatch metadata	D1.3 Section reference	Dublin Core	DCMI Metadata Terms	CIDOC
Person or Organisation primarily responsible for making the content of the resource (the Creator, Artist, Author) or responsible for making the resource available (Publisher) or responsible for making contributions to the content of the resource (Contributor) or otherwise related to the Creation (e.g. Discoverer, Conservator, Archiver etc.).		1.2			<E39.Actor>
Preferrably from a controlled list of names (ULAN), however if this does not include the necessary actor a MultiMatch specific identifier must be used.	Identifier		Identifier	dc:identifier	P131F.is_identified_by <E82.Actor_Appellation ULAN>ID: 500115588
Name of the Actor, either a person name (can be based on ULAN (the preferred name) or on the controlled vocabularies of the content providers) or an organisation name (in the case of an anonymous artist or when the person name is unknown) . Names in ULAN may include given names, pseudonyms, variant spellings, names in multiple languages, and names that have changed over time (e.g., married names).	Name	1.2	Identifier	dc:identifier	P131F.is_identified_by <E82.Actor_Appellation ULAN>Gogh, Vincent van
The source of information from which this Actor was derived. For example the Web site/pages or content providers.	Source		Source	dc:source	N/A

Name of the organisation that the Actor as a person is affiliated with. Controlled vocabulary from ULAN ((Related Corporate bodies) combined with those from the content providers. E.g. Magnum for a specific group of photographers or the name of a monastery for an unknown monk who has written a specific manuscript.	Affiliated Organisation		N/A	N/A	P107B.is_current_or_former_member_of<E74.Group>
A closed list to indicate the main roles of the Actor. Minimum, relevant Actors that can be associated to the Dublin Core Roles, i.e. Creator (primarily responsible for making the content of the resource); Publisher (responsible for making the resource available); Contributor (a "catch all" for other roles, that is responsible for making contributions to the content of the resource). These categories will be derived from a closed list that consists of at least two levels. E.g. Creator.Painter. The narrower type level can be derived from the controlled vocabulary form ULAN (Roles element) in combination with the controlled vocabularies on types of actors in the catalogues of the content providers. Also the MARC list can be applied: http://www.loc.gov/marc/sourcecode/relator/relatorlist.html Roles are the major professional roles or activities performed by the artist throughout his or her lifetime (e.g., artist, architect, sculptor). For a corporate body, roles include the major activities or purpose of the firm, institution, or other corporate body (e.g., studio, manufactory, workshop). This information is part of the Profile presentation (Category).	Main Role	1.2	Type	dc:type	P2F.has_type<E55.Type>
Date of birth of the Actor. Via ULAN or indexed material. This date is part of the Profile presentation.	Date of birth	1.2 page 10	Coverage	dc:temporal	P98B.was_born<E67.Birth>P117F.occurs_during<E2.Tempral_Entity>
If applicable, the date of death of the Actor. Derived from ULAN or indexed material. This date is part of the Profile presentation.	Date of death	1.2 page 10	Coverage	dc:temporal	P100B.died_in<E69.Death>P117F.occurs_during<E2.Tempral_Entity>
Place of birth of the Actor. Can be derived from the Born element from the ULAN concerned or from the indexed material. Combination with TGN.	Place of birth	1.2 page 10	Coverage	dcterms:spatial	P98B.was_born<E67.Birth>P7F.took_place_at<E53.Place ULAN>

If applicable, the name of the place of death of the Actor. Can be derived from the Died element from the ULAN concerned or from the indexed material. Combination with TGN.	Place of death		Coverage	dcterms:spatial	P100B.died_in<E69.Death>P7F.took_place_at<E53.Place ULAN>
The topic of the Actor, in other words: the characterization of the Actor in keywords (presented as a tag cloud) based on indexed material. "An author's tag cloud with those keywords most representative of the author according to the indexed material " It is possible to be free text and also can be controlled, if based on the catalogue descriptions of the content providers.	Tags	1.2 page 10	Subject	dc:subject	P3F.has_note
A free text note about the Actor that gives additional information not recorded in other elements. Can be derived from the Note element from ULAN and from Description elements from the content providers catalogues.	Description		Description	dc:description	P3F.has_note
This element points to the identifier of related Creations. For example that are created by OR contributed to OR published by this Actor.	Related Creations	1.2 page 10	Relation	dc:relation	P14B.performed<E65.Creation>P1F.is_identified_by<E42.Object_Identifier> or P14B.performed<E7.Activity><P2F.has_type>Contribution Publication ...<P1F.is_identified_by<E42.Object_Identifier>
Role of Actor in this relation. Closed list to be derived from some controlled vocabulary, e.g. MARC.	Related Creations - Type		N/A	N/A	P2F.has_type<E55.Type>
Date associated with the related creation and this actor, e.g. Creation Date, Publication Date.	Related Creations - Date		N/A	N/A	P117F.occurs_during

<p>This element points to the identifier of the Actors related to the Actor being described. A search for a Actor name with additional search terms (e.g. painter), should a.o. render the following specific search result presentation: "2. A graphical depiction of the author's network of relationships to other authors, the different strengths of such relationships, and the concepts/keywords/terms that characterize such relations (see Figure 6) The idea of the author's networks is to allow the user to discover and explore those authors closely related with the queried author to improve his knowledge about the author's context. In this way, relations may be extracted using common terms or concepts shared by MultiMatch indexed authors."</p>	<p>Related Actors</p>	<p>1.2 page 10</p>	<p>Relation</p>	<p>dc:relation</p>	<p>For family relations: P98B.was_born<E67.Birth>P96F.by_mother P97F.from_father<E21.Person>P96B.gave_birth P97B.was_father_for<E67.Birth></p> <p>For other relations: There is no direct way to express relations but it is possible either through connection to a E39.Actor, E74.Group or E7.Activity with other related Actors P131F.is_identified_by <E82.Actor_Appellation ULAN>ID: 500115588</p>
<p>Type of the Actor-Actor relation. Controlled vocabulary derived from ULAN (the following Related People or Corporate Bodies: associates; partners; parents; siblings; children; teachers and students.) combined, if applicable, with types from the content providers.</p>	<p>Related Actors - Type</p>		<p>N/A</p>	<p>N/A</p>	<p>P2F.has_type<E55.Type></p>
<p>This element points to content where the Actor can be seen or heard (depicted). In case of Video Creations: Link to the set of representative keyframes where the Actor can be seen. Link to the Video scenes (timecodes) where the Actor can be seen. In case of Audio and Image Creations: Link to the part of the Transcript of speech where the Actor is speaking. Link to an image of a portrait of the Actor. The portrait is part of the Profile presentation.</p>	<p>Depiction</p>	<p>1.2 page 10</p>	<p>Relation</p>	<p>dc:relation</p>	<p>P62B.is_depicted_by</p>
<p>URIs that link to related Websites that contain some reasonable amount of reference to this Actor.</p>	<p>Related Websites</p>		<p>Relation</p>	<p>dc:relation</p>	<p>P67B.is_referred_to_by <P67.1.has_type<E55.Type>> <E73.Information_Object></p>
<p>URIs that link to related Web pages (Wikipedia articles and to other harvested web content) that contain some reasonable amount of reference to this Actor. E.g. an interview with this Actor or a review on the techniques applied by this Actor.</p>	<p>Related Web pages</p>		<p>Relation</p>	<p>dc:relation</p>	<p>P67B.is_referred_to_by <P67.1.has_type<E55.Type>> <E73.Information_Object></p>

Annex 8.4: The format for Creation

Explanation	MultiMatch metadata	D1.3 Section reference	Dublin Core	DCMI Metadata Terms	CIDOC
In the context of MultiMatch, the Creation metadata format describes the physical cultural heritage object, i.e. it is - a unique man-made object; - the result of a distinct intellectual or creative creation process which can be a Text, Still Image, Video, Audio or Physical Object; - an Endeavour in terms of the FRBR reference model, in that it represents a Work in its abstract term and also the actual Manifestation of that work.		1.2			As a Creation covers all the type of FRBR Endeavour it is a subclass of both <E73.Information_Object> and <E24.Physical_Man-Made_Thing>
Unique identifier for the resource within MultiMatch. Might be a URL, ISBN for Books, DOI etcetera.	Identifier - MultiMatch		Identifier	dc:identifier	P131F.is_identified_by <E41.Appellation>
Unique identifier provided by the institute that archives the creation. This is the record number of the description from an indexed catalogue (not from an object management information system). This information is the link to the Catalogue.	Identifier - Source		Identifier	dc:identifier	P131F.is_identified_by <E41.Appellation>
Official, original title of the creation.	Title	1.2	Title	dc:title	P102F.has_title <E35.Title>
The source of information from which this Creation was derived. For example the Web site/pages or content providers.	Source		Source	dc:source	N/A

Identifier of actor who creates or modifies the Creation. Multiple values possible.	Related Actor	1.2	Creator Contributor Publisher Relation	dc:creator dc:contributor dc:publisher dc:relation	For Creation: P94B.was_created_by <E65.Creation> For other activities Types/Roles: P17B.motivated <E65.Activity> then for Actor identifier: P14F.carried_out_by <E39.Actor>
The role of the Actor in relation to the Creation being described. A closed list of roles. The minimum being related to the Dublin Core actor roles, i.e. creation, contribution and publication.	Related Actor - Type		N/A	N/A	P2F.has_type <E65.Creation> or [Other_Activity_Type]
The date of the Creation-Actor relationship. E.g. creation date. This might be an exact date in the form of yyyy-mm-dd or a period indicated by a start date and an end date. If Creation date is uncertain or unknown, it should be presented to the user that the date or period is estimated. ISO standard for date and for period notation.	Related Actor - Date	1.2	Date (only for Creation Date)	dc:date (only for Creation Date) dc:issued (only for Publication Date)	P117F.occurs_during
Terms or phrases that describe, identify or interpret the Creation and what it depicts or expresses. "A search for a Creation title with optionally additional search terms, should a.o. render the following specific search result presentation: "3. An artwork's tag cloud with those keywords most representative of the artwork according to the indexed material.	Tags	1.2 page 11	Subject	dc:subject	P3F.has_note

Topic of the creation, described by keywords, key phrases, or classification codes. Controlled vocabularies of the content providers.	Subject		Subject	dc:subject	P2F.has_type<E55.Type>
A free text note about the Creation that gives additional information not recorded in other elements. Can be used to store or link to low-level descriptors of the Still Image, Audio or Video Creation (MPEG-7, TRS, etc.). Can contain the original content provider (legacy) metadata (as XML document) Can be used to indicate the Audience and Educational level of the Creation.	Description		Description	dcterms:abstract	P3F.has_note
Search element for Audio and Video Creations. The actual speech recognition transcripts will be indexed for full-text searching. This element will need a language property.	Description – Transcript 1)	2.1	N/A	N/A	N/A
Place where Creation was created, based on indexed content. Controlled vocabulary: TGN in combination with geographic vocabularies of content providers.	Location		Coverage	dcterms:spatial	P94B.was_created_by <E65.Creation> P7F.took_place_at <E53.Place>

<p>Place where the Creation is archived or physically stored.</p> <p>Note that this only applies to physical objects, if it represents the location of the original unique creation (video, audio, text, painting, etc.)</p> <p>Controlled vocabulary: TGN in combination with geographic vocabularies of content providers.</p>	Archive Location		N/A	N/A	<p><E19.Physical_Object> P55F.has_current_location<E53.Place></p>
<p>A free text description of the physical medium or dimensions of the Creation, including the description of the materials used in the construction and the dimensions of the Creation. Dimensions can include size and duration.</p> <p>Only for digital Creations the file type should be included. E.g. MPEG-2 for video documents.</p> <p>Note, that the format of the digital representations are not described here.</p>	Format	2.1	Format	dc:format	<p>P2F.has_type<E55.Type> or P43F.has_dimension<E54.Dimension> For <E19.Physical_Object> P45F.consists_of<E57.Material></p>
<p>Closed list with categories of creations for browsing and filtering purposes to indicate the creation type. The list will consist of at least two levels. The first level: Text, Audio, Still Image, Video, Physical Object, Other.</p> <p>No cultural heritage object types will be excluded.</p> <p>AAT and DC vocabulary can be used to specify the Creation types per broad type, in combination with the vocabularies of the content providers.</p> <p>E.g. Text: book, poem, newspaper article; Still Image: painting, etche; Physical object: sculpture, table; Other: monument, historic building. Or more specific: several types of books, depending on the indexed material.</p>	Type		DCMI – Type : Image; Moving Image; Still Image; Physical Object; Sound; Text	dc:type	<p>P2F.has_type<E55.Type></p>

Identifier of related Creations. Special attention for other creations that depict or describe this creation. For example: a monument or a painting that can be seen in a video, a painting that can be seen in another painting, a painting that is described in a book.	Related Creations		Relation	dc:relation	P67F.refers_to <E71.Man-Made_Thing>
The type of relation between the related Creation and the Creation being described. Closed list. Minimum: is created by same creator (if the creation being described is created by the same creator as the related creation); is depicted in (if the creation being described is depicted in the related creation). Further examples: is a preparatory creation for a later work; is copied after; is referenced by (the other creation); forms a serie of works together with.	Related Creations -Type		N/A	N/A	P67.1.has_type <E55.Type>
Link to the digital representations of the Creation, including image(s) of the Creation in whole or part, (if applicable) a link to set(s) of representative keyframes, to the text of the Creation or to the recording of a reading of (a selection of) the text of the Creation.	Link to Digital representations	1.2	Relation	dc:relation	P67B.is_referred_to_by <P67.1.has_type<E55.Type>> <E73.Information_Object> or P62B.is_depicted_by <E24.Man-Made_Physical_Object>
URIs that link to related Websites that contain some reasonable amount of reference to this Actor.	Related Websites	1.2	Relation	dc:relation	P67B.is_referred_to_by <P67.1.has_type<E55.Type>> <E73.Information_Object>
URIs that link to related Web pages (Wikipedia articles and to other harvested web content) that contain some reasonable amount of reference to this Actor. E.g. an interview with this Actor or a review on the techniques applied by this Actor.	Related Web pages		Relation	dc:relation	P67B.is_referred_to_by <P67.1.has_type<E55.Type>> <E73.Information_Object>
Name of the cultural heritage institute and/or (contact) person that is the rights holder.	Rights Holder		Rights	dcterms:rightsHolder	P104F.is_subject_to <E30.Right>

Yes or No or Unknown. VRA acknowledges 4 values for the Rights - type: copyrighted; publicDomain; undetermined; other.	Copyrighted		Rights	dcterms:dateCopyrighted	P104F.is_subject_to <E30.Right>
Information about any justifications, conditions, or restraints on use, contact or licensing information. This element can be filled from the mapped metadata elements of the content providers. For the crawled sources the default text might be "Unknown, contact the archiving institute concerned."	License conditions		Rights	dcterms:license	P104F.is_subject_to <E30.Right>
Information about who can access the creation or an indication of its security status, includes information regarding access or restrictions based on privacy, security or other regulations. It is not sure if we will have this kind of information, but DC Terms has an element for Access Rights.	Access Rights		Rights	dcterms:accessRights	P104F.is_subject_to <E30.Right>
Language of the creation. Only applicable for the following Creation types: Text, Sound and Video.	Language 2)		Language	dc:language	P2F.has_type <E56.Language>

- 1) The Description-Transcript property is only applicable for the sub-entities Video and Audio.
- 2) The Language property is only applicable for the sub-entities: Text, Video and Audio.

Annex 8.5: The format for Digital Representation

Explanation	MultiMatch metadata	D1.3 Section reference	Dublin Core	DCMI Metadata Terms	CIDOC	Alinari	Cervantes
A digital representation is a visual surrogate or reproduction of a Creation.		2.1			<E73.Information_Object>		This information concerns the Table of Contents of the books.

Unique identifier, stored in different databases, depending on file type. Image number in GIFT. Identifier of Video or Audio content files.	Identifier - MultiMatch		Identifier	dc:identifier	P131F.is_identified_by <E41.Appellation>		
Image number from content provider or crawled website.	Identifier - Source		Identifier	dc:identifier	P131F.is_identified_by <E41.Appellation>	<imagecode cod>	
URL of the web pages, video files, image databases, etc. where the image comes from.	Representation source	1.2	Source	dc:source	P67F.refers_to<P67.1.has_type<E55.Type>[Source]><E73.Information_Object>	default "www.alinari.com"	URL in dc:identifier
Closed list to indicate if the Creation is represented as a whole or as a part. Whole; Part	Type		Type	dc:type	P2F.has_type<E55.Type>	This might be automatically extracted from <titolo>. "Detail of" = Part. If "Detail of" is not present in <titolo>, then = Whole.	default "Part", as it concerns Table of Content.
The first prototype will cover the following file types: - MIME types: plain, html, xml. For text documents. - MIME types: mpeg, xwav. For audio documents. - BMP, JPG, GIF, TIFF and PNG. For still image documents. - MPEG-1, MPEG-4, MPEG-7. For video documents. This closed list possibly needs extension for the second prototype.	Format	1.2	Format	dc:format	P2F.has_type<E55.Type>	default "jpg"	default "html"
The size of the file in kB or MB.	File size	1.2 page 14	Format	dcterms:extent	P43F.has_dimension <E54.Dimension>		
Link to the MultiMatch Identifier of the Creation that is represented by this Digital representation.	Creation represented		Relation	dc:relation			
This element contains the identifier of the Actor depicted in this representation.	Actor represented		Relation	dc:relation			

Name of the cultural heritage institute and/or (contact) person that is the rights holder for this digital representation.	Rights Holder		Rights	dcterms:rightsHolder	P104F.is_subject_to <E30.Right>	<fotografo>	default "Cervantes"
Yes or No or Unknown.	Copyrighted		Rights	dcterms:dateCopyrighted	P104F.is_subject_to <E30.Right>	default "Yes"	
Link to low-level features of Image or Video representing the Creation, e.g. MPEG-7 descriptors. "Standard low-level visual features such as colour distributions, textures and identifiable edges in the content.(1.3)"	Low-level features	1.3	N/A	N/A	N/A		

Annex 8.6: The format for Event

Explanation	MultiMatch metadata	D1.3 Section reference	Dublin Core	DCMI Metadata Terms	CIDOC
An Event. The only type of Event currently envisaged is an Exhibition of Creations.			Type.Event	dc:Type.Event	<E5.Event> In practice it may be more likely that an Event will be more closely related to an <E7.Activity> which is carried out by an actor. P2F.has_type [DCMIType.Event]
A type of event. Closed list to be derived from indexed material.	Type		Type	dc:type	P2F.has_type[Event_Type]
The title of the event described.	Title		Title	dc:title	P102F.has_title <E35.Title>
A unique MultiMatch identifier for the event described.	Identifier		Identifier	dc:identifier	P131F.is_identified_by <E41.Appellation>
Theme of the event in keywords. For Exhibitions this relates to the type of Creations exhibited and thus may be derived from the metadata associated with the linked Creations, i.e. Artistic Periods, Artists, Creation.Type.	Tags		Subject	dc:subject	P3F.has_note
Starting date of the event described.	Date - Start		Date	dc:date	P116F.starts <E2.Temporal_Entity>
Ending date of the event described.	Date - End		N/A	N/A	P115F.finishes <E2.Temporal_Entity>
The place where the event takes place. A link to TGN can help the geographical search of the user (e.g. search for events in the neighbourhood of Dublin OR in the North of Spain that are exhibiting paintings).	Location		Location	dc:location	E53.Place

A free text note about the Event that gives additional information not recorded in other elements. For example for Exhibitions information that clarifies when the event can be visited: opening hours, opening days or closing days.	Description		Description	dc:description	P3F.has_note
The Creators related to this event.	Related Creators		Relation	dc:relation	P11F.had_participant <E39.Actor>
The Creations related to this event	Related Creations		Relation	dc:relation	P67F.refers_to <E71.Man-Made_Thing>
URI of the Website describing the event.	Related Website		Relation	dc:relation	P67B.is_referred_to_by <E73.Information_Object>
URI of the Webpage describing the event.	Related Webpage		Relation	dc:relation	P67B.is_referred_to_by <E73.Information_Object>
If applicable, a link to the catalogue that describes the exhibited creations.	Related Catalogue 1)		Relation	dc:relation	P16F.used_specific_object <E70.Thing>

1) This property is only applicable for the sub-entity Exhibition

Annex 9: MultiMatch Metadata xsd schema

```
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  targetNamespace="http://multimatch.org/schema/multimatchMetadata-1.0"
  xmlns:mmns="http://multimatch.org/schema/multimatchMetadata-1.0" version="1.0" id="mmns">
  <!--
  The top level structure of the MultiMatch Entity
  -->
  <xsd:element name="MM_Entity">
    <xsd:complexType>
      <xsd:sequence>
        <xsd:element ref="mmns:Actor" minOccurs="0" maxOccurs="unbounded"/>
        <xsd:element ref="mmns:Collection" minOccurs="0" maxOccurs="unbounded"/>
        <xsd:element ref="mmns:Creation" minOccurs="0" maxOccurs="unbounded"/>
        <xsd:element ref="mmns:DigitalRepresentation" minOccurs="0" maxOccurs="unbounded"/>
        <xsd:element ref="mmns:Event" minOccurs="0" maxOccurs="unbounded"/>
        <xsd:element ref="mmns:Webpage" minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
    </xsd:complexType>
  </xsd:element>
  <!--
  The substructure of the MultiMatch Entity
  -->
  <xsd:element name="Collection">
    <xsd:complexType>
      <xsd:choice>
        <xsd:element ref="mmns:Catalogue"/>
        <xsd:element ref="mmns:Website"/>
      </xsd:choice>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="Creation">
    <xsd:complexType>
      <xsd:choice>
        <xsd:element ref="mmns:Audio"/>
        <xsd:element ref="mmns:PhysicalObject"/>
        <xsd:element ref="mmns:StillImage"/>
        <xsd:element ref="mmns:Text"/>
        <xsd:element ref="mmns:Video"/>
      </xsd:choice>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="DigitalRepresentation">
    <xsd:complexType>
      <xsd:choice>
        <xsd:element ref="mmns:DigitalAudio"/>
        <xsd:element ref="mmns:DigitalImage"/>
        <xsd:element ref="mmns:DigitalText"/>
        <xsd:element ref="mmns:DigitalVideo"/>
      </xsd:choice>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="Actor">
    <xsd:complexType>
      <xsd:choice>
        <xsd:element ref="mmns:Person"/>
        <xsd:element ref="mmns:Organisation"/>
      </xsd:choice>
    </xsd:complexType>
  </xsd:element>
</xsd:schema>
```

```

    </xsd:choice>
  </xsd:complexType>
</xsd:element>
<!--

```

The definition of the subelements of the main MultiMatch elements.
 Note that these are defined by groups in line with XML Schema best practice recommendations.

see <http://www.xfront.com/BestPracticesHomepage.html>

```

-->
  <xsd:element name="Webpage">
    <xsd:complexType>
      <xsd:group ref="mmns:WebpageElements"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="Website">
    <xsd:complexType>
      <xsd:group ref="mmns:WebsiteElements"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="Person">
    <xsd:complexType>
      <xsd:group ref="mmns:PersonElements"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="Organisation">
    <xsd:complexType>
      <xsd:group ref="mmns:OrganisationElements"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="Catalogue">
    <xsd:complexType>
      <xsd:group ref="mmns:CatalogueElements"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="StillImage">
    <xsd:complexType>
      <xsd:group ref="mmns:CreationElements"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="Text">
    <xsd:complexType>
      <xsd:group ref="mmns:TextElements"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="Video">
    <xsd:complexType>
      <xsd:group ref="mmns:AudioVideoElements"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="Audio">
    <xsd:complexType>
      <xsd:group ref="mmns:AudioVideoElements"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="PhysicalObject">
    <xsd:complexType>
      <xsd:group ref="mmns:CreationElements"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="DigitalText">
    <xsd:complexType>

```

```

    <xsd:group ref="mmns:DigitalRepresentationElements"/>
  </xsd:complexType>
</xsd:element>
  <xsd:element name="DigitalAudio">
    <xsd:complexType>
      <xsd:group ref="mmns:DigitalRepresentationElements"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="DigitalVideo">
    <xsd:complexType>
      <xsd:group ref="mmns:DigitalRepresentationElements"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="DigitalImage">
    <xsd:complexType>
      <xsd:group ref="mmns:DigitalRepresentationElements"/>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="Event">
    <xsd:complexType>
      <xsd:group ref="mmns:EventElements"/>
    </xsd:complexType>
  </xsd:element>
<!--

```

The definition of the grouping of subelements of the main MultiMatch elements.

```

-->
  <xsd:group name="CatalogueElements">
    <xsd:sequence>
      <xsd:element name="Identifier" type="mmns:CatalogueIdentifier"></xsd:element>
      <xsd:element name="Publisher" type="mmns:ActorIdentifierRef"></xsd:element>
      <xsd:element name="Description">
        <xsd:complexType>
          <xsd:attribute name="language" type="xsd:language"></xsd:attribute>
        </xsd:complexType>
      </xsd:element>
    </xsd:sequence>
  </xsd:group>
  <xsd:group name="WebsiteElements">
    <xsd:sequence>
      <xsd:element name="Identifier" type="mmns:WebsiteIdentifier"></xsd:element>
      <xsd:element name="Type" type="xsd:string"></xsd:element>
      <xsd:element name="Title" type="xsd:string"></xsd:element>
      <xsd:element name="DateCaptured" type="xsd:date"></xsd:element>
      <xsd:element name="Tags" type="xsd:string"></xsd:element>
      <xsd:element name="Description">
        <xsd:complexType>
          <xsd:attribute name="language" type="xsd:language"></xsd:attribute>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Language" type="xsd:language"></xsd:element>
      <xsd:element name="Publisher" type="mmns:ActorIdentifierRef"></xsd:element>
      <xsd:element name="RelatedCreator" type="mmns:ActorIdentifierRef"></xsd:element>
      <xsd:element name="RelatedCreation" type="mmns:CreationIdentifierRef"></xsd:element>
      <xsd:element name="RelatedWebpage" type="mmns:WebpageIdentifierRef"></xsd:element>
      <xsd:element name="RelatedWebsite" type="mmns:WebsiteIdentifierRef"></xsd:element>
    </xsd:sequence>
  </xsd:group>
  <xsd:group name="WebpageElements">

```

```

<xsd:sequence>
  <xsd:element name="Identifier" type="mmns:WebpageIdentifier"></xsd:element>
  <xsd:element name="Type" type="xsd:string"></xsd:element>
  <xsd:element name="Title" type="xsd:string"></xsd:element>
  <xsd:element name="Source" type="mmns:WebsiteIdentifierRef"></xsd:element>
  <xsd:element name="DateCaptured" type="xsd:date"></xsd:element>
  <xsd:element name="DateModified" type="xsd:date"></xsd:element>
  <xsd:element name="Publisher" type="mmns:ActorIdentifierRef"></xsd:element>
  <xsd:element name="Subject" type="xsd:string"></xsd:element>
  <xsd:element name="Tags" type="xsd:string"></xsd:element>
  <xsd:element name="Description">
    <xsd:complexType>
      <xsd:attribute name="language" type="xsd:language"></xsd:attribute>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="RelatedCreator" type="mmns:ActorIdentifierRef"></xsd:element>
  <xsd:element name="RelatedCreation" type="mmns:CreationIdentifierRef"></xsd:element>
  <xsd:element name="Language" type="xsd:language"></xsd:element>
  <xsd:element name="Links" type="mmns:WebpageIdentifierRef"></xsd:element>
</xsd:sequence>
</xsd:group>
<xsd:group name="ActorElements">
  <xsd:sequence>
    <xsd:element name="Identifier" type="mmns:ActorIdentifier"></xsd:element>
    <xsd:element name="Type" type="xsd:string"></xsd:element>
    <xsd:element name="Name" type="xsd:string"></xsd:element>
    <xsd:element name="Source" type="xsd:string"></xsd:element>
    <xsd:element name="MainRole" type="xsd:string"></xsd:element>
    <xsd:element name="Tags" type="xsd:string"></xsd:element>
    <xsd:element name="Description">
      <xsd:complexType>
        <xsd:attribute name="language" type="xsd:language"></xsd:attribute>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="RelatedActor">
      <xsd:complexType>
        <xsd:group ref="mmns:RelatedActorElements"/>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="RelatedCreation">
      <xsd:complexType>
        <xsd:sequence>
          <xsd:group ref="mmns:RelatedCreationElements"/>
          <xsd:element name="Date">
            <xsd:complexType>
              <xsd:sequence>
                <xsd:group ref="mmns:DateElements"/>
              </xsd:sequence>
            </xsd:complexType>
          </xsd:element>
        </xsd:sequence>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="Depiction" type="xsd:IDREF"></xsd:element>
    <xsd:element name="RelatedWebpage" type="mmns:WebpageIdentifierRef"></xsd:element>
    <xsd:element name="RelatedWebsite" type="mmns:WebsiteIdentifierRef"></xsd:element>
  </xsd:sequence>
</xsd:group>
<xsd:group name="PersonElements">

```

```

<xsd:sequence>
  <xsd:group ref="mmns:ActorElements"/>
  <xsd:element name="DateOfBirth">
    <xsd:complexType>
      <xsd:sequence>
        <xsd:group ref="mmns:DateElements"/>
      </xsd:sequence>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="PlaceOfBirth" type="xsd:string"></xsd:element>
  <xsd:element name="DateOfDeath">
    <xsd:complexType>
      <xsd:sequence>
        <xsd:group ref="mmns:DateElements"/>
      </xsd:sequence>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="PlaceOfDeath" type="xsd:string"></xsd:element>
  <xsd:element name="Affiliation" type="mmns:ActorIdentifierRef"></xsd:element>
</xsd:sequence>
</xsd:group>
<xsd:group name="OrganisationElements">
  <xsd:sequence>
    <xsd:group ref="mmns:ActorElements"/>
  </xsd:sequence>
</xsd:group>
<xsd:group name="CreationElements">
  <xsd:sequence>
    <xsd:element name="Identifier" type="mmns:CreationIdentifier"></xsd:element>
    <xsd:element name="Type" type="xsd:string"></xsd:element>
    <xsd:element name="Source" type="xsd:string"></xsd:element>
    <xsd:element name="SourceIdentifier" type="xsd:string"></xsd:element>
    <xsd:element name="Title">
      <xsd:complexType>
        <xsd:attribute name="language" type="xsd:language"></xsd:attribute>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="Subject" type="xsd:string"></xsd:element>
    <xsd:element name="Tags" type="xsd:string"></xsd:element>
    <xsd:element name="Location" type="xsd:string"></xsd:element>
    <xsd:element name="ArchiveLocation" type="xsd:string"></xsd:element>
    <xsd:element name="Format" type="xsd:string"></xsd:element>
    <xsd:element name="RightsHolder" type="xsd:string"></xsd:element>
    <xsd:element name="Copyright" type="xsd:string"></xsd:element>
    <xsd:element name="LicenseCondition" type="xsd:string"></xsd:element>
    <xsd:element name="DigitalRepresentation"
type="mmns:DigitalRepresentationIdentifierRef"></xsd:element>
    <xsd:element name="Description">
      <xsd:complexType>
        <xsd:attribute name="language" type="xsd:language"></xsd:attribute>
      </xsd:complexType>
    </xsd:element>
    <xsd:element name="RelatedActor">
      <xsd:complexType>
        <xsd:sequence>
          <xsd:group ref="mmns:RelatedActorElements"/>
          <xsd:element name="Date">
            <xsd:complexType>
              <xsd:sequence>

```

```

                <xsd:group ref="mmns:DateElements"/>
            </xsd:sequence>
        </xsd:complexType>
    </xsd:element>
</xsd:sequence>
</xsd:complexType>
</xsd:element>
<xsd:element name="RelatedCreation">
    <xsd:complexType>
        <xsd:group ref="mmns:RelatedCreationElements"/>
    </xsd:complexType>
</xsd:element>
<xsd:element name="RelatedWebpage" type="mmns:WebpageIdentifierRef"></xsd:element>
<xsd:element name="RelatedWebsite" type="mmns:WebsiteIdentifierRef"></xsd:element>
</xsd:sequence>
</xsd:group>
<xsd:group name="TextElements">
    <xsd:sequence>
        <xsd:group ref="mmns:CreationElements"/>
        <xsd:element name="Language" type="xsd:language"></xsd:element>
    </xsd:sequence>
</xsd:group>
<xsd:group name="AudioVideoElements">
    <xsd:sequence>
        <xsd:group ref="mmns:CreationElements"/>
        <xsd:element name="Language" type="xsd:language"></xsd:element>
        <xsd:element name="Transcription">
            <xsd:complexType>
                <xsd:attribute name="language" type="xsd:language"></xsd:attribute>
            </xsd:complexType>
        </xsd:element>
    </xsd:sequence>
</xsd:group>
<xsd:group name="DigitalRepresentationElements">
    <xsd:sequence>
        <xsd:element name="Identifier" type="mmns:DigitalRepresentationIdentifier"></xsd:element>
        <xsd:element name="Type" type="xsd:string"></xsd:element>
        <xsd:element name="Source" type="xsd:string"></xsd:element>
        <xsd:element name="SourceIdentifier" type="xsd:string"></xsd:element>
        <xsd:element name="Format" type="xsd:string"></xsd:element>
        <xsd:element name="RightsHolder" type="xsd:string"></xsd:element>
        <xsd:element name="Copyright" type="xsd:string"></xsd:element>
        <xsd:element name="FileSize" type="xsd:integer"></xsd:element>
        <xsd:element name="DepictedActor" type="mmns:ActorIdentifierRef"></xsd:element>
        <xsd:element name="DepictedCreation" type="mmns:CreationIdentifierRef"></xsd:element>
    </xsd:sequence>
</xsd:group>
<xsd:group name="EventElements">
    <xsd:sequence>
        <xsd:element name="Identifier" type="mmns:WebsiteIdentifier"></xsd:element>
        <xsd:element name="Type" type="xsd:string"></xsd:element>
        <xsd:element name="Title" type="xsd:string"></xsd:element>
        <xsd:element name="Tags" type="xsd:string"></xsd:element>
        <xsd:element name="DateStart" type="xsd:date"></xsd:element>
        <xsd:element name="DateEnd" type="xsd:date"></xsd:element>
        <xsd:element name="Location" type="xsd:string"></xsd:element>
        <xsd:element name="Description">
            <xsd:complexType>
                <xsd:attribute name="language" type="xsd:language"></xsd:attribute>
            </xsd:complexType>
        </xsd:element>
    </xsd:sequence>
</xsd:group>

```

```

        </xsd:complexType>
    </xsd:element>
    <xsd:element name="Language" type="xsd:language"></xsd:element>
    <xsd:element name="RelatedCreator" type="mmns:ActorIdentifierRef"></xsd:element>
    <xsd:element name="RelatedCreation" type="mmns:CreationIdentifierRef"></xsd:element>
    <xsd:element name="RelatedWebpage" type="mmns:WebpageIdentifierRef"></xsd:element>
    <xsd:element name="RelatedWebsite" type="mmns:WebsiteIdentifierRef"></xsd:element>
    <xsd:element name="RelatedCatalogue" type="mmns:CatalogueIdentifierRef"></xsd:element>
</xsd:sequence>
</xsd:group>
<!--

```

Definition of Actor and Creation typed relations.

-->

```

<xsd:group name="RelatedActorElements">
    <xsd:sequence>
        <xsd:element name="Identifier" type="mmns:ActorIdentifierRef"></xsd:element>
        <xsd:element name="Type" type="xsd:string"></xsd:element>
    </xsd:sequence>
</xsd:group>
<xsd:group name="RelatedCreationElements">
    <xsd:sequence>
        <xsd:element name="Identifier" type="mmns:CreationIdentifierRef"></xsd:element>
        <xsd:element name="Type" type="xsd:string"></xsd:element>
    </xsd:sequence>
</xsd:group>
<xsd:group name="DateElements">
    <xsd:sequence>
        <xsd:element name="DisplayDate" type="xsd:string"></xsd:element>
        <xsd:element name="EarliestDate" type="xsd:date"></xsd:element>
        <xsd:element name="LatestDate" type="xsd:date"></xsd:element>
    </xsd:sequence>
</xsd:group>
<!--

```

Identifiers and reference to identifiers used in the metadata.

Note that best practice is to use unique, key and keyref definition and not ID, IDREF.

ID are in the scope of the entire document and thus all IDs must be unique, i.e. A WebsiteIdentifier and WebpageIdentifier can not have the same value.

-->

```

<xsd:simpleType name="ActorIdentifier">
    <xsd:restriction base="xsd:ID"/>
</xsd:simpleType>
<xsd:simpleType name="ActorIdentifierRef">
    <xsd:restriction base="xsd:IDREF"/>
</xsd:simpleType>
<xsd:simpleType name="CatalogueIdentifier">
    <xsd:restriction base="xsd:ID"/>
</xsd:simpleType>
<xsd:simpleType name="CatalogueIdentifierRef">
    <xsd:restriction base="xsd:IDREF"/>
</xsd:simpleType>
<xsd:simpleType name="WebsiteIdentifier">
    <xsd:restriction base="xsd:ID"/>
</xsd:simpleType>
<xsd:simpleType name="WebsiteIdentifierRef">
    <xsd:restriction base="xsd:IDREF"/>
</xsd:simpleType>
<xsd:simpleType name="CreationIdentifier">
    <xsd:restriction base="xsd:ID"/>
</xsd:simpleType>

```

```

<xsd:simpleType name="CreationIdentifierRef">
  <xsd:restriction base="xsd:IDREF"/>
</xsd:simpleType>
<xsd:simpleType name="WebpageIdentifier">
  <xsd:restriction base="xsd:ID"/>
</xsd:simpleType>
<xsd:simpleType name="WebpageIdentifierRef">
  <xsd:restriction base="xsd:IDREF"/>
</xsd:simpleType>
<xsd:simpleType name="DigitalRepresentationIdentifier">
  <xsd:restriction base="xsd:ID"/>
</xsd:simpleType>
<xsd:simpleType name="DigitalRepresentationIdentifierRef">
  <xsd:restriction base="xsd:IDREF"/>
</xsd:simpleType>
<!--
An initial definition of the MultiMatch Date format which allows for the
definition of approximate dates, e.g. circa 1942, circa 1985-12, 1920-29,
1910-1912, circa mid 1900-1999, as well as exact dates 0634-05-17 AD
Note that this still refers to a (possible uncertain) point in time
and not a range.
-->
<xsd:simpleType name="DateType">
  <xsd:restriction base="xsd:string">
    <xsd:pattern value="(before |after)?(circa)?(early |mid |late )?[0-9][0-9][0-9][0-9](-
[0-9][0-9](-?[0-9][0-9])?)?( BC| AD)?"/>
  </xsd:restriction>
</xsd:simpleType>
</xsd:schema>

```

Annex 10: Example illustrating the application of the MultiMatch metadata schema

In this Annex an example is presented of the way in which the MultiMatch metadata schema could be populated. The starting point for this example is the Dutch painter Vincent van Gogh (Actor1). Searching the catalogue of Sound and Vision and the websites of the museums Van Gogh and Rijksmuseum (both in Amsterdam) provided the following MultiMatch records:

- three Actor records: describing Vincent van Gogh himself (Actor1) and two related Creators. With Actor – Actor relations of the following type: is assistant of; is child of; collaborated with. With Actor1 – Creation relations of the type: is depicted in; is discussed in.
- five Creation records: two records describing a two video creations about several works of art created by Vincent van Gogh, one record describing a video creation about the life of Van Gogh and two records describing two works of art by Van Gogh, that are depicted in Creation1 and Creation2. With Creation – Creation relations of the type: is depicted in; is inspired by; is part of.
- two Digital representation records that are related to two of the video creations as well as to Actor1 (is represented in)

Note, that the presentation is in human-readable form, while for MultiMatch the semantic web encoding will probably be in OWL or in RDF format.

Explanation on the syntax in the following tables:

- the name of the source is entered between []
- the language property is entered between ()
- multiple values of the metadata elements are indicated with bullet points.

Actor MM00001

Metadata (sub)element	Example
Identifier	URI to ULAN ID 500115588 [ULAN]
Actor Name	Gogh, Vincent van [ULAN]
Affiliated Organisation	NA
Actor Main Roles	- Creator.Painter [ULAN] (en) - Creator.Draftsman [ULAN] (en)
Date of birth	1853
Date of death	1890
Birth place	Zundert (nl) (North Brabant (en), Netherlands (en)) [ULAN; is actually derived from TGN].
Death place	Auvers-sur-Oise (Val-d'Oise, Île-de-France, France) [ULAN; is actually derived from TGN] (fr)
Actor Tags [automatic information extraction from first prototype content] (...several...)
Description	- Except for some brief periods of formal instruction, van Gogh was self-taught; he collected prints and reproductions to study and copy, especially those of Millet. His life and work are legendary in the history of art, making him the quintessential misunderstood, tormented, even insane artist, who sold only one work in his lifetime but whose paintings achieved record auction sales prices after his death. Van Gogh was active as an artist for only ten years, during which time he produced around 1000 watercolors, drawings and sketches and nearly 1250 paintings. His styles included an early dark, Realist style and a later colourful, intense,

	<p>expressionistic style. Almost more than on his oeuvre, his fame has been based on the extensive, diary-like correspondence he maintained, in particular with his brother, Theo. [ULAN] (en)</p> <p>- Vincent van Gogh werd geboren in het Brabantse Zundert, waar zijn vader dominee was. Zijn ooms zaten in de kunsthandel en daar begon ook Vincents loopbaan, bij Goupil & Co. in Den Haag. Na enkele jaren in de kunsthandel koos Van Gogh een andere richting. Hij werd hulppredikant in Engeland en later zendingswerker in België. In 1880 besloot hij kunstenaar te worden. Hij werkte enige maanden in Den Haag bij de schilder Anton Mauve, een aangetrouwd familielid wiens werk hij zeer bewonderde. In Den Haag ontmoette hij ook Breitner, met wie hij regelmatig op stap ging om buiten te tekenen. Daarna werkte Van Gogh vooral in Nuenen in Brabant, waar hij het boerenleven weergaf in sombere kleuren, zoals in de beroemde 'Aardappeleters'. Na een korte studie in Antwerpen (winter 1885-1886), vertrok Van Gogh naar Parijs. Daar schilderde hij een groot aantal zelfportretten. Via zijn broer Theo, die kunsthandelaar was, ontmoette hij veel Franse kunstenaars, zoals Toulouse-Lautrec en Paul Gauguin. Mede onder invloed van het impressionisme en het pointillisme Pointillisme Pointillisme is een werkwijze in de schilderkunst waarbij de verf in kleine stippen van verschillende kleuren naast elkaar op het doek wordt gezet. Voor het oog vervloeien de stippen - op een afstand gezien - min of meer tot een geheel. Deze techniek, al eerder gebruikt door de impressionisten, werd consequent en systematisch toegepast door kunstenaars als George Seurat en Paul Signac. Deze Franse schilders werden pointillisten of ook wel 'neo-impressionisten' genoemd. Hun manier van werken had veel invloed op andere Europese kunstenaars, zoals de Belg Theo van Rysselberghe en de Nederlandse schilder Jan Toorop. ontwikkelde hij zijn kenmerkende stijl met losse strekjes verf en heldere, felle kleuren. In 1888 verhuisde Van Gogh naar het Zuid-Franse stadje Arles, waar hij de omgeving en de bevolking schilderde. Vanwege geestelijke problemen werd de kunstenaar enkele malen opgenomen in een inrichting. Tussen de crises door bleef hij koortsachtig doorwerken, tot aan zijn dood in 1890. [Rijksmuseum website] (nl)</p>
Actor Related Creations	<ol style="list-style-type: none"> 1. De aardappeleters, see Creation MM0000004 2. Zonnebloemen 3. Beeldenstorm, see Creation MM0000001 4. Beeldenstorm, see Creation MM0000002 5. Mooiste plek van Nederland, see Creation MM0000003.
Actor Related Creations - Role	<ul style="list-style-type: none"> - 1. Creator - 2. Creator - 3. is depicted in; is discussed in - 4. is depicted in; is discussed in Creator - 5. is depicted in; is discussed in
Actor Related Creations - Date	<ul style="list-style-type: none"> - 1. 1885 - 4. 2006 - 5. 2006 - 6. 2003
Related Actors	<ol style="list-style-type: none"> 1. Cormon, Fernand [ULAN], see Actor MM00002 2. Gauguin, Paul [ULAN], see Actor MM00003 <p>Further related People listed in this ULAN record:</p> <ul style="list-style-type: none"> - assistant of Mauve, Anton 1881-1882³⁷ (Dutch artist, 1838-1888) [500009041] - child of Gogh-Carbentus, Anna Cornelia van (Dutch, subject of portraits³⁸, 1819-

³⁷ This is an example of a date element that describes this relationship.

	1907) [500089487] - collaborated with Breitner, George Hendrik (Dutch artist, 1857-1923) [500006861] - cousin of Mauve, Anton (Dutch artist, 1838-1888) [500009041]
Related Actors – Type	- 1. student of [ULAN] - 2. collaborated with [automatic extraction Rijksmuseum website]
Link to depiction	N/A
Related Websites	http://www3.vangoghmuseum.nl/vgm/index.jsp?lang=nl
Related Web pages	http://en.wikipedia.org/wiki/Category:Vincent_van_Gogh

³⁸ Although this Person is not an Actor in the context of MultiMatch, it is interesting to study the possibility of automatic extraction of this information into a relation between two artists (Actor A and Actor B) and between the portrayed artist (Actor B) and creations of the creator being described (Actor A).

Actor MM00002

Metadata (sub)element	Example
Identifier	URI to ULAN ID 500115385 [ULAN]
Actor Name	Cormon, Fernand [ULAN]
Affiliated Organisation	NA
Actor Main Roles	- Creator.Painter [ULAN] (en) - Creator.Teacher [ULAN] (en)
Date of birth	1845 [ULAN]
Date of death	1924 [ULAN]
Birth place	Paris (Ville de Paris department, Île-de-France, France) [ULAN; is actually derived from TGN] (fr)
Death place	Paris (Ville de Paris department, Île-de-France, France) [ULAN; is actually derived from TGN] (fr)
Description	Cormon was best known as a painter of historical and religious subjects in academic style, although he also produced some portraits. [ULAN] (en)

Actor MM00003

Metadata (sub)element	Example
Identifier	URI to ULAN ID 500011421 [ULAN]
Actor Name	Gauguin, Paul [ULAN]
Affiliated Organisation	NA
Actor Main Roles	- Creator.Painter [ULAN] (en) - Creator.Printmaker [ULAN] (en)
Date of birth	1848 [ULAN]
Date of death	1903 [ULAN]
Birth place	Paris (Ville de Paris department, Île-de-France, France) [ULAN; is actually derived from TGN] (fr)
Death place	Atuona (Hiva Oa, French Polynesia) [ULAN; is actually derived from TGN] (en)
Description	He was one of the leading French painters of the Postimpressionist period. He is noted for his imaginative subjects and expressive use of color, in attempts to capture a more primitive emotion in his works. He professed an appreciation of exotic peoples, whom he believed to be innocent of modern civilization's woes. [ULAN] (en)

Creation MM0000001

Metadata (sub)element	Definition
Creation identifier MultiMatch	MM00000001

Creation identifier institute	24278
Title	BEELDENSTORM (expr_id:24278) [iMMix]
Source	Catalogue of Sound and Vision as published on https://videotheek.surfnet.nl , also named iMMix.
Related Actor	<ul style="list-style-type: none"> - 1. AVRO [iMMix] - 2. Kuyper, Bas [iMMix] - 3. Os, Henk van [iMMix] - 4. Rawie, Marijke [iMMix] - 5. Stokkermans, Wendy [iMMix] - 6. Verbunt, Camille [iMMix]
Related Actor -Type	- 3. Contributor.Presenter (en) ; Maker.Presentator (nl)
Related Actor -Actor	Identifier of the Actor: unknown
Related Actor - Date	- 3. 2006-02-25 [iMMix]
Related Actor - Date - Type	- 3. Creation - 3. Broadcasting
Tags	NA
Subject	<ul style="list-style-type: none"> - portretten [iMMix] (nl) - zonnebloemen [iMMix] (nl) - schilderkunst [iMMix] (nl) - schilderijen [iMMix] (nl) - kunstschilders [iMMix] (nl)
Description	<p>BG_37102-out.wmv Teleblik expressie_id:#24278#</p> <p>SHOTS: - klein paneel met een voorstelling van verlepte zonnebloemen (1887) van Van Gogh (Rijksmuseum Vincent van Gogh in Amsterdam); - vier verlepte zonnebloemen van Van Gogh (Rijksmuseum Kröller-Müller in Otterlo); - div. voorstellingen van vaas met zonnebloemen van Van Gogh (Rijksmuseum Vincent van Gogh, Neue Pinakothek in München en National Gallery in Londen); - portret van Van Gogh die zonnebloemen schildert (1888) van Paul Gauguin; - portret van Madame Roulin "La Berceuse" (1889) van Van Gogh; - schets in een brief van Van Gogh van een drieluik, waarbij hij het portret van Madame Roulin laat flankeren door twee zonnebloemenschilderijen; - omslag van de catalogus van de eerste Van Gogh-tentoonstelling in Amsterdam in 1892 met een houtsnede van een zonnebloem, ontworpen door Richard Roland Holst.</p> <p>Serie wekelijkse programma's waarin Henk van Os, hoogleraar Kunst en Samenleving (Universiteit van Amsterdam), voorwerpen uit de Nederlandse musea bespreekt. In deze aflevering behandelt hij een aantal zelfportretten en studies van zelfportretten van de 19e-eeuwse kunstschilder Vincent van Gogh (1853-1890) in het Rijksmuseum Vincent van Gogh in Amsterdam. Van Os gaat oa in op Van Goghs techniek, zijn kleurgebruik en de wijze waarop andere kunstschilders, onder wie Paul Signac, hem mogelijkwijs hebben beïnvloed. [iMMix] (nl)</p>
Creation Location	Hilversum
Archive Location	Hilversum

Format Type	
Related Creations	1. Vier verlepte zonnebloemen 2. Vaas met zonnebloemen 3. Beeldenstorm, see Creation MM0000005
Related Creations – Type	1. is depicted in 2. is depicted in 3. is part of
Link to Digital representations	see DigRep MM00000001
Related Websites	
Related Web pages	
Copyrighted	Yes
Language	nl

Creation MM0000002

Metadata (sub)element	Definition
Creation identifier MultiMatch	MM00000002
Creation identifier institute	24279
Title	BEELDENSTORM (expr_id:24279) [iMMix]
Source	Catalogue of Sound and Vision as published on https://videotheek.surfnet.nl , also named iMMix.
Related Actor	1. AVRO [iMMix] 2. Kuyper, Bas [iMMix] 3. Os, Henk van [iMMix] 4. Rawie, Marijke [iMMix] 5. Stokkermans, Wendy [iMMix] 6. Verbunt, Camille [iMMix]
Related Actor –Type	3. Contributor.Presenter (en) ; Maker.Presentator (nl)
Related Actor -Actor	Identifier of the Actor: unknown ³⁹
Related Actor – Date	3. 2006-02-25 [iMMix]
Related Actor - Date – Type	3. Creation 3. Broadcasting
Tags	NA
Subject	- schilderkunst [iMMix] (nl) - schilderijen [iMMix] (nl) - kunstschilders [iMMix] (nl)

³⁹ Henk van Os is not described in ULAN. To be decided: how to process the Actors from the three catalogues that do not have a ULAN record.

Description	BG_37103-out.wmv Teleblik expressie_id:#24279# SHOTS: - zelfportret (1887) van Van Gogh met aan de achterzijde een studie/schets voor de "De Aardappeleters" en enkele andere studies/zelfportretten van Van Gogh; - portret van een meisje met loshangend haar (1885) van Van Gogh. [iMMix] (nl)
Creation Location	Hilversum
Archive Location	Hilversum
Format	
Type	Video.Documentary
Related Creations	1. De aardappeleters, see Creation MM0000004 2. portret van meisje met loshangend haar 3. Beeldenstorm, see Creation MM0000005
Related Creations – Type	1. is depicted in 2. is depicted in 3. is part of
Link to Digital representations	Ipv verwijzing: see DigRep MM00000002
Related Websites	
Related Web pages	
Copyrighted	Yes
Language	nl

Creation MM0000003

Metadata (sub)element	Definition
Creation identifier MultiMatch	MM00000003
Creation identifier institute	24279
Title	DE MOOISTE PLEK VAN NEDERLAND (4) (Uitz. 04-01-2003) [iMMix]
Source	Catalogue of Sound and Vision as published on https://videotheek.surfnet.nl , also named iMMix.
Related Actor	- 1. NCRV [iMMix] - 2. Palazzina TV Productions [iMMix] - 3. Zimmerman, Taco [iMMix] - 4. Kooi, Roel [iMMix] - 5. Rennings, Paula [iMMix] - 6. Lutz, Victorine [iMMix] - 7. Mourits, Monique [iMMix] - 8. Gogh, Vincent van
Related Actor -Type	- 8. is discussed in
Related Actor -Actor	- 8. 500115588
Related Actor - Date	- 1. 2003-01-04 [iMMix]

Related Actor - Date - Type	- 1. Creation - 1. Broadcasting
Tags	NA
Subject	- 1. Natuur [iMMix] (nl) - 2. landschappen [iMMix] (nl) - 3. natuurgebieden [iMMix] (nl) - 4. Gogh, Vincent van [iMMix] (nl), see Actor MM00001
Description	17:10:21 VINCENT IN HET VEEN Reportage waarin kunsthistoricus van Staatsbosbeheer Marcel van Ool gevolgd wordt naar hoogveengebied het Bargerveen. Hij legt uit hoe hij denkt dat Vincent van Gogh in zijn tijd dit landschap ervaren heeft en hoe deze nalatenschappen van oude kunst van belang zijn voor het reconstrueren van het oorspronkelijke landschap. Van Ool laat afgravingen van hompen turf zien. Van Ool vertelt al wandelend door het gebied dat het een van de laatste restjes natuur in Nederland is waar je al ronddwalend een stukje levendige geschiedenis meemaakt zonder dat het door een museale instelling is neergezet. Hij leest voor uit werk van Van Gogh. SHOTS : div. natuur van het Bargerveen; div. kunstwerken van Van Gogh (kort). [iMMix] (nl)
Creation Location	Hilversum
Archive Location	Hilversum
Format	File type: Windows Media
Type	Video.Tvprogramme
Related Creations	not specified, depiction is too short according to documentalist.
Related Creations – Type	
Link to Digital representations	http://videotheek.surfnet.nl/play_proxy/mmc/14210/BG_31162-out.wmv
Related Websites	
Related Web pages	
Copyrighted	Yes
Language	nl

Creation MM0000004

Metadata (sub)element	Definition
Creation identifier MultiMatch	MM00000004
Creation identifier institute	F 82 [Van Goghmuseum website]
Title	De aardappeleters
Source	http://www3.vangoghmuseum.nl/vgm/index.jsp?page=95&lang=nl
Related Actor	- 1. 500115588 Gogh, Vincent van , see Actor MM00001

Related Actor -Type	- 1. Creation
Related Actor - Date	- 1. 1885 [Van Goghmuseum website]
Related Actor - Date - Type	- 1. Creation
Tags	NA
Subject	- 1. Natuur [iMMix] (nl) - 2. landschappen [iMMix] (nl) - 3. natuurgebieden [iMMix] (nl) - 4. Gogh, Vincent van [iMMix] (nl)
Description	<i>De aardappeleters</i> was Van Goghs eerste grote schilderij met meerdere figuren. De kunstenaar maakte dit werk in april/mei 1885 en zag het als een soort meesterproef. Hij koos een moeilijke compositie om te bewijzen dat hij op weg was een goede figuurschilder te worden. Bovendien wilde hij een echte ‘boerenschilder’ zijn, net als de Franse meester Jean-François Millet die hij zeer bewonderde. Met <i>De aardappeleters</i> probeerde Van Gogh een heel realistisch ‘boerenschilderij’ te maken, zonder de werkelijkheid te idealiseren of - zoals andere schilders volgens hem deden - ‘zoetsappig’ te maken. Hij wilde benadrukken dat deze mensen ‘met de handen die ze in de schotel steken, zelf de aarde hebben omgespit en [...] dat zij hun eten zo eerlijk verdiend hebben’. Voor de gezichten streefde hij naar ‘de kleur van een goed stoffige aardappel, ongeschild natuurlijk’. [Van Goghmuseum website] (nl)
Creation Location	unknown
Archive Location	Amsterdam
Format	Olieverf op doek, 82 X 114 cm [Van Goghmuseum website] (nl)
Type	StillImage.Painting
Related Creations	1. Boerengezin aan tafel (Creator: Jozef Israëls) (1882) [Van Goghmuseum website] (nl)
Related Creations – Type	1. inspired by
Link to Digital representations	http://www3.vangoghmuseum.nl/vgm/mmbase/images/7464
Related Websites	
Related Web pages	http://www3.vangoghmuseum.nl/vgm/index.jsp?page=1303&collection=1294&lang=nl
Copyrighted	Yes
Language	NA

Creation MM00000005

Metadata (sub)element	Definition
Creation identifier MultiMatch	MM00000005
Creation identifier institute	180527

Title	Beeldenstorm [iMMix]
Source	Catalogue of Sound and Vision as published on https://videotheek.surfnet.nl , also named iMMix.
Related Actor	1. AVRO [iMMix] 2. Os, Henk van {presentatie} [iMMix] 3. Rawie, Marijke {eindredactie} [iMMix] 4. Popping, Fancy {productie} [iMMix] 5. Timmer, Eline {regie} [iMMix]
Related Actor – Type	2. Contributor.Presenter (en) ; Maker.Presentator (nl)
Related Actor - Actor	Identifier of the Actor: unknown
Related Actor – Date	1. 1997 – 2006 [iMMix]
Related Actor - Date – Type	1. Creation 1. Broadcasting
Tags	NA
Subject	- schilderkunst [iMMix] (nl) - schilderijen [iMMix] (nl) - kunstschilders [iMMix] (nl)
Description	Serie wekelijkse programma's waarin Henk van Os, hoogleraar Kunst en Samenleving (Universiteit van Amsterdam), voorwerpen uit de Nederlandse musea bespreekt. [iMMix] (nl)
Creation Location	Hilversum
Archive Location	Hilversum
Format	
Type	Video.Documentary
Related Creations	1. Episode 1 [fictitious], see Creation MM0000001 2. Episode 2 [fictitious], see Creation MM0000002
Related Creations – Type	- 1. has part - 2. has part
Link to Digital representations	
Related Websites	http://www.avromuseumtv.nl/zoeken/default.aspx?trefwoord=beeldenstorm
Related Web pages	
Copyrighted	Yes
Language	nl

DigRep MM00000001

Identifier - MultiMatch	DigRep MM00000001
Identifier - Source	BG_37102
Representation source	http://videotheek.surfnet.nl/
Type	Part
Format	Windows Media
File size	Bitrate 56 Kbps; Duration: 6 min 49 sec
Creation represented	Beeldenstorm 24278, see Creation MM0000001
Actor represented	Gogh, Vincent van , see Actor MM00001
Rights Holder	AVRO

Copyrighted	Yes
Low-level features	http://videotheek.surfnet.nl/play_proxy/mmc/18418/BG_37102-out.wmv

DigRep MM00000002

Identifier - MultiMatch	DigRep MM00000002
Identifier - Source	BG_37103
Representation source	http://videotheek.surfnet.nl/
Type	Part
Format	Windows Media
File size	Bitrate 56 Kbps; Durance: 6 min 46 sec
Creation represented	Beeldenstorm 24279, see Creation MM00000002
Actor represented	Gogh, Vincent van , see Actor MM000001
Rights Holder	AVRO
Copyrighted	Yes
Low-level features	http://videotheek.surfnet.nl/play_proxy/mmc/19542/BG_37103-out.wmv

Annex 11: TEL sample record

```
<?xml version="1.0" encoding="utf-8"?>
<dc:dc xmlns:bn="http://www.bs.dk/standards/MarcXchange" xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/" xmlns:lib="http://dublincore.org/usage/meetings/2003/06/dclib-
encodingschemes.html" xmlns:mods="http://www.loc.gov/mods"
xmlns:tel="http://krait.kb.nl/coop/tel/handbook/telterms.html"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<tel:recordId
xsi:type="dcterms:URI">http://opac.porbase.org/ipac20/ipac.jsp?profile=porbase&uri=full=3100024@!1
000194@!0&ri=1&aspect=basic_search&menu=search&source=192.168.0.17@!porbas
e&ipp=20&staffonly=&term=&index=&uindex=&aspect=basic_search&me
nu=search&ri=1</tel:recordId>
<dc:identifier xsi:type="lib:ISBN">972-9209-63-4</dc:identifier>
<dc:identifier>Cota (Call-Number): BN-B.A. 16950 V.</dc:identifier>
<dc:identifier>Cota (Call-Number): BN-B.A. 16950 V.</dc:identifier>
<dc:identifier>Cota (Call-Number): BPMP-7c 000023</dc:identifier>
<dc:identifier>URN:LDN:PT:127765/98</dc:identifier>
<dc:title>C. Teles: itiner&#xE2;ncias</dc:title>
<dcterms:alternative>C&#xE2;ndido Teles</dcterms:alternative>
<dc:creator>Universidade da Beira Interior</dc:creator>
<dc:subject>Teles, C&#xE2;ndido</dc:subject>
<dc:subject xsi:type="dcterms:UDC">75 Teles, C&#xE2;ndido (083.81), por</dc:subject>
<dc:subject xsi:type="dcterms:UDC">061.4, por</dc:subject>
<dc:description>Monografia</dc:description>
<dc:description>XXV anivers&#xE1;rio do ensino superior na Covilh&#xE3;, 1998</dc:description>
<dc:publisher>Univ. da Beira Interior</dc:publisher>
<dc:publisher>Covilh&#xE3;</dc:publisher>
<dc:date>1998</dc:date>
<dc:type>material textual, impresso</dc:type>
<dc:format>47 p., il., 30 cm</dc:format>
<dc:language xsi:type="dcterms:ISO 639-2">por</dc:language>
</dc:dc>
```