Mapping Encoded Archival Description to CIDOC CRM

Lina Bountouri and Manolis Gergatsoulis

Database & Information Systems Group (DBIS),
Laboratory on Digital Libraries and Electronic Publishing,
Department of Archives and Library Science, Ionian University,
Ioannou Theotoki 72, 49100 Corfu, Greece.
{boudouri, manolis}@ionio.gr

Abstract. In this paper we analyze the semantics of the archival description, expressed through the Encoded Archival Description. Through this analysis it is concluded that an EAD document is a hierarchy of documentation elements and attributes and that through this documentation the archive is semantically expressed through three different hierarchies (hierarchy of physical objects, hierarchy of information objects and hierarchy of linguistic objects). The semantic views of the archive as well as their interrelationships are mapped to the CIDOC CRM.
Keywords: Metadata interoperability, Encoded Archival Description, Ontologies, CIDOC CRM, Mappings.

1 Introduction

Cultural heritage institutions, archives, libraries, and museums host and develop various collections with heterogeneous types of material, often described by different metadata schemas. Managing metadata as an integrated set of objects is vital for information retrieval and (meta)data exchange. To achieve this, *interoperability* techniques have been applied. One of the widely approved and implemented techniques is the *Ontology-Based Integration*. Ontologies play a vital role in semantic interoperability and integration scenarios, and they are preferred in regard to other schemas, due to their ability to conceptualize particular domains of interest and express their rich semantics in a formal manner. One of their main roles in an interoperability scenario is to act as the *mediated schema* between heterogeneous information systems [14, 4].

This paper builds upon an ontology-based metadata integration architecture, which considers the *CIDOC Conceptual Reference Model* (*CIDOC CRM*) ontology [3] as the *mediator*. The proposed architecture considers a set of data sources each of them providing information encoded by a different metadata schema (e.g. EAD, VRA, DC, MODS, etc). Each schema is semantically mapped to the CIDOC CRM based mediator, which may also retain its own database of metadata encoded in CIDOC CRM format. Various integration scenarios can be built on this architecture.

The main research result of this paper is the mapping of the *Encoded Archival Description* (*EAD*) [9] to CIDOC CRM. In order to create this mapping, we firstly analyzed the main concepts of the archive and of its components parts, as well as the main concepts of the archival description, which are the hierarchical structure and the inheritance of information between the hierarchical levels of description. These concepts (being expressed through EAD) have to be mapped to the ontology so as to promote the semantic integration. Part of the mapping procedure was to properly define these highly complex semantic structures in order to be expressed by the CIDOC CRM. Furthermore, the EAD descriptive fields must be also mapped to the ontology. This research work is the first complete effort to define the semantic mappings of the EAD to the CIDOC CRM.

2 Preliminaries

2.1 CIDOC Conceptual Reference Model

The *CIDOC CRM* is a core ontology, which consists of a hierarchy of 86 *entities* (or *classes*) and 137 *properties*. A *class* (also called *entity*) groups items (called *class instances*) that share one or more common characteristics. A class may be the *domain* or the *range* of *properties*, which are binary relations between classes. An *instance of a property* is a relation between an instance of its domain and an instance of its range. A property can be interpreted in both directions (active and passive voice), with two distinct but related interpretations. A *subclass* is a class that specializes another class (its *superclass*). A class may have one or more immediate superclasses. When a class A is a subclass of a class B then all the instances of A are also instances of B. A subclass inherits the properties declared on its superclasses without exception (*strict inheritance*) in addition to having none, one or more properties of its own.

A *subproperty* is a property that specializes another property (its *superproperty*). If a property P is a subproperty of a property Q then a) all instances of P are also instances of Q, b) the domain of P is the same or a subclass of the domain of Q, and c) the range of P is the same or a subclass of the range of Q. Some properties are associated with an additional property (called *property of property*) whose domain contains the property instances and whose range is the class E55 Type. Properties of properties are used to specialize the meaning of their parent properties. A sample of CIDOC CRM properties is shown in Table 1.

CIDOC CRM expresses semantics as a sequence of path(s) of the form entity-property-entity. It is an event-based model and its main notions are the temporal entities. As a consequence, the presence of CIDOC CRM entities, such as actors, dates, places and objects, implies their participation to an event or activity [11].

2.2 Encoded Archival Description

The *archival description* documents the *archive*, which is a complex set of materials sharing common provenance, regardless of form or medium. The description involves a hierarchical and progressive documentation, beginning from the

Property id & Name	Entity - Domain	Entity - Range
P1 is identified by (identifies)	E1 CRM Entity	E41 Appellation
P2 has type (is type of)	E1 CRM Entity	E55 Type
P14 carried out by (performed)	E7 Activity	E39 Actor
P67 refers to (is referred to by)	E89 Propositional Object	E1 CRM Entity
P70 documents (is documented in)	E31 Document	E1 CRM Entity
P71 lists (is listed in)	E32 Authority Document	E55 Type
P102 has title (is title of)	E71 Man-Made Thing	E35 Title
P106 is composed of (forms part of)	E90 Symbolic Object	E90 Symbolic Object
P108 has produced (was produced by)	E12 Production	E24 Physical Man-Made Thing
P128 carries (is carried by)	E24 Physical Man-Made Thing	E73 Information Object

Table 1. A sample of CIDOC CRM properties.

archive, and proceeding with its sub-components, the sub-components of sub-components, and so on, often reaching the item level (e.g. a map). In parallel, it supports the inheritance of information between the hierarchical levels. *Finding aids* materialize archival descriptions and the *EAD* [9, 8] is the most widely used schema that supports the creation of electronic finding aids. An *EAD document*, starting from the ead root element, consists of three elements: the *EAD Header* (eadheader), which is the mandatory element including the metadata for the EAD document, the *Front Matter* (frontmatter), which carries optional information for the printed finding aid (if any), and the mandatory *Archival Description* (archdesc), which provides information on the archive's content and context of creation, such as:

- core identification information (incorporated in the did element), e.g. the archive's creator (origination) and title (unittitle),
- administrative and supplemental information that facilitate the use of the archival materials, such as the biography or history (bioghist), and
- description of the components, bundled in a wrapper element called dsc that encodes the hierarchical groupings of the archival components being described. An archival component is an easily recognizable archival entity, characterized by an attribute level as series, subseries, file, item etc, and it may be in any level within the hierarchical structure of the description. Components are deployed as nested elements, called either c or c01 to c12.

Example 1 presents an archival description on the level of fonds. Basic descriptive identification information for the archive, such as the title (unittitle), the creation date (unitdate), the identifier of the archive (unitid) and its creator (origination), is given inside the did element. Administrative and supplemental information is provided through the bioghist and controlaccess elements. Description of subordinate components is presented inside the dsc element, where two components are provided through c01 elements (both on the level of series) and include basic identification information, such as unittitle, unitdate, etc.

Example 1. In this example a fragment of an EAD document is presented:

```
<ead>
 <eadheader>...</eadheader>
 <archdesc level="fonds">
  <did>
    <unitid countrycode="GR" repositorycode="IU">ARC.14</unitid>
    <unittitle>Ionian University Archive</unittitle>
    <unitdate>1984 - 2007</unitdate>
    <origination>
      <corpname>Ionian University</corpname>
    </origination>
  </did>
  <bioghist>
    <p>The Ionian University was founded in 1984...</p>
  </bioghist>
  <controlaccess>
    <corpname>Ionian University</corpname>
  </controlaccess>
  <dsc>
   <c01 level="series">
    <did>
      <unitid countrycode="GR" repositorycode="IU">ARC.14/1</unitid>
      <unittitle>R. C. Archives</unittitle>
      <unitdate>1998 - 2007</unitdate>
      <origination>
        <corpname>I. U. Research Committee</corpname>
      </origination>
    </did>
   </c01>
   <c01 level="series">
    <did>
      <unitid countrycode="GR" repositorycode="IU">ARC.14/2</unitid>
      <unittitle>I. U. Library Archives</unittitle>
      <unitdate>1998 - 2000</unitdate>
      <origination>
        <corpname>I. U. Library</corpname>
      </origination>
    </did>
   </c01>
  </dsc>
 </archdesc>
</ead>
```

3 The archive and the archival description: the main concepts

According to [6] "*an ontology is a specification of a conceptualization*". More specifically, the CIDOC CRM ontology is the specification of the Cultural Heritage conceptualization. Based on that fact, a necessary step that must be taken

before the mapping of a metadata schema to a domain ontology is to capture its concepts, aiming to map them to the ontology. In general terms, the concepts of a metadata schema are related to:

- the semantics of the description (in this case, the semantics of the archival description),
- the semantics of the information resource they describe (in this case, the semantics of the archive), and
- the semantics of its descriptive fields (in this case, the descriptive fields - elements and attributes - of the EAD).

The main semantic concepts of an archive, expressed through its description, are [13]:

- the archive is a *physical object* that acts as evidence for the functions/activities of the human or of the corporate body that created it, and
- the archive is an *information object* that includes information in different *formats* and *languages*.

The basic characteristic of the archive and of the archival description is the hierarchical and multilevel tree-based structure including also the principal of inheritance of information. An archive usually consists of a large number of components, which form the hierarchical relationship from the upper level of description (e.g. the archive) to the lower levels of description (e.g. the subfonds, the series, the files etc).

As far as the hierarchical structure is concerned, since an archive follows it, its semantic concepts are also expressed through this structure. As a result, an archive as a set of physical objects may contain one or more subfonds, which are also a set of physical objects and they may also contain one or more series, which are also a set of physical objects. In parallel, an archive as a set of information objects consists of one or more information objects, for instance the subfonds, which in turn consists of one or more information objects, such as the series etc.

The archival description is expressed in a machine readable way through the EAD. The EAD includes - apart from the archival description - the metadata of the EAD document and of the archival description. To express this documentation, an EAD document is structured as a tree having as root the element ead, which includes three subelements: the eadheader, the frontmatter and the archdesc.

Analytically, the root element ead includes the whole EAD document. The element eadheader includes the metadata of the machine readable archival description and the element frontmatter includes information for the creation, publication and use of the finding aid. Finally, the archdesc element includes the description of the archive and of its components (c01-c12 and c) defining also the hierarchical and multilevel tree-based structure, according to Figure 1.

In this figure, an illustrative structure of an archive is expressed through the EAD and in particular through the archdesc and its subelements c01-c05 for the components. Note that the description of the archive is expressed though

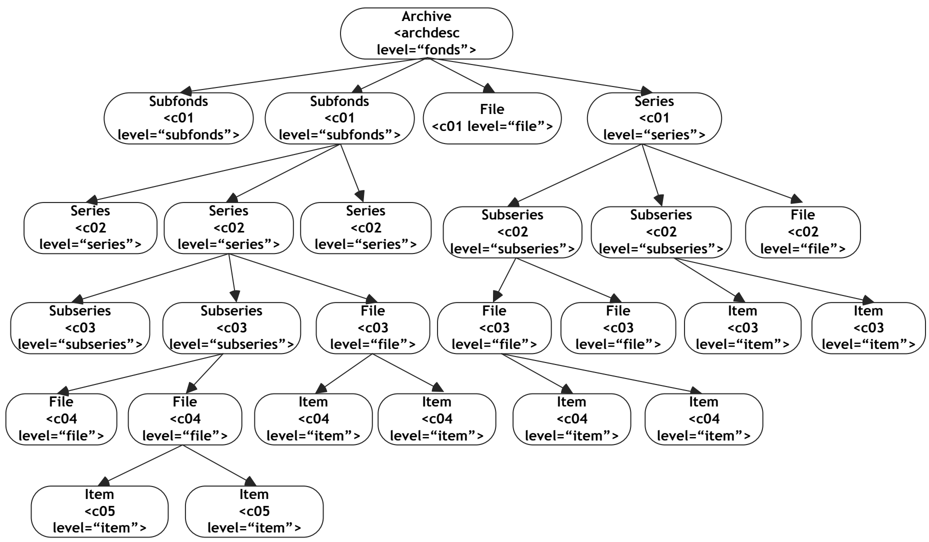

Fig. 1. The illustrative structure of an archive expressed through EAD.

the element archdesc declaring also the level of description which is the fonds (level="fonds"). The first level components (being one level lower than the archive) are expressed through the element c01, defining also the level of description for every archival entity that each c01 represents, for instance the subfonds (level="subfonds"), the series (level="series") and the file (level="file"). Lower levels may follow.

It is important to notice that for every archival entity various XML elements and attributes are implemented as the descriptive fields, in which archivists can provide all the necessary information for the archive and its components.

Consequently, in order to define the semantic mapping of the EAD to the CIDOC CRM, the following concepts must be mapped:

- the tree-based hierarchical structure of the archive and of the archival description, which is expressed through the archdesc, c01-c12 and c elements, and the inheritance property of the archival description,
- the semantic views of the archive, and
- the descriptive fields, which are expressed through the XML subelements and attributes of the archdesc, c01-c12 and c elements.

In this paper, emphasis is given to the mapping of the subelements' and attributes' semantics for the archdesc, c01-c12 and c elements, given that they encode the documentation of the archive.

4 The archive and the archival description: the mapping of the main concepts

4.1 The EAD document as a hierarchy of documentation elements and attributes

As already mentioned, the ead root element includes the whole documentation of the EAD document. The documentation concept is expressed in CIDOC CRM through the class E31 Document, which includes instances that are immaterial objects defining and documenting the reality, such as the sentences of a text, the databases etc. As a result, the ead element is mapped to this class, creating and mapping the whole EAD document to an instance of this class.

Respectively, the eadheader, frontmatter and archdesc elements are also mapped to instances of the class E31 Document, since: a) the eadheader semantically includes the documentation of the machine readable archival description, b) the frontmatter includes the documentation of the printed finding aid, and c) the archdesc includes the documentation of the archive. Provided that the c01-c12 and c elements "carry" the documentation of the archival components, they are also mapped to instances of the E31 Document class.

The aforementioned instances of the E31 Document class express the semantics of the main EAD elements that form the basic structure of an EAD document. What is more, the archdesc, c01-c12 and c elements express at the same time the structure of the archival description, which is one of the main archival characteristics that must be mapped to the ontology. The hierarchical structure between the instances of the E31 Document class representing the ead and the eadheader, frontmatter, archdesc, c01-c12 and c elements is expressed in the CIDOC CRM ontology starting by the instance of the E31 Document representing the ead element. From this point, three new paths begin leading to three instances of the E31 Document class representing respectively the mapping of eadheader, frontmatter and archdesc. The instance of the E31 Document class representing the root element ead is linked through the P106 is composed of property to the instances of these three classes.

Correspondingly, the instances of the E31 Document representing the archival components (c01-c12 and c) are linked between them as part of the tree-based hierarchical structure via the P106 is composed of property. The tree structure obtained by mapping the EAD structure to the ontology is named as the "*Hierarchy of Documentation Elements and Attributes*" ("*HDEA*") and it is pictured in Figure 2.

4.2 The archive as a hierarchy of physical objects

An archive is a physical object, since it is a physical product of a person, a family or of a corporate body [13]. In addition, an archive as a physical object has an internal, well defined structure. In other words, an archive physically includes its components parts, which in turn include other components parts and so forth. Therefore, these archival physical objects also follow the hierarchical and

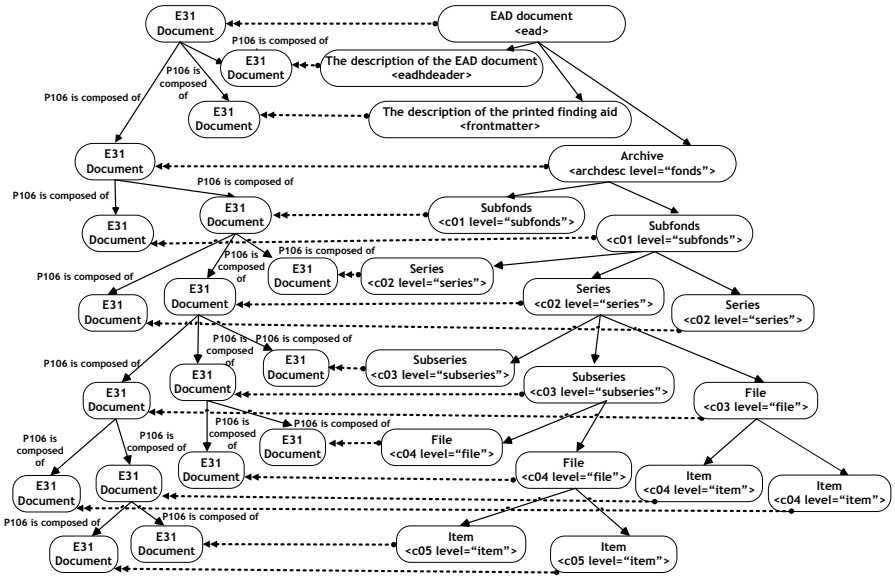

Fig. 2. Hierarchy of Documentation Elements and Attributes (HDEA).

multilevel structure. The structure of the archive as a physical object is hence expressed though the archdesc, c01-c12 and c.

In CIDOC CRM, the E22 Man-Made Object class defines the instances of the physical objects that have been created by human activity. According to this definition, every physical object expressed in EAD through the archdesc, c01-c12 and c elements is mapped to an instance of this class.

Moreover, in order to map their in between hierarchical relationship, these instances are linked via the P46 is composed of property. As it is presented in Figure 3, the tree structure obtained by mapping the archive and its components as a set of physical objects to the ontology is named as the "*Hierarchy of Physical Objects*" ("*HPO*").

4.3 The archive as a hierarchy of information objects

An archive is also an information object, since it carries information in one or more languages. An archive serves different purposes (for instance information purposes) and it is not only an evidence of the activity that produced it [13]. Both the archive and its component parts carry information. In detail, an archive contains information on its components as a set; an archival component (e.g. a subfonds) contains information on its components as a set and so on. For that reason, the informational aspect of the archive and of its components follow the hierarchical and multilevel tree structure.

To map to the CIDOC CRM ontology the concept of the archive as an object carrying information, the E73 Information Object class is used. This class includes

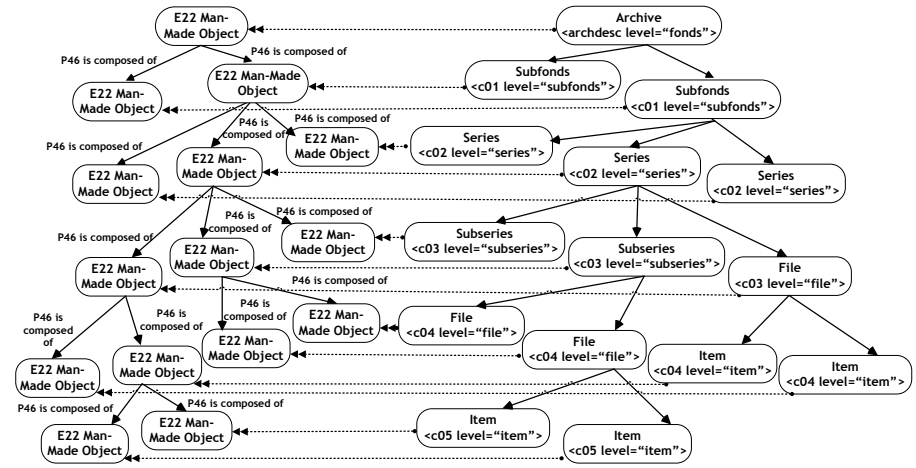

Fig. 3. Hierarchy of Physical Objects (HPO).

instances for the immaterial objects, which can be carried through any carrier. This semantic analysis comes to fully express the informational aspect of the archive, which is indeed immaterial and independent of any medium carrier [7].

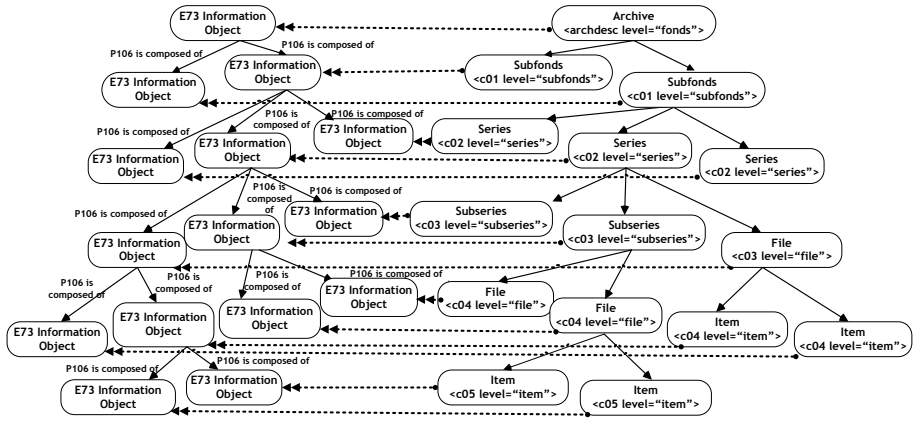

Fig. 4. Hierarchy of Information Objects (HIO).

In this context, the archdesc, c01-c12 and c elements are mapped to instances of the E73 Information Object. The expression of the hierarchical structure between these instances is defined through the P106 is composed of property. The tree structure obtained by mapping the archive and its components as a set of information objects to the ontology (Figure 4) is named as the "*Hierarchy of*

Information Objects" ("*HIO*") and it maps the semantics and the structure of the archive as an information object.

4.4 The archive as a hierarchy of linguistic objects

As mentioned in Section 4.3, an archive carries information in one or more languages, hence it is also a linguistic object. In CIDOC CRM, the E33 Linguistic Object class contains instances of information that can be expressed in one or more languages. Consequently, the semantic combination of the E73 Information Object and of the E33 Linguistic Object classes covers the semantic view of the archive as an information and linguistic object. Aiming to express these semantics, the archdesc, c01-c12 and c elements are mapped to instances of the E33 Linguistic Object class.

The expression of the hierarchical structure between these instances is defined through the P106 is composed of property, creating a hierarchy that maps the semantics and the tree structure obtained by mapping the archive and its components as a set of linguistic objects to the ontology. This tree is named as the "*Hierarchy of Linguistic Objects*" ("*HLO*") and it is pictured in Figure 5.

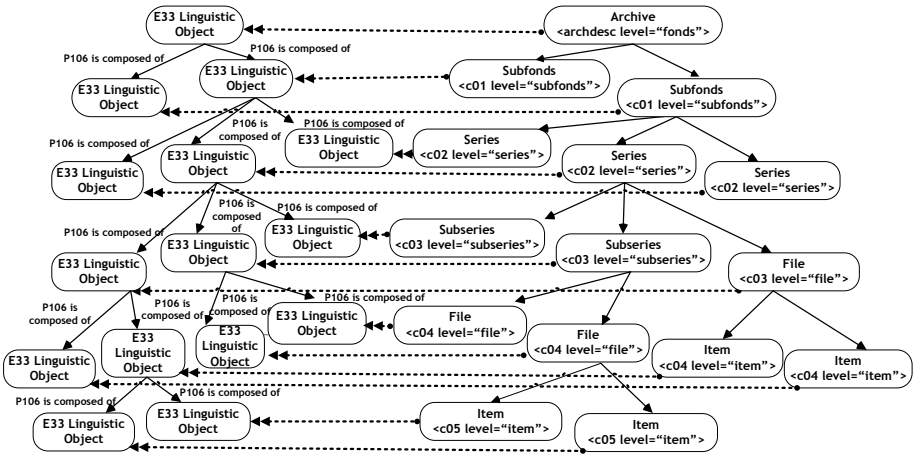

Fig. 5. Hierarchy of Linguistic Objects (HLO).

5 The relationships between the semantic views of an archive

Based on the mapping of the semantic views of the archive to the CIDOC CRM ontology, we conclude that the archive and its components are mapped to three

different CIDOC CRM hierarchies (*HPO*, *HIO* and *HLO*), each of them representing a different structured semantic view of the archive. Besides, the archival description is mapped to another hierarchy of the ontology (*HDEA*).

Note that these four hierarchies have the same structure and they only differ in terms of the names of the classes appearing in the tree nodes. It is clear that these hierarchies refer to the same object (the archive), which is documented by the archival description. Moreover, based on the analysis of Section 4, it is concluded that these hierarchies are semantically related to each other. Hence, it is necessary to: a) relate the four hierarchies with the tree of the EAD (and in particular with the archdesc, c01-c12 and c elements), since it is the metadata schema that expresses the archival description, and b) to associate these four hierarchies, since they all refer to the same object, the archive.

As the archdesc, c01-c12 and c elements are firstly referred to the archival description, which incorporates the semantic views of the archive, the *HDEA* is the starting point for the association of the different views. In detail, the *HDEA* refers to the archive as a physical object. Furthermore, the archive as a physical object carries information. Moreover, the analysis of this information can produce additional information for the archive. An illustrative example is the abstract of the archive's content as well as the controlled access points. Finally, an archive can also be a carrier of linguistic content, since the information it carries is usually expressed via written and/or oral speech, independently of the medium that carries this content.

In order to show an example of the hierarchies' association, the node <c01 level="subfonds"> of the EAD structure is chosen, and more specifically the node that contains three archival series in the Figure 1. This node is mapped to an instance of the E31 Document class expressing the documentation of this specific node that represents a subfonds. In detail, this instance documents its corresponding node in the *HPO* (see Figure 3), which is an instance of the E22 Man-Made Object class that represented the subfonds as a physical object. This relationship is expressed in the CIDOC CRM ontology through the P70 documents property that has as a domain the instances of the E31 Document class and as range the instances of the E1 CRM Entity class. For that reason, it can associate the instance of the E31 Document class to its corresponding instance of the E22 Man-Made Object class (since E22 Man-Made Object is a subclass of E1 CRM Entity).

To continue with, the subfonds as a physical object carries information and thus it is also an information object, hence an instance of the E73 Information Object maps the <c01 level="subfonds"> node in the *HIO* of the Figure 4. The relationship between these two instances (i.e. the instance of E22 Man-Made Object and E73 Information Object representing the same element (<c01 level="subfonds">) can be expressed through the P128 carries property, which has as domain the E24 Physical Man-Made Thing class and as range the E73 Information Object class. For that reason, it relates the instance of the E22 Man-Made Object class (which is a subclass of the E24 Physical Man-Made Thing

class) to the component of the archive documented in the instance of the E73 Information Object class.

The component of the archive documented in the <c01 level="subfonds"> element may also be an information object that carries information in one or more languages and this semantic view can be expressed as an instance of the E33 Linguistic Object, being in the same position in the *HLO* as it is in the *HIO* (see Figure 5). The relationship between these two instances is expressed in the CIDOC CRM ontology through the P67 refers to property, which has as a domain the E89 Propositional Object and as a range the E1 CRM Entity, hence linking the instance of the E73 Information Object (which is a subclass of the E89 Propositional Object) to its corresponding instance of the E33 Linguistic Object (which is a subclass of the E1 CRM Entity).

As a consequence, these four hierarchies are linked in a way that allows the expression of their in between relationship inside the CIDOC CRM ontology. This "chain of relationships" is expressed through the following CIDOC CRM path:

> E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P128 carries → E73 Information Object → P67 refers to → E33 Linguistic Object

This path declares that an EAD document includes the archival description (E31 Document → P106 is composed of →), which is the documentation (E31 Document → P70 documents) of a physical object that has been created by human activity (E22 Man-Made Object) and that carries (P128 carries) information which is immaterial and can be carried by any physical medium (E73 Information Object). To finish, the information carried by the archive can be expressed in one or more languages (P76 refers to → E33 Linguistic Object).

This "chain of relationships" expresses in the CIDOC CRM ontology the semantics for every archival unit (encoded in archdesc, c01-c12 and c) defining a horizontal relationship between them in every descriptive level. Therefore, the instances representing the archival units and being expressed in a vertical relationship inside the four hierarchies (*HDEA*, *HPO*, *HIO* and *HLO*) are also interconnected horizontally so as to express the relationship between the different semantic hierarchies of the archive and its description (see Figure 6).

6 Associating the EAD descriptive fields with the semantic hierarchies

Besides the mapping of the archdesc, c01-c12 and c elements studied in the previous sections, the mappings for the EAD descriptive fields that include the information for the content and context of the archive are also provided.

With the intention of defining the mappings of these elements/attributes to the CIDOC CRM, we are based on their semantics as they appear in the EAD Tag Library [8] and the published best practices and implementation guidelines for the EAD (for example the [10]). Derived from this investigation, we associate

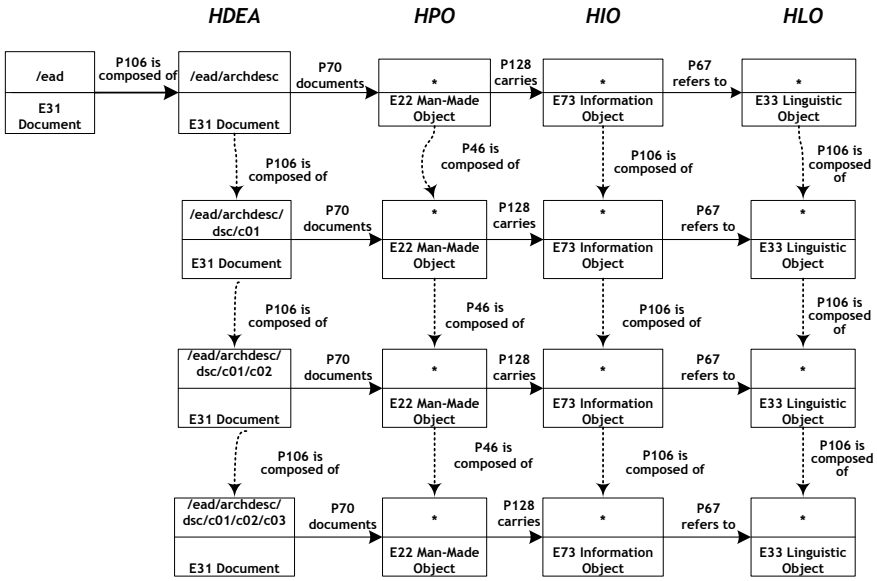

Fig. 6. Parallel Hierarchies.

the information content of the elements and attributes to one or more of the semantic hierarchies. More specifically, to map these elements/attributes to the CIDOC CRM the following steps are followed:

- Step 1: Associate the element/attribute to the semantic hierarchy(ies).
- Step 2: Select an appropriate CIDOC CRM class to map the element/attribute.
- Step 3: Associate the class selected in Step 2 (by constructing the appropriate paths) with the proper semantic hierarchy(ies) selected in Step 1.

6.1 Association of the element/attribute to the semantic hierarchy(ies)

A class' instance (on which an element/attribute is mapped) may be associated to: a) the *HDEA*, when it provides information for the archival documentation, b) the *HPO*, when it provides information for the archive as a physical object, c) the *HIO*, when it provides information for the archive as an information object, and d) the *HLO*, when it provides information for the archive as a linguistic object.

In Table 2 several EAD elements and attributes and the semantic hierarchy(ies) they are associated with are presented. In order to come up with this proposal the semantics of every node, and based on them, its association with one or more hierarchies are defined. In the following paragraphs, examples of nodes associated to the four hierarchies are presented.

More analytically, the information that refers to the EAD document and the archival description is semantically associated with the *HDEA*. For instance, the attributes of the archdesc element are referred to the *HDEA*, provided that they encode information for the archival description. Illustrative examples are the attributes audience and relatedencoding[1]:

- audience: This attribute provides information to help controlling whether the information contained in the element (to which the audience is attached) should be available to all viewers or only to the repository staff.
- relatedencoding: This attribute defines a descriptive encoding system, such as MARC 21, to which certain EAD elements can be mapped using the encodinganalog attribute.

Hence, given their meaning, both attributes are associated with the *HDEA*.

Subnode of the archdesc or c01-c12	HDEA	HPO	HIO	HLO
@audience	x			
@level	x	x		
@otherlevel	x	x		
@relatedencoding	x			
@type	x			
accessrestrict		x	x	
altformavail			x	
arrangement		x	x	
bioghist		x		
controlaccess			x	
fileplan		x	x	
phystech		x	x	
relatedmaterial			x	
scopecontent			x	
separatedmaterial		x		
userestrict			x	
did/unittitle			x	
did/note			x	
did/physloc		x		
did/unitdate		x		
did/langmaterial				x
did/unitid		x	x	
did/origination		x		

Table 2. The association of some EAD nodes with the semantic hierarchies.

The nodes that refer to the archive as a physical object have as their point of reference the *HPO* and, as a consequence, the E22 Man-Made Object class. These nodes are mostly part of the did wrapper element or they are part of the administrative and supplemental information for the archive. Illustrative examples are the creator of the archive (origination), its date of creation (unitdate), its

[1] Note that this attribute is an attribute of the archdesc and not of the c01-c12 and c.

physical location (physloc) etc. An example of an element associated with the *HPO* is the following:

- origination: This element provides information about the individual organization responsible for the creation, accumulation, or assembly of the described materials. The activities of creating, accumulating or assembling the archival material are all associated with its physical substance. Thus, its association with the *HPO* is obvious.

Furthermore, most of the administrative and supplemental information included in the archdesc, c01-c12 and c elements refers to the informational aspect of the archival material, which is expressed by the instance of the E73 Information Object. This information is provided from the archive and sometimes it comes up after its content analysis, such as the scope and content of the archive (scopecontent), its custodial history (custodhist) etc. What is more, certain subelements of the did wrapper element (such as the unittitle and abstract) refer to the E73 Information Object. For example:

- unittitle: This element declares the title of the archival unit, which is a name either given by the archivist or expressed by the archival unit. Thus, the unittitle is an information provided by the archival unit or by the archivist (after its context and content analysis), hence it is associated with the *HIO*.

The archive is also a linguistic object, since it can carry verbal or oral speech. For this reason, there are nodes that are associated with the *HLO*. Currently in EAD, there is only one element referred to this semantic hierarchy, the langmaterial, provided that this element includes a prose statement enumerating the language(s) of the archival materials found in the unit being described.

It is important to notice that - while analyzing the semantics of certain subnodes of the archdesc, c01-c12 and c elements we conclude that they are associated with more than one of the four hierarchies and this fact arises from their semantics. For example, the unitid element defines the identifier of the archival unit, which is a unique reference point for it or a control number, such as the accession number or the classification number, and sometimes it secondarily provides location information. Hence, this element refers to the descriptive unit as a physical object (when it identifies the archival unit to its accession or its location), nevertheless it is also information given by the archivist in order to uniquely identify the item. Thus, the unitid is associated both to the *HPO* and the *HIO*.

6.2 Selection of a CIDOC CRM class to map the elements/attributes and its association with the semantic hierarchy(ies)

In Section 6.1, we presented how an element/attribute is associated with the appropriate semantic hierarchy(ies) based on the semantics of this element/attribute. The next step that must be followed is to map this node to an appropriate class.

Then, this class must be connected to the appropriate node of the semantic hierarchy (i.e. the node that corresponds to the archival component to which the node refers to) through an appropriate constructed CIDOC CRM path. This path consists of a single CIDOC CRM property; often it includes several properties and intermediate classes.

The presentation of the mappings of the EAD nodes is beyond the scope of this paper. Nonetheless, in the following paragraphs, some examples are presented to show the above mentioned paths. As you will see below, the relatedencoding is mapped to a CIDOC CRM path that includes several properties and intermediate classes, while the langmaterial is mapped to a CIDOC CRM path that consists of a single CIDOC CRM property"

- relatedencoding: The relatedencoding attribute includes values that define the descriptive encoding system to which the EAD elements can be mapped and, as already mentioned, it is associated with the *HDEA*. It is semantically mapped to the E55 Type, which is also semantically associated with the E32 Authority Document in order to define that the E55 Type instances are taken from an authoritative vocabulary named "relatedencoding". The EAD path (/ead/archdesc/@relatedencoding) is mapped to the following CIDOC CRM path: E31 Document → P106 is composed of → E31 Document → P2 has type → E55 Type → P71 lists in → E32 Authority Document{=relatedencoding}, declaring that the EAD documentation (E31 Document) consists of (P106 is composed of) the documentation of the archive (E31 Document), which has a specific type (P2 has type → E55 Type) and that this type is characterized (P71 lists in) as relatedencoding (E32 Authority Document{=relatedencoding}).
- langmaterial: This element encodes the language(s) in which the archive is written or expressed and it is mapped to an instance of the E56 Language, which comprises the natural languages. Based on its semantics, it is associated with the *HLO*. The EAD path (/ead/archdesc/did/langmaterial/language) is mapped to the following CIDOC CRM path: E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P128 carries → E73 Information Object → P67 refers to → E33 Linguistic Object → P72 has language → E56 Language. This path expresses that the EAD documentation (E31 Document) consists of (P106 is composed of) the archive's documentation (E31 Document) that documents (P70 documents) a physical object (E22 Man-Made Object), which carries (P128 carries) an information object (E73 Information Object). Additionally, the archive is a linguistic object (P67 refers to → E33 Linguistic Object), which is expressed in one or more languages (→ P72 has language → E56 Language).

7 Discussion and related work

The key problem of integrating XML metadata schemas is an issue of great concern to the international research community. However, in most integration efforts no emphasis is given to the mapping of the semantics and of the documentation's targets of an XML metadata schema, even though these characteristics

shape the area of the metadata in archives, libraries and museums and that are indeed based on documentation policies followed for many years.

According to the literature, there are many XML (meta)data mapping to the CIDOC CRM ontology efforts, since this ontology is considered one of the most appropriate models in integration architectures. An example is the work of the *STAR* project [5], in which access to digital archaeological sources is enhanced through the mapping of them to an extension of the CIDOC CRM. Furthermore, the issue of mapping the Cultural Heritage metadata schemas to the ontology is also explored in the *BRICKS* project [15].

A well documented research proposal in relation to the mapping of the EAD semantics is presented in [16]. This mapping of EAD to CIDOC CRM ontology differs from the proposed mapping of our research work on the following points:

- this mapping refers to the first version of the EAD,
- the different semantic views of the archive and of the archival description are not defined and analyzed, hence not mapped to the ontology, and
- the EAD is considered as a format for describing the whole and there is no reference in mapping its hierarchical structure.

In general, the semantics of the metadata and of the information sources they describe are not taken into account while creating their mappings to an ontology. In [1] the mapping of the XML metadata schema of the Cultural Heritage domain to an ontology (which is similar to the CIDOC CRM) is proposed, nevertheless there is no reference to the importance of the metadata semantics.

The proposed mapping of the EAD to the CIDOC CRM ontology is targeted not only to capture the syntactic rules, but also to express the rich semantics of the EAD and of the information source it describes. The main goal is to be able to use this mapping in various integration scenarios that implement the CIDOC CRM as the mediated schema. It should be also noted that other mappings work of our team have been proposed, for schemas such as the DC and the VRA to the CIDOC CRM ontology (respectively presented in [12, 2]).

To conclude, we are currently working on the issue of the inheritance of information. Note that the inheritance of information between the hierarchically linked descriptive levels is one of the main characteristics of the archival description. Thus, specific techniques are needed in order to take into account this characteristic, during the mapping of an EAD document to the CIDOC CRM, otherwise considerable information may be lost.

References

1. B. Amann, C. Beeri, I. Fundulaki, and M. Scholl. Ontology-Based Integration of XML Web Resources. In I. Horrocks and J. Hendler, editors, *The 1st Int. Semantic Web Conference, Sardinia, Italy, June 912, 2002 Proceedings*, volume 2342 of *LNCS*, pages 117–131. Springer, 2002.
2. C. Kakali and I. Lourdi and T. Stasinopoulou and L. Bountouri and C. Papatheodorou and M. Doerr and M. Gergatsoulis. Integrating Dublin Core metadata

for cultural heritage collections using ontologies. In *Proc. of the Int. Conference on Dublin Core and Metadata Applications (DC 2007), Singapore, 27 - 31 August*, pages 128–139, 2007.

3. CIDOC CRM Special Interest Group. Definition of the CIDOC Conceptual Reference Model, version 5.0.2. Technical report, January 2010.
4. I.F. Cruz and H. Xiao. The Role of Ontologies in Data Integration. *Journal of Engineering Intelligent Systems*, 13(4):245–252, 2005.
5. D. Tudhope and C. Binding and K. May. Semantic interoperability issues from a case study in archaeology. In S. Kollias and J. Cousins, editor, *Semantic Interoperability in the European Digital Library, Proc. of the 1st Int. Workshop SIEDL 2008, associated with 5th ESWC*, pages 88–99, 2008.
6. T. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.
7. International Council on Archives. Committee on Descriptive Standards. *ISAD(G): General International Standard Archival Description*. ICA, 2nd edition, 2000.
8. Library of Congress. Encoded Archival Description Tag Library Version 2002. `http://www.loc.gov/ead/tglib/index.html`, 2002.
9. Library of Congress. Encoded Archival Description: Version 2002. `http://www.loc.gov/ead/`, 2002.
10. Library of Congress. Library of Congress Encoded Archival Description Best Practices. `http://www.loc.gov/rr/ead/lcp/lcp.pdf`, 2008.
11. M. Doerr. The CIDOC CRM: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, 24:75–92, 2003.
12. M. Gergatsoulis and L. Bountouri and P. Gaitanou and C. Papatheodorou. Mapping Cultural Metadata Schemas to CIDOC Conceptual Reference Model. In S. Konstantopoulos and S. Perantonis and V. Karkaletsis and C.D. Spyropoulos and G. Vouros, editor, *Artificial Intelligence: Theories, Models and Applications*, volume 6040 of *LNCS*, pages 321–326. Springer, 2010.
13. M.J. Fox and P.L. Wilkerson. *Introduction to Archival Organization and Description: Access to Cultural Heritage*. Getty Publications, 1999.
14. N. Noy. Semantic Integration: a Survey of Ontology-Based Approaches. *SIGMOD Record*, 33(4):65–70, 2004.
15. P. Nussbaumer and B. Haslhofer. CIDOC CRM in Action: Experiences and Challenges. In L. Kovacs, N. Fuhr, and C. Meghini, editors, *Research and Advanced Technology for Digital Libraries. ECDL 2007, Budapest, Hungary, September 16-21, 2007. Proc.*, volume 4657 of *LNCS*, pages 532–533. Springer, 2007.
16. M. Theodoridou and M. Doerr. Mapping of the Encoded Archival Description DTD Element Set to the CIDOC CRM. Technical Report 289, June 2001.