Mapping Encoded Archival Description to CIDOC CRM

Lina Bountouri and Manolis Gergatsoulis

Database & Information Systems Group
Laboratory of Digital Libraries and Electronic Publishing

Department of Archives and Library Science

Ionian University, Corfu, Greece.

{boudouri, manolis}@ionio.gr

1st Workshop on Digital Information Management
30 March 2011

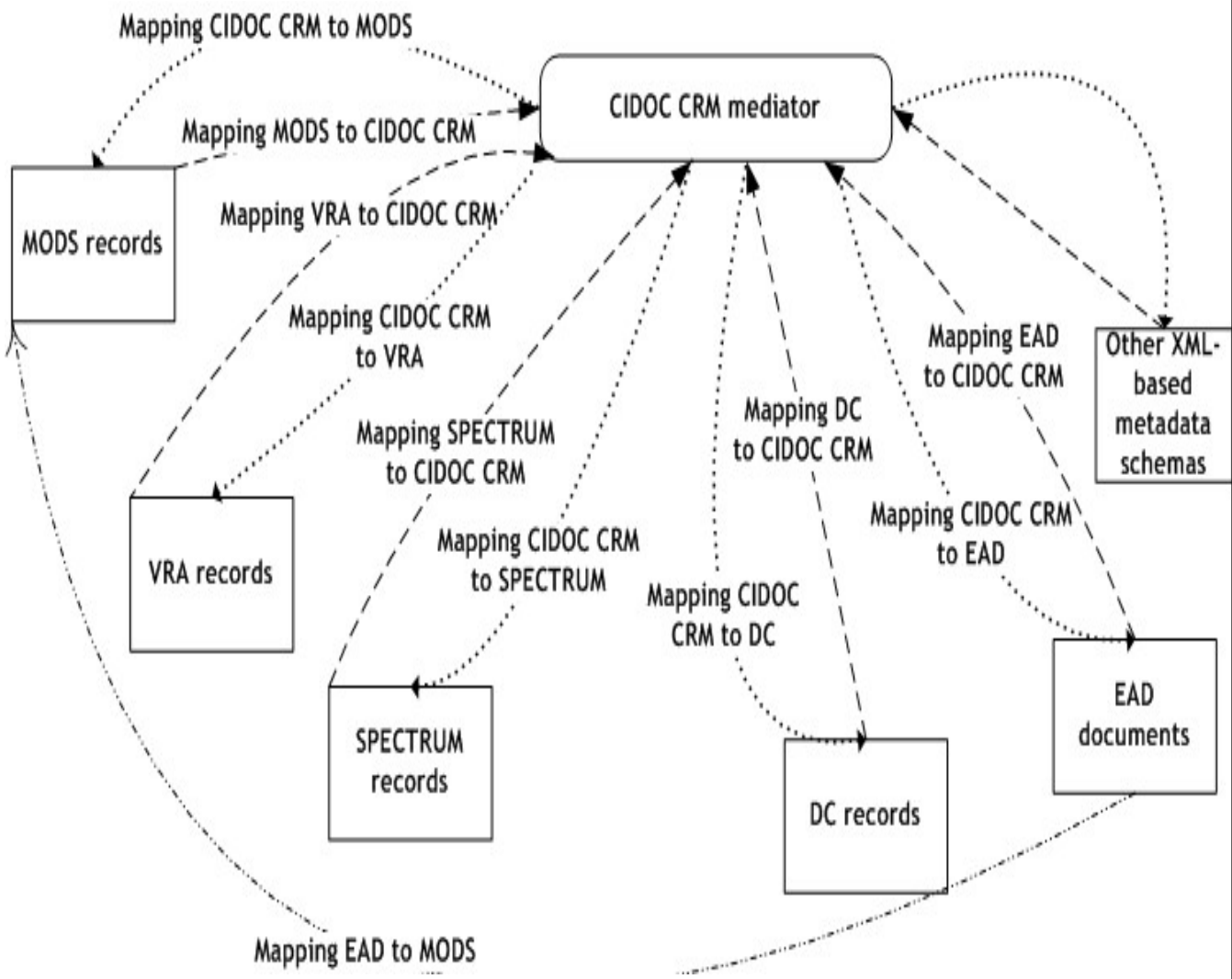
DBIS
database & information systems group
ionian university

Metadata Interoperability

- Cultural Heritage (CH) institutions, such as archives, libraries and museums, host and develop various types of material often described by different metadata schemas
- Need for *Metadata Interoperability*
  - i.e. for metadata exchange, Information Retrieval
- Various methods have been implemented
  - Crosswalks, Application Profiles, *Ontology based Integration*, etc
- Ontologies can act as the *mediated schema* between heterogeneous sources
  - The mapping of the metadata schemas to the ontology is a necessary step, so as to promote their integration
- Most of the mapping efforts do not focus on the semantics of the metadata and on how they will be expressed to the ontology

The proposed integration architecture

- Our research team has:
  - defined ***the semantic mappings*** of various metadata schemas (***EAD***, DC, VRA) to ***CIDOC CRM***,
  - defined the *Mapping Description Language* to formally describe the mappings,
  - defined an algorithm to transform XPath queries to queries expressed in terms of the ontology, and
  - proposed a data transfer logic between the local sources and the ontology.



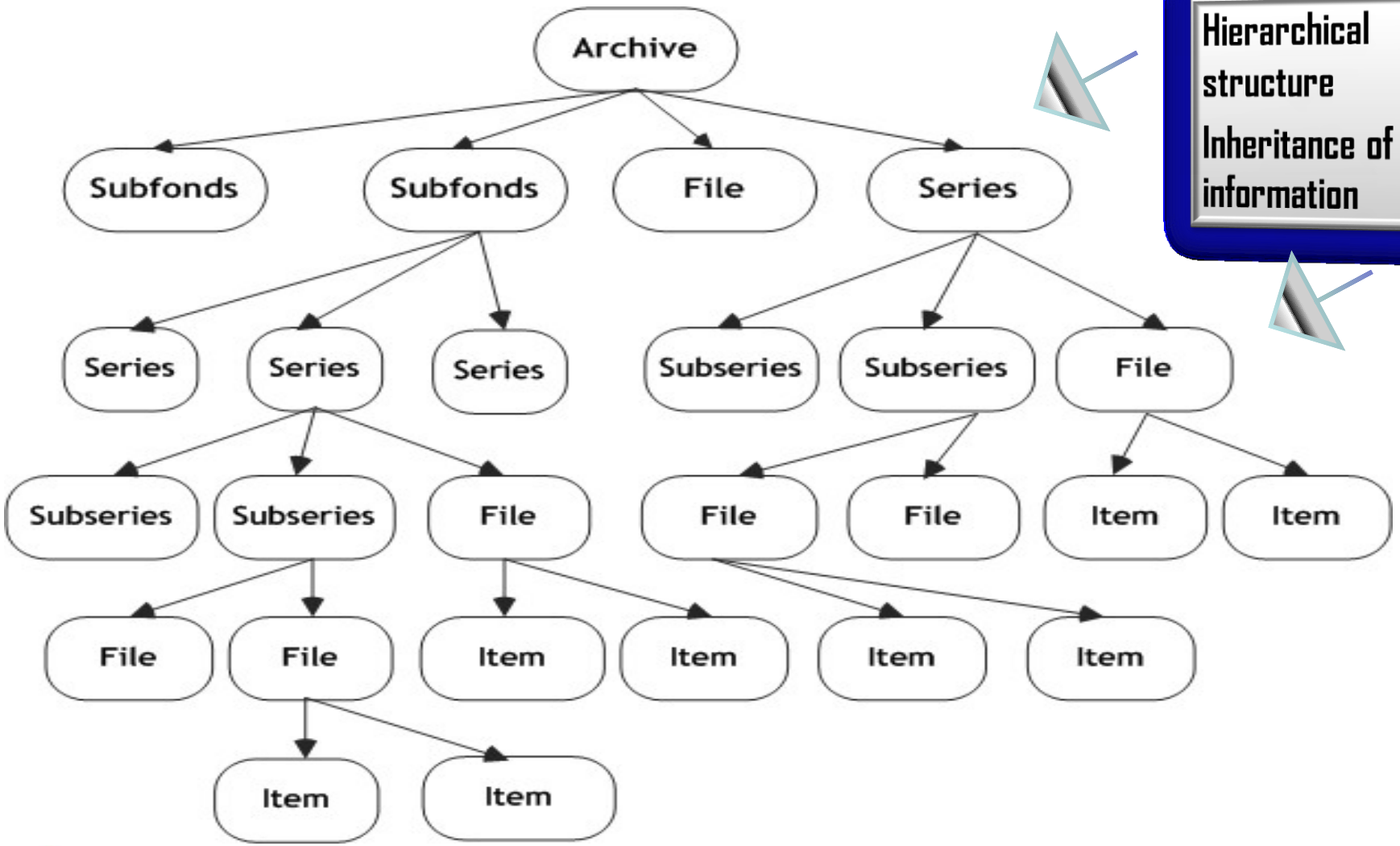
CIDOC Conceptual Reference Model (CRM)

- It is a formal ontology for the CH domain, consisting of a hierarchy of 86 *entities* (named also *classes*) and 137 *properties* and expressing semantics as a sequence of class/property path(s)
  - Entities group items, called *class instances*, sharing 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
- It can act as the mediator for diverse CH metadata, since it is the specification of the CH domain's conceptualization

Encoded Archival Description (EAD)

- *Finding aids* materialize the *archival description*
- The archival description is based on a *hierarchical tree-based structure* and on the *inheritance of information*
- *EAD* is the metadata schema to encode *electronic finding aids*
- Root element: **<ead>** (subelements: **<eadheader>**, **<frontmatter>**, **<archdesc>**)
- **<eadheader>**: metadata for the EAD document
- **<frontmatter>**: metadata for the printed finding aid
- **<archdesc>**: information for the archive's structure, content and context of creation
  - Core identification information
  - Administrative and supplemental information
  - Description of the archival components

An illustrative archival structure



Mapping Description Language (MDL) (1)

- The ***Mapping Description Language (MDL)*** is a formal language that expresses the mapping rules between a source schema and a target schema, based on a *path-oriented approach*.
  - We map the paths of the source schema to paths of the target schema
- Metadata are XML-based, hence source paths are *XPath location paths* enriched with *variables* and *stars*
- The ***MDL rules*** have the following syntax:
  - *Left part*: extension of XPath
  - *Right part*: sequence of CIDOC CRM class/property path(s)
  - *variables*: declare and refer to branching points
  - *stars*: declare the transfer of value from the XML element/attribute to the corresponding class' instance

Mapping Description Language (MDL) (2)

R ::= Left '- -' Right

Left ::= A_{Path} | V_{Path}

A_{Path} ::= ϵ | '/' R_{Path}

R_{Path} ::= L | L `*' | L `{' V_i `}' | L `*' `{' $V_i$ `}'

V_{Path} ::= '$' V_i '/' R_{Path} | `$' $V_i$ `{' V_i `}'

Right ::= E_t | E_e '→' O | `$' $V_c$ '→' O | `$' V_p '→' P_p '→' E_{t55}

O ::= P_e '→' E_t | P_e '→' E_e '→' O

E_e ::= E | E `{' $V_c$ `}'

E_t ::= E | E `{' $V_c$ `}' | E `{=' String `}'

E_{t55}::= E55 | E55 `{' $V_c$ `}' | E55 `{=' String `}'

P_e ::= P | P `{' $V_p$ `}'

Mapping Rules in MDL: the EAD to CIDOC CRM case

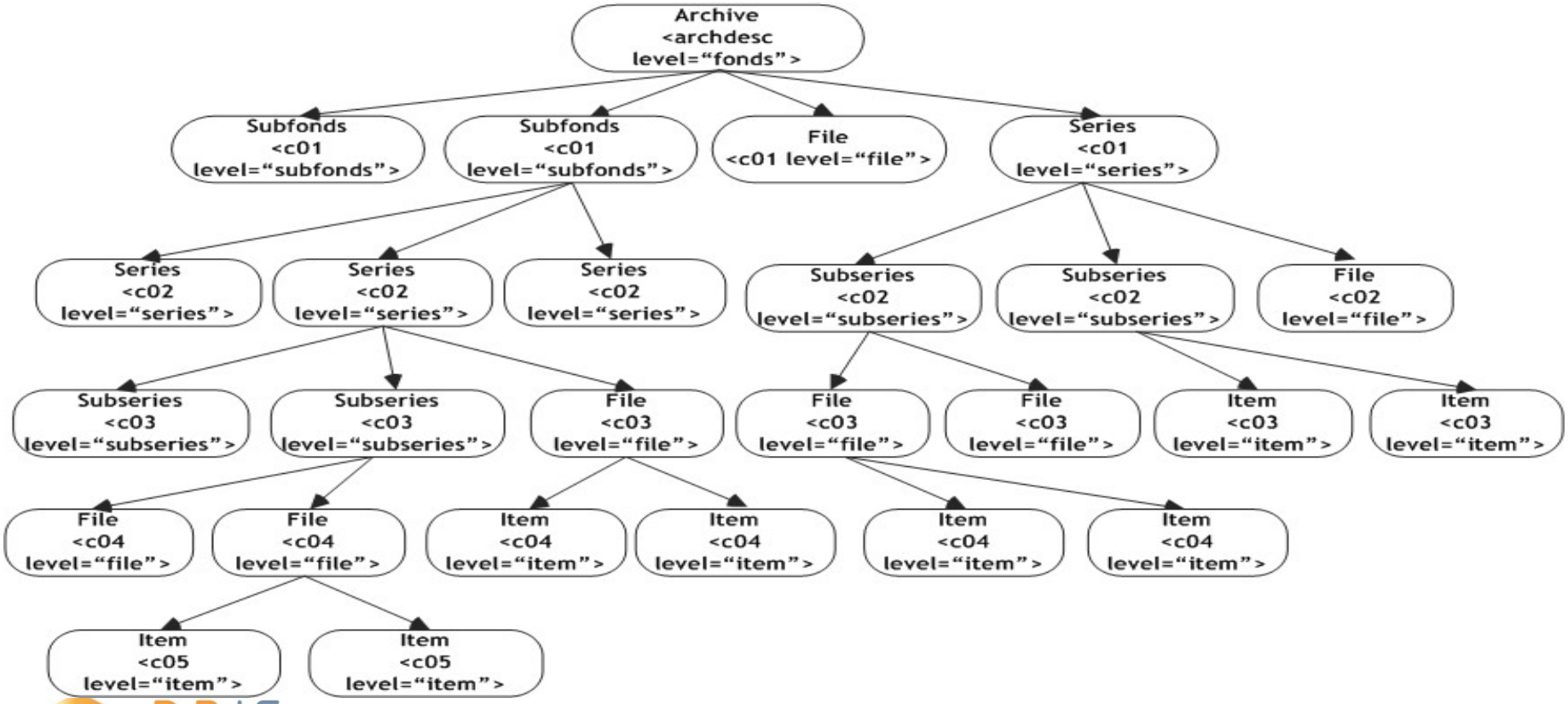
Rule number	XPath location paths	CIDOC CRM paths
R1	/ead{X0}	E31{D0}
R2	$X0/archdesc{X2}	$D0->P106->E31{D2} ->P70->E22{A0} ->P128->E73{I0}
R3	$X2/@level*{Y2}	$A0->P2->E55{A01}
R4	$Y2	$A01->P71->E32 (= level)
R5	$X2/did/unitid*	$A0->P1->E42
R6	$X2/did/unittitle*	$I0->P102->E35->P1->E41
R7	$X2/did/origination{X22}	$A0->P108b->E12{A03}
R8	$X22/corpname*	$A03->P14->E40->P1->E41
R9	$X2/controlaccess/corpname*	$I0->P67->E40->P1->E41
R10	$X2/dsc/c01{X24}	$D2->P106->E31{D3}->P70->E22{A1} ->P128->E73{I1}
R11	$X24/@level*{Z2}	$A1->P2->E55{A11}
R12	$Z2	$A11->P71->E32 (= level)
R13	$X24/did/unittitle*	$I1->P102->E35->P1->E41

The archive and the archival description: the main concepts (1/3)

- A necessary step that must be taken before the mapping is the "capturing" of the EAD' semantics, aiming to map them to the ontology
- The main semantic views of an archive are:
  - An archive is a physical object that acts as evidence for the functions/activities of the human or the corporate body that created it
  - An archive is an information object that includes information in different formats and languages
- Since the archive and its description follow a hierarchical tree-based structure, its semantic concepts are also expressed through that structure
  - For instance, an archive as a set of physical objects may contain one or more subfonds, which are a set of physical objects and they may also contain one or more archival series, which are also a set of physical objects etc.

The archive and the archival description: the main concepts (2/3)

- The description of the archive is expressed through the EAD structured as a tree
  - `<ead>: <eadheader>, <frontmatter>` and `<archdesc>`
  - `<archdesc>: <dsc>: <c01>-<c12>`



The archive and the archival description: the main concepts (3/3)

- 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 (expressed through the <archdesc>, <c01>-<c12> and <c>)
  - the semantic views of the archive, and
  - the descriptive fields (expressed through the XML subelements and attributes of the <archdesc>, <c01>-<c12> and <c>)

The EAD as a hierarchy of documentation elements and attributes

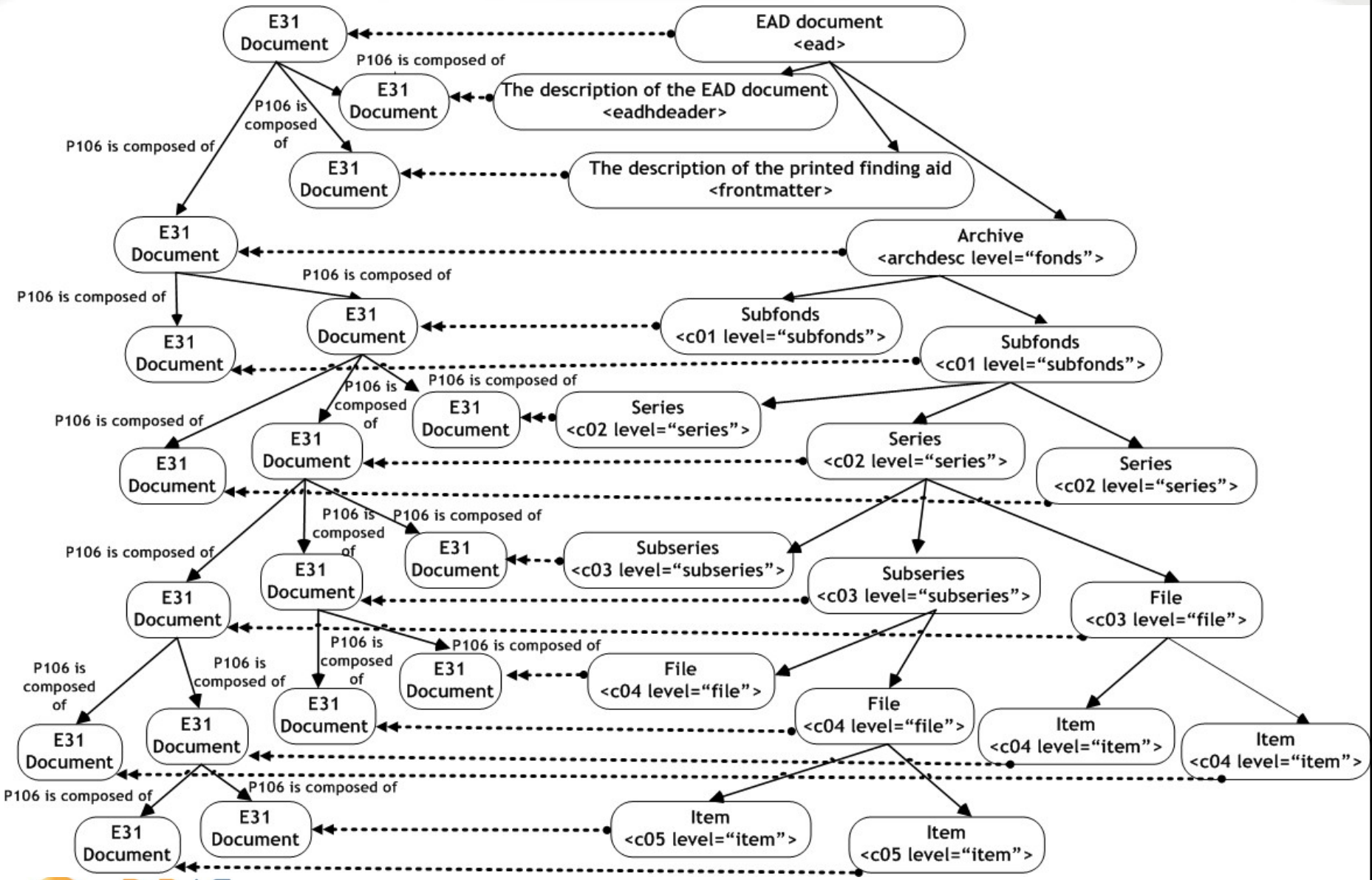
- The <ead> includes the documentation of the whole EAD document
- This concept is expressed in CIDOC CRM through the E31 Document

➢ *<ead> = E31 Document*

- *"This class comprises identifiable immaterial items that make propositions about reality. These propositions may be expressed in text, graphics, images, audiograms, videograms or by other similar means."*

- Resp. the <eadheader>, <frontmatter>, <archdesc> and <c01> -<c12> are also mapped to this class, since:
  - <eadheader>: is the documentation of the machine readable archival description
  - <frontmatter>: is the documentation for the printed finding aid
  - <archdesc> and <c01> -<c12> : is the documentation of the archive and of the components

The tree of the archival documentation



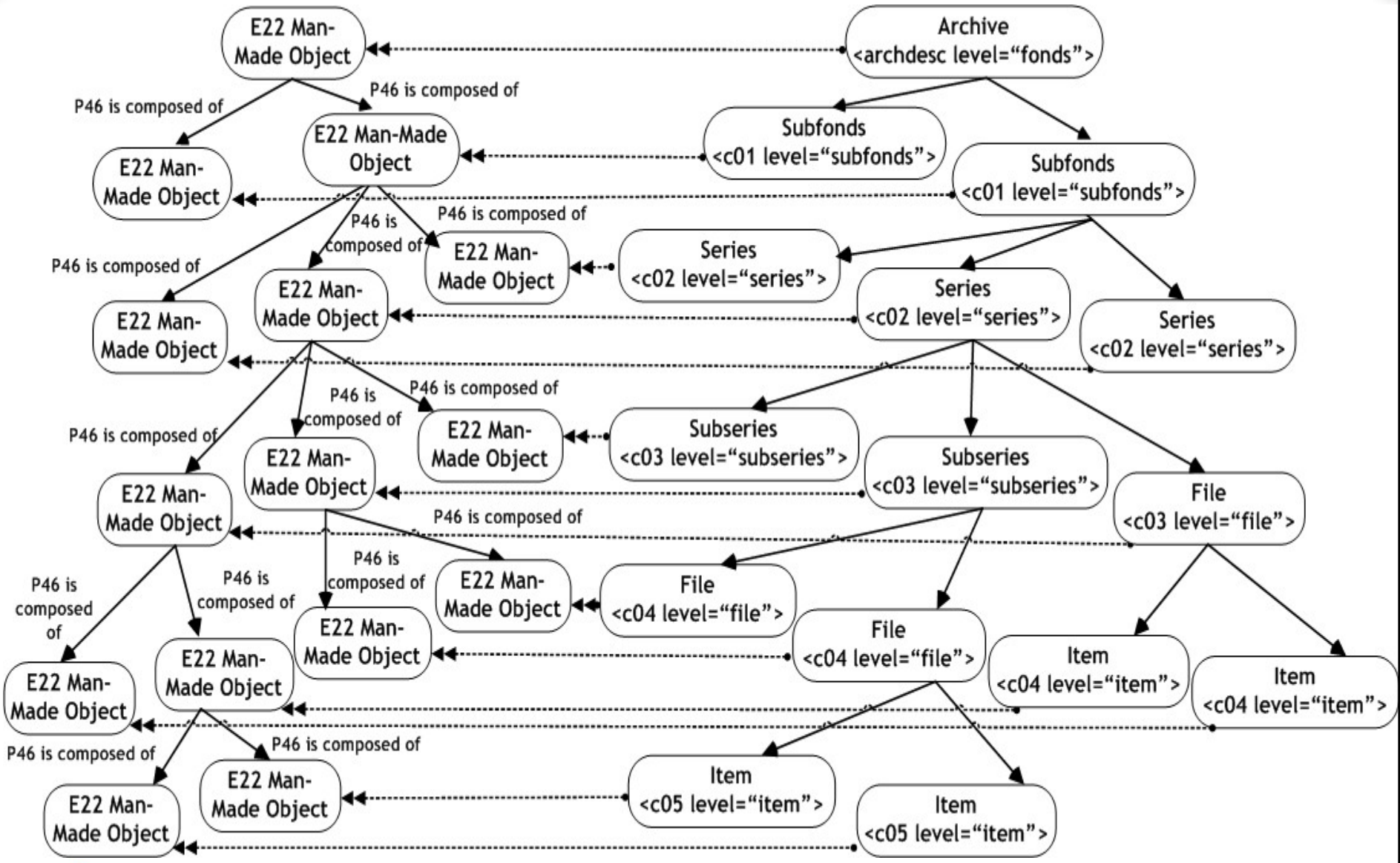
The archive as a hierarchy of physical objects

- The archive as a physical object has a hierarchical structure and it includes its components parts which are physical objects and which in turn include other components parts which are physical objects and so forth
  - Hence, these archival physical objects also follow the hierarchical and multilevel structure
- The `<archdesc>` and `<c01>-<c12>` express:
  - the archive and of its components as physical objects, and
  - their structure as physical objects
- Every physical object or set of physical objects of the archive is expressed in CIDOC CRM through the `E22 Man-Made Object` class

➢ *`<archdesc>` **and** `<c01>-<c12>` = `E22 Man-Made Object`*

- *"This class comprises physical objects purposely created by human activity."*

The tree of the archive as a physical object



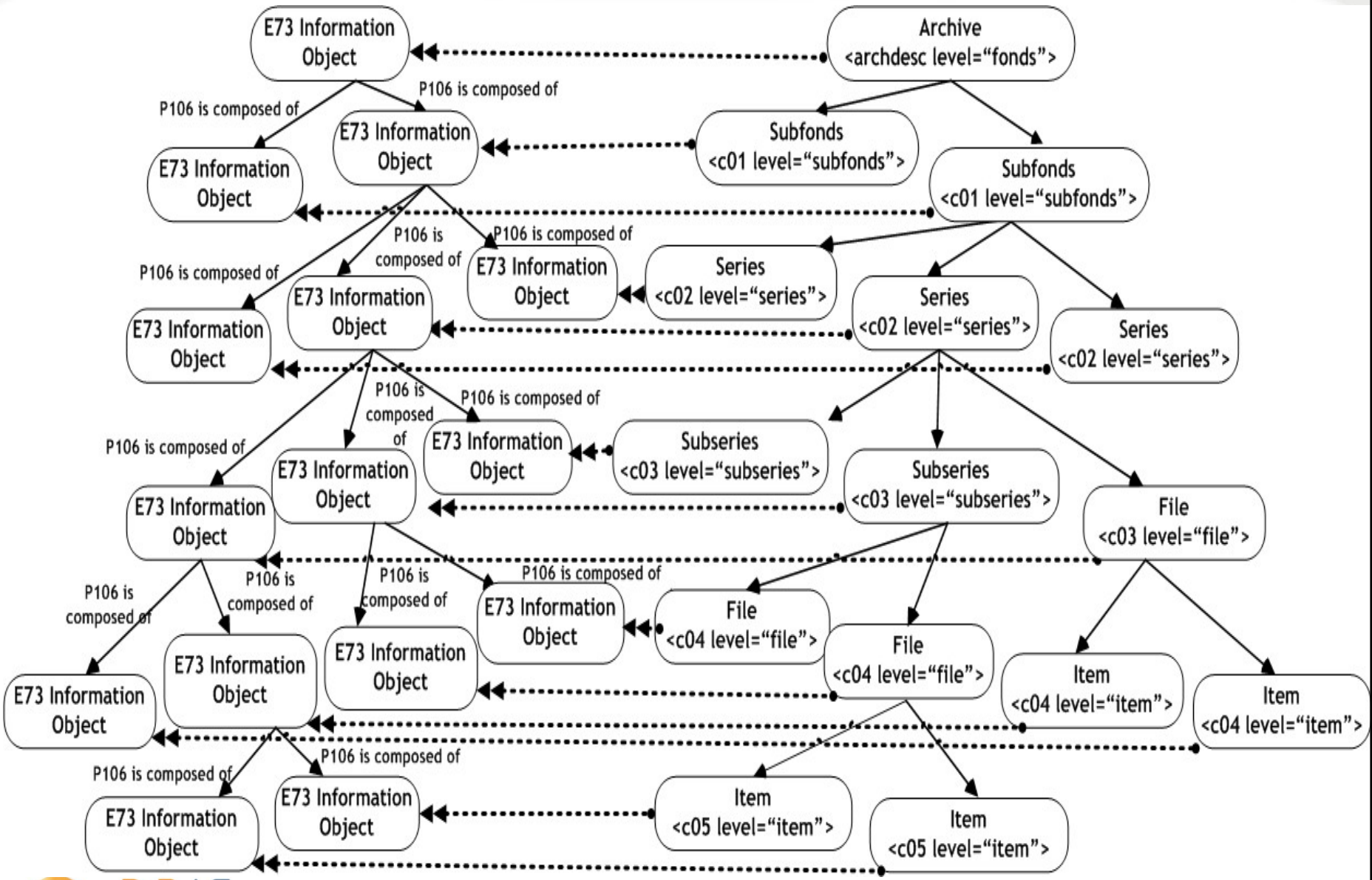
The archive as a hierarchy of information objects

- An archive contains information for its components as a whole and these components contain information in turn for their components as a whole and so on
  - Hence, these archival information objects follow the hierarchical and multilevel tree structure
- The `<archdesc>` and `<c01>-<c12>` express:
  - the archive and of its components as information objects, and
  - their structure as information objects
- Every information object or set of information objects of the archive is expressed in CIDOC CRM through the `E73 Information Object` class

➢ ***<archdesc> and <c01>-<c12> = E73 Information Object***

  - *"This class comprises identifiable immaterial items, such as a poems, jokes, data sets, images, texts, multimedia objects, procedural prescriptions, computer program code, algorithm or mathematical formulae, that have an objectively recognizable structure and are documented as single units. An E73 Information Object does not depend on a specific physical carrier, which can include human memory, and it can exist on one or more carriers simultaneously."*

The tree of the archive as an information object



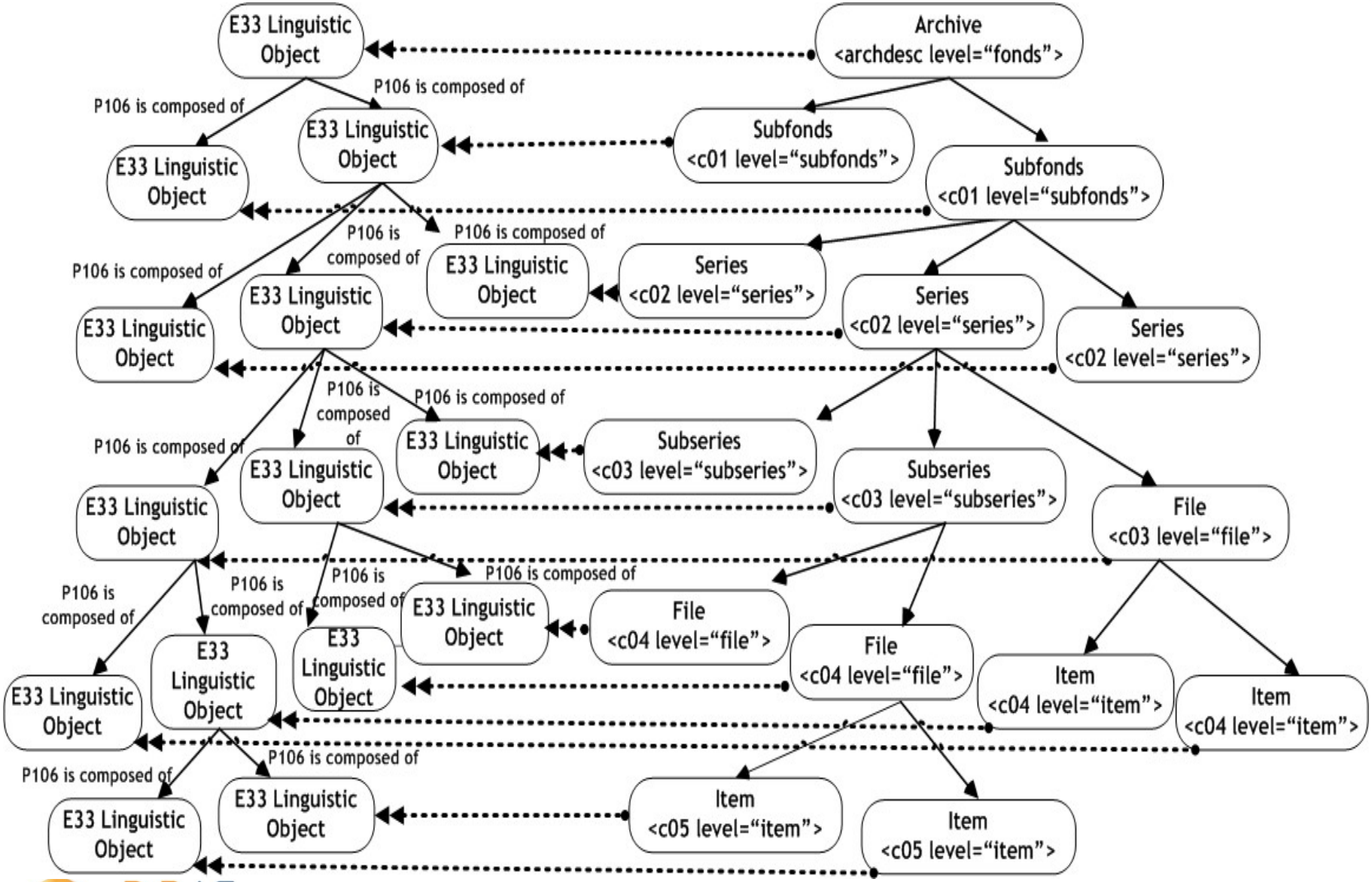
The archive as a hierarchy of linguistic objects

- An archive contains information in one or more languages for its components as a whole and these components contain information in one or more languages in turn for their components as a whole and so on
- The linguistic aspect of the archive and of its components follows the hierarchical and multilevel structure
- The combination of the `E73 Information Object` and of the `E33 Linguistic Object` covers the semantic view of the archive as an information and linguistic object
- The `<archdesc>` and `<c01>-<c12>` express:
  - the archive and of its components as linguistic objects, and
  - their structure as physical objects
- Every linguistic object or set of linguistic objects in EAD is expressed in CIDOC CRM through the `E33 Linguistic Object` class

➢ *`<archdesc>` **and** `<c01>-<c12>` = `E33 Linguistic Object`*

  - *"This class comprises identifiable expressions in natural language or languages."*

The tree of the archive as a linguistic object



Defining the relationships between the semantic views of the archive (1/3)

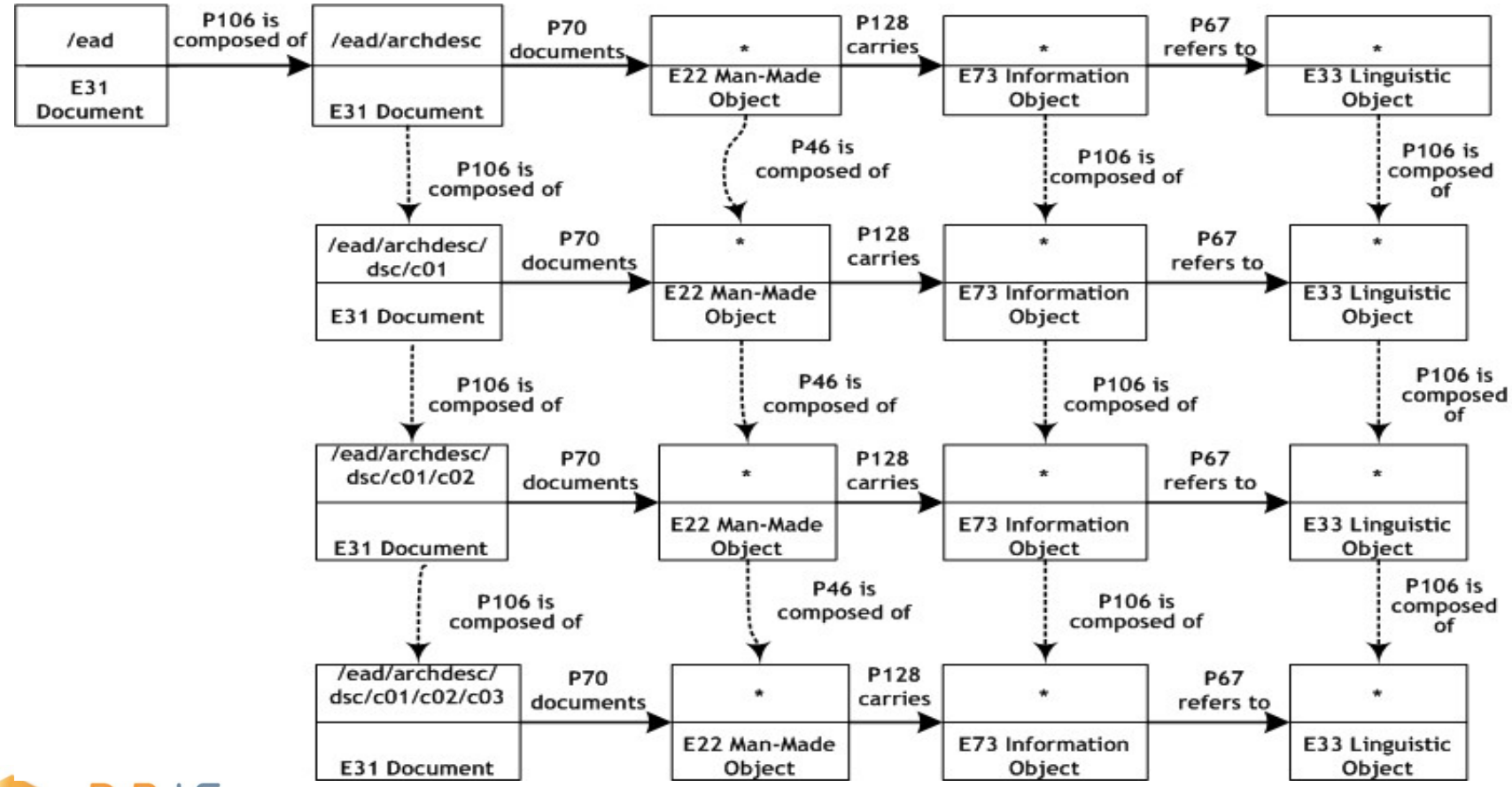
- The archive and its components are equally mapped to *three different CIDOC CRM trees* each of them declaring their structure and their different semantic views
- The archival description is mapped to the "*tree of the archival documentation*" in the ontology
- These four trees have the same structure and they only differ in terms of the names of the classes
- How they are associated?
  - The <archdesc>, <c01>-<c12> and <c> are firstly referred to the archival description, which incorporates the semantic views of the archive, hence the "*tree of the archival documentation*" is the starting point for the association of the different views
  - The documentation refers to the archive as a physical object ("*tree of the archive as a physical object*")
  - The archive as a physical object carries information, which can lead – after being analyzed and processed – to the export of added information ("*tree of the archive as an information object*")
  - The archive can also be a carrier of linguistic content, since the information it carries can be expressed via written or oral speech, independently of the medium that carries this content ("*tree of the archive as a linguistic object*")

Defining the relationships between the semantic views of the archive (2/3)

- We conclude that since these trees refer to the same object (the archive), they are semantically related to each other
- It is necessary to:
  - relate these four trees with the tree of the EAD schema (and in particular with the `<archdesc>`, `<c01>`-`<c12>` and `<c>`) and of the archival description, given that the archival documentation is referred to all the different views of the archive, and
  - to associate these four trees, since they all refer to the same object, the archive

Defining the relationships between the semantic views of the archive (3/3)

- As a consequence, these four trees are linked (through the mapping rules) in a way that allows the expression of the afore-mentioned analysis inside the CIDOC CRM ontology



Associating the metadata with the trees of archival description and the semantic views (1/2)

- The `E31 Document`, `E22 Man-Made Object`, `E73 Information Object` and `E33 Linguistic Object` classes (part of the CIDOC CRM path that maps the `<archdesc>`,`<c01>-<c12>` and `<c>`) are associated with classes' variables in the MDL rules a) to indicate the point where new CIDOC CRM paths start and b) to allow the association of `<archdesc>`, `<c01>-<c12>` and `<c>` subnodes to the various semantic views of the archive
- These subnodes may be associated to:
  - `the E31 Document`, when they provide information for the archival documentation,
  - the `E22 Man-Made Object`, when they provide information for the archive as a physical object,
  - the `E73 Information Object`, when they provide information for the archive as an information object, and
  - the `E33 Linguistic Object`, when they provide information for the archive as a linguistic object
- For example, the mapping of the creator of the archive (corporate body):
  - The location path `/ead/archdesc/did/origination/corpname` is mapped to the CIDOC CRM path: `E31 Document -> P106 is composed of -> E31 Document -> P70 documents -> E22 Man-Made Object -> P108b was produced by -> E12 Production -> P14 carried out by -> E40 Legal Body-> P1 is identified by -> E41 Appellation`

Associating the metadata with the trees of archival description and the semantic views (2/2)

- Certain subnodes may be associated with one or more of the four trees, based on their semantics' analysis
- For example, the `<unitid>` defines the identifier of the archival descriptive unit, which is a unique reference point for it or a control number, such as the accession number or the call number. Hence, it refers:
  - to the descriptive unit as a physical object, when it identifies the unit to its accession or its physical position (``*tree of the archive as a physical object*")
  - to the descriptive unit as an information object, given that it is an information given by the archivist in order to uniquely identify the descriptive item (``*tree of the archive as an information object*")

Related work (1/2)

- Various efforts to map the XML (meta)data in the area of CH to the CIDOC CRM
  - The STAR project [1]
  - The BRICKS project [2]
  - A well documented research proposal is presented in [3]
    - Differs from our proposed mapping on the following points: a) this mapping refers to the first edition of the EAD metadata schema , b) the different semantic views of the archive and of the archival description are not defined and analyzed, hence they are not mapped to the ontology, and c) the EAD is considered as a format for describing the whole and there is no reference to mapping of its hierarchical structure.

Related work (2/2)

1. D. Tudhope and C. Binding and K. May. Semantic interoperability issues from a case study in archaeology. In Stefanos Kollias and Jill Cousins, editors, *Semantic Interoperability in the European Digital Library, Proceedings of the First International Workshop SIEDL 2008, associated with 5th European Semantic Web Conference, Tenerife*, pages 88-99, 2008.
2. 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. 11th European Conference, ECDL 2007, Budapest, Hungary, September 16-21, 2007. Proceedings,* volume 4657 of *LNCS*, pages 532-533. Springer Berlin / Heidelberg, 2007.
3. M. Theodoridou and M. Doerr. Mapping of the Encoded Archival Description DTD Element Set to the CIDOC CRM. Technical Report 289, June 2001.

Discussion

- 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 the information source it describes. The main goal is to be able to use this mapping in various integration scenarios through the ontology CIDOC CRM
- The proposed mapping is the first complete research effort to map all the semantic views of the archive to a domain ontology

Using the mapping…

- The proposed mapping has been used in the mechanisms presented in the following research papers:
  - M. Gergatsoulis, L. Bountouri, P. Gaitanou, and C. Papatheodorou. Query Transformation in a CIDOC CRM Based Cultural Metadata Integration Environment. In Mounia Lalmas, Joemon M. Jose, Andreas Rauber, Fabrizio Sebastiani, and Ingo Frommholz, editors, *Research and Advanced Technology for Digital Libraries, 14th European Conference, ECDL 2010, Glasgow, UK, September 6-10, 2010. Proceedings, volume 6273 of Lecture Notes in Computer Science*, pages 38-45. Springer, 2010.
  - M. Gergatsoulis, L. Bountouri, P. Gaitanou, and C. Papatheodorou. Mapping Cultural Metadata Schemas to CIDOC Conceptual Reference Model. In Stasinos Konstantopoulos, Stavros Perantonis, Vangelis Karkaletsis, Constantine D. Spyropoulos, and George Vouros, editors, *Artificial Intelligence: Theories, Models and Applications, volume 6040 of Lecture Notes in Computer Science*, pages 321-326. Springer, 2010.
  - L. Bountouri, M. Gergatsoulis, and C. Papatheodorou. Integrating Cultural Heritage Information Sources. *ERCIM - DIS (Data and Information Spaces) Workshop, 27 May, Paris, France, 2009*, 2009.