Mapping EAD to CIDOC CRM

Lina Bountouri (1), Manolis Gergatsoulis (1),

Christos Papatheodorou (1,2)

{boudouri, manolis, papatheodor}@ionio.gr

1. Database and Information systems group, Department of Archive and Library Sciences, Ionian University, Greece

2. Digital Curation Unit, Institute for the Management of Information Systems, Athena Research Centre, Greece

DBIS
database & information systems group
ionian university

Digital Curation Unit
IMIS - Athena Research Centre

Problem statement

- Growing number of heterogeneous Cultural Heritage (CH) resources
- Growing number of metadata schemas (EAD, MODS, DC APs, TEI, VRA, MARCs)
- Need for Metadata Interoperability (MI)
- Positive effects of the Semantic Web (SW) to deal with MI problems
  - SW promotes Semantic Integration
    - Part of data integration oriented to solve semantic heterogeneity problems *"by using conceptual representations of the data and of their relationships to eliminate possible heterogeneities"* (Cruz and Xiao, 2005)

Metadata Interoperability Approach

- Ontology – based Metadata Integration Architecture
  - Use of CIDOC Conceptual Reference Model (CRM) as the mediated schema to integrate CH metadata sources
  - Mapping Encoded Archival Description (EAD) metadata sources to CIDOC CRM
  - Mapping EAD queries CIDOC CRM queries

Ontology – Based Metadata Integration

- Ontologies: SW infrastructure, promoting Semantic Integration needs
  - One of their main roles: mediated schema in an integration scenario
- Ontology-based integration architecture based on CIDOC CRM
  - Conceptual model for CH resources
  - Intended to facilitate the integration, mediation and interchange of heterogeneous CH information
  - Consists of 86 classes and 137 properties
  - Classes are connected through properties

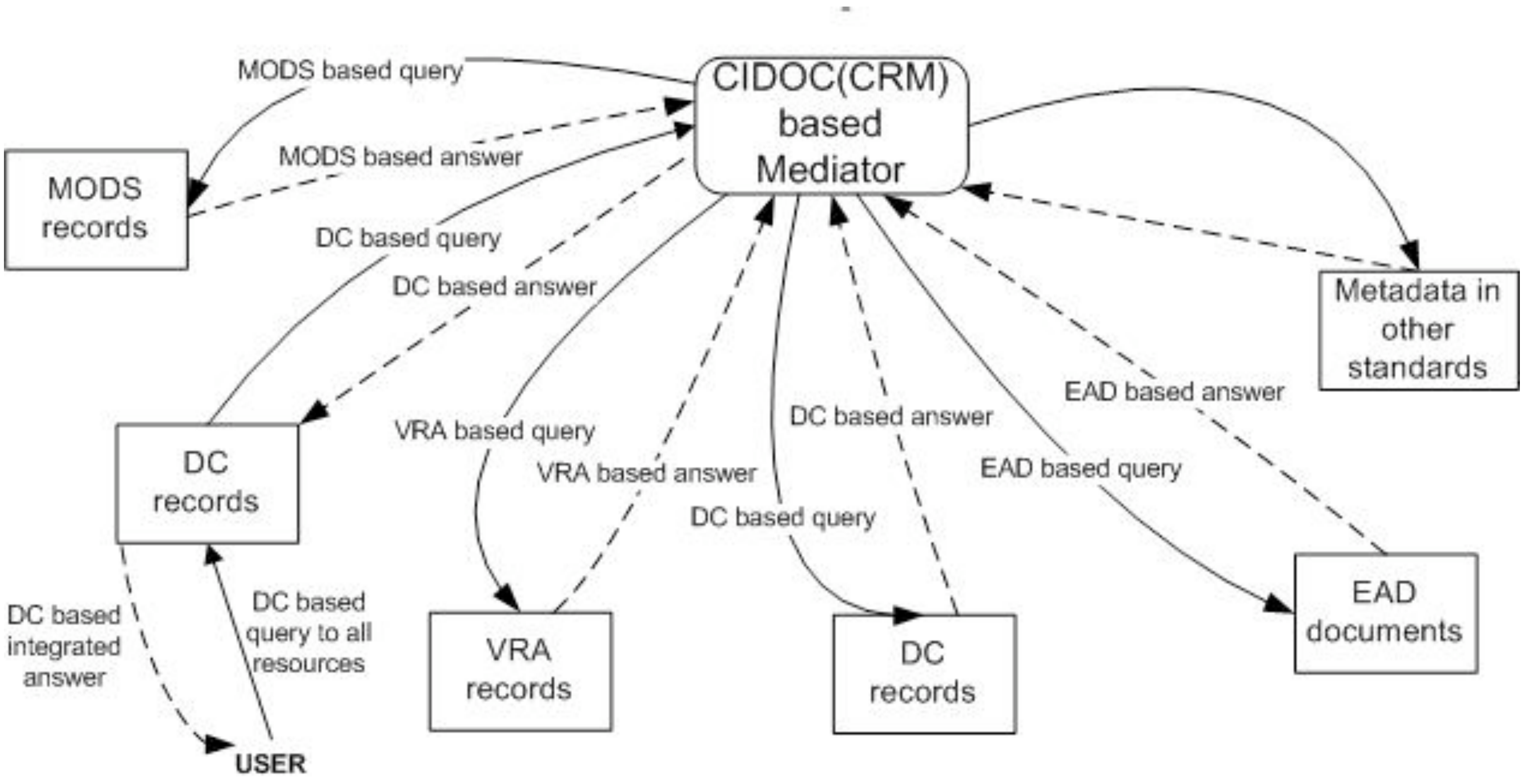
CIDOC CRM

- Classes
  - Group items that share one or more common characteristics acting as the criteria to categorize the items that belong to the class
  - Classes may be interlinked through IsA subclasses relationships
    - Subclasses inherit all the characteristics of their superclasses
- Properties
  - Define a relationship between two classes (domain and range)
  - Can be interpreted in both directions (active and passive voice), with two distinct, but related interpretations
  - May themselves have properties that relate to other classes (specializing the meaning of the property)
  - Properties may be interlinked through IsA subproperties relationships

Ontology – Based Metadata Integration – Proposed Architecture

- Integration scenario
  - Metadata sources (DC, VRA, EAD, MODS etc) are mapped to CIDOC CRM and vice versa
  - Users can execute queries to a local data source depending on the restrictions of the local metadata schema
  - Local query engine promotes the query to the mediator which translates the query to suitable forms, using the appropriate mappings, and forwards them to be answered by the other sources

Ontology – Based Metadata Integration – Proposed Architecture



database & information systems group
ionian university

Digital Curation Unit
IMIS - Athena Research Centre

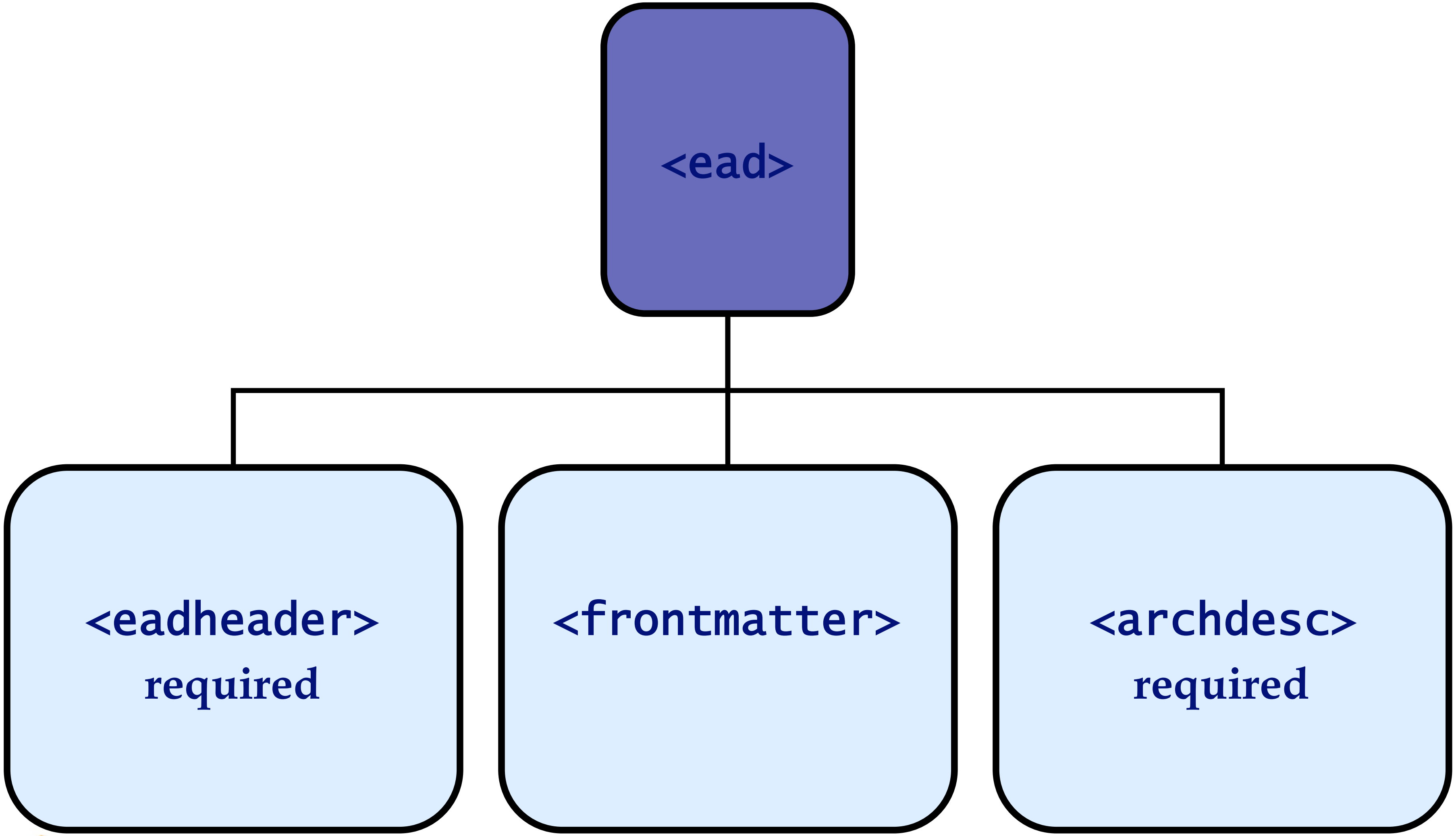
A Metadata Example: Encoded Archival Description [1]

- International metadata standard for encoding archival finding aids
- Finding aids: tools for (metadata of) the archival description
- Archival description documents the archive (complex set of materials that share common provenance)
  - Hierarchical and progressive documentation
  - Begins with the description of the whole
  - Defines and describes the sub-components of the archive, the sub-components of subcomponents, and so on

A Metadata Example: Encoded Archival Description [2]

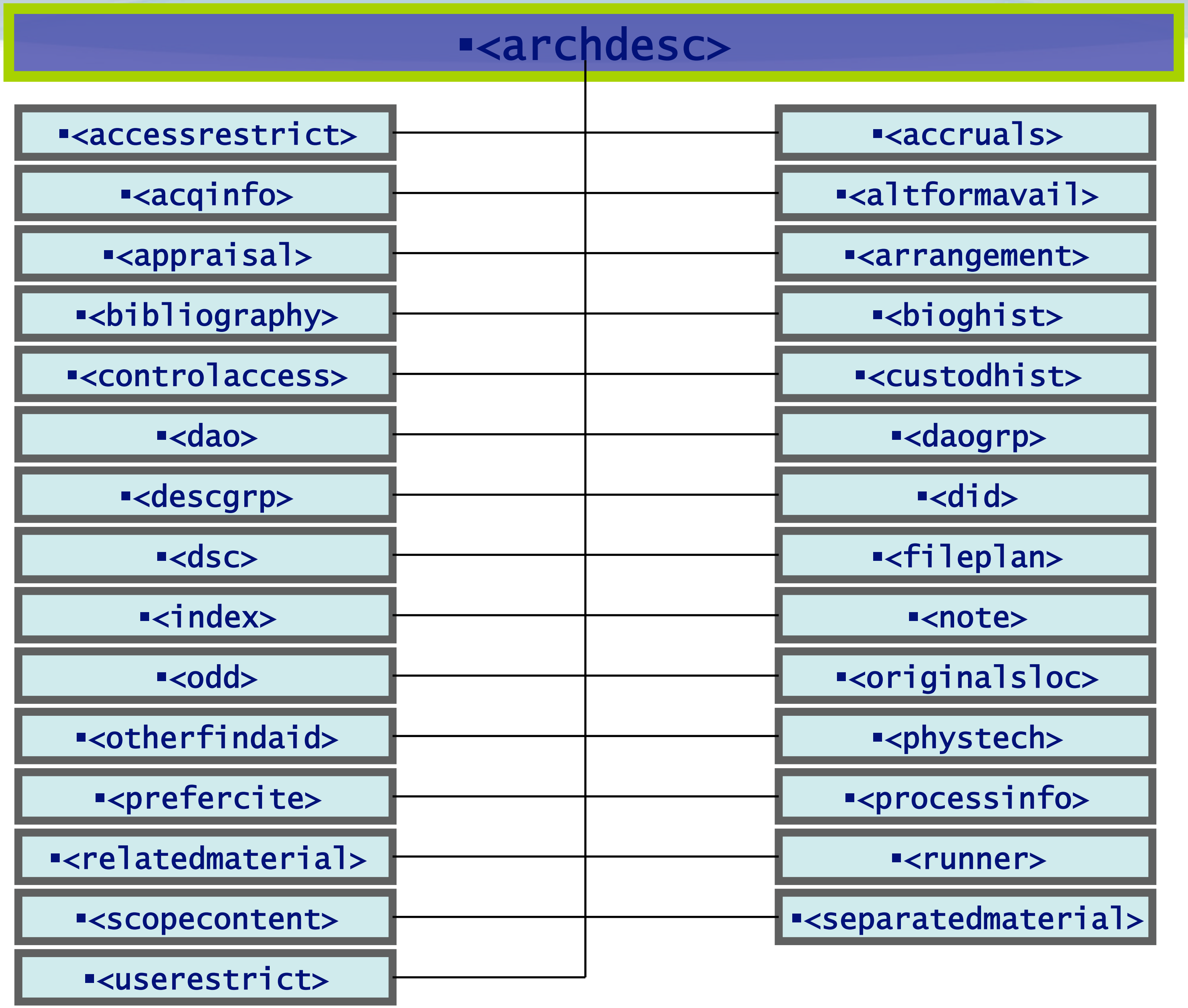
- XML language: flexible and tree structure based
  - Allows EAD to introduce a machine readable form of the archives' multi-level structure
- EAD metadata are mainly encapsulated in three parts
  - <eadheader>: information on the archival description, such as the creator of the EAD document, the date(s) of encoding etc
  - <frontmatter>: information on the creation, publication and/or use of the finding aid rather than information about the materials being described
  - <archdesc>: information on the archive itself, such as the title, the date(s) of creation and the origination of an archive etc.

A Metadata Example: Encoded Archival Description [3]



A Metadata Example: Encoded Archival Description [4]

- Archival Description (<archdesc>) consists of three main categories of information:
  - Descriptive identification information (included in the element <did>), such as:
    - Title of the archive (<unittitle>),
    - Dates of production/creation (<unitdate>),
    - Creator of the archive (<origination>) etc
  - Administrative and supplemental information, such as:
    - Scope and content of the archive (<scopecontent>),
    - Acquisition information (<acqinfo>),
    - Access points (<controlaccess>) etc
  - Description of subordinate components (included in the element <dsc>):
    - For every component the elements for the descriptive identification information administrative and supplemental information can be repeated to provide information for the specific archival component



▪<archdesc>

▪<accessrestrict>

▪<accruals>

▪<acqinfo>

▪<altformavail>

▪<appraisal>

▪<arrangement>

▪<bibliography>

▪<bioghist>

▪<controlaccess>

▪<custodhist>

▪<dao>

▪<daogrp>

▪<descgrp>

▪<did>

▪<dsc>

▪<fileplan>

▪<index>

▪<note>

▪<odd>

▪<originalsloc>

▪<otherfindaid>

▪<phystech>

▪<prefercite>

▪<processinfo>

▪<relatedmaterial>

▪<runner>

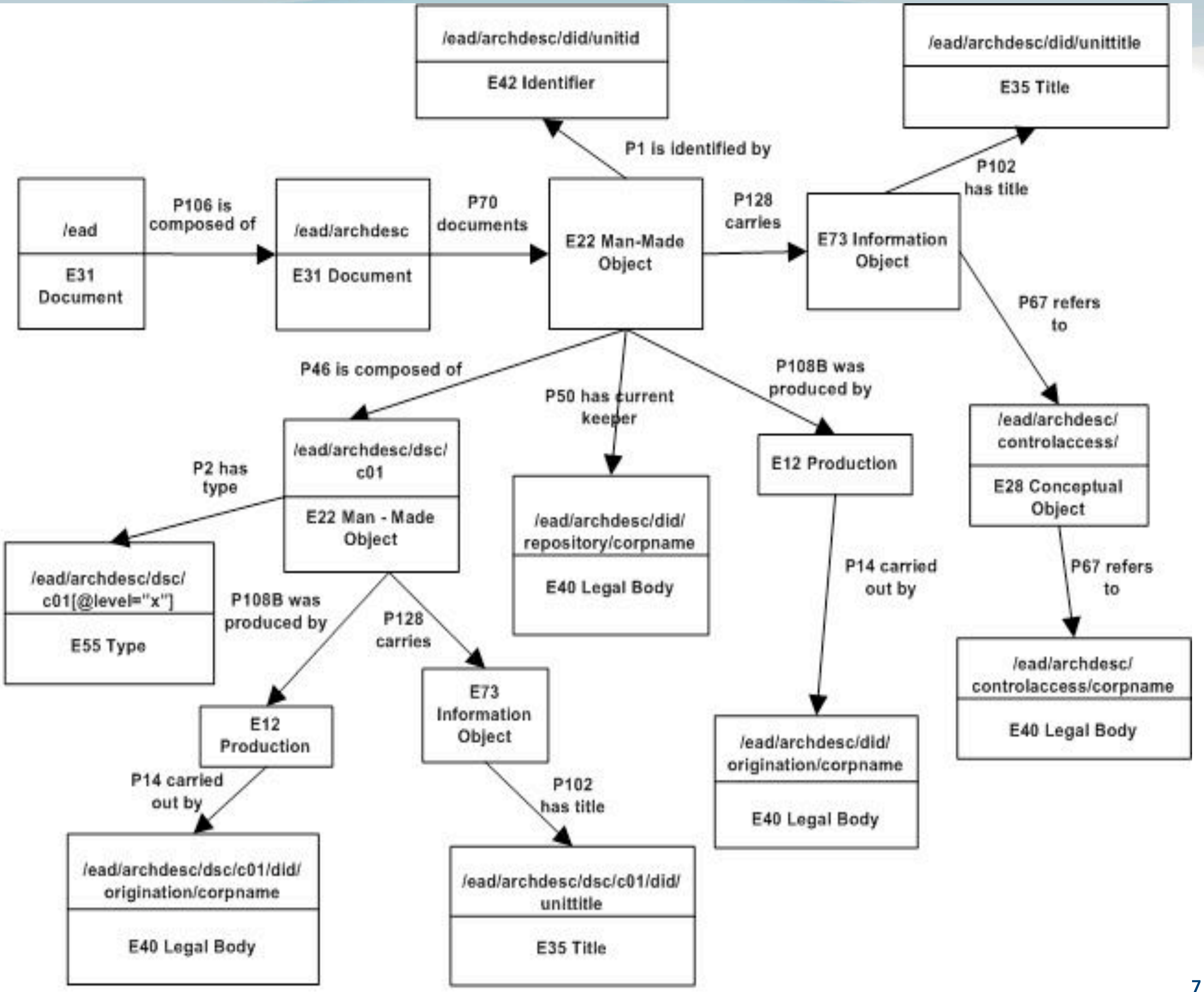
▪<scopecontent>

▪<separatedmaterial>

▪<userestrict>

Mapping EAD to CIDOC: Mapping Methodology

- Mapping methodology based on Path-Oriented Approach
  - Mapping EAD paths to CIDOC CRM paths and vice versa
- EAD Xpath
  - Sequence of EAD nodes
  - Starting from the schema root element <ead> separated by the slash symbol (/)
  - For example, the path /ead/archdesc/did/unittitle documents the title of the archive
- CIDOC CRM path
  - Sequence of class - property – class
  - E31 Document -> P106 is composed of -> E31 Document -> P70 documents -> E22 Man-Made Object -> P128 carries -> E73 Information Object -> P102 has title -> E35 Title



/ead/archdesc/did/unitid

E42 Identifier

/ead/archdesc/did/unittitle

E35 Title

P1 is identified by

P102 has title

/ead

E31 Document

P106 is composed of

/ead/archdesc

E31 Document

P70 documents

E22 Man-Made Object

P128 carries

E73 Information Object

P67 refers to

P46 is composed of

P50 has current keeper

P108B was produced by

/ead/archdesc/controlaccess/

E28 Conceptual Object

/ead/archdesc/dsc/c01

E22 Man - Made Object

P2 has type

/ead/archdesc/did/repository/corpname

E40 Legal Body

E12 Production

/ead/archdesc/dsc/c01[@level="x"]

E55 Type

P108B was produced by

P128 carries

P14 carried out by

P67 refers to

/ead/archdesc/controlaccess/corpname

E40 Legal Body

E12 Production

E73 Information Object

P14 carried out by

P102 has title

/ead/archdesc/did/origination/corpname

E40 Legal Body

/ead/archdesc/dsc/c01/did/origination/corpname

E40 Legal Body

/ead/archdesc/dsc/c01/did/unittitle

E35 Title

Mapping EAD to CIDOC: The Results

- Semantic richness of the ontology becomes obvious, since it allows the explicit definition of the notions implied in EAD
  - The /ead/archdesc path 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, which is semantically analyzed as:
    - The EAD document (E31 Document) comprises the following (P106 is composed of):
      - Identifiable immaterial items that make propositions about reality (E31 Document) and document (P70 documents)
      - The archive as a physical object created by human activity (E22 Man-Made Object) that carries (P128 carries)
      - Immaterial items that include human memory and do not depend on any particular physical carrier (E73 Information Object)

Querying EAD metadata using XPath

- Xpath is used to identify specific parts of XML documents by allowing the processing of values
- XPath denotes the XML nodes by position, relative position, type, content, and several other criteria
- EAD is an XML based standard
  - Queries over EAD documents could be expressed in terms of XPath

Querying EAD metadata using XPath

- Query 1: “Find the title of the archive”.
  - EAD XPath: /ead/archdesc/did/unittitle
- Query 2: “Find the creator (corporate name of the originator) of the series titled “I.U. Library Archives””.
  - EAD XPath: /ead/archdesc/dsc/c01[@level=“series”]/did[unittitle=“I.U. Library Archives”]/origination/corpname

Querying CIDOC CRM

- The RQL-like syntax: “select-from-where” set of clauses
  - “select” clause: the variables to be answered are inserted
  - “from” clause: data path expressions are used based on the triple syntax of CIDOC CRM paths (class - property - class)
  - For data filtering: “where” clause for string pattern matching
- The reuse of a particular variable in more than one data path expressions introduces joins between the triples

Querying CIDOC CRM

- Query 1: “Find the title of the archive”.
- Corresponding CIDOC CRM path: E31 Document -> P106 is composed of -> E31 Document -> P70 documents -> E22 Man-Made Object -> P128 carries -> E73 Information Object -> P102 has title -> E35 Title
- RQL-like syntax:
  - select X5 from
    {X1;E31_Document}P106_is_composed_of{X2;E31_Document},
    {X2;E31_Document}P70_documents{X3;E22_Man-Made_Object},
    {X3;E22_Man-Made_Object}P128_carries{X4;
    E73_Information_Object},
    {X4;E73_Information_Object}P102_has_title{X5;E35_Title}

Querying CIDOC CRM

- Query 2: “Find the creator (corporate name of the originator) of the series titled “I.U. Library Archives”.
- Corresponding CIDOC CRM path (Query 2): E31 Document -> P106 is composed of -> E31 Document -> P70 documents -> E22 Man-Made Object -> P46 is composed of -> E22 Man-Made Object (P2 has type -> E55 Type=“series”) (P128 carries -> E73 Information Object -> P102 has title -> E35 Title=“I.U. Library Archives”) -> P108B was produced by -> E12 Production -> P14 carried out by -> E40 Legal Body
- RQL-like syntax:
  - select X9 from {X1;E31_Document}P106_is_composed_of{X2;E31_Document}, {X2;E31_Document}P70_documents{X3;E22_Man-Made_Object}, {X3;E22_Man-Made_Object}P46_is_composed_of{X4;E22_Man-Made_Object}, {X4;E22_Man-Made_Object}P2_has_type{X5;E55_Type}, {X4;E22_ManMade_Object}P128_carries {X6;E73_Information_Object}, {X6;E73_Information_Object}P102_has_title{X7;E35_Title}, {X4;E22_Man-Made_Object}P108B_was_produced_by{X8;E12_Production}, {X8;E12_Production}P14_carried_out_by{X9;E40_Legal_Body} where X5=“series‘” where X7=“I.U. Library Archives”

Conclusion

- Deep conceptual work by CH metadata specialists is required for the definition of mappings
- The semantic richness of CIDOC CRM provides a stable point of reference for heterogeneous data
- Current research work focuses on
  - Creating mappings of new metadata standards (i.e. VRA, MODS etc) to CIDOC CRM
  - Exploring a PROLOG implementation for the execution of queries in CIDOC CRM