



Digital Scholarly Editions
Initial Training Network

Grant Agreement No:
317436



Metadata

Georg Vogeler



DIXIT Camp 2, Graz, 14-19 September 2014

What is metadata?

- „data about data“
 - data about containers of data = structural metadata
 - data about the content represented by data = descriptive metadata
- functions:
 - descriptive
 - administrative
 - technical
 - use

teiHeader

- See the slides of James Cummings

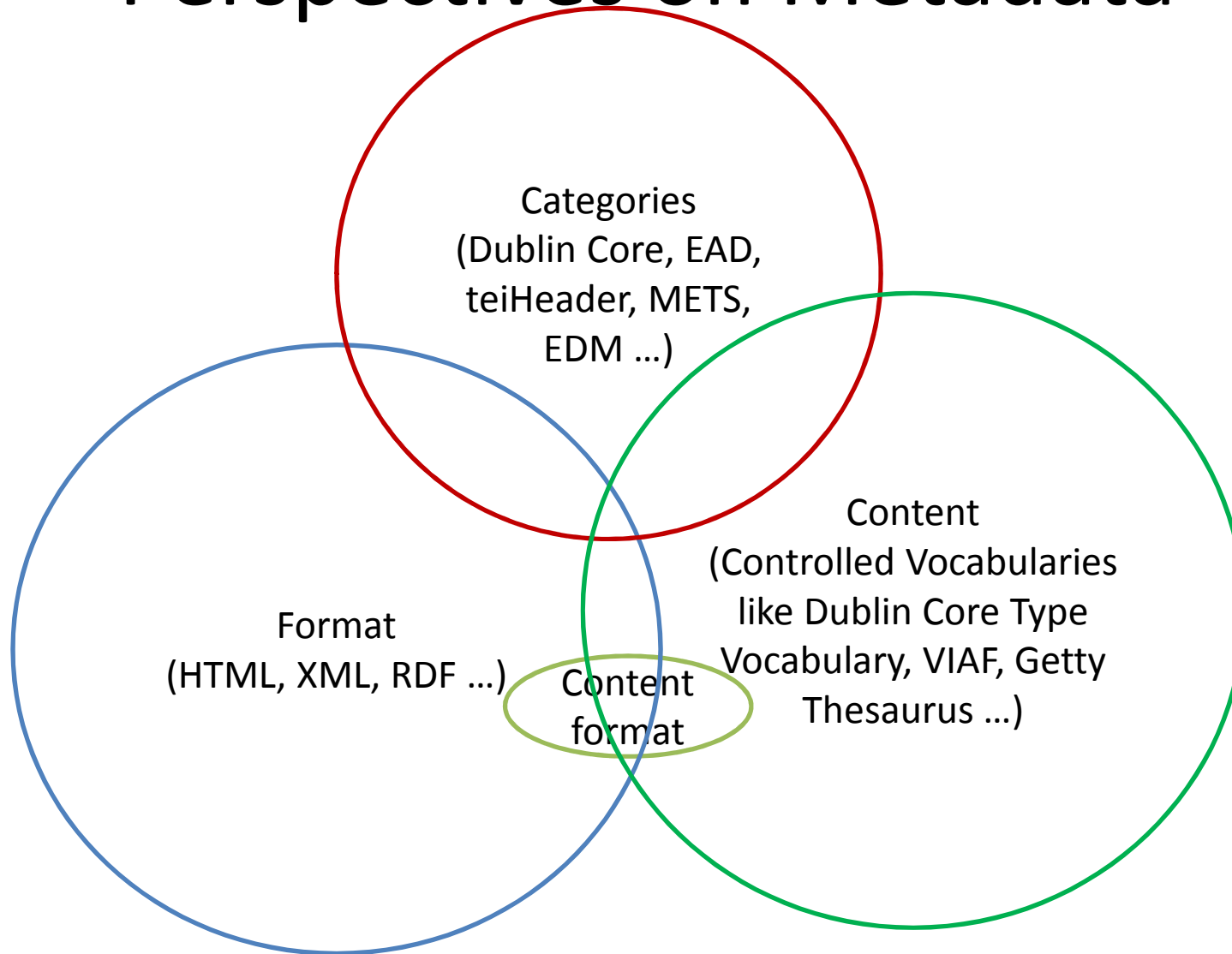
Others ...

- Dublin Core
- Machine-Readable Cataloging (MARC)
- Metadata Object Description Schema (MODS)
- Encoded Archival Description (EAD)
- Lightweight Information Describing Objects (LIDO)
- Collective Description of Works of Art (CDWA) / Visual Research Association (VRA)
- European Broadcasting Union Core Metadata (EBUCore)
- Europeana Metadata Model (EDM)
- IPTC Header
- Resource Description Framework (RDF)
- Metadata Encoding & Transmission Standard (METS)
- ...

Others ...

- **Dublin Core**
- Machine-Readable Cataloging (MARC)
- Metadata Object Description Schema (MODS)
- **Encoded Archival Description (EAD)**
- Lightweight Information Describing Objects (LIDO)
- Collective Description of Works of Art (CDWA) / Visual Research Association (VRA)
- European Broadcasting Union Core Metadata (EBUCore)
- **Europeana Metadata Model (EDM)**
- IPTC Header
- Resource Description Framework (RDF)
- **Metadata Encoding & Transmission Standard (METS)**
- ...

Perspectives on Metadata



EAD

(Encoded Archival Description)

<http://www.loc.gov/ead/>

Georg Vogeler

The World of Archives

- Archives preserve the stored information/ documentation of an institution/person and make them accessible (for internal use and for historical research)
 - Traditionally on paper
 - Organised as originally produced ("principle of provenance")

CCXLI. Mompelgart: das Fürstenthum.
CCXLI. Militare.

*Institut Compagnie
Bistums Regiment
Militar de Recuten, d'armes*

CCXLII. Munition, und Pulver-Mühle.

CCXLIII. Musterung, auch besichtigung der Wöhren, Auf-
legung und zuzug Hochstift Basel: Unterthanen.

CCXLIV. Mülhausen.

CCXLV. Münsterthal: die Landschaft und Probsteij.

<i>Probsteij, Herrschafft, ist nachher zu neu, und alten Abtheilung gleich.</i>	Moutier, Münster.	Maire.
	Roche.	Ambourg.
	Belpraon.	in Leffenbach Maire.
	Grandval.	Maire.
	Creminie.	Maire.
	Courçelle.	Maire.
	Echert.	Maire.
	Perfite.	Maire.

Champoz. Maire

Court. Maire

Sornetaim.

Sousboz. Maire

Chetellat & Fornet dellous. Maire

Monible.

Elaj, oder Seehof, sine voto, weite,
und ist Catholisch. Item auf dem Münster-
berg seind viel te. Hof.

*Als die Herrschafft im Münsterthal, ob und
unter dem selben verziehen in öffentlichen An-
gelegenheiten, in der Ordnung, wie selbe-
hier gesetzt, außert seyt etwan 6 Jahren hat
Court, Perfite und Champoz den vorzug vor
Grandval, Creminie, Courçelle und Echert.*

Tavanne: Sayfelden. Maire

Reconvelier und Chindon. Maire

Saule. Maire

Saicourt & Fuet. weite

Loveresse.

Mallerajj. Maire

Bevillard. Maire

Pontenet.

Sourvelier. Maire

*Stifte: Herrschafft Tavanne und Mallerajj
waren vor sechsen ein St. erchthum.*

Handwritten notes at the bottom of the page, including dates like '14. September 2014' and numbers like '328'.

Finding Aid / Mean of Reference

- "A document, published or unpublished, listing or describing a body of record/archives thereby establishing administrative and intellectual control over them by a records centre/archives, making them more readily accessible and comprehensible to the user. "
- include
 - guides,
 - inventories,
 - catalogues,
 - calendars,
 - lists,
 - indexes,
 - location indexes/registers and, for machine-readable records/archives, software documentation. Also called

EAD

- Society of American Archivists
- Version 1.0 1998
- Version 2002
- Currently a new version in preparation
<<http://www2.archivists.org/groups/technical-subcommittee-on-encoded-archival-description-ead>>
- Tries to implement ISAD(G)

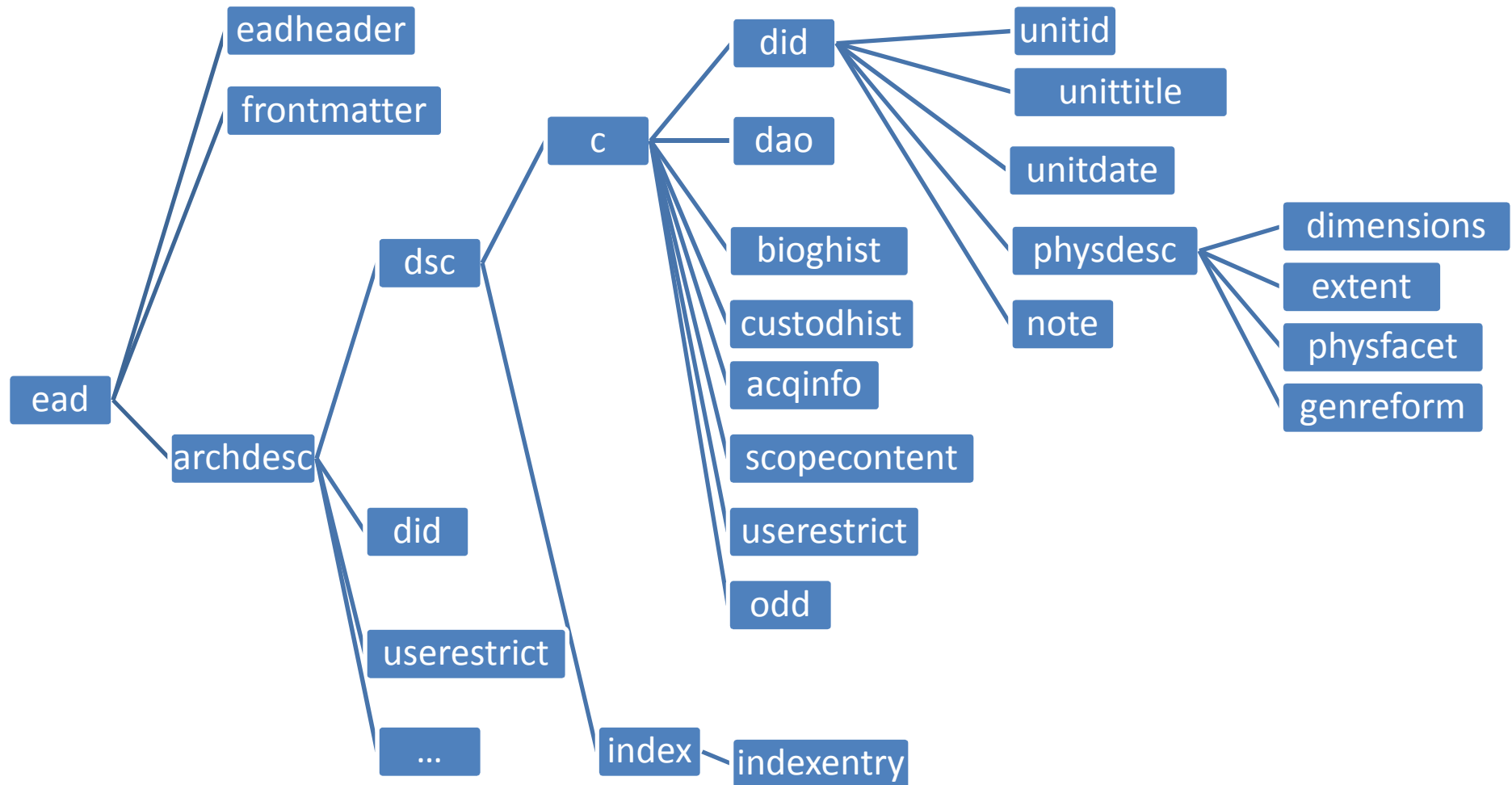
ISAD(G)

- International Standard Archival Description (General)
<<http://www.icacds.org.uk/eng/ISAD%28G%29.pdf>>
- International Council on Archives
- 2. ed. 2000

ISAD(G)

- mandatory elements:
 - Reference code
 - Title
 - Name of Creator
 - Dates of Creation
 - Extent of the Unit of Description
 - Level of description
 - and 21 others
- for
- Identification
 - context
 - content and structure
 - conditions of access and use
 - allied material

Structure



Basic structure

- Content root: `<archdesc>`, with general information on the finding aid
- (logical) structure of the archival collection: `<dsc>`
 - `<head>`, `<p>` ...
 - `@type`:
 - analyticcover, combined, in-depth, othertype
- Hierarchy of „components“ (`<c>`, or `<c01>` to `<c12>`)

Hierarchy of components

- Nesting `<c>`
- or numbering the hierarchical level: `<c01>` to `<c12>`

@LEVEL

- *collection*
- *fonds*
- *class*
- *recordgrp*
- *series*
- *subfonds*
- *subgrp*
- *subseries*
- *file*
- *item*
- *Otherlevel*

Component <c>

- Document identifier <did>
- Digital Archival Object (images, digital records)
<dao>, <daogrp>
- Restrictions in use and access <accessrestrict>, <userrestrict>
- Content <bioghist>, <scopecontent>, <index>
- History of the fond in the archives:
<custodhist>, <acqinfo>
- Other information: <odd>

Document Identifier <did>

- <unitid> : identification of the component (obligatory)
- <unittitle>, <unitdate>: elementary content description, at least one is obligatory
 - unitdate@type: *bulk* or *inclusive*
 - unitdate@normal: standardised
- <physdesc>: physical description
 - <dimensions>, <physfacet>, <genreform>, <extent>
- <note>

Attributes

- @encodinganalog
 - Reference to other metadata standards
- @audience
 - Internal / external
- @label

Component cont'd

- Preferred citation <prefercite>
- Fond management <accruals>, <arrangement>, <processinfo>
- content <controlaccess>, <appraisal>,
- Place of storage <originalsloc>, <separatedmaterial>
- Alternative form <altformavail>, <otherfindingaid>

<did> cont'd

- <container>, <physloc>, <materialspec>
= other information on the physical object
- <langmaterial>, <abstract>
= other information on the content of the documents
- <origination>, <repository>
= other information on the institutions involved in the process of archival processing
- <head>

General elements

- <head>, <p>, <table>, <address>, <blockquote>, <list>, <chronlist>, <note>
- <abbr>, <expan>
- <emph>
- <archref>
- <bibl>, <bibref>
- <exptr>, <ref>, <ptr>

Content

- verbal description: `<scopecontent>`
- tagging in the single component description:
`<subject>` (free), `<controlaccess>` (controlled vocabulary)
An index referring to single entities of the description: `<index>`, `<indexentry>`
- As part of any verbal descriptions: `<name>`,
`<function>`, `<occupation>`, `<persname>`,
`<famname>`, `<corpname>`, `<geogname>`

Authority File

- ISAD(G) refers to ISAAR (CPF) (International Standard Archival Authority Record for Corporate Bodies, Persons, and Families)
<http://www.ica.org/10203/standards/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition.html>
- merged since 2011 with EAC (Encoded Archival Context)
<http://eac.staatsbibliothek-berlin.de/>

ISAD(G) and EAD

- Identification
 - context
(in particular information on the producer of the document collection)
 - content and structure
 - conditions of access and use
 - allied material
- <did>
 - <bioghist>, (<custodhist>)
 - <scopecontent, controlaccess>
 - <userrestrict>
 - <dao>

EAD describes original material

... but what is missing?

Dublin Core

www.dcmi.org

Georg Vogeler

History

- 1994 WWW-Conference: foundation of the DCMI
- 1995 at Dublin/Ohio: special conference org. by OCLC, NCSA
- 1998 DC Element Set Version 1.0, 1999: 1.1, continued up to <http://dublincore.org/documents/2010/10/11/dces/> without major modifications.
- Addition of Dublin Core Qualifiers (2000ff., „schemes“ 2002ff.) and Vocabularies (2000ff.)
- 2003 dcmi-type vocabulary/elements/schemes integrated into dcmi-terms

core elements

<http://www.dublincore.org/documents/dces/>

1. identifier
2. creator
3. contributor
4. publisher
5. rights
6. source
7. relation
8. title
9. subject
10. description
11. coverage
12. date
13. language
14. format
15. type

Dublin Core - Serialisation

- Text namespace declaration
label (content)

- HTML `<head profile="http://dublincore.org/documents/2008/08/04/dc-html/">`
`<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />`
`<meta name="DC.title" content="Services to Government" />`

- XML `<dc:descriptionSet xml:base="http://purl.org/dc/terms/" xmlns:dc="http://purl.org/dc/xmlns/2008/09/01/dc-ds-xml/">`
`<dc:statement`
`dc:propertyURI="http://purl.org/dc/terms/title">`
`<dc:literalValueString>DCMI Home`
`Page</dc:literalValueString>`
`</dc:statement>`

- RDF/
Turtle `<resource-name> <dc:elementName> "Value" .`

Qualified Dublin Core

(aka dcterms)

- Added rules for content and structure of metadata
 - E.g. date format
 - List delimiters for subjects
 - Enlarged tag-set:
- Audience
- Provenance
- Accrual
- Specifications of existing categories

Element refinements

- principle:
All Elements can be mapped to one of the 15 core elements.
- <http://dublincore.org/documents/dcmi-terms/>
 - List
 - Notes on how to organize the categories as semantic web resource

no need to refine

- `dcterms:contributor = dc:contributor`
- `dcterms:creator = dc:creator`
- `dcterms:publisher = dc:publisher`

dc:identifier

- dcterms:identifier
 - like an DOI, URN
- dc:terms:bibliographicCitation
 - full citation

dc:title

- `dcterms:title`
 - the very title
- `dcterms:alternative`
 - additional titles
 - original title

dc:rights

- `dcterms:accessRights`
 - who can access the data/security status
- `dcterms:license`
 - the text of the license
- `dcterms:rights`
 - general information on the rights
- `dcterms:rightsHolder`
 - to know who owns the rights

dc:date

- dcterms:date
- dcterms:created
- dcterms:dateAccepted
- dcterms:dateCopyrighted
- dcterms:dateSubmitted
- dcterms:available
- dcterms:modified
- dcterms:issued
- dcterms:valid

dc:description

- `dcterms:description`
 - general account of the resource
- `dcterms:abstract`
 - description of the content of the resource
- `dcterms:tableOfContent`
 - structure of the content

dc:coverage

- dcterms:coverage
- dcterms:spatial
- dcterms:temporal

dc:relation

- dcterms:relation
- dcterms:source
- dcterms:conformsTo
- dcterms:hasFormat / dcterms:isFormatOf
- dcterms:hasPart / dcterms:isPartOf
- dcterms:references / dcterms:isReferencedBy
- dcterms:replaces / dcterms:isReplacedBy
- dcterms:requires / dcterms:isRequiredBy
- dcterms:hasVersion / dcterms:isVersionOf

dc:type

- Preferably following the DCMI Type Vocabulary
<<http://purl.org/dc/dcmitype/>>, <<http://dublincore.org/documents/dcmi-terms/#H7>>
- Collection
- Dataset
- Event
- Image
 - MovingImage
 - StillImage
- PhysicalObject
- Sound
- Text
- Service
- Software

dc:format

- dcterms:extent
 - description of the „size“ (e.g. pages, KB)
- dcterms:format
 - preferred MIME Types, like text/xml, image/jpg
(<http://www.iana.org/assignments/media-types/>)
- dcterms:medium
 - physical medium

No need to refine

- `dcterms:contributor` = `dc:contributor`
- `dcterms:creator` = `dc:creator`
- `dcterms:publisher` = `dc:publisher`

Only dcterms:

- dcterms:provenance
 - Change of ownership and custody
- dcterms:Accrual:
 - dcterms:accrualMethod
 - dcterms:accrualPeriodicity
 - dcterms:accrualPolicy
- dcterms:audience:
 - terms:educationLevel
 - terms:instructionalMethod
 - terms:mediator

Mapping TEI to DC

Dublin Core

- identifier
- creator
- contributor
- publisher
- rights
- source
- relation
- title
- subject
- description
- coverage
- date
- language
- format
- type

teiHeader (e.g.)

- fileDesc/publicationStmt/idno
- fileDesc/titleStmt/author
- fileDesc/titleStmt/respStmt/*
- fileDesc/publicationStmt/publisher
- fileDesc/publicationStmt/availability/licence
- fileDesc/sourceDesc//*
- fileDesc/seriesStmt/*
- fileDesc/titleStmt/title
- profileDesc/textClass/*
- profileDesc/abstract/*
- profileDesc/textClass/*
- fileDesc/publicationStmt/date
- profileDesc/langUsage
- "text/xml"
- (<http://dublincore.org/documents/dcmi-type-vocabulary/#H7>)

Exercise I

Create Dublin Core Metadata for
<http://diglib.hab.de/edoc/ed000166/start.htm>

- With Dublin Core Elements (dc)
- With Dublin Core Qualified (dcterms)

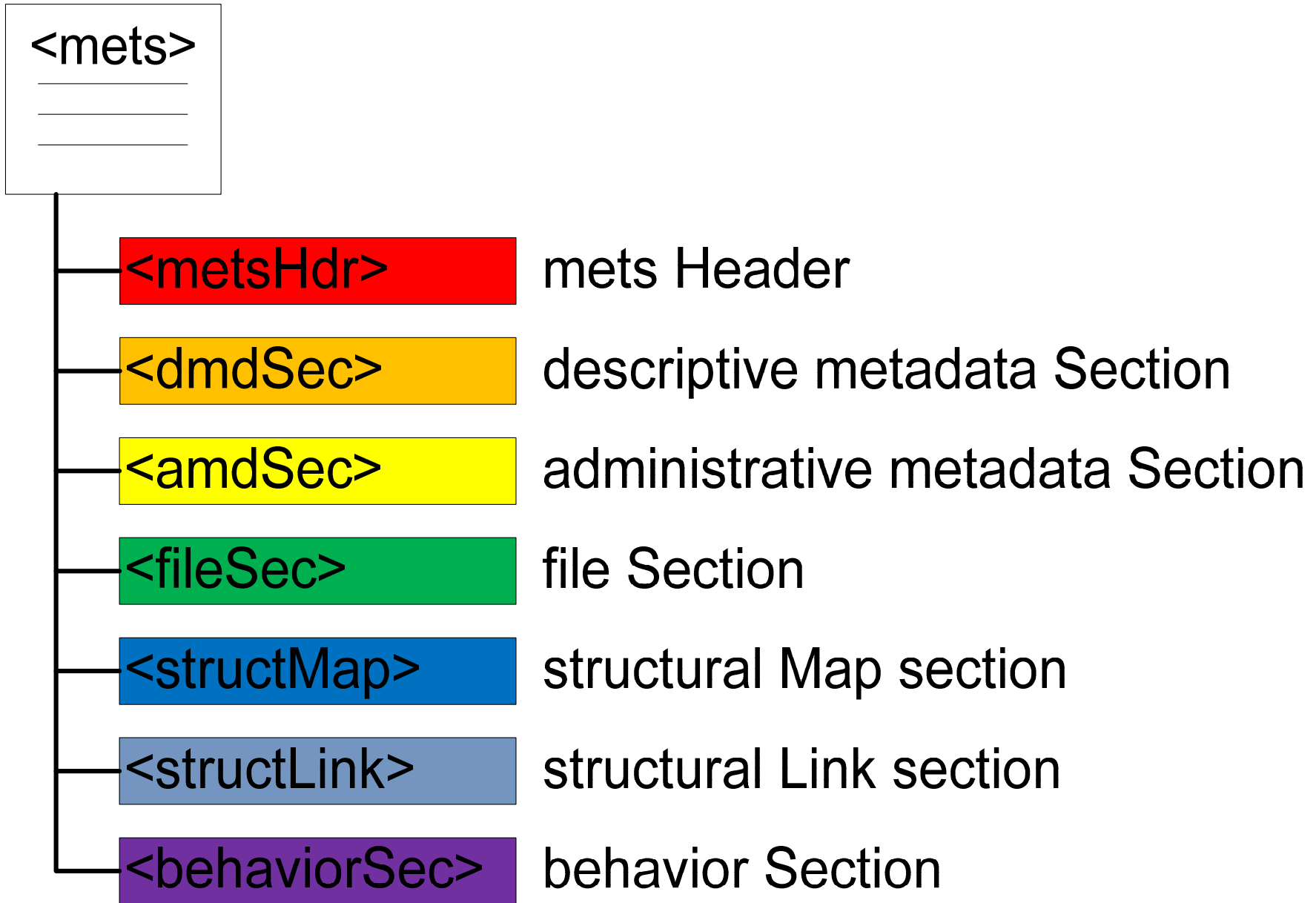
Metadata Encoding & Transmission Standard (METS)

<http://www.loc.gov/standards/mets/>

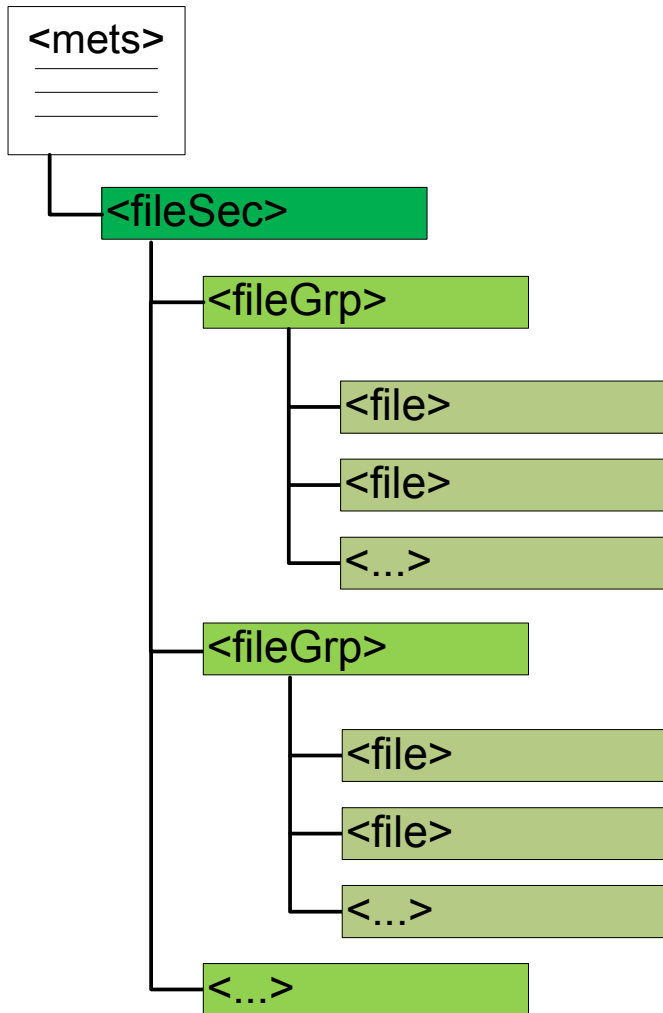
Georg Vogeler

METS

- Metadata Encoding & Transmission Standard
<<http://www.loc.gov/standards/mets/>>
- Purpose: Description of a digital object consisting of several digital parts:
 - Metadata
 - Structure
 - Content files
 - Behaviour



fileSec



lists all files of the project:

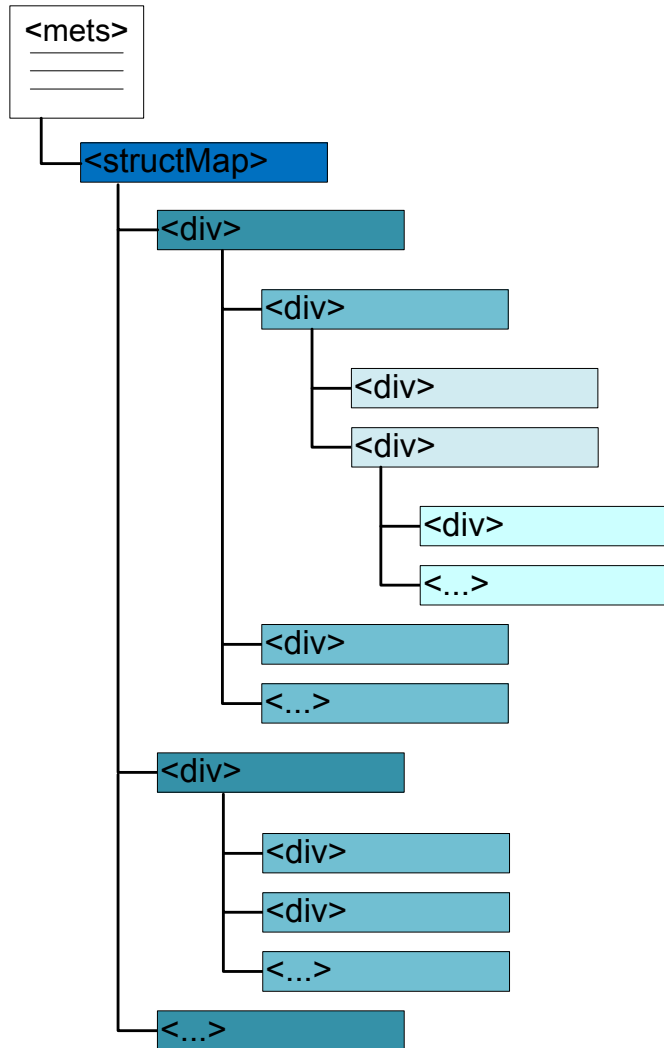
- **<fileGrp>**: e.g. by format, resolution, type
 - **<file>**
- **@USE**: e.g. distinguish between archival version, image to be displayed online or a thumbnail

fileSec

- For each `<file>`
 - `@ID`: obligatory, necessary for the reference in other sections
 - `@MIMETYPE`: typ/subtype e.g. image/tiff, image/jpeg
 - `<FLocat>` `@xlink:href @LOCTYPE` (URL, URN, PURL, HANDLE): the very link to the file

```
<fileGrp USE="master">
  <file ID="FID1" MIMETYPE="image/tiff" >
    <FLocat
xlink:href="http://nma.berkeley.edu/ark:/28722/bk
0001c3x3g" LOCTYPE="URL"/>
  </file>
</fileGrp>
<fileGrp USE="thumbnail">
  <file ID="FID187" MIMETYPE="image/gif">
    <FLocat
xlink:href="http://nma.berkeley.edu/ark:/28722/bk
0001c4t5h" LOCTYPE="URL"/>
  </file>
</fileGrp>
```

structMap

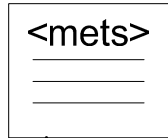


- Description how these file can be organized:
- Hierarchy of sections: `<div>`
- Different types of structures in separate `<structMap>` with `@TYPE` (like *physical*, *logical*)

structMap

- `<div>` **@TYPE** (like chapter, article, page, track, segment, section)
- Includes a list of references to the image of the section: **fptr@FILEID**
 - `<area>` (parts of page)
 - `<seq>` (images overlap)
 - `<par>` (alternative images)
- **@LABEL** (name for the division), **@ORDER** (numerical sort order), **@ORDERLABEL** (information to be displayed when sorting)

```
<structMap>
  <div TYPE="text" LABEL="Mary Refugio
Carpenter diary - 1861">
  <div TYPE="page" LABEL="double page">
    <fptr FILEID="FID1"/>
    <fptr FILEID="FID187"/>
  </div>
</div>
</structMap>
```



dmdSec

<dmdSec> descriptive metadata Section

The **dmdSec** has to have a **@ID**.

Metadata per the object as a whole

- **<mdWrap>** : „metadata wrapper“
 - **@MDTYPE**: format of the metadata (TEIHDR, TEI, EAD ...)
 - **<xmlData>** : integrate an XML file:
 - E.g. MODS (Metadata Object Description Schema), Dublin Core, TEI
 - **<binData>** : integrate binary data
- **<mdRef>**: refer to external metadata
 - **@XPTR**, **@LOCTYPE** (*ARK, URN, URL, PURL, HANDLE, DOI, OTHER*), **@MDTYPE**

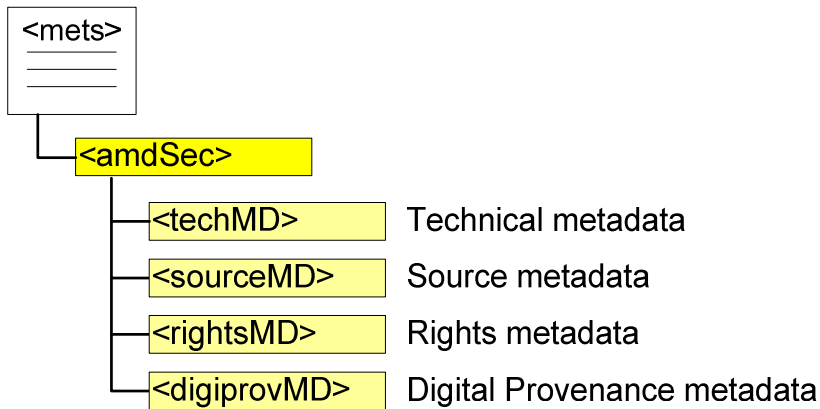
<mets>

@MDTYPE

<dmdSec> descriptive metadata Section

- MARC | MODS | EAD | DC | NISOIMG | LC-AV
| VRA | TEIHDR | DDI | FGDC | LOM | PREMIS
| PREMIS:OBJECT | PREMIS:AGENT |
PREMIS:RIGHTS | PREMIS:EVENT | TEXTMD |
METSRIGHTS | ISO 19115:2003 NAP | OTHER

```
<dmdSec ID="DM1">
  <mdWrap MDTYPE="MODS" LABEL="Mary
  Refugio Carpenter diary - 1861">
    <xmlData>
      <mods:mods>
        <mods:titleInfo>
          <mods:title>Mary Refugio Carpenter
diary - 1861</mods:title>
        </mods:titleInfo>
      </mods:mods>
    </xmlData>
  </mdWrap>
</dmdSec>
```



amdSec

Administrative metadata

- Rights **<rightsMD>**
- Technical metadata **<techMD>**
- source **<sourceMD>**
- Preservation actions to the digital representatives **<digiprovMD>**

behaviour

- **interfaceDef**
 - URL for the description of the interface as text or WSDL etc.
- **mechanism**
 - points to code which can process the data to which the METS file refers

metsHdr

- metsDocumentID
- agent (@role, @type)
 - name, note

Gives the name of CREATORS of the METS file, EDITORS of the metadata contained in the METS field, ARCHIVISTS responsible for the digitized object, people/institutions who are occupied with PRESERVATION, DISSEMINATION, curation of the digitized object ...

Can be INDIVIDUALs or ORGANIZATIONs (@type)

Example

- Jacques de Fonteny's Livre d'Enigmes - Manuscript of an Early 17th-Century Para-Emblematic, Illustrated Sonnet Sequence. Ed. by Gerhard F. Strasser, assist. by Eva Christina Glaser. Wolfenbüttel: Herzog August Bibliothek, 2012 (Editiones Electronicae Guelferbytanae 9)
<http://diglib.hab.de/edoc/ed000166/start.htm>

The situation:

- Manuscript with images (emblems) and poems on the images
- Transcription in Original French, text in modernized French
- Extensive commentary on the images and text

Solution

- Sets of files
 - Images
 - Transcription
 - Text
 - Comments
- Aggregated via METS
 - <http://diglib.hab.de/edoc/ed000166/mets.xml>

Excercise II

Take the METS file from

<http://diglib.hab.de/edoc/ed000166/mets.xml>

It should help you to answer the following questions:

- How can you access the images?
- How are the two transcriptions linked?
- How is the ToC/index on the left of the starting page `<http://diglib.hab.de/edoc/ed000166/start.htm>` created?

Europeana Data Model (EDM)

<http://pro.europeana.eu/edm-documentation>

Georg Vogeler

Europeana

- 1997 Gabriel, 2005 The European Library
- 28. April 2005: political initiative
- 2005 EDLNet
- 20. Nov. 2008: went online ... and crashed
- 2008 Dec. restart
- 2009 Feb. Europeana 1.0
- 2010 more than 10 Mio objects
- 2011 Dec. Europeana 2.0

Functionalities

- <http://www.europeana.eu>
- Thumbnail
- Facetted search
- Timeline
- Multilingual

ESE

- 15 DC Core with dcterms-Extension
- ese:isShownBy; ese:isShownAt
 - ese:object
- Data source:
 - ese:country
 - ese:dataProvider
 - ese:provider
- ese:language
- Resource:
 - ese:rights
 - ese:type (TEXT, IMAGE, SOUND or VIDEO)
- ese:uri
- ese:userTag
- ese:year
- ese:unstored

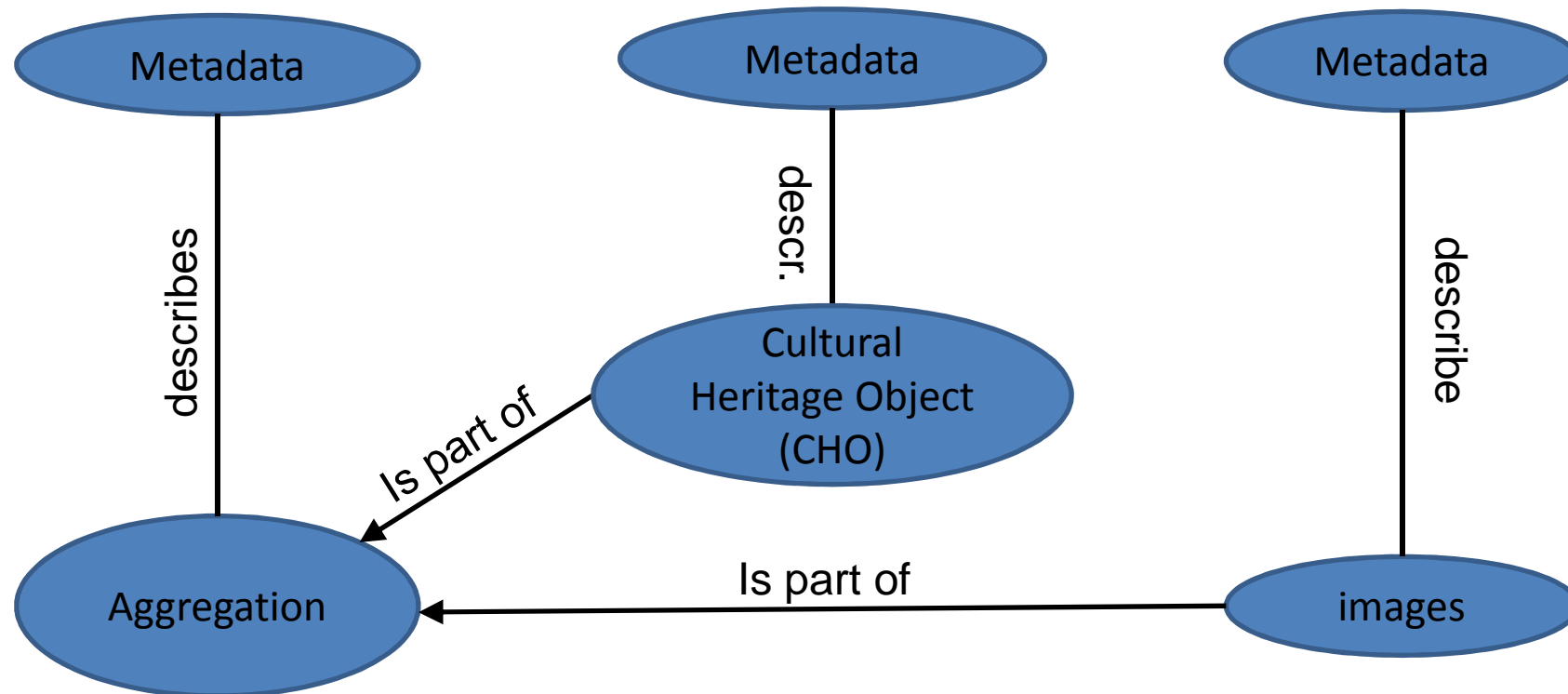
Purpose

- **Simple search**
- **Facet**
- **Timeline**
- **Advanced search**
- **Full search result display**

Search Example

- <http://www.europeana.eu/portal/search.html?query=Haus-+Hof-+und+Staatsarchiv+Wien+Urkunde>

Europeana Data Model (EDM)



Main objects of the EDM

- ore:Aggregation
 - ore:aggregates ...
- =
- EDM resource:
 - edm:hasView digital representation (=class *edm:WebResource*)
 - edm:aggregatedCHO literal or URI

Metadata of the EDM resource

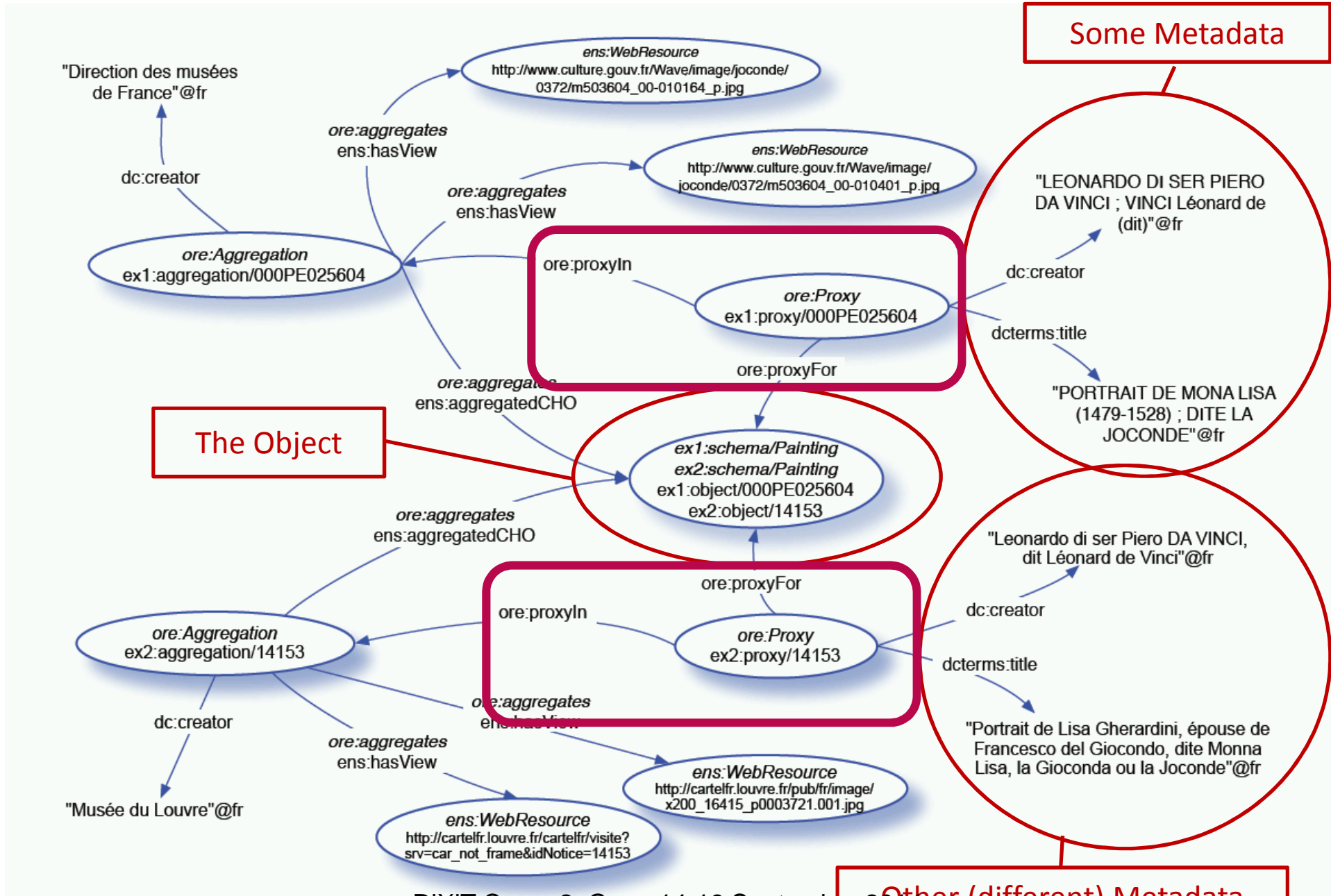
- URL of the link to the object
 - `edm:landingPage` = webpage for the object
 - `edm:isShownAt` = site with the object in its full information context
 - `edm:isShownBy` = site with a visual representation in the best possible resolution
 - `edm:hasView` = alternative views

Metadata of the EDM resource

- URL of the link to the object
- Metadata of the aggregation (created by, when, etc.)
 - E.g. `dc:creator`, `dc:date`

Metadata of the EDM resource

- URL of the link to the object
- Metadata of the aggregation (created by, when, etc.)
- Metadata of the objects as "proxy"
 - several metadata aggregations can refer via `edm:proxyFor` to the same CHO.



Other (different) Metadata

Metadata: object centric

- edm:isRelatedTo
 - edm:hasMet
 - edm:currentLocation
 - dcterms:created
 - dcterms:creator
 - edm:Place
 - skos:Concept
 - edm:TimeSpan
 - edm:hasType
 - dcterms:medium
 - dcterms:title
 - etc.

Metadata: event oriented

- edm:Event
 - *edm:wasPresentAt* edm:Agent , skos:Concept .
 - edm:happendAt edm:Place .
 - edm:occurredAt edm:TimeSpan .

Context categories

- who (`edm:Agent`)
 - where (`edm:Place`)
 - when (`edm:TimeSpan`)
 - what (`skos:Concept`)
-
- Allows references to more complex definitions (like alternative names for persons) => **"controlled vocabularies"**

EDM descriptive metadata properties

- Dublin Core
- From ESE
- owl:sameAs
- EDM specific:
 - incorporates
 - isDerivativesOf
 - isNextInSequence
 - isRepresentationOf
 - isSimilarTo
 - isSuccessorOf
 - realizes

Core Properties

- edm:ProvidedCHO:
 - dc:creator
 - edm:NextInSequence
- edm:WebResource
 - dc:format
- ore:Aggregation
 - edm:aggregatedCHO
 - edm:isShownBy
 - edm:DataProvider

Example

```
<ore:Aggregation
about="http://data.europeana.eu/item/03486/BibliographicResource_1000129064
725">
  <edm:shownAt
    resource="http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:bvb:12-
    bsb00006780-2"/>
  <edm:aggregatedCHO>
    <edm:ProvidedCHO about="http://www.wikidata.org/wiki/Q184742">
      <dc:creator>Publius Ovidius Naso</dc:creator>
      <dc:title>Metamorphoses</dc:title>
      <dcterms:created>1 n. Chr. - 8 n. Chr.</dcterms:created>
      <owl:sameAs resource="http://d-nb.info/gnd/4123895-3"/>
    </edm:ProvidedCHO>
  </edm:aggregatedCHO>
  <dc:creator>Georg Vogeler</dc:creator>
  <dc:date>2014-09-13</dc:date>
</ore:Aggregation>
```

EDM for digital scholarly editions

- allows the conversion of your metadata into an RDF structure
- distinguishes between the Permalink showing your edition, images of your source or whatever, and an abstract permalink representing the object you are editing
- is the standard to publish metadata in the Europeana
- is the format how you can access data of the Europeana in a structured way

Exercise III

- Create an EDM record for **<http://diglib.hab.de/edoc/ed000166/start.htm>** with:
 - An **ore:Aggregation**
 - referring to the URL of the digital edition
 - describing the digital edition as an **edm:ProvidedCHO**
 - giving some elementary metadata who created the aggregation and when
 - Add metadata on the digital edition as object oriented metadata
 - as part of the **edm:ProvidedCHO**
 - as **ore:Proxy**