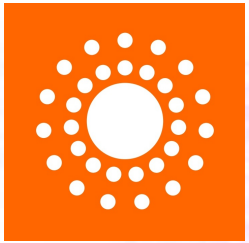
CSC

ICT Solutions for Brilliant Minds

Metadata for research data

Jessica Parland-von Essen & Johan Kylander

Dublin Core



- Dublin Core was created in 1995 in a cross-disciplinary workshop
  - In Dublin, Ohio
- It was created for resource description with interoperability in mind
  - It is independent of resource types
  - Basically anything can be described using Dublin core
- Dublin Core is managed by the Dublin Core Metadata Initiative (DCMI)
  - An open cross disciplinary organization

LOOK AND FEEL: What is it like?

- It is not a system, but a vocabulary defining a set of terms
- Originally Dublin Core consisted of 15 metadata elements, called Simple Dublin Core:
  - Title
  - Creator
  - Subject
  - Description
  - Publisher
  - Contributor
  - Date
  - Type
  - Format
  - Identifier
  - Source
  - Language
  - Relation
  - Coverage
  - Rights

LOOK AND FEEL: What is it like?

- The elements are very basic and are intended for common semantic understanding
  - They seldom need to be explained, for example *Title* or *Description*
- A controlled vocabulary for DCMI Types is published and has been highly influential in categorizing resource types:
  - Collection
  - Dataset
  - Event
  - Image
  - InteractiveResource
  - MovingImage
  - PhysicalObject
  - Service
  - Software
  - Sound
  - StillImage
  - Text

LOOK AND FEEL: What is it like?

- Qualified Dublin Core expanded the term set with three new elements (Audience, Provenance and RightsHolder) and adding element refinements:
  - For example *spatial* and *temporal* as refinements for *coverage*
- In 2012 DCMI Metadata Terms replaced the two sets
  - It is expressed in RDF describing relationships between terms and their history

```
<rdf:Description rdf:about="http://purl.org/dc/terms/title">
    <rdfs:label xml:lang="en">Title</rdfs:label>
    <rdfs:comment xml:lang="en">A name given to the resource.</rdfs:comment>
    <rdfs:isDefinedBy rdf:resource="http://purl.org/dc/terms/"/>
    <dcterms:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2008-01-14</dcterms:issued>
    <dcterms:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2010-10-11</dcterms:modified>
    <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
    <dcterms:hasVersion rdf:resource="http://dublincore.org/usage/terms/history/#titleT-002"/>
    <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
    <rdfs:subPropertyOf rdf:resource="http://purl.org/dc/elements/1.1/title"/>
</rdf:Description>
```

LOOK AND FEEL: What is it like?

```
−<metadata xsi:schemaLocation="http://example.org/myapp/ http://example.org/myapp/schema.xsd">
    <dc:title>UKOLN</dc:title>
    <dcterms:alternative>UK Office for Library and Information Networking</dcterms:alternative>
  −<dc:subject>
      national centre, network information support, library community, awareness, research, information services,public library
      networking, bibliographic management, distributed library systems, metadata, resource discovery, conferences,lectures,
      workshops
    </dc:subject>
    <dc:subject xsi:type="dcterms:DDC">062</dc:subject>
    <dc:subject xsi:type="dcterms:UDC">061(410)</dc:subject>
  −<dc:description>
      UKOLN is a national focus of expertise in digital information management. It provides policy, research and awareness services to
      the UK library, information and cultural heritage communities. UKOLN is based at the University of Bath.
    </dc:description>
  −<dc:description xml:lang="fr">
      UKOLN est un centre national d'expertise dans la gestion de l'information digitale.
    </dc:description>
    <dc:publisher>UKOLN, University of Bath</dc:publisher>
    <dcterms:isPartOf xsi:type="dcterms:URI">http://www.bath.ac.uk/</dcterms:isPartOf>
    <dc:identifier xsi:type="dcterms:URI">http://www.ukoln.ac.uk/</dc:identifier>
    <dcterms:modified xsi:type="dcterms:W3CDTF">2001-07-18</dcterms:modified>
    <dc:format xsi:type="dcterms:IMT">text/html</dc:format>
    <dcterms:extent>14 Kbytes</dcterms:extent>
  </metadata>
```

LOOK AND FEEL: What is it like?

- Dublin Core is not a technically difficult standard
- It is used by librarians, archivists and other people describing resources
- No mandatory elements exist

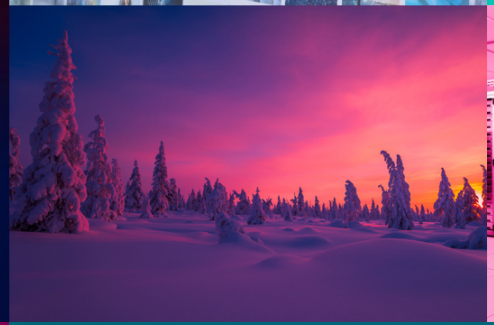
USING Dublin Core

- http://dublincore.org/
- Dublin Core is very widely used since it is simple and it can be used to describe almost anything
  - Specifically with interoperability between catalogs in mind
  - Very common when describing resources on the internet
- It is endorsed in IETF RFC 5013
- It is for example used by the Open Archives Initiative in the Protocol for Metadata Harvesting OAI-PMH
- Was the basis for the Europeana Element Set (ESE)
- And many more, like Darwin core

USING Dublin Core

- Dublin Core is very easy to adopt and fosters information exchange
- Its weaknesses are mainly due to its simplicity
  - The element set is too narrow to document resources at a detailed level
  - Its structure is flat, so it can't really be used to describe complex hierarchical structures
  - The terms are open for interpretation, for example *Date* can really mean a lot of different dates

Descriptive metadata for research datasets



RESEARCH DATA

F FINDABLE
- Described in relevant catalog with enough detail
- Landing page with globally unique identifier

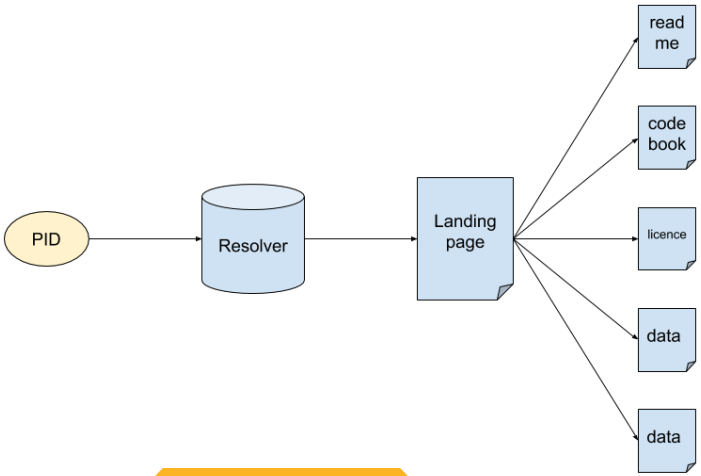
A ACCESSIBLE
- Can be retrieved over the internet
- Versioning and lifecycle documented
- Tombstone page if data is deleted

I INTEROPERABLE
- Common, documented, and open formats

R RE-USABLE
- Well documented and intelligible
- Rights clearly stated

Persistent identifiers

IMMUTABLE DATASETS



DYNAMIC DATASETS

a) Cite a specific slice or subset (the set of updates to the dataset made during a particular period of time or to a particular area of the dataset).

b) Cite a specific snapshot (a copy of the entire dataset made at a specific time).

c) Cite the continuously updated dataset, but add Access Date and Time to the citation. (Does not necessarily ensure reproducibility.)

d) Cite a query, time-stamped for re-execution against a versioned database.

DOI requires that the landing page and tombstone page to be kept forever

Research data and metadata

- The FAIRsharing.org contains more than 1200 standards
- Different fields of science and different kinds of data need different kinds of metadata
- Always try to use or conform to existing standards like schemas or vocabularies as much as possible
- Keep track of master metadata and keep it separate from aggregated metadata or self-reported information
- Choose the right format for each context
- Schema.org helps web users

FAIRsharing.org
standards, databases, policies

About DataCite

- DataCite is a global non-profit organisation
- Provides persistent identifiers (DOIs) for research data
- DOI is Digital Object Identifier, ISO 26324:2012
- DataCite is a generic, common metadata can be embedded into METS and other schemas
- DOI often used as persistent identifier for articles and often required for data citation
- CSC coordinator of national DataCite consortium

LOOK AND FEEL: What is it like?

- Data catalog metadata similar to a library catalog (or Dublin Core) but adapted for data citation
- Mandatory info: Identifier (=DOI), Creator, Title, Publisher, publicationYear, resourceType
- Close cooperation with ORCID and OpenAIRE opens up for linking data

Recommended and Optional properties in DataCite

- Subject (with sub-properties)
- Contributor (with sub-properties)
- Date (with sub-property)
- Language
- Alternate identifier (with sub-property)
- Related identifier (with sub-properties)
- Size
- Format
- Version
- Rights
- Language
- Description (with sub-property)
- Geo location (with sub-properties)
- Funding reference (with sub-properties)

DataCite

- Good quality generic metadata suitable for datasets
- Used by Metax APIs, Fairdata PAS and Eudat B2Share
- It doesn't allow much flexibility, for instance variables not possible to describe well
- Still developing and good API and cooperation Freya etc
- Useful due to the strong brand of DOI, widespread use by publishers and CRIS systems
- Available from CSC : https://research.csc.fi/support-for-persistent-identifiers

Preservation metadata

PREMIS

PREMIS PRESERVATION METADATA MAINTENANCE ACTIVITY

- PREMIS: *Preservation Metadata: Implementation Strategies*
- Was created for describing information that relate to archived digital objects , with the understanding that digital assets will become obsolete and their preservation requires cataloging information about the objects themselves (technical aspects) and the environment in which these assets are accessible (which programs they run on)

CONTEXT: Who created and why?

- It is managed by an international group called the PREMIS Maintenance Activity (PREMIS Editorial Committee) under the Library of congress
- The first version of the data dictionary was published in 2005
- The current version of the data dictionary is 3.0 (published in 2015)
  - PREMIS 2.2 is still widely used (such as in the Finnish National Digital Preservation Services)

CONTEXT: Who created and why?

- PREMIS is used for describing the technical aspects of digital objects and as such can be used as a part of a METS document

LOOK AND FEEL: What is it like?

- It is a data dictionary and an XML schema for implementation
- Built on the OAIS reference model
  - A model for digital preservation systems
  - OAIS defines archival concepts needed for digital preservation actions and access
  - PREMIS translates the OAIS framework into implementable semantic units (with some changes in the vocabulary)
- PREMIS defines four basic interrelated semantic units:
  - (Digital) Object, Event, Agent, Rights
  - (Intellectual entity, Environment)

LOOK AND FEEL: What is it like?

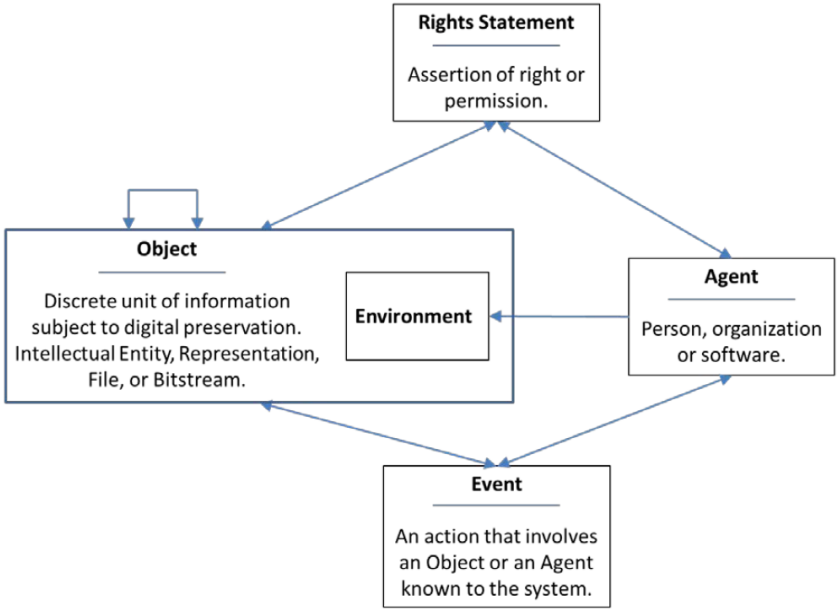


Figure 1: The PREMIS Data Model

LOOK AND FEEL: What is it like?

- Object: a unit of information subject to digital preservation.
  - Typically a file (can also be a bitstream, an intellectual entity, or a description of an environment)
  - Elements are for example identifier, fixity information, file format information, information about the creating application and relationships to other PREMIS entities

LOOK AND FEEL: What is it like?

```
▼<premis:object xsi:type="premis:file">
  ▼<premis:objectIdentifier>
     <premis:objectIdentifierType>urn</premis:objectIdentifierType>
     <premis:objectIdentifierValue> URN:NBN:fi-fe201215081520 </premis:objectIdentifierValue>
   </premis:objectIdentifier>
  ▼<premis:objectCharacteristics>
     <premis:compositionLevel>0</premis:compositionLevel>
    ▼<premis:fixity>
       <premis:messageDigestAlgorithm>MD5</premis:messageDigestAlgorithm>
       <premis:messageDigest> aa4bddaacf5ed1ca92b30826af257a1b </premis:messageDigest>
     </premis:fixity>
    ▼<premis:format>
      ▼<premis:formatDesignation>
         <premis:formatName>image/png</premis:formatName>
         <premis:formatVersion>1.2</premis:formatVersion>
       </premis:formatDesignation>
      ▼<premis:formatRegistry>
         <premis:formatRegistryName>PRONOM</premis:formatRegistryName>
         <premis:formatRegistryKey>fmt/13</premis:formatRegistryKey>
       </premis:formatRegistry>
     </premis:format>
    ▼<premis:creatingApplication>
       <premis:dateCreatedByApplication> 2011-02-15T15:43:03 </premis:dateCreatedByApplication>
     </premis:creatingApplication>
   </premis:objectCharacteristics>
 </premis:object>
```

LOOK AND FEEL: What is it like?

- Event: an action that involves at least one object
  - Used for describing preservational activities (such as fixity checks, virus scans, ingestion processes, replication activities …)
  - Elements are for example type of event, date and time, the outcome (e.g. "success" or "failure") and relationships to other PREMIS entities

```
▼<premis:event>
  ▼<premis:eventIdentifier>
     <premis:eventIdentifierType>uuid</premis:eventIdentifierType>
     <premis:eventIdentifierValue> 1c002c38-837d-437f-8af2-de8c8864b5b1 </premis:eventIdentifierValue>
   </premis:eventIdentifier>
   <premis:eventType>creation</premis:eventType>
   <premis:eventDateTime>2011-03-15T11:12:13</premis:eventDateTime>
  ▼<premis:eventOutcomeInformation>
     <premis:eventOutcome>success</premis:eventOutcome>
   </premis:eventOutcomeInformation>
  ▼<premis:linkingAgentIdentifier>
     <premis:linkingAgentIdentifierType> local </premis:linkingAgentIdentifierType>
     <premis:linkingAgentIdentifierValue> gimp-1 </premis:linkingAgentIdentifierValue>
   </premis:linkingAgentIdentifier>
 </premis:event>
```

LOOK AND FEEL: What is it like?

- Agent: a person, organization or software associated with preservational activities
  - Describes software or persons performing digital preservation actions (PREMIS events)

```
▼<premis:agent>
  ▼<premis:agentIdentifier>
      <premis:agentIdentifierType>local</premis:agentIdentifierType>
      <premis:agentIdentifierValue>gimp-1</premis:agentIdentifierValue>
    </premis:agentIdentifier>
    <premis:agentName>Gimp v. 2.9.2</premis:agentName>
    <premis:agentType>software</premis:agentType>
  </premis:agent>
```

LOOK AND FEEL: What is it like?

- Rights Statement: a set of rights and permissions concerning PREMIS objects
  - Typically describes what kind of actions are permitted when preserving the objects
- The use of controlled vocabularies is encouraged
  - For example event types
- It focuses on machine readable technical information and therefore usually is generated automatically

USING PREMIS

- http://www.loc.gov/standards/premis/
- It is important for the Finnish National Digital Preservation Services at CSC because it is created specifically with that type of services in mind
  - It describes generic (not file format dependent technical metadata) technical metadata for digital objects and their preservational history
  - It facilitates interoperability between digital preservation repositories
- It is a widely used standard in the international digital preservation community

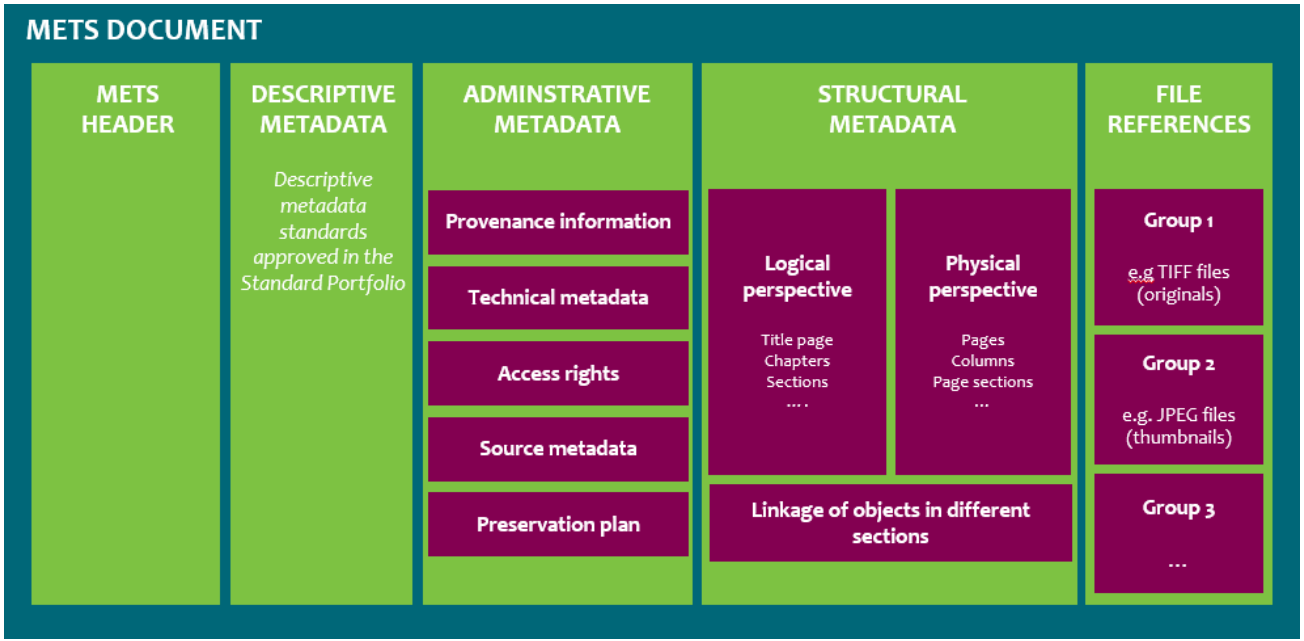
METS

- *Metadata Encoding and Transmission Standard*
- It is used for conveying the metadata necessary for managing digital objects and transmitting them between repositories
- It is maintained by the Library of Congress and developed as an initiative of the Digital Library Federation
- It combines descriptive, administrative, technical and structural metadata

LOOK AND FEEL: What is it like?

- METS is expressed in XML
- It is composed of main sections that can contain (or reference to) metadata in different standards:
  - e.g. metsHdr, dmdSec, amdSec, fileSec and structMap
- As METS is quite flexible, it supports creating METS Profiles that define mandatory contents of the METS document
  - The Finnish National Digital Preservation Services have defined such profiles for cultural heritage materials and research data materials

LOOK AND FEEL: What is it like?



LOOK AND FEEL: What is it like?

- Each dmdSec and section in the amdSec has an unique ID and contains elements for either wrapping XML data or pointers to external metadata references
  - Below a descriptive metadata section wrapping a simple Dublin Core description

```
▼<mets:dmdSec ID="dmd-dc" CREATED="2011-03-31T15:40:00">
  ▼<mets:mdWrap MDTYPE="DC" MDTYPEVERSION="1.1">
    ▼<mets:xmlData>
       <dc:identifier>urn:nbn:fi-fd2011-0000126</dc:identifier>
       <dc:title>Picture of flower</dc:title>
       <dc:creator> Patty Photographer</dc:creator>
       <dc:date>2011-01-15</dc:date>
       <dc:type>Image</dc:type>
     </mets:xmlData>
   </mets:mdWrap>
 </mets:dmdSec>
```

- The ID is used for referencing within the METS document

LOOK AND FEEL: What is it like?

- The metsHdr is a header section describing the document, its intended usage and the agent responsible for the document
- The dmdSec is a repeatable section containing descriptive metadata
  - e.g. DataCite and Dublin Core and many more
- The amdSec is a section containing administrative and technical metadata in subsections:
  - techMD, digiprovMD, rightsMD, sourceMD

LOOK AND FEEL: What is it like?

- techMD contains information about the digital objects
- digiprovMD is the provenance metadata and contains information on preservation related actions
- sourceMD describes the original source that the digital objects are derived from
- rightsMD contain information about intellectual property rights and licenses
- Technical and administrative metadata can be expressed for example in PREMIS

LOOK AND FEEL: What is it like?

- The fileSec contains a list of files and their bitstreams
  - It contains the paths to all digital objects described in the METS document
  - It also contains links to the administrative metadata sections
  - Below an example of a single file in the fileSec with links to administrative metadata sections describing the technical aspects and the digital provenance of the file:

```
▼<mets:fileSec>
  ▼<mets:fileGrp>
    ▼<mets:file ID="image-0005" ADMID="tech-001 tech-002 dp-001 dp-002 dp-003">
       <mets:FLocat LOCTYPE="URL" xlink:type="simple" xlink:href="file://img005-master.png"/>
     </mets:file>
   </mets:fileGrp>
 </mets:fileSec>
```

LOOK AND FEEL: What is it like?

- The structMap is the structural map of the digital assets being described in the document
- It outlines a physical or logical structure that includes all the digital objects in the METS document
  - For example a digitized book consisting of several TIFF files can be arranged describing the order of the TIFFs according to the book's physical page structure starting from the cover
  - It contains links to the file section and the descriptive metadata sections
- structLink and behaviorSec are quite rare

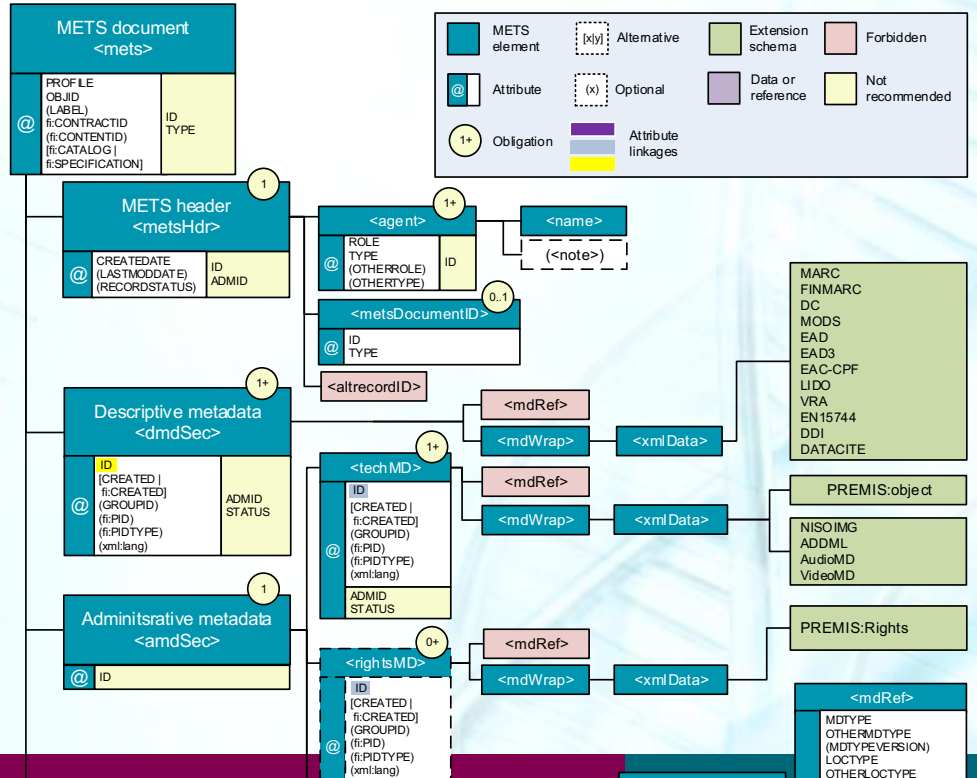
USING METS

- http://www.loc.gov/standards/mets/mets-home.html
- METS is used when transmitting digital assets and is important for the interoperability between digital repositories since it contains the metadata necessary for management of resources
- It is intended for automatic processing
- It is a common internationally used format for documenting OAIS information packages
  - It is used in the Finnish National Digital Preservation services for describing information packages

USING METS

- The METS document can easily grow too big, which makes it difficult and slow to process
  - If all metadata is in one document
  - If the document describes a large resource consisting of several thousand digital objects

USING METS


METS document
<mets>
@
PROFILE
OBJID
(LABEL)
fi:CONTRACTID
(fi:CONTENTID)
[fi:CATALOG |
fi:SPECIFICATION]
ID
TYPE
METS element
Alternative
Extension schema
Forbidden
Attribute
Optional
Data or reference
Not recommended
Obligation
Attribute linkages
1
METS header
<metsHdr>
@
CREATEDATE
(LASTMODDATE)
(RECORDSTATUS)
ID
ADMID
1+
<agent>
@
ROLE
TYPE
(OTHERROLE)
(OTHERTYPE)
ID
<name>
(<note>)
0..1
<metsDocumentID>
@
ID
TYPE
<altrecordID>
1+
Descriptive metadata
<dmdSec>
@
ID
[CREATED |
fi:CREATED]
(GROUPID)
(fi:PID)
(fi:PIDTYPE)
(xml:lang)
ADMID
STATUS
<mdRef>
<mdWrap>
<xmlData>
MARC
FINMARC
DC
MODS
EAD
EAD3
EAC-CPF
LIDO
VRA
EN15744
DDI
DATACITE
1+
<techMD>
@
ID
[CREATED |
fi:CREATED]
(GROUPID)
(fi:PID)
(fi:PIDTYPE)
(xml:lang)
ADMID
STATUS
<mdRef>
<mdWrap>
<xmlData>
PREMIS:object
NISOIMG
ADDML
AudioMD
VideoMD
1
Adminitsrative metadata
<amdSec>
@
ID
0+
<rightsMD>
@
ID
[CREATED |
fi:CREATED]
(GROUPID)
(fi:PID)
(fi:PIDTYPE)
(xml:lang)
<mdRef>
<mdWrap>
<xmlData>
PREMIS:Rights
<mdRef>
MDTYPE
OTHERMDTYPE
(MDTYPEVERSION)
LOCTYPE
OTHERLOCTYPE

Johan Kylander

Jessica Parland-von Essen


facebook.com/CSCfi


twitter.com/CSCfi


youtube.com/CSCfi


linkedin.com/company/csc---it-center-for-science


github.com/CSCfi