TIMBUS

TIMELESS BUSINESS

Digital Preservation Metadata - Hands-On Exercise

Angela Dappert
Digital Preservation Coalition
angela@dpconline.org

Learning Goals

- Methodology of metadata definition
- Unless your situation is identical to someone else's there is no instantaneous metadata definition – see why!

Hands-On Exercise

TIMELESS BUSINESS

- Imaginary eJournal submission (inspired by the Elsevier ScienceServer specification)
- You want to collect this content-type in your repository to ensure long-term access.
- It is the first time that you see this publisher’s format and you start to think about your metadata needs.

Hands-On Exercise

TIMELESS BUSINESS

Goal:

- Store metadata in the repository with the content to create complete, self-descriptive units
- Specify metadata profiles for archival information packages (AIP)

Creating Metadata Profiles

TIMELESS BUSINESS

1. Which objects / entities do we describe?
    a. Which?
    b. How many?
2. Which metadata do we need?
    a. Which do we need?
    b. Which do we get?
3. Which standard do we use for which metadata?
4. How do we implement them?

Creating Metadata Profiles for eJournals

Answers are based on analysis of the

- Concepts in the domain
- Sources of objects and metadata
- Technical properties of the repository
- Use Cases
  - Functions supported (what is Metadata for?)
  - Workflow (how is Metadata used?)

Creating Metadata Profiles for eJournals

Answers are based on analysis of the

- Concepts in the domain
- Sources of objects and metadata
- Technical properties of the repository
- Use Cases
  - Functions supported (what is Metadata for?)
  - Workflow (how is Metadata used?)

Hands-On Exercise

- What sorts of digital objects need to be described?
- What are the relationships between them?
- What descriptive metadata can you find?
- Can you tell what events the objects have undergone?
- What technical metadata can you find?
- What information can you find that supports fixity, integrity and authenticity?
- What rights information can you find?

Don’t fret over details!

If you get bored have a look at a PREMIS example:

- Object descriptions http://timbusproject.net/component/docman/doc_download/135-metadata-exercise-objects
- Event and Agent descriptions http://timbusproject.net/component/docman/doc_download/136-metadata-exercise-event
- Questions http://timbusproject.net/component/docman/doc_download/134-metadata-exercise-questions

Creating Metadata Profiles

1. Which objects do we describe?
   a. Which?
   b. How many?

- What sorts of domain objects are you wanting to preserve?
- Do you want to describe intellectual entities, representations, files, bitstreams?

Question 1a: Which objects do we describe?

- For eJournals:
  - Journal
  - Issue
  - Article
  - Representation
  - File
  - Submission

D-Lib® Magazine

ISSN: 1082-9873

Question 1a: Which objects do we describe?

- For eJournals:
  - Journal
  - Issue
  - Article
  - Representation
  - File
  - Submission

January/February 2009

Vol. 15 No. 1/2
doi:10.1045/dlib.magazine

Question 1a: Which objects do we describe?

- For eJournals:
  - Journal
  - Issue
  - Article
  - Representation
  - File
  - Submission

A Policy Checklist for Enabling Persistence of Identifiers
Nick Nicholas, Nigel Ward, and Kerry Blinco, *Link Affiliates*
doi:10.1045/january2009-nicholas

Question 1a: Which objects do we describe?

- For eJournals:
  - Journal
  - Issue
  - Article
  - Representation
  - File
  - Submission

- provider specific
- XML
- HTML
- PDF

not identical content

Question 1a: Which objects do we describe?

- For eJournals:
  - Journal
  - Issue
  - Article
  - Representation
  - File
  - Submission

Thumb.jpg in the XML representation

Question 1a: Which objects do we describe?

- For eJournals:
  - Journal
  - Issue
  - Article
  - Representation
  - File
  - Submission

packages contain all the content files, metadata, manifests; for convenience, records provenance information (events) that are shared by many files

Question 1b : How many objects do we implement?

- For eJournals:

write-once architecture => split objects into chunks which are updated together. This avoids, for example, creating new generations of journal objects with every submission of a new issue.

Creating Metadata Profiles

2. Which metadata do we need?
   a. Which do we need?
   b. Which do we get?

Question 2a: Which metadata do we need?

- Which functions are supported by the system and what information do they need?
  - Can the repository demonstrate the fixity, integrity, authenticity of archived materials?
  - What preservation strategies (migration, normalization, emulation, cannonicalization, etc.) will the system implement; how will it use metadata in this process?
- Which relationships exist between objects?
- Which events, agents, rights do we describe?
  - Which of these events change the objects or their metadata?

Question 2a: Which metadata do we need?

For eJournals:

- Which functions are supported by the system and what information do they need?

  Preservation, technical requirements, resource discovery, management information, reading room access, …

Preservation metadata does not exist in isolation!

Question 2a: Which metadata do we need?

For eJournals:

- Which relationships exist between objects?
  - generation, part-of, host, migrated-from, series, preceding, manifestation-of, …
- Which events, agents, rights do we describe?
  - Accession, validation, virus check, uncompress, metadata extraction, format identification, migration, …

Question 2a: Which metadata do we need?

- Outside the repository (index)
- In the repository
  - metadata bundled with the content files
  - metadata embedded in the content files

Question 2b: Which metadata can we get?

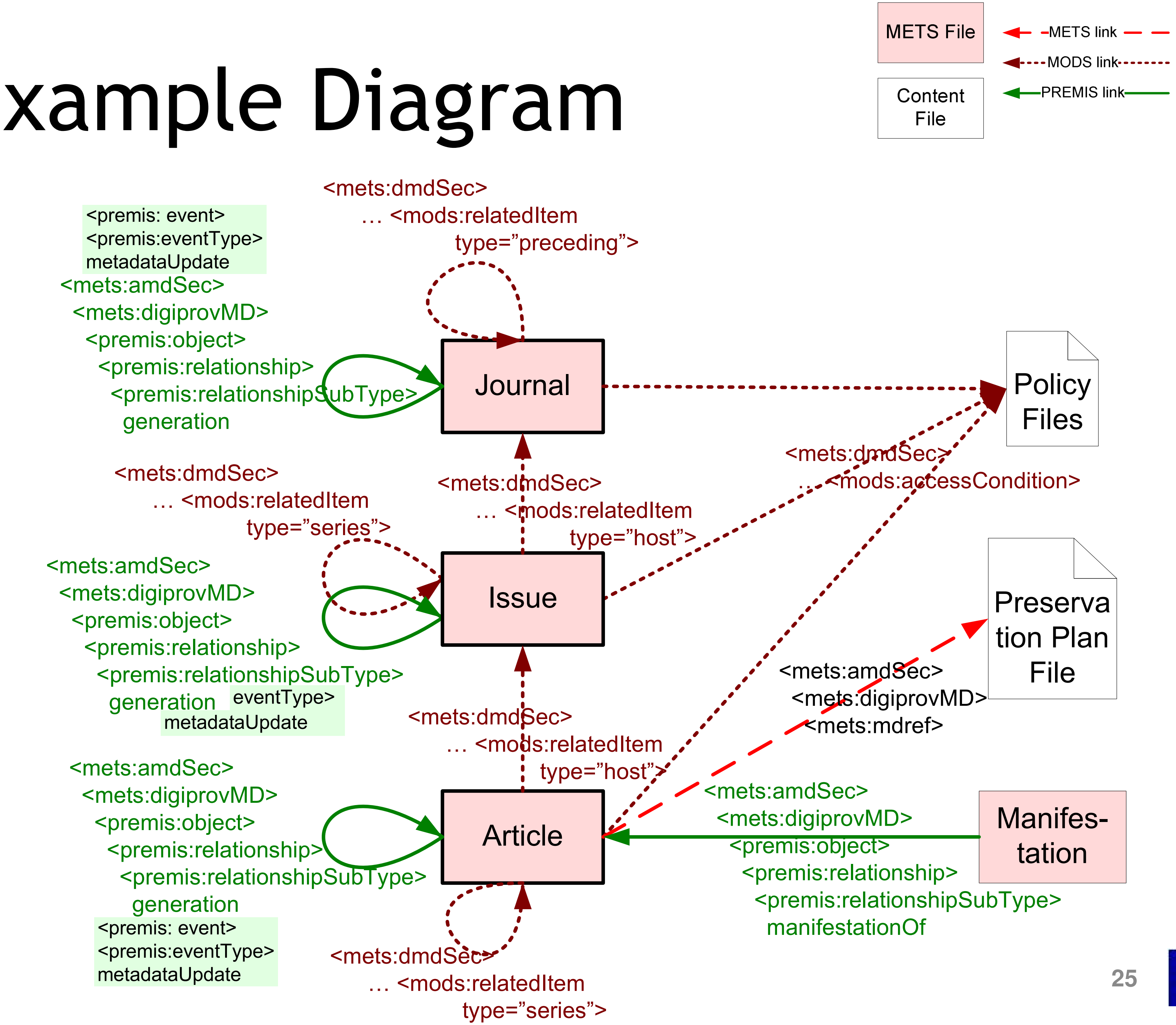
TIMELESS BUSINESS

<u>For eJournals</u>:

- Many suppliers of eJournals to one repository
- Formats of metadata and content are out of the control of the repository
- Translators to the internal metadata format need to be written
- To guide the writing of translators, the metadata profiles need to be very precise so that the translators will produce high-quality, uniform metadata

- Thanks

Example Diagram



METS File

Content File

METS link

MODS link

PREMIS link

<premis: event>
<premis:eventType>
metadataUpdate

<mets:amdSec>
<mets:digiprovMD>
<premis:object>
<premis:relationship>
<premis:relationshipSubType>
generation

<mets:dmdSec>
… <mods:relatedItem
type=”preceding”>

Journal

Policy Files

<mets:dmdSec>
… <mods:relatedItem
type=”series”>

<mets:dmdSec>
… <mods:relatedItem
type=”host”>

<mets:dmdSec>
… <mods:accessCondition>

<mets:amdSec>
<mets:digiprovMD>
<premis:object>
<premis:relationship>
<premis:relationshipSubType>
generation

eventType>
metadataUpdate

Issue

Preservation Plan File

<mets:amdSec>
<mets:digiprovMD>
<mets:mdref>

<mets:dmdSec>
… <mods:relatedItem
type=”host”>

<mets:amdSec>
<mets:digiprovMD>
<premis:object>
<premis:relationship>
<premis:relationshipSubType>
generation

<premis: event>
<premis:eventType>
metadataUpdate

Article

<mets:amdSec>
<mets:digiprovMD>
<premis:object>
<premis:relationship>
<premis:relationshipSubType>
manifestationOf

Manifes-tation

<mets:dmdSec>
… <mods:relatedItem
type=”series”>