PREMIS, METS and preservation metadata: emerging trends and future directions

Eld Zierau
The Royal Library of Denmark
Introduction

My background
- Masters in Computer Science in 1989
- At the Royal Library of Denmark since 2007
  - Strategy and design of preservation systems
  - Creation of preservation policies and strategies
  - Policies of using preservation metadata
- PhD in Digital Preservation in 2011

Currently at the Royal Library
- SIFD – the digital library
  Management, dissemination and preservation
- Packaging and re-packaging for Bit Repository
  WARC, METS, PREMIS
- Framework for OAIS and Distributed Digital Preservation
Contents of this presentation

- Practices (at The Royal Library)
  - Strategies and policies
  - Putting it into practice

- Challenges
  - Expressing preservation levels and intellectual entities
  - Preserving preservation metadata
  - Expressing preservation levels and intellectual entities over time

- An example on the bit level
  - Risks mitigated in bit preservation
  - Bit integrity/safety, confidentiality and availability

- Types of preservation Levels
  - How to express them – also over time

- Identification of intellectual entities
  - How to express them – also over time

- Summary
Preservation Strategies

- **Logical preservation**
  - Migration
  - Emulation
  - Technology preservation

- **Bit preservation**
  0101100010001000

Digital preservation
- Logical preservation
- Bit preservation
Currently at The Royal Library

- **Strategy and policies**
  - Bit preservation
  - Logical preservation

- **Putting it into practice**
  - The chosen Metadata Standards
  - The Digital Library infrastructure
  - The Danish Bit Repository Framework
Metadata Standards and use

Inspired by the Australian way
http://www.dlib.org/dlib/march08/pearce/03pearce.html
So what to do

- PREMIS does not do it all (Richard)
  e.g. other standards for technical metadata
- PREMIS – do not re-invent (Richard)
- Sustainability, Community (Huw Jones, Dave)
- Rights ”packaging” (All)
  as it suits your organisation
- Material – representations (Angela, Robert)
  AIP be careful – different levels, lot of information
- Events when useful on different levels (audit trail)
- All sorts of agents, challenges with description

- METS why – and on what? (Steffen)
- Large METS (Steffen, Huw Jones)
- Tree-structure – not net (as e.g. web or emails)
Digital Library infrastructure

Preservation
- Prefer static
- Simplicity
- Standards

Dissemination
- Prefer dynamic
- New technology
- Add value
- Fast access

Ingest

Management

Access

Representations with metadata
Challenges with metadata

- Preserving preservation metadata (if …)
- Expressing preservation levels
- Expressing preservation levels over time

We need to look more closely on bit preservation to define levels and levels over time

- Expressing Intellectual Entity (identifiers)
- Expressing Intellectual Entity (identifiers) over time
A General View of a Bit Repository

Elements in bit preservation:
- Number of copies
- Independence between copies
- Frequency of integrity checks
Integrity – Bit error

Risk: Bits can change value
1. Error has occurred in Backup
2. File is corrupted
3. Error is not discovered
4. Cannot determine which file is intact

Solutions:
1. No backup. All are copies of data
2. Vote on which copy that is the right one
Bit error – System Layer

Solutions:
1. System layer checks and follow-up on basis of comparing copies
2. Minimum three voters, *optimize by checksums*
Integrity – Bit error

Risk: Bits can change value
1. Error has occurred in Backup
2. File is corrupted
3. Error is not discovered
4. Cannot determine which file is intact

File in form of bit stream

Solutions:
1. No backup. All are copies of data
2. Vote on which copy that is the right one
3. Introduce checksums of files to discover errors
Integrity – Bit error

Risk: The same error occurs for more copies
1. Same hardware
2. Same software
3. Same vendor
Integrity – Bit error

Risk: The same error occurs for more copies
1. Same hardware
2. Same software
3. Same vendor

Solutions:
1. Different hardware solutions
2. Different vendors
3. Different software (OS, interpreters, etc.)
Integrity – Disasters

Risk: All copies are damaged at the same time
1. Natural disasters
2. Attacks in connection with war or terror
Integrity – Disasters

Risk: All copies are damaged at the same time
1. Natural disasters
2. Attacks in connection with war or terror

Solutions:
1. Different geographical locations
Integrity – Organisation

Risk: Errors/mistakes are made by the same person/org.
1. The same person has access and has delete rights
2. The same person makes procedural mistakes
Integrity – Organisation

Risk: Errors/mistakes are made by the same person/org.
1. The same person has access and has delete rights
2. The same person makes procedural mistakes

Solutions:
1. Different organisations
Confidentiality

Risk: Unauthorised gets access to confidential data
1. Unauthorised gets access to Bit Repository
2. Unauthorised gets access to data from Bit Repository

Solutions:
1. Authentication of users of pillars with copies
2. Encryption internally on pillar
3. Hardware secured in locked rooms
Confidentiality – System Layer

Bit Repository (BR)

System Layer

Pillar Layer

Likewise on System layer
Availability

Risk: Cannot get access as required
1. Cannot get any response on request
2. Processing not possible in reality

Solutions:
1. Specialised pillar with distributed architecture
Availability

Solutions:
1. Redirection if access to a pillar is down
2. Distributed requests to different pillars
3. Scaling
4. Diversity,
5. …

Depends on what is required
Bit Repository Offering Solutions

- Media
- Data safety
- Access speed
- On-line
- Off-line
- Organisational placement
- Geographical placement
- …
## Bit Safety

<table>
<thead>
<tr>
<th>Value</th>
<th>Comment for preservationLevelType = BitSafety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>Maximum bit safety</td>
</tr>
<tr>
<td>VeryHigh</td>
<td>Very high bit safety</td>
</tr>
<tr>
<td>High</td>
<td>High bit safety</td>
</tr>
<tr>
<td>Medium</td>
<td>Medium bit safety</td>
</tr>
<tr>
<td>Low</td>
<td>Low bit safety</td>
</tr>
<tr>
<td>VeryLow</td>
<td>Very low bit safety</td>
</tr>
<tr>
<td>Min</td>
<td>Minimum bit safety</td>
</tr>
</tbody>
</table>

from [http://id.kb.dk/vocabulary/](http://id.kb.dk/vocabulary/)
Bit Safety

<table>
<thead>
<tr>
<th>Value</th>
<th>Comment for preservationLevelType = BitSafety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>Maximum bit safety</td>
</tr>
<tr>
<td>VeryHigh</td>
<td>Very high bit safety</td>
</tr>
<tr>
<td>High</td>
<td>High bit safety</td>
</tr>
<tr>
<td>Medium</td>
<td>Medium bit safety</td>
</tr>
<tr>
<td>Low</td>
<td>Low bit safety</td>
</tr>
<tr>
<td>VeryLow</td>
<td>Very low bit safety</td>
</tr>
<tr>
<td>Min</td>
<td>Minimum bit safety</td>
</tr>
</tbody>
</table>

**Policy:**
As high bit safety that we can get

**Strategy 2013:**
10 copies spread over 3 continents, both optical and magnetic medias, checked every …

**Strategy 2050:**
8 copies; at least 2 on Mars, at least two written to DNA, checked every …
## Confidentiality

<table>
<thead>
<tr>
<th>Value</th>
<th>Comment for preservationLevelType = Confidentiality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>Maximum confidentiality</td>
</tr>
<tr>
<td>VeryHigh</td>
<td>Very high confidentiality</td>
</tr>
<tr>
<td>High</td>
<td>High confidentiality</td>
</tr>
<tr>
<td>Medium</td>
<td>Medium confidentiality</td>
</tr>
<tr>
<td>Low</td>
<td>Low confidentiality</td>
</tr>
<tr>
<td>VeryLow</td>
<td>Very low confidentiality</td>
</tr>
<tr>
<td>Min</td>
<td>Minimum confidentiality</td>
</tr>
</tbody>
</table>

from [http://id.kb.dk/vocabulary/](http://id.kb.dk/vocabulary/)
## Confidentiality

<table>
<thead>
<tr>
<th>Value</th>
<th>Comment for preservationLevelType = Confidentiality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>Maximum confidentiality</td>
</tr>
<tr>
<td>VeryHigh</td>
<td>Very high confidentiality</td>
</tr>
<tr>
<td><strong>High</strong></td>
<td><strong>High confidentiality</strong></td>
</tr>
<tr>
<td>Medium</td>
<td>Medium confidentiality</td>
</tr>
<tr>
<td>Low</td>
<td>Low confidentiality</td>
</tr>
<tr>
<td>VeryLow</td>
<td>Very low confidentiality</td>
</tr>
<tr>
<td>Min</td>
<td>Minimum confidentiality</td>
</tr>
</tbody>
</table>

### Policy:
Only restricted access, where it is as hard as it can get for others when skipping encryption

### Strategy 2013:
No more than 2 copies, that are secured on off-line medias ...

### Strategy 2050:
??? ...
## Availability

<table>
<thead>
<tr>
<th>Value</th>
<th>Comment for preservationLevelType = Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>Maximum availability</td>
</tr>
<tr>
<td>VeryHigh</td>
<td>Very high availability</td>
</tr>
<tr>
<td>High</td>
<td>High availability</td>
</tr>
<tr>
<td>Medium</td>
<td>Medium availability</td>
</tr>
<tr>
<td>Low</td>
<td>Low availability</td>
</tr>
<tr>
<td>VeryLow</td>
<td>Very low availability</td>
</tr>
<tr>
<td>Min</td>
<td>Minimum availability</td>
</tr>
</tbody>
</table>
## Logical Preservation

<table>
<thead>
<tr>
<th>Value</th>
<th>Comment for preservationLevelType = logicalStrategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migration</td>
<td>Migration of digital material to keep data interpretable</td>
</tr>
<tr>
<td>Emulation</td>
<td>Emulation of digital material to keep data interpretable</td>
</tr>
<tr>
<td>Technical</td>
<td>Technology preservation to keep data interpretable</td>
</tr>
</tbody>
</table>

...  

*Angela had more*
## Preservation Level information

<table>
<thead>
<tr>
<th>preservationLevelType</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bit safety</td>
<td>Bit preservation</td>
</tr>
<tr>
<td>Confidentiality</td>
<td>Bit preservation</td>
</tr>
<tr>
<td>Availability</td>
<td>Bit preservation</td>
</tr>
<tr>
<td>Logical Preservation Strategy</td>
<td>Logical Preservation</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

- **Policy**
  - With institution preservation policies
  - Express values - Same over time

- **Strategy**
  - Requirements for fulfilment with current technologies ...
Preservation Level in metadata

<digiprovMD CREATED="2013-01-18T19:28:01.456+01:00" ID="Premis1">
  <mdWrap MDTYPE="PREMIS">
    <xmlData>
      <preservationLevel xmlns:xlink="http://www.w3.org.1999/xlink" xsi:…">
        <preservationLevelValue>bitSafetyHigh</preservationLevelValue>
        <preservationLevelDateAssigned>2013-01-18T19:28:01.458+01:00
      </preservationLevelDateAssigned>
    </preservationLevel>
    <preservationLevel xmlns:xlink="http://www.w3.org.1999/xlink" xsi:…">
      <premis:preservationLevelValue>logicalStrategyMigration</premis:preservationLevelValue>
      <preservationLevelDateAssigned>2013-01-18T19:28:01.459+01:00
    </preservationLevelDateAssigned>
    <preservationLevel>
      <preservationLevelValue>confidentialityLow</preservationLevelValue>
      <preservationLevelDateAssigned>2013-01-18T19:28:01.460+01:00
    </preservationLevelDateAssigned>
  </xmlData>
</mdWrap>
</digiprovMD>
Identification in the future

Angela and Robert said: what is an intellectual entity?

As for representations

Go back finding earlier versions (Richard Gartner)
Preservation messages (Huw Jones)
Functions (Eld)
Search rather than identification (tagging, …)?
Packaging – METS (Richard Gartner) -- WARC?

Fragments for commenting (Careful with preservation!!!!)
Preservation Level in metadata

What is an intellectual entity?
Summary

PREMIS, METS and preservation metadata: emerging trends and future directions

- “Choosing” preservation metadata standards
  - Which and how

- Preserving preservation metadata
  - Which and how

- Some challenges for the future
  - Definition of preservation levels and intellectual entities
  - Expressing preservation levels and intellectual entities
  - Expressing preservation levels and intellectual entities over time

*It depends*
Questions and Comments