Using OAIS reference model for storing and preserving large amounts of scientific data

Yonny CARDENAS
cardenas@cc.in2p3.fr

Journées Calcul Données
15 December 2021
Motivations

- Research institutions or projects can produce large amounts of digital data
  - size dataset from some terabytes to petabytes
- Precious dataset
  - non-reproducible data
  - reproducible with disproportionate cost or effort
- When a project become inactive, how about the data?
  - data become orphan and/or obsolete over time 😞
  - nothing can be done !
  - What will be the future of the data?
- Principles F.A.I.R.
  - not specifically cite preservation as a requirement
  - management of data cycle life
  - early planification: DMP
- Key question for the «current» and «future» projects
## Backup vs Preservation

<table>
<thead>
<tr>
<th><strong>Backup</strong></th>
<th><strong>Preservation</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• operational continuity</td>
<td>• patrimony</td>
</tr>
<tr>
<td>• data in production</td>
<td>• precious or finished data</td>
</tr>
<tr>
<td>• data modifiable, in progress</td>
<td>• frozen, validated data</td>
</tr>
<tr>
<td>• all data is potentially concerned</td>
<td>• only selected data</td>
</tr>
<tr>
<td>• frequency: many times (versions)</td>
<td>• frequency: one time</td>
</tr>
<tr>
<td>• short-term content retention (hours, days, weeks, months)</td>
<td>• long-term content retention (years, decades, …)</td>
</tr>
<tr>
<td>• use proprietary technologies</td>
<td>• use open and free technologies</td>
</tr>
<tr>
<td>• strong dependencies</td>
<td>• weak dependencies</td>
</tr>
<tr>
<td>• low interoperability</td>
<td>• full interoperability</td>
</tr>
<tr>
<td>• create and restitution time must be (very)short</td>
<td>• create and restitution time are not critical</td>
</tr>
<tr>
<td>• automatic process (humanless)</td>
<td>• semi-automatic (curation)</td>
</tr>
<tr>
<td>• automatic data removing</td>
<td>• only manual data removing</td>
</tr>
<tr>
<td>• Internal (operational)</td>
<td>• external: dissemination</td>
</tr>
</tbody>
</table>
Digital Preservation: some principles

- predict the scene of a great disaster
  - recover information directly from storage support (e.g. tape)

- information packages
  - wrapped: data object + metadata + administrative information
  - complain with standard specifications
  - self-contained and self-described
  - human-readable and machine-actionable

- strong reduction of technical dependencies
  - the minimum possible
  - use standard, open and widely known technologies
  - archive cannot depend of archive management software (disposable)
  - proprietary solutions are forbidden

- several copies
  - minimal two, three recommended
  - on different technologies (e.g. tape and disk)
  - on different locations (a copy more than 300 km away)
Preservation: OAIS reference model

Producer  OAIS  Consumer

Management
Preservation: OAIS reference model
Common Specification for Information Packages

CISP information package structure

- **IP_Name**
  - **METS.xml**
    - **descriptive**
      - **EAD.xml**
      - **DC.xml**
    - **preservation**
  - **metadata**
    - **preservation**
  - **representations**
    - **representationID**
      - **METS.xml**
      - **metadata**
        - **preservation**
      - **data**
        - **File1**
        - **File2**
      - **documentation**
        - **DOC-format-spec.pdf**
  - **schemas**
    - **EAD.xsd**
    - **DC.xsd**
  - **documentation**
    - **UserManual.pdf**
    - **SubmissionProcess.pdf**
Preserving research data

I want preserve and disseminate my data!
Preserving research data

DATA

Metadata

DMP
Cycle live data and Preservation

DATA

Documentation

Metadata

DMP
Preserving process

DATA

METADATA

DMP

checkum

METS

AIP
(Archival Information Package)

METS: Metadata Encoding and Transmission Standard
Preservation + Dissemination
Using metadata and DMP

- METS
  - Metadata
    - Descriptive
- Dublin Core
  - Metadata
    - Descriptive
- DMP
  - Metadata
    - Administrative
- OTHER
  - e.g. Metadata
    - Technical
- Documentation
- AIP
- Catalogues
Summary

- target large datasets (from several terabytes to petabytes)
- light and flexible implementation of OAIS reference model
  - data curation based on F.A.I.R. process (producer: research project)
  - not addressed to administrative documents (not probatory value)
  - simplified procedures: e.g. ingest (not SEDA protocol)
  - accept all data formats (not migration)
- compliant with european specifications E-ARK/CSIP
  - interoperability
- implemented at CC-IN2P3 using existing infrastructure
  - IRODS
  - Tape library
Thanks to

Céline Guyon
Présidente de l'Association des Archivistes Français (AAF)

Laurent Duplouy
Chef du service multimédias du département de l’audiovisuel
Bibliothèque Nationale de France (BNF)

Lorène Béchard
Responsable fonctionnelle du système d'archivage électronique
CINES

Thomas Leibovici and Patrice Lucas
Phobos: Parallel Heterogeneous Object Store
CEA
Using OAIS reference model for storing and preserving large amounts of scientific data

Yonny CARDENAS
cardenas@cc.in2p3.fr

Journées Calcul Données
15 December 2021
DMP Common Standard Model