Digital Preservation in Finland

Juha Lehtonen
CSC – IT Center for Science Ltd.
12.9.2014
Mission

CSC, as part of the Finnish national research structure, develops and offers high-quality information technology services.

Vision 2015

CSC – Pioneer in the Sustainable Development of ICT Services
CSC at a Glance

CSC – IT Center for Science Ltd.
Owned by Ministry of Education and Culture of Finland
Operates on a non-profit principle

Short history:
- Founded in 1971 as a technical support unit for Univac 1108
- Connected Finland to the Internet in 1988
- Reorganized as a company, CSC – Scientific Computing Ltd. in 1993
- Facilities in Espoo, close to Otaniemi campus (of 15,000 students and 16,000 technology professionals) and Kajaani
- Staff 260
- Turnover 2013: 31.2 million euros
CSC’s Services

- Funet Services
- Computing Services
- Application Services
- Data Services for Science and Culture
- Information Management Services

Universities
Polytechnics
Ministries
Public sector
Research centers
Companies
Terminology

DP system

- Hardware and software for implementing the digital preservation

DP solution

- DP system and the organization maintaining and organizing the digital preservation

DP service

- Part of the DP solution visible for users
  - NDL DP service for cultural heritage (in production)
  - ATT DP service for research data (being planned)
National Digital Library of Finland


The objective of the project is to make the digital data repositories of archives, libraries and museums available to the public now and in the future.

The project additionally promotes interoperability of processes and IT systems in Finnish memory organizations.

The NDL project includes:

- **Common user interface Finna** for the information resources of libraries, archives and museums.
- **Digitisation** of the most essential cultural heritage materials of libraries, archives and museums.
- Development of a **digital preservation solution** for digital cultural heritage.
- National Digital Library works as an **aggregator** for the European Digital Library Europeana.
Digital Preservation in NDL

The Digital preservation (DP) is a system of services which is offered for partner organisations:

- under the administration of the Ministry of Education and Culture
- that preserve cultural heritage

These organisations will transfer the materials intended for long-term or permanent preservation to the NDL’s DP service.

Even in the DP system, the ownership of materials will remain with the organisations which stored them.

Bit level preservation started 2014.

The aim is to have full functionality of the DP system in use by 2016.

Current capacity of the DP system is 0.5 PB which is stored on 3 different media types (disk and 2 tapes).
Enterprise Architecture for NDL

**Public Interface**
- Metadata
- Object request
- External Services:
  - Ontology services
  - Authentication Service
  - Integration Platform
  - Reachability Information
  - Geographical Information
  - Online Payment System

**Digital Preservation**
- Metadata
- Submission package
- Dissemination package
- Support Services
- Standard Portfolio

**Library, Archive and Museum Systems**

**Europeana and other 3rd party services**
Development of NDL DP System

High-quality digital preservation

Continuos processes
- Support services
- Maintenance of specifications
- Management
- Cooperation with ATT DP

Distributed locations
Preserving the intelligibility
Preservation planning and development of preservation actions
Bit-level preservation

Phase I
2011-2013
Phase II
2014-2016
Standard Portfolio

- Early adoption of common standards
- Description of standards with justifications
  - Metadata, interfaces, …
  - File formats, …
- Enables provision of consistent services, combining data, and developing metadata
  - Semantic commensurable
NDL DP Specifications

Recommended file formats
Acceptable file formats for transfer
Administrative and structural metadata
Descriptive metadata
Standard portfolio
NDL METS profiles
BACK-END SYSTEM
SUBMISSION INFORMATION PACKAGES (SIP)
DIGITAL PRESERVATION
Some Metadata Standards

METS

Descriptive metadata:
- MARC21, FINMARC, DC, MODS, EAC-CPF, EAD, LIDO, VRA Core, Film identification (EN15744), DDI
- Additionally, other formats may be used:
  but only additionally and with proper schema

Technical metadata:
- PREMIS:OBJECT, MIX, textMD, audioMD, videoMD

Provenance metadata:
- PREMIS:EVENT, PREMIS:AGENT

Rights metadata:
- Limitations for DP system: PREMIS:RIGHTS
- Copyright metadata: Currently agreed case by case
Submission Information Package

- METS document
- Digital signature
- Digital objects

Siirtopaketti-88564
mets.xml
varmiste.sig
native_imgs
master_imgs
img_001.tiff
img_002.tiff
img_003.tiff
Mandatory Metadata

- Descriptive metadata
- Provenance metadata
  - premis:event
- File format and version
  - premis:format
- Fixity
  - premis:fixity
- Some other technical metadata for some file formats
- Struct map
  - mets:structMap
Mandatory Metadata

- ID for the SIP
  - @OBJID

- IDs for digital objects
  - premis:objectIdentifier

- Creation (and last modification) time of the SIP
  - @CREATEDATE, @LASTMODDATE

- Creation time for metadata sections and files
  - Metadata: @CREATED, @ndl:CREATED
  - NDL attribute extension for EDTF time format
  - Files: premis:dateCreatedByApplication

- Metadata type of wrapped metadata
  - @MDTYPE

- Some other fields
Recommended Metadata

- Technical metadata
  - Partly mandatory
- Rights metadata
- Event history
- Persistent IDs for the metadata sections
  - NDL attribute extension, e.g. techMD@ndl:PID, techMD@ndl:PIDTYPE
- Some other fields
Some Restrictions in NDL METS

- Only one admSec section allowed
- No structLink section allowed
- No behaviourSec section allowed
- No embedded binary data allowed
- No digital object references outside the SIP allowed (everything needs to be in the SIP)
- No mdRef (metadata references outside the METS file) allowed, except mandatory for the preservation plan
- Various controlled vocabularies given
NDL METS Validation

XML validation

- Various metadata formats
- NDL Schema Catalog
  - Forces to use local copies of schema
  - XML Catalogs v. 1.1, OASIS Standard, 2005
    - http://www.oasis-open.org

Schematron rules for more complex validation

- ISO/IEC 19757-3:2006
- Validates via XSLT conversion
- e.g. “For a file, which format is image/tiff, validate that there exists a technical metadata section done with MIX, and that the file and the metadata section are linked together in METS.”
Public Specifications

Specifications available at (in Finnish)
- Standard portfolio: 24.2.2014
- Administrative and structural metadata and packaging material v1.4 (14.4.2014)
- File formats: v1.3 (14.4.2014)

Specifications are updated regularly especially in the early stages
- …but those will be never ready…

METS and SIP specifications were piloted with 10 memory organizations (archives, libraries, museums) in 2012.
## Amount of Data to be Preserved (2011)

<table>
<thead>
<tr>
<th></th>
<th>Number of files (millions)</th>
<th>Size (TB)</th>
<th>Number of files (millions)</th>
<th>Size (TB)</th>
<th>Number of files (millions)</th>
<th>Size (TB)</th>
<th>Number of files (millions)</th>
<th>Size (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2010</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Documents</td>
<td>11,6</td>
<td>328</td>
<td>15,4</td>
<td>394</td>
<td>25,6</td>
<td>646</td>
<td>48,7</td>
<td>1301</td>
</tr>
<tr>
<td>Still Images</td>
<td>1,7</td>
<td>18</td>
<td>2,1</td>
<td>30</td>
<td>3,9</td>
<td>68</td>
<td>6,1</td>
<td>120</td>
</tr>
<tr>
<td>Digital Video</td>
<td>0,1</td>
<td>495</td>
<td>0,2</td>
<td>1143</td>
<td>0,8</td>
<td>3055</td>
<td>1,2</td>
<td>8020</td>
</tr>
<tr>
<td>Sound</td>
<td>1,2</td>
<td>606</td>
<td>1,5</td>
<td>771</td>
<td>2,4</td>
<td>1418</td>
<td>3,7</td>
<td>2176</td>
</tr>
<tr>
<td>References</td>
<td>19,5</td>
<td>1,2</td>
<td>21</td>
<td>1,5</td>
<td>27</td>
<td>2,4</td>
<td>34</td>
<td>3,4</td>
</tr>
<tr>
<td>Web Archive</td>
<td>496</td>
<td>20</td>
<td>646</td>
<td>27</td>
<td>1396</td>
<td>59</td>
<td>2300</td>
<td>97</td>
</tr>
<tr>
<td>Radio and TV Archive</td>
<td>0,8</td>
<td>95</td>
<td>1,2</td>
<td>142</td>
<td>2,9</td>
<td>327</td>
<td>5,0</td>
<td>558</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>530</td>
<td>1563</td>
<td>687</td>
<td>2509</td>
<td>1458</td>
<td>5575</td>
<td>2400</td>
<td>12275</td>
</tr>
</tbody>
</table>
Software Architecture of DP System

Front end applications
- Transfer
- Validation
- Reporting

Back end applications
- Information management
- Archival Storage
- Data management
- Monitoring

Databases

INGEST
ADMINISTRATION
ACCESS
Thank you

http://kdk.fi/en