Practical Experiences with Ingesting Materials for Long-Term Preservation

Esa-Pekka Keskitalo <firstname.lastname@helsinki.fi>
20.10.2011
Digital Preservation Summit 2011, Hamburg
Overview

- About the National Library of Finland
- Digitization at the NLF
- Logistics
- Post-processing
- Metadata
- Preservation
- National Digital Library Initiative
National Library of Finland

- The National Library of Finland ensures the availability of the published national heritage.
- The Library develops library and information services together with the Finnish library network and other actors of the modern information society
- Circa 230 employees, budget 29 million €
Organization and its Challenges

**NLF**

- Admin
- Centre of digitization and conservation (in Mikkeli)
- Research Library (Collections, services to customers)
- National Library Network Services (digital preservation, information systems)

Digitization process

Map showing locations:
- National Library of Finland
- Preservation and Digitization Centre
- Collections
- Mikkeli
- 230 km

**THE NATIONAL LIBRARY OF FINLAND**
Digitization at the NLF

- Based on a Strategy,
- And Policies
  - Preservation Policy
  - Collections Policy
  - Digitization Policy
  - Cataloguing Policy (under revision)
- Aligned with the National Digital Library Initiative
- Internationally acknowledged best practices, EU guidelines, standards

NLF Strategy (will be revised in early 2012): http://www.nationallibrary.fi/infoe/organization/nationallibrarystrategy_20062015_summary.html
Criteria for Selection

- Critical masses; coherent collections
  - For the NLF, digitization of single items is a non-strategic service
- Preservation: digitization for protection
- Demand
- Contents: national or international value
Challenges

- Long distance between locations: reliable tracking needed for large batches
- Gaps in cataloguing → integration of cataloguing and digitization
- Complex system of collections
- Cooperation over organizational boundaries
- Preservation vs Dissemination
Improving Logistics

Old Process

Decentralized workflow
- Material assembled at the collections
- Scanning at a remote location
- Conversion at another location
- Material returned to the collections

New Process

Seamless overview and control
- Batch created at the collections
- Gapless Status & Location Tracking at a glance
- All have access to the same web interface (project managers, scan operators, QA operators etc.)
- Finally, batch disassembled at collections
Need to track the items on every stage
Remote Item Tracking

<table>
<thead>
<tr>
<th>Batch Name</th>
<th>Created By</th>
<th>Created On</th>
<th>Current Location</th>
<th>Current Status</th>
<th>Days in Production</th>
<th>Physical Units</th>
<th>To be scanned</th>
<th>To be processed</th>
<th>Cataloging needed</th>
<th>In Conservation</th>
<th>Ready/Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEOLOG-2009001</td>
<td>DWlasn</td>
<td>05.01.2010</td>
<td>Helsinki</td>
<td>Unloaded</td>
<td>37</td>
<td>51</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TEOLOG-2009002</td>
<td>DWlasn</td>
<td>13.01.2010</td>
<td>Mikkel</td>
<td>Processing</td>
<td>38</td>
<td>52</td>
<td>2</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>TEOLOG-2009003</td>
<td>DWlasn</td>
<td>13.01.2010</td>
<td>Mikkel</td>
<td>Shipping back from production</td>
<td>42</td>
<td>47</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TEOLOG-2009004</td>
<td>DWlasn</td>
<td>14.01.2010</td>
<td>Mikkel</td>
<td>Processing</td>
<td>41</td>
<td>61</td>
<td>27</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>TEOLOG-2009005</td>
<td>DWlasn</td>
<td>15.01.2010</td>
<td>Mikkel</td>
<td>Processing</td>
<td>41</td>
<td>57</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>34</td>
</tr>
<tr>
<td>TEOLOG-2009006</td>
<td>DWlasn</td>
<td>15.01.2010</td>
<td>Mikkel</td>
<td>Processing</td>
<td>41</td>
<td>63</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>4</td>
<td>50</td>
</tr>
<tr>
<td>TEOLOG-2009007</td>
<td>DWlasn</td>
<td>18.01.2010</td>
<td>Mikkel</td>
<td>Processing</td>
<td>37</td>
<td>43</td>
<td>37</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>TEOLOG-2009008</td>
<td>DWlasn</td>
<td>19.01.2010</td>
<td>Mikkel</td>
<td>Processing</td>
<td>37</td>
<td>52</td>
<td>44</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>TEOLOG-2009009</td>
<td>DWlasn</td>
<td>19.01.2010</td>
<td>Mikkel</td>
<td>Processing</td>
<td>37</td>
<td>67</td>
<td>67</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TEOLOG-2009010</td>
<td>DWlasn</td>
<td>19.01.2010</td>
<td>Mikkel</td>
<td>Processing</td>
<td>37</td>
<td>64</td>
<td>20</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>28</td>
</tr>
</tbody>
</table>
Item Tracking: Entities

1) Project
   - Location in collections
   - Genre (monograph, newspaper, etc.)
   - Can be tracking only (no digitization or conversion involved)

2) Batch
   - Collection of items (books, newspapers, manuscripts, etc.)
   - Assembled for transportation to digitization center

3) Physical Unit
   - Single item (e.g. book, microfilm reel, box of parchments)
   - Unique ID / Barcode

4) Logical Unit
   - Smallest unit in item tracking
   - can be book chapter, newspaper issue, part of a series, …
Post-Processing

- DocWorks (CCS)
- Identifying language and font type
- Cropping, deskewing, file format validation etc.
- Identification of structural elements
  - Labour intensive – how long can you go?
- OCR (when applicable); not proofread, but you can help: http://www.digitalkoot.fi/en/splash
- Identification of structural elements with DocWorks (CCS)
Crowdsourcing

In Mole Bridge you write the words where the computer failed. Watch the tutorial!

In Mole Hunt you try to identify words the computer has misread. Watch the tutorial!
METS – wrapping for preservation

- XML Schema for creating a document that explains and structures a digital object and its metadata
- [http://www.loc.gov/standards/mets/](http://www.loc.gov/standards/mets/)
- Information may be embedded or be referenced to
- Uses:
  - Transmit objects to others
  - Rendering
METS profiles

- Different content types require different profiles
  - Monographs
  - Ephemera
  - Periodicals
  - Parchment Fragments
  - Etc.
METS Package Contents

- METS with MARC, MODS, PREMIS and MIX
- For each page:
  - master image in JPEG2000 (lossless)
  - access image in JPEG
  - thumbnail
  - OCRd text in ALTO XML
- Single PDF with hidden text layer
- Whole SIP compressed to a zip file
PREMIS in METS

• PREMIS distributed within different METS sections
• Common amdSec for the intellectual entity
  – Events and Agents in digiprovMDs
  – Rights in rightsMD
• amdSec for each file
  – PREMIS Object in techMD
    ▪ objectCharacteristicsExtension containing MIX 2.0
  – PREMIS Events in digiprovMDs
  – Agents referenced from the common amdSec
National Digital Library (NDL) Initiative

- www.kdk.fi
- Improves access to digital materials…
- and preservation of digital materials
- Comprises libraries, archives, and museums
- Until end of 2013
NDL User Interface

- Common User Interface Infrastructure
- Customizable by organization, region, community
- Integrated contents and services
- Ex Libris Primo
- Interoperability of content systems need a lot of work
- Work for NDL benefits
- Into production in 2012
NDL and Long-Term Preservation

- Work in progress
- Aim: a shared / centralized long-term preservation system
- Technology run by CSC – IT Centre for Science (non-profit company owned by the Ministry of Education and Culture)
- Different models of service for different organizations
NDL Specifications for Preservation

- Accepted file formats and their profiles
- Accepted METS profiles
- Specifications for metadata, e.g.
  - Minimum requirements
  - Expression of access rights and restrictions
  - Identifiers
- Transfer procedures between systems
Links

- http://digi.kansalliskirjasto.fi
- https://www.doria.fi/handle/10024/69173