One process to rule them all? The role of a repository platform in the management of digitized cultural heritage at the National Library of Finland

Open Repositories, DSpace User Group, Charlottetown, PEI, July 12, 2013

Jyrki Ilva (jyrki.ilva@helsinki.fi) & Esa-Pekka Keskitalo (esa-peka.keskitalo@helsinki.fi)

http://urn.fi/URN:NBN:fi-fe201306304112

This map: http://urn.fi/URN:NBN:fi-fe2012120510135
Cultural Heritage & Institutional Repositories

- The National Library of Finland provides repository services for nearly 40 Finnish organisations.
- The National Library also has large collections of cultural heritage materials.
- DSpace is used for both.

Cultural heritage and current research: is there synergy or is there a conflict?
Digitized Cultural Heritage

- Newspapers
- Journals
- Books
- Maps
- Manuscripts
- Sound
- Ephemera
DSpace and cultural heritage materials

- DSpace replaced an unsuccessful proprietary software in 2006.
- DSpace is flexible enough to be used for many kinds of materials, including digitized cultural heritage.
- There is another platform for newspapers and periodicals.
DSpace at the National Library

- National Library has 7 public DSpace instances,
  - 3 contain cultural heritage materials: Doria, Fragmenta membranea, Fenno-Ugrica
- Located in the same virtual server infrastructure
  - DSpace code is managed with SVN
- Some special development has been made, e.g.
  - HTML5-based viewer tool for METS files (not directly integrated with DSpace)
  - SimpleREST interface for DSpace
  - Simplestats statistics tool for DSpace
A generic system vs. a specialized one?

- The funding for digitization projects comes in bits and pieces and usually ends at some point.
  - A new dedicated platform for each new project is not feasible
  - Both initial adoption costs and long-term management costs
- Minimizing the number of platforms, using them for many purposes
  - Cultural heritage materials and related processes are heterogeneous – one must be prepared to make adjustments
One process vs. many processes?

- Would it be possible to develop one common process for the handling of all of the digitized materials?
  - The National Library has been developing processes that would produce standardized METS files suitable for long-term preservation.
  - There are other parts of the process that have been (or could be) standardized and automated as well.

- However, there are many kinds of different materials that require special solutions at some points of the process.
Fragmenta membranea
(http://fragmenta.kansalliskirjasto.fi)

- A collection of medieval parchment fragments digitized by the National Library
- Contains most of what survives of the medieval literature circulated in Finland before the Reformation
- After the Reformation the parchment leaves confiscated by the Crown were reused as covers for bailiff’s account books
Working with scholars

- The metadata was created by a group of scholars specializing on the identification of medieval manuscripts
- Due to special characteristics of the material and the metadata, the user interface of Fragmenta has been heavily modified
Making the parchments browsable

- Digitized parchments contain a lot of detailed information (=big file size necessary)
- To make it easier to browse the documents at various levels of magnification, a special HTML5-based viewing tool was added to the user interface
DSpace is not a standalone system

- Metadata for cultural heritage materials mostly comes from library databases
  - Aleph and/or Voyager, metadata in MARC21
- We use DocWorks in the digitization and production METS files.
- METS files contain both the metadata (as MARCXML) and full-text files
- METS contents converted into the ingest format used by Dspace.
- METS files are stored separately for preservation
Original

Catalogue

Digitization Package

Published Item

Archival Package
DSpace and other systems

- Processes should be modular
  - DSpace does not need to do everything
  - Some things *can* be done within DSpace, but are better done elsewhere
- System managers would generally like to do things in their "own" system
  - But you need to look at the larger picture
Where is the master data and how is it updated?

- **Fennica**, the Finnish national bibliography, contains the master data on Finnish books
  - Can handle only metadata, no full-text files
  - Only a small number of people can update the data
  - Rules created for printed matter
- No support for multilingual metadata – would be useful
  - In some cases multilingual fields have been added in metadata conversions, but can be automated only to some extent
- We may have to wait for the next-generation library system and the adoption of new description formats
Master data and interoperability

- In an ideal world all of the master data would be in one place and any changes made to it would be propagated to other systems.
- So far works only on a very modest scale between the DocWorks system of the Digitization Centre and DSpace.
- Ephemera collection in Doria: Corrected METS data is re-sent to the repository, and a new record automatically replaces the old version.
For whom are we building our services?

- Libraries and other memory organizations have their own inherent motives for the digitization and long-term preservation of their assets.
  - However, the services produced by the libraries don’t always serve the short-term needs of the scholars or other potential end users.
- We need to maintain a dialogue with our stakeholders and try to find out what kind of materials or services they really need from us.
  - How to balance the digitization done for internal reasons with projects that are done in co-operation with the scholars?
Fenno-Ugrica

(http://fennougrica.kansalliskirjasto.fi)

- A special repository for the researchers of Finno-Ugric languages
- Contains books and newspapers published in Finno-Ugric languages in the Soviet Union in the 1920s-1930s
- Materials digitized in co-operation with the National Library of Russia (St. Petersburg)
- The Digitization Project of Kindred Languages has been funded by the Kone Foundation
OCR correction and annotation tools

- The materials are digitized for the needs of linguistic research – near-perfect quality required
- Poor printing quality & obscure languages => the results of automatic OCR far from optimal
- The National Library has been working on an HTML5-based OCR correction and annotation tool for scholars
Work in progress

- OCR correction and annotation of the corrected text contained in a PDF file works fine in most cases
- Can deal with texts that have several separate columns
- Works with both Latin and Cyrillic letters
- The synchronization of the results back to the repository still requires work
- Integration with DSpace?
Re-use of the data / metadata

- The National Library has decided to use CC0 license for the metadata of its digital collections and for the out-of-copyright materials they contain.
- Some of the materials are already available via Europeana and Finna.
Future challenges: Newspapers and journals?

- The National Library has digitized all Finnish newspapers up to 1910 and most of the other periodicals up to 1944.
- Millions of pages freely available, hugely popular service.
- Crowdsourcing solutions being developed for the correction of OCR’d text and the annotation of the content.
Future challenges (2)

- The technical platform for newspapers and journals is getting old
- Needs to be replaced at some point
- One option would be to build on some repository platform