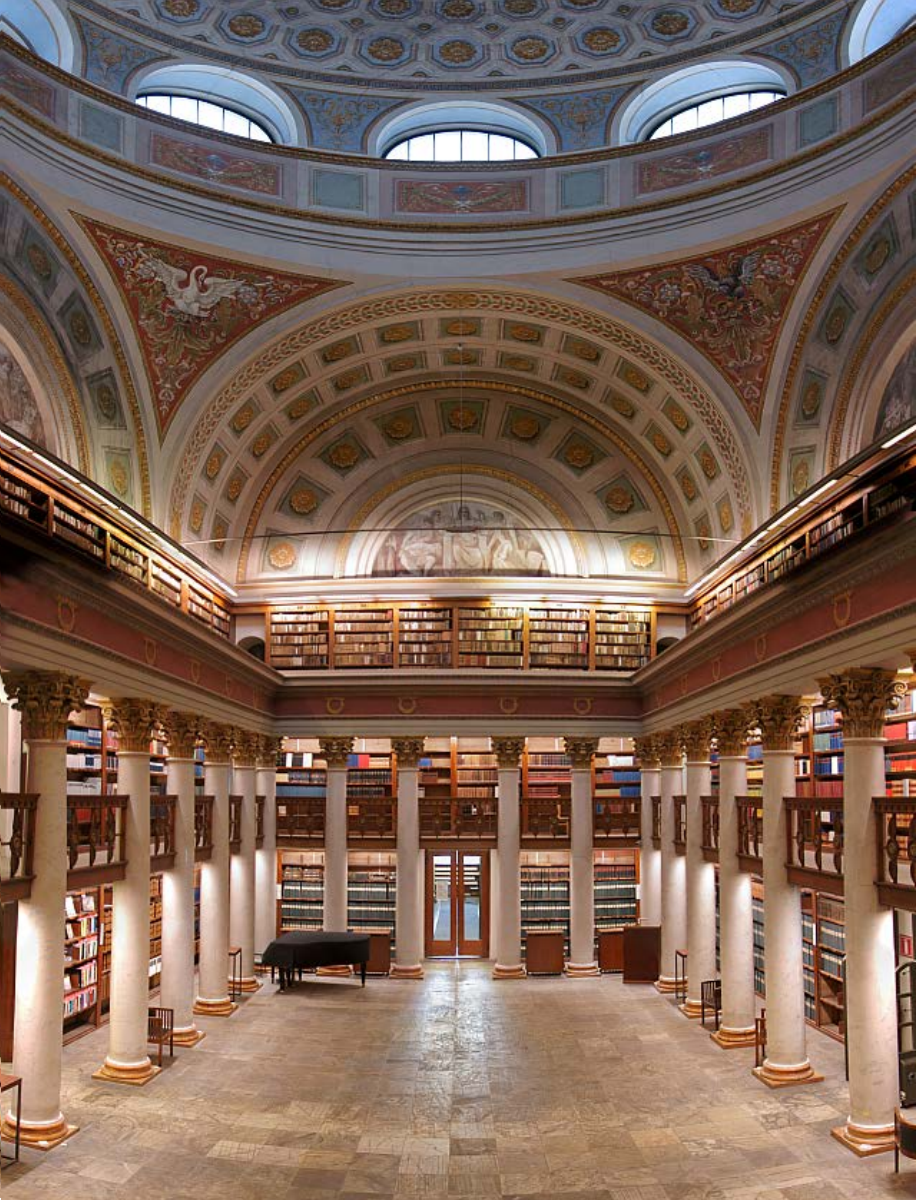
One process to rule them all? The role of a repository platform in the management of digitized cultural heritage at the National Library of Finland

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

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

THE NATIONAL LIBRARY OF FINLAND

Digitized Cultural Heritage

- Newspapers
- Journals
- Books
- Maps
- Manuscripts
- Sound
- Ephemera

NATIONAL LIBRARY NETWORK SERVICES

THE NATIONAL LIBRARY OF FINLAND

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

THE NATIONAL LIBRARY OF FINLAND

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

Monthly download statistics

Parent hierarchy: The entire DSpace / Kansalliskirjasto /

	1 / 2013	2 / 2013	3 / 2013	4 / 2013	5 / 2013	6 / 2013	Total
Pienpainatteet (5329 items, 83411 bitstreams, 128456 megabytes)	95150	24305	16357	13265	13913	13442	**176432**
Eläinsuojelu (383 items, 5279 bitstreams, 5768 megabytes)	6577	331	1912	886	804	1004	**11514**
Elämisentaito (63 items, 1465 bitstreams, 1118 megabytes)	6152	1139	1671	1862	308	2007	**13139**
Kansanhuolto (892 items, 7196 bitstreams, 14197 megabytes)	9621	1002	737	1483	2166	1088	**16097**
Kyläjärjestykset (31 items, 527 bitstreams, 746 megabytes)	1027	444	108	110	422	40	**2151**
Matkailumainokset (760 items, 17820 bitstreams, 22029 megabytes)	14675	5449	2149	2815	3111	2396	**30595**
Moottoriajoneuvojen esitteet ja hinnastot (1914 items, 37035 bitstreams, 67792 megabytes)	34473	12116	4695	3293	2952	2336	**59865**
Naisasia (341 items, 4547 bitstreams, 4955 megabytes)	6779	360	2632	668	489	602	**11530**
Olympialaiset (72 items, 1291 bitstreams, 1774 megabytes)	2736	742	136	398	589	194	**4795**
Pyöräily (208 items, 3639 bitstreams, 4405 megabytes)	5179	2032	722	721	1451	2220	**12325**
Ratsastus (135 items, 2166 bitstreams, 2852 megabytes)	2195	168	417	217	508	409	**3914**
Rauhanasia (140 items, 954 bitstreams, 1147 megabytes)	1740	119	739	124	602	375	**3699**
Ruokalistoja ja ohjelmalehtisiä (320 items, 277 bitstreams, 575 megabytes)	1370	286	340	328	288	456	**3068**
Soutu (70 items, 1215 bitstreams, 1093 megabytes)	2626	117	99	360	223	315	**3740**

Top 10 downloads for this community (Pienpainatteet)

	1 / 2013	2 / 2013	3 / 2013	4 / 2013	5 / 2013	6 / 2013	Total
Kuinka hyvä aviomies voi kesyttää pahan vaimonsa?	1085	626	308	1407	164	1803	**5393**

A generic system vs. a specialicized 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

16 40
THE NATIONAL LIBRARY OF FINLAND

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

NATIONAL LIBRARY NETWORK SERVICES

FRAGMENTA MEMBRANEA

Suomeksi In English

Homepage > Collection List

Search Help

+ more options

The Fragmenta membranea collection

The Fragmenta membranea collection at the National library of Finland contains the vast majority of the remains of books written and used in the eastern parts of medieval Sweden, the Diocese of Turku. The Fragmenta membranea database contains 9,319 digitized parchment leaves meaning 18,638 pages which come from approximately 1,500 different medieval manuscripts.

The leaves in the collection come from books that were confiscated by the Crown in the sixteenth and seventeenth centuries. These were then unbound and the loose leaves were reused by the central administration as covers for bailiffs' account books. Because of this most of the leaves have not only medieval content but also notes made by the bailiffs and the administration in the sixteenth and seventeenth centuries. The parchment covers were removed from the account books in the 19th century. The account books themselves are currently preserved in the National Archives whereas the covers constitute the Fragmenta membranea collection.

>> Introductory article

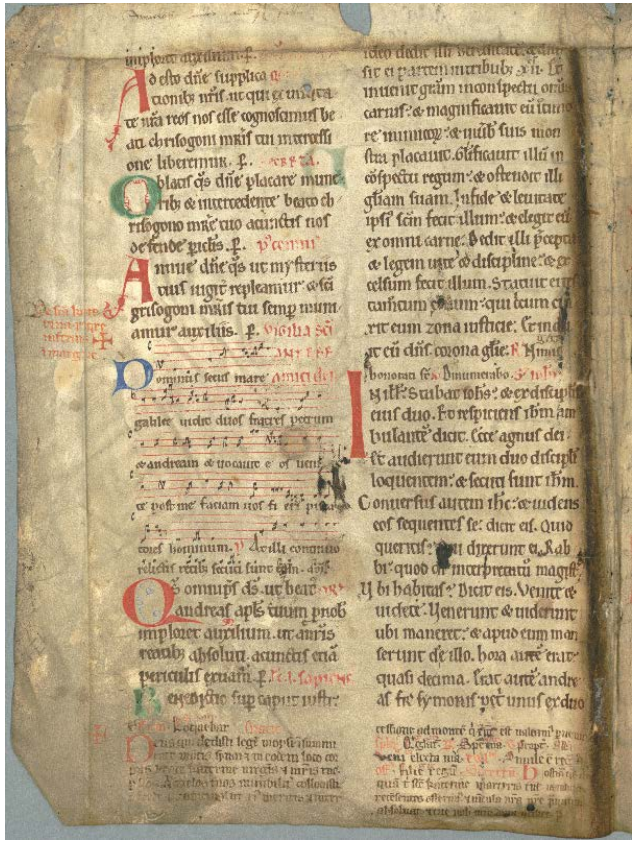
browse by

shelfmark
author
date
contents
origin
collection hierarchy

Browse Collection

- **Fragmenta membranea [1569]**
  - F.m.I (Missalia) [369]
  - F.m.IIa (Gradualia) [129]
  - F.m. IIb (Lectionaria) [7]
  - F.m.III (Breviaria) [190]
  - F.m.IV (Antiphonaria) [203]
  - F.m. IVb [3]
  - F.m.Va (Biblia Sacra) [53]
  - F.m.Vb (Theologiae Auctores) [129]
  - F.m.Vc (Varia) [27]
  - F.m.VII (Varia liturgica & hagiographica) [137]
  - DIG (Luetteloimaton aineisto) [318]
  - Catalog [4]

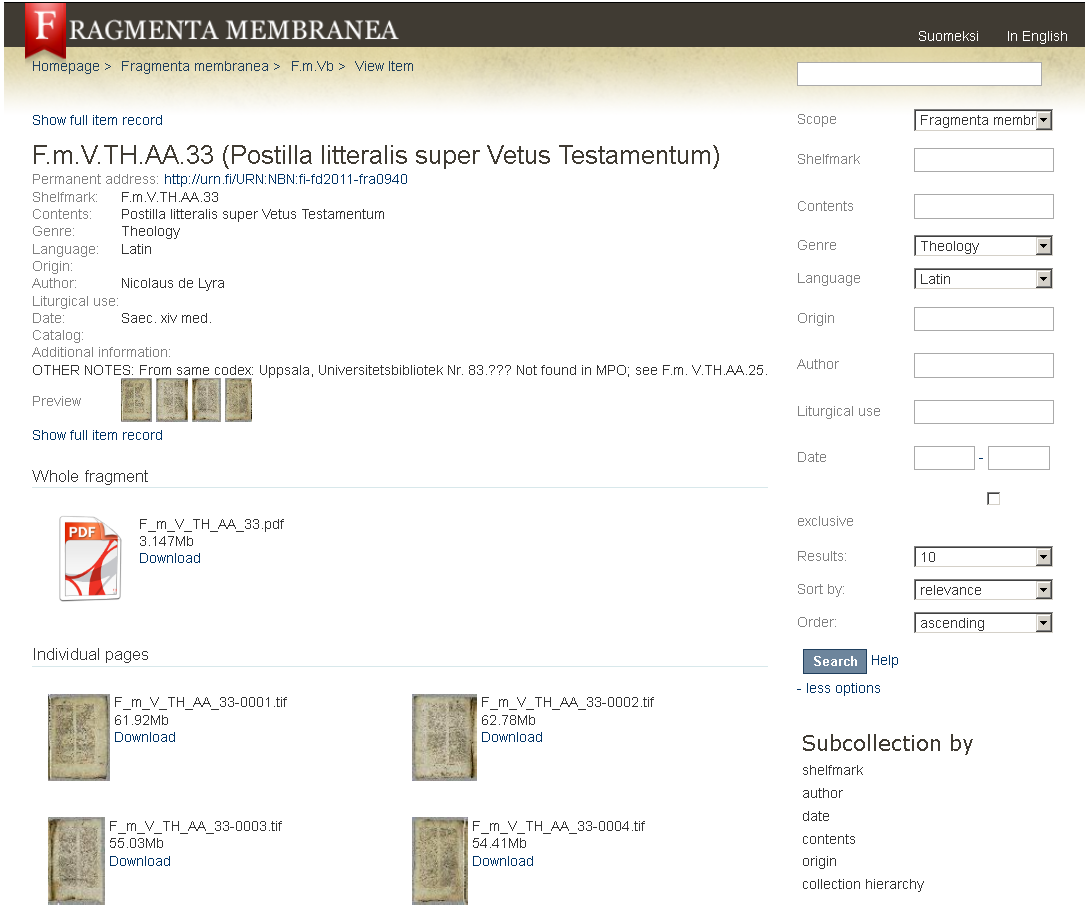
doria-oa@helsinki.fi



THE NATIONAL LIBRARY OF FINLAND

Working with scholars

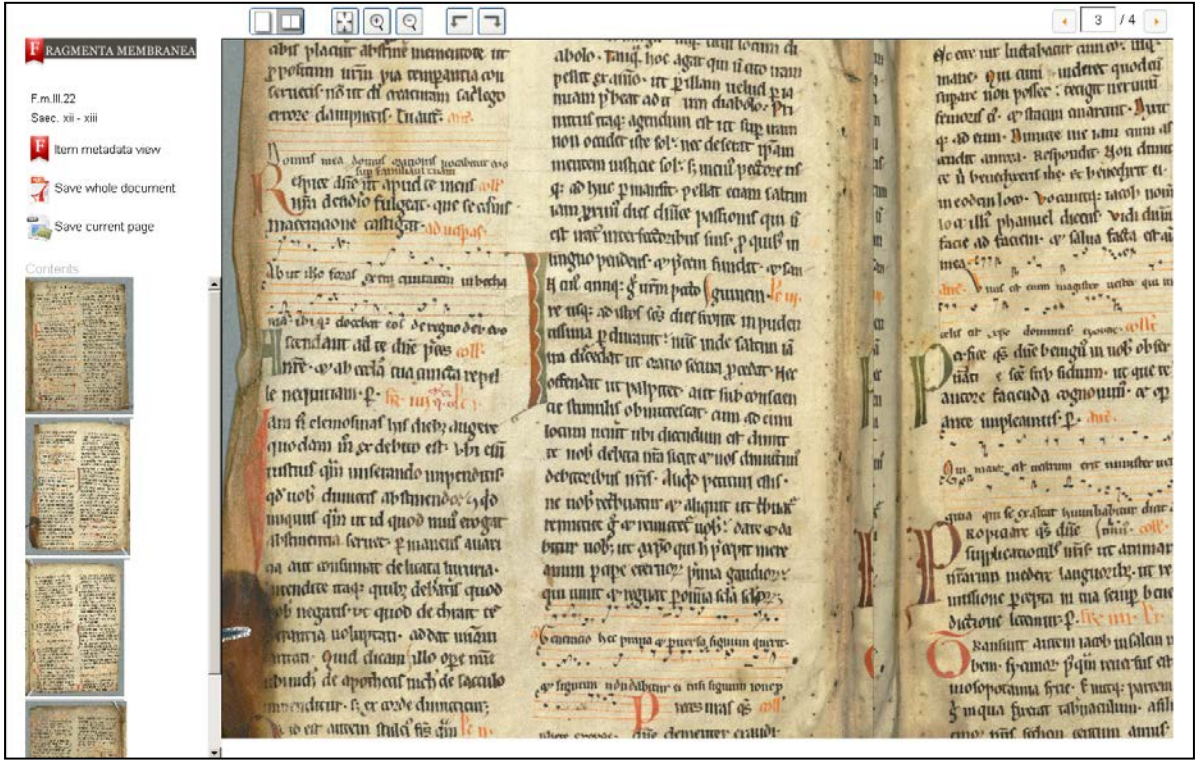
- The metadata was created by a group of scholars specializing on the identification of medieval manuscripts
- Due to special characteristics of the material and the metadata, the user interface of Fragmenta has been heavily modified



THE NATIONAL LIBRARY OF FINLAND

Making the parchments browsable

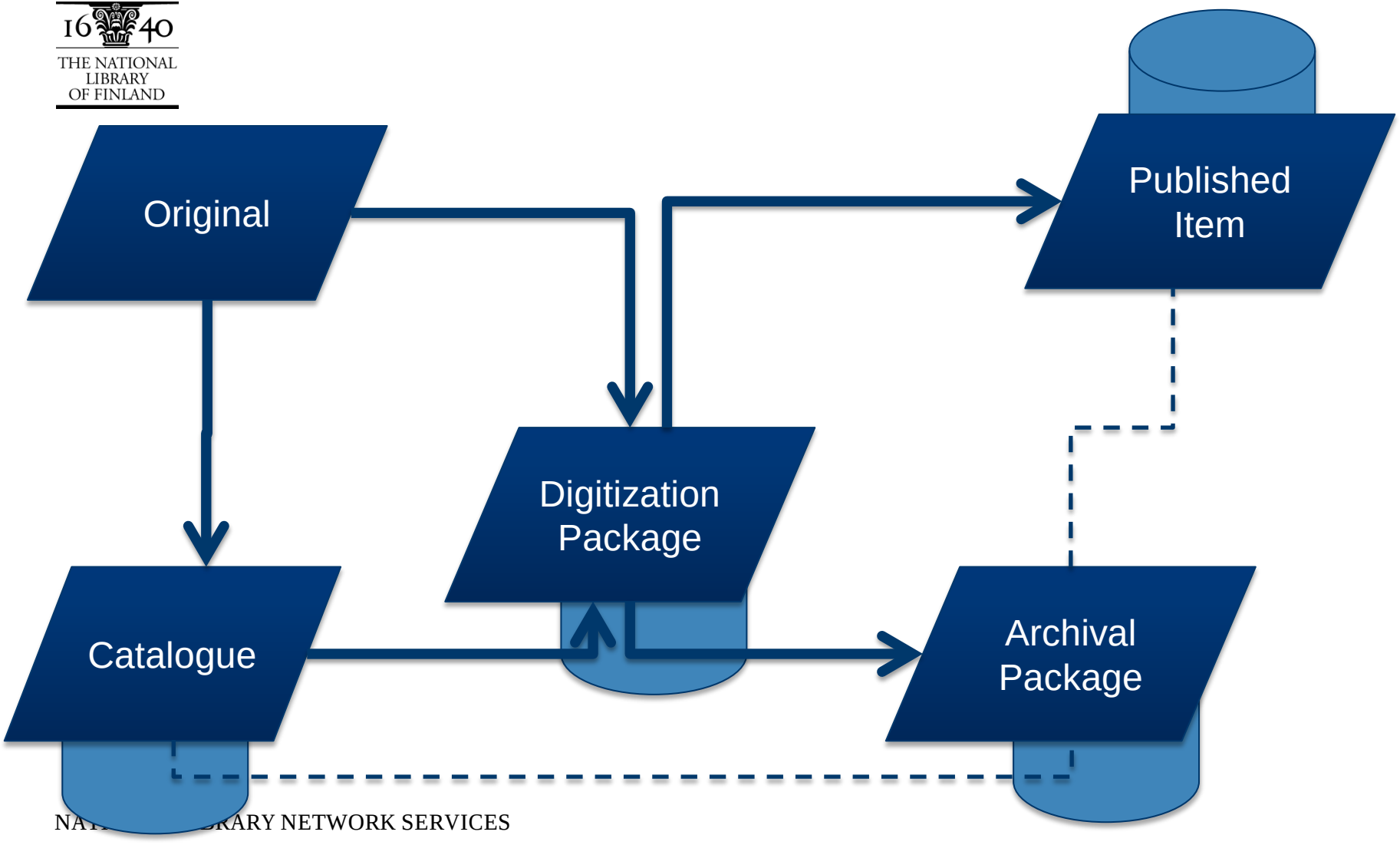
- Digitized parchments contain a lot of detailed information (=big file size necessary)
- To make it easier to browse the documents at various levels of magnification, a special HTML5-based viewing tool was added to the user interface



THE NATIONAL LIBRARY OF FINLAND

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
Published Item
Digitization Package
Catalogue
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

THE NATIONAL LIBRARY OF FINLAND

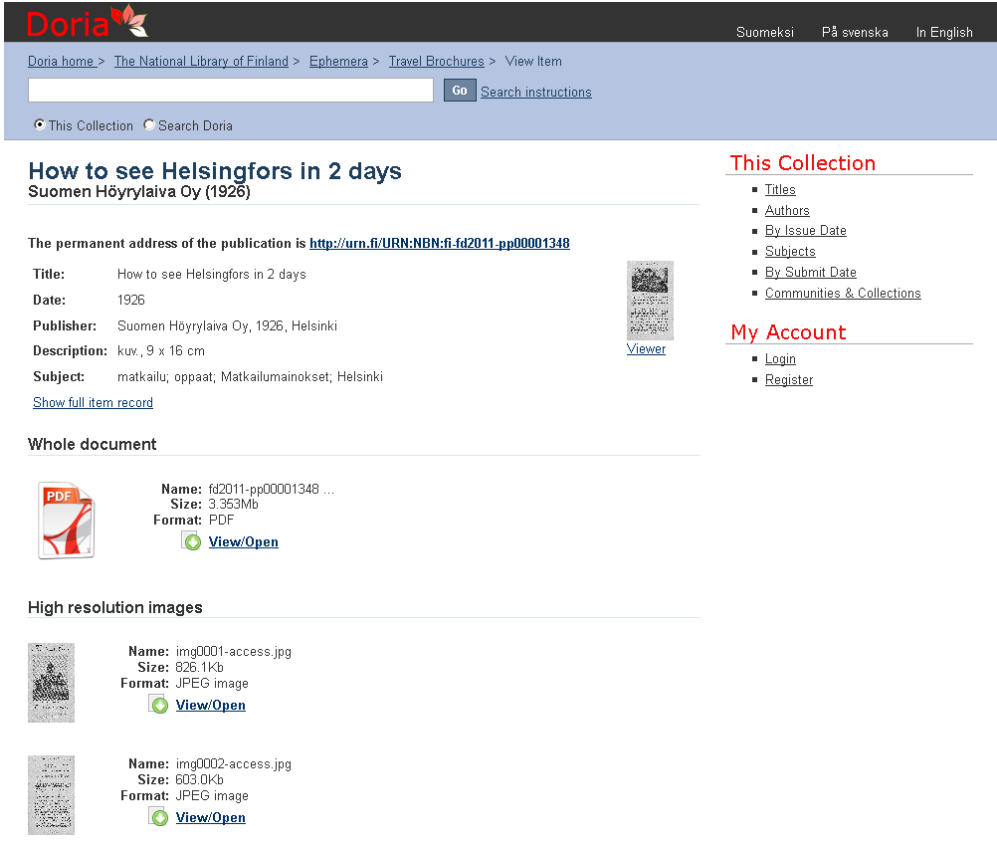
Where is the master data and how is it updated?

- 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

THE NATIONAL LIBRARY OF FINLAND

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?

THE NATIONAL LIBRARY OF FINLAND

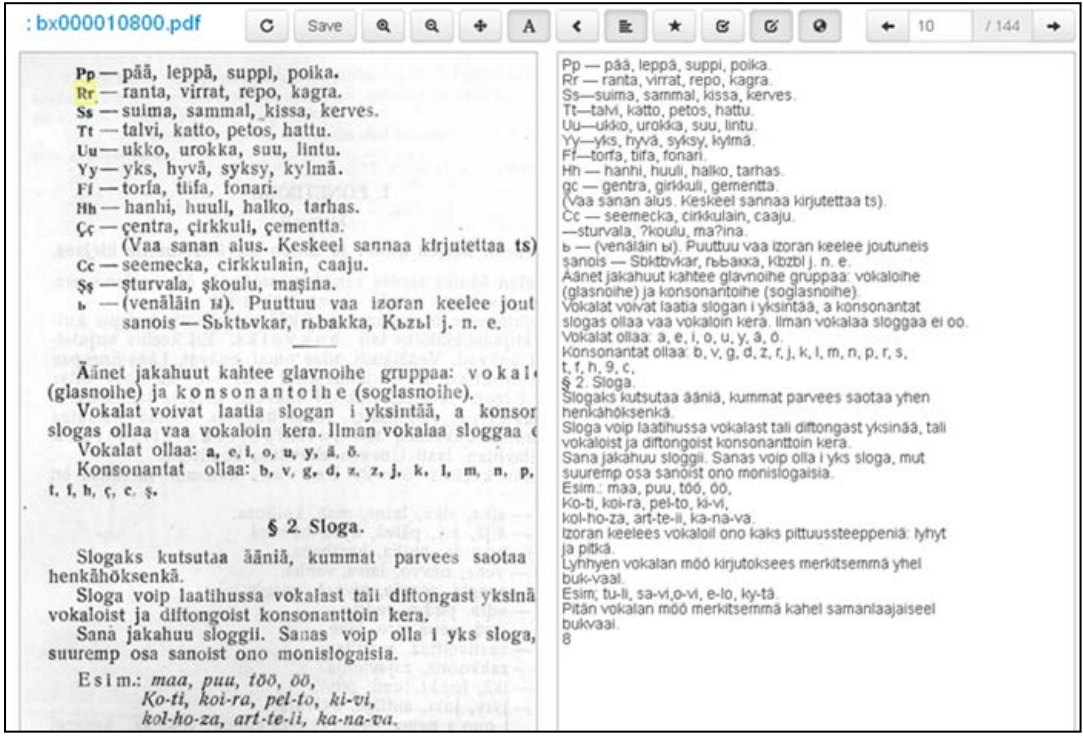
Fenno-Ugrica

(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

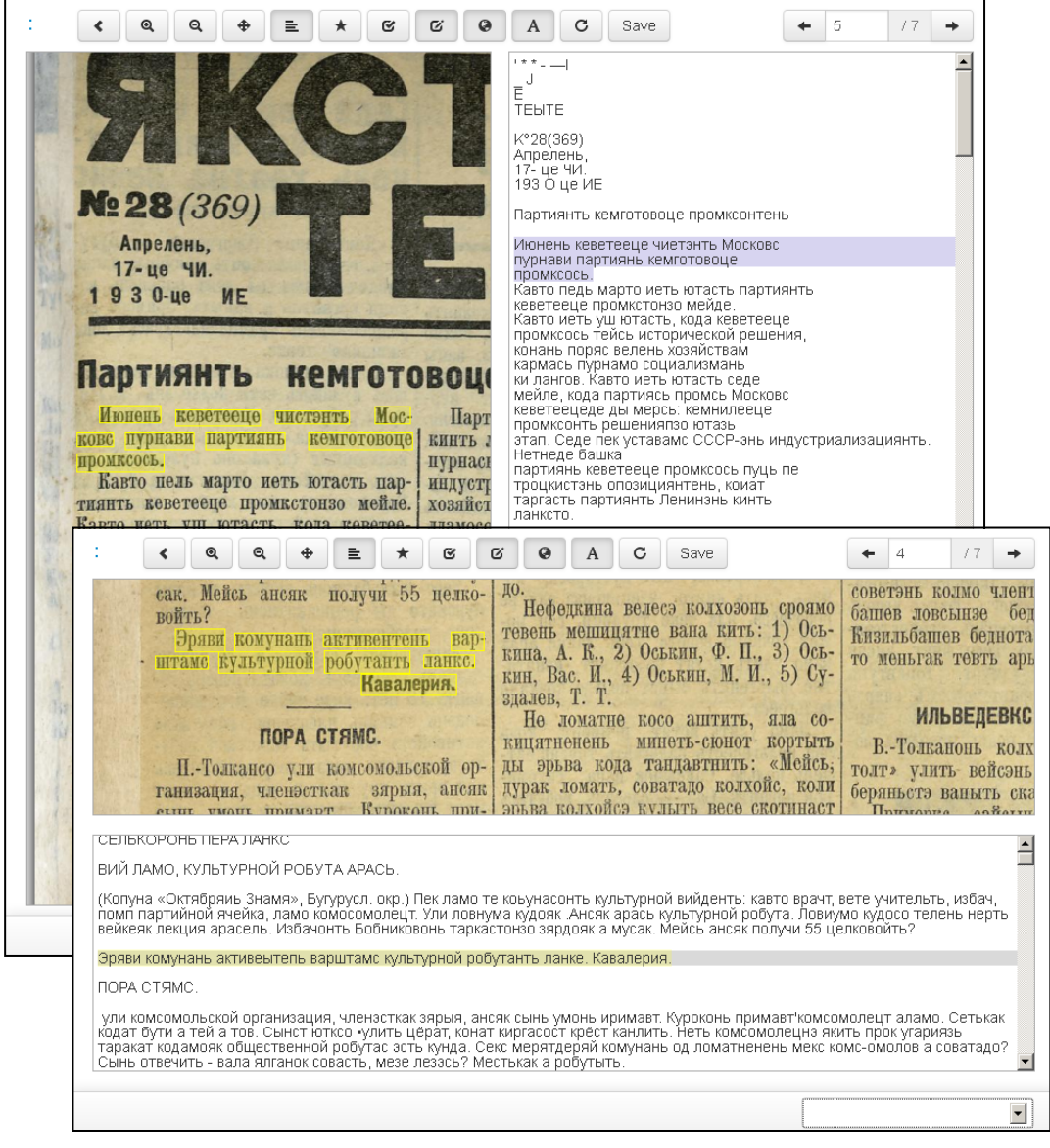
- The materials are digitized for the needs of linguistic research – near-perfect quality required
- Poor printing quality & obscure languages => the results of automatic OCR far from optimal
- The National Library has been working on an HTML5-based OCR correction and annotation tool for scholars



THE NATIONAL LIBRARY OF FINLAND

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?



THE NATIONAL LIBRARY OF FINLAND

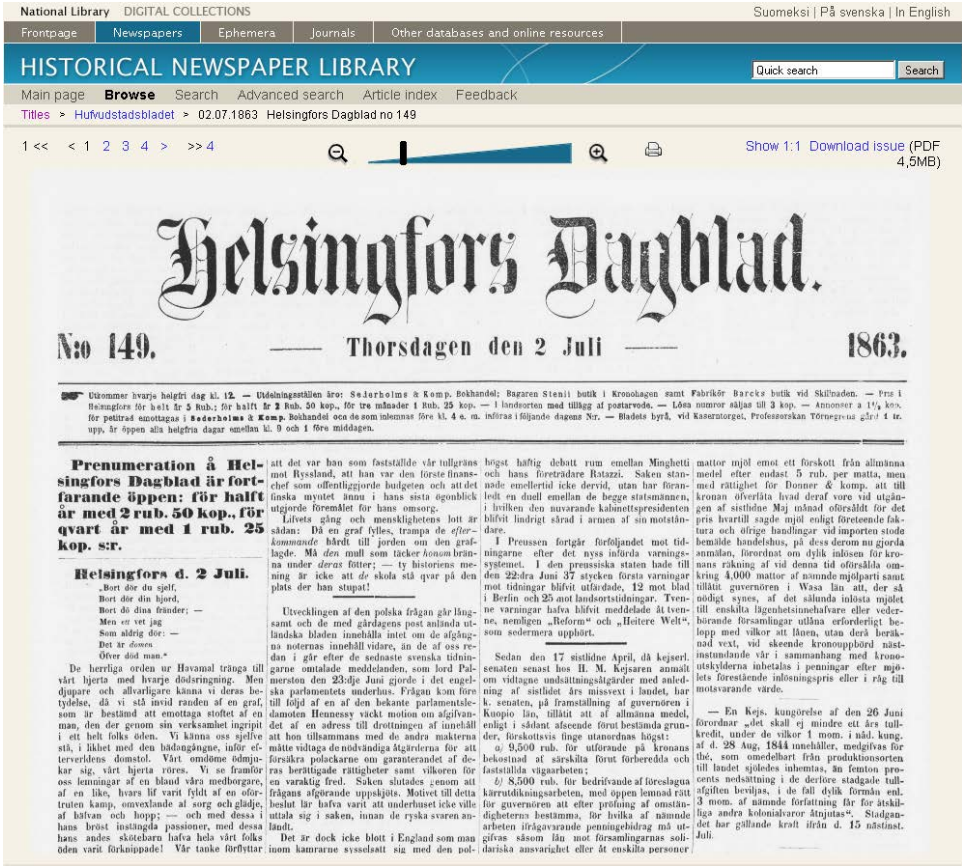
Re-use of the data / metadata

- 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

THE NATIONAL LIBRARY OF FINLAND

Future challenges: Newspapers and journals?

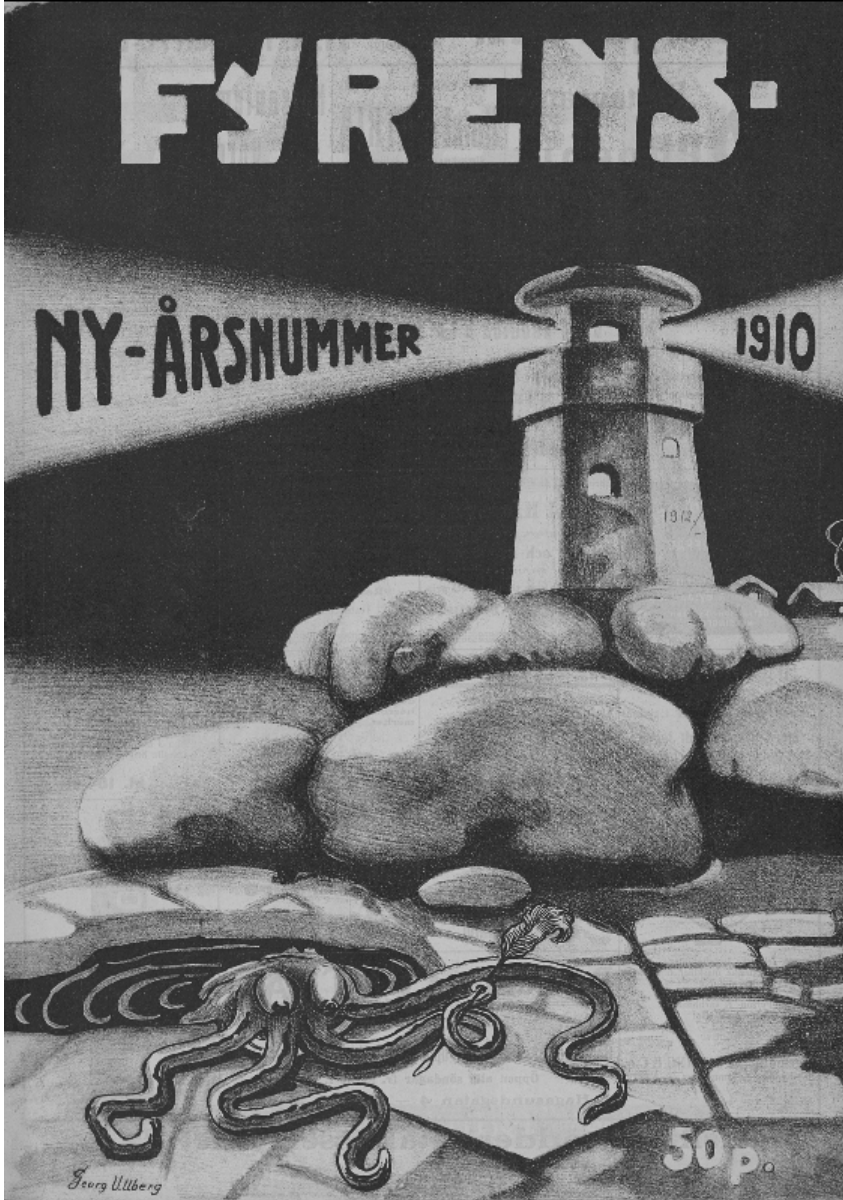
- The National Library has digitized all Finnish newspapers up to 1910 and most of the other periodicals up to 1944
- Millions of pages freely available, hugely popular service
- Crowdsourcing solutions being developed for the correction of OCR'd text and the annotation of the content



THE NATIONAL LIBRARY OF FINLAND

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



1640
THE NATIONAL
LIBRARY
OF FINLAND