



# 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

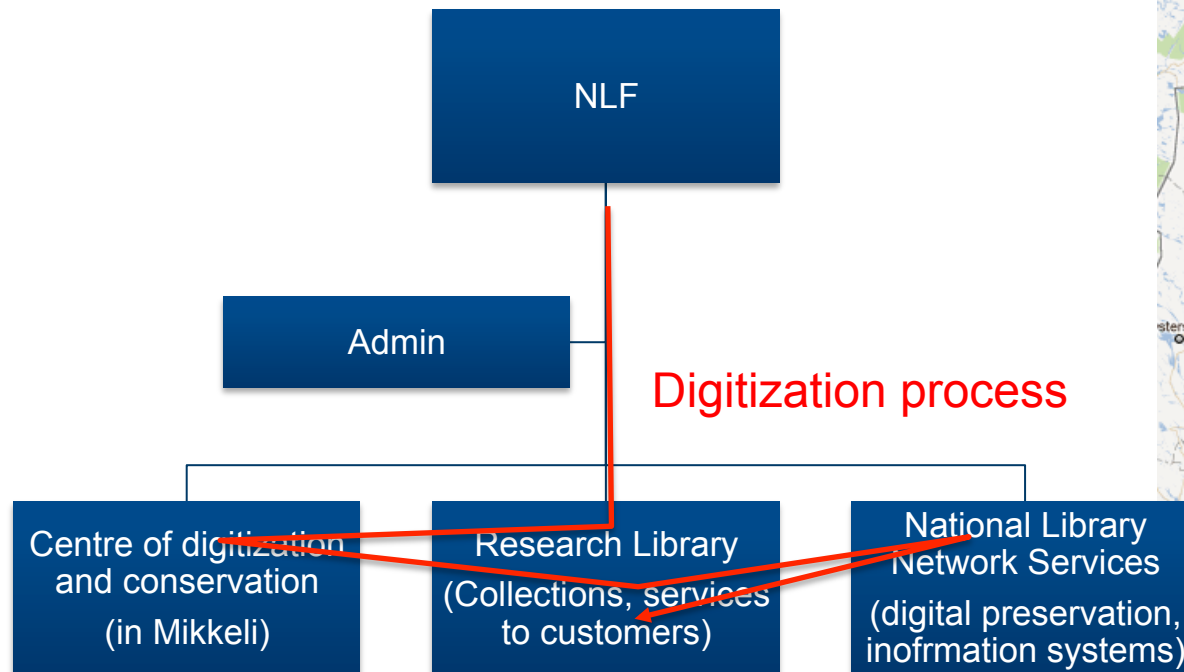
- About the National Library of Finland
- Digitization at the NLF
- Logistics
- Post-processing
- Metadata
- Preservation
- National Digital Library Initiative



# 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



NATIONAL LIBRARY NETWORK SERVICES

## 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](http://www.nationallibrary.fi/infoe/organization/nationallibrarystrategy_20062015_summary.html)  
Preservation Policy & Digitization Policy:  
<http://www.nationallibrary.fi/libraries/dimiko.html>  
Collection Policy:  
<http://www.nationallibrary.fi/services/kokoelmat.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

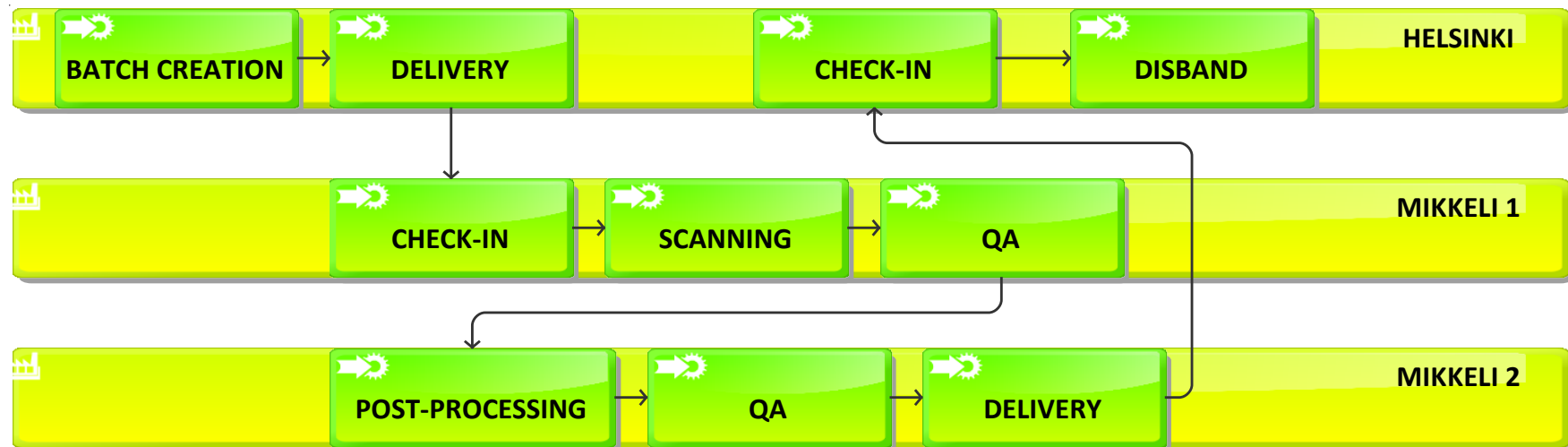
## New Process

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

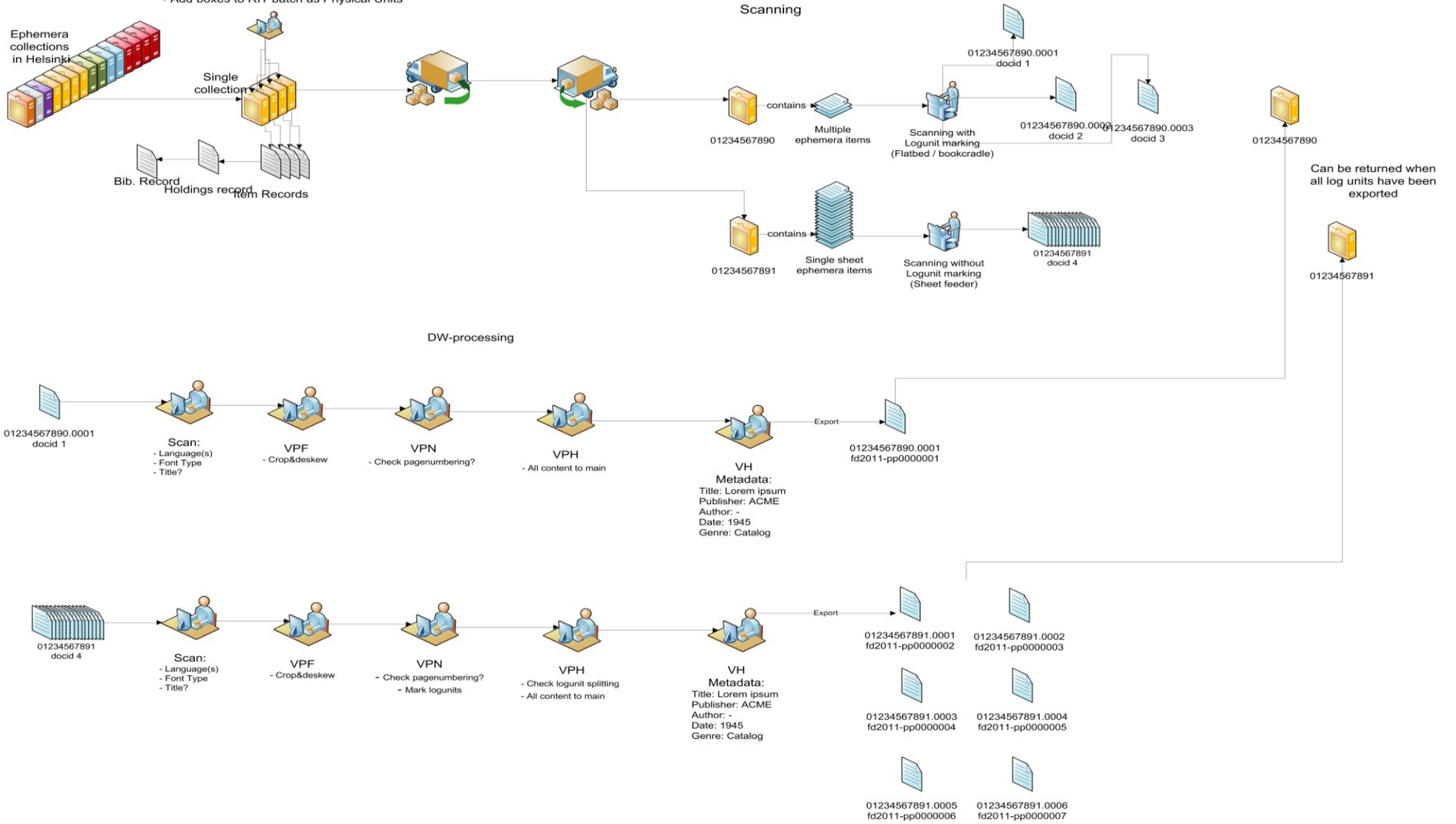


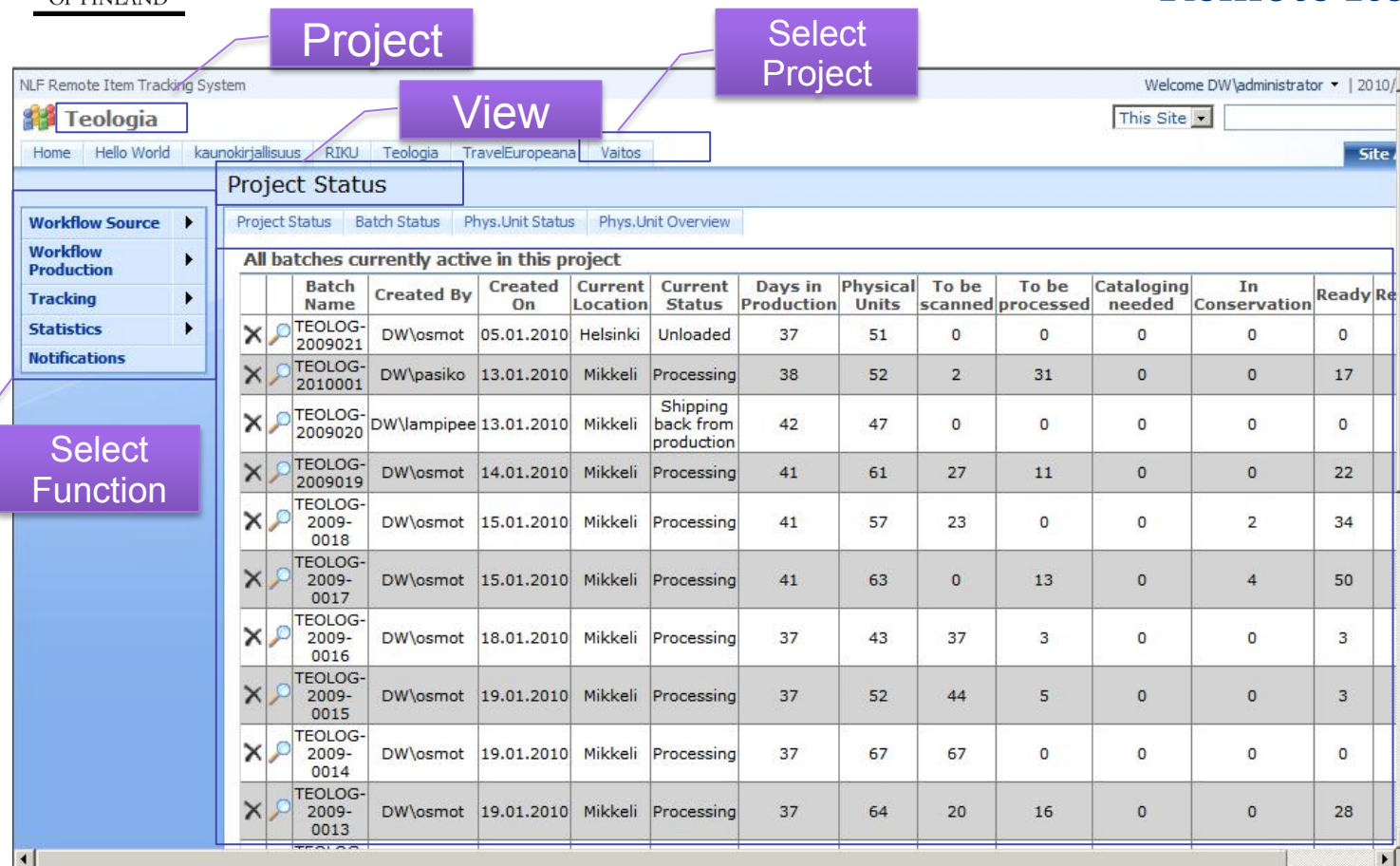
## Need to track the items on every stage



# Ephemera processing

- Helsinki:
- create Collection level bibliographical record
  - Create holding record
  - Create item records
  - Add barcodes to item records
  - Add boxes to RIT batch as Physical Units





The screenshot shows the 'Project Status' view of the NLF Remote Item Tracking System. The interface includes a navigation menu on the left, a main data table, and a top navigation bar. Callouts point to specific elements: 'Project' points to the 'Project Status' tab, 'View' points to the 'Project Status' view, 'Select Project' points to the 'Batch Name' column, and 'Select Function' points to the 'Workflow Source' menu.

**Project Status**

Project Status | Batch Status | Phys.Unit Status | Phys.Unit Overview

All batches currently active in this project

	Batch Name	Created By	Created On	Current Location	Current Status	Days in Production	Physical Units	To be scanned	To be processed	Cataloging needed	In Conservation	Ready Re
X	TEOLOG-2009021	DW\osmot	05.01.2010	Helsinki	Unloaded	37	51	0	0	0	0	0
X	TEOLOG-2010001	DW\pasiko	13.01.2010	Mikkeli	Processing	38	52	2	31	0	0	17
X	TEOLOG-2009020	DW\lampipee	13.01.2010	Mikkeli	Shipping back from production	42	47	0	0	0	0	0
X	TEOLOG-2009019	DW\osmot	14.01.2010	Mikkeli	Processing	41	61	27	11	0	0	22
X	TEOLOG-2009-0018	DW\osmot	15.01.2010	Mikkeli	Processing	41	57	23	0	0	2	34
X	TEOLOG-2009-0017	DW\osmot	15.01.2010	Mikkeli	Processing	41	63	0	13	0	4	50
X	TEOLOG-2009-0016	DW\osmot	18.01.2010	Mikkeli	Processing	37	43	37	3	0	0	3
X	TEOLOG-2009-0015	DW\osmot	19.01.2010	Mikkeli	Processing	37	52	44	5	0	0	3
X	TEOLOG-2009-0014	DW\osmot	19.01.2010	Mikkeli	Processing	37	67	67	0	0	0	0
X	TEOLOG-2009-0013	DW\osmot	19.01.2010	Mikkeli	Processing	37	64	20	16	0	0	28



The screenshot displays the Ephemera web application interface. The main window is titled "Ephemera" and shows a navigation menu on the left with options like "Workflow Source", "Workflow Production", "Tracking", "Statistics", "Notifications", "Lists", and "Documents". The main content area is divided into several sections:

- Edit Batch:** This section includes a navigation bar with "Create new Batch", "Edit Batch", "Prepare Batch", "Batch Shipment", "Receive returned Batch", and "Verify Batch". Below this, there is a text input field for "Enter/Scan the Batch Barcode" and a dropdown menu showing "EPHEME-2011-0004". There are buttons for "Add Phys. Units" and "Finalize Batch". Below the input fields, a table header is visible with columns: "ID", "Added", "Quality Level", "Comment Codes", "Comment", "Scanning Mode", "Color Mode", "Resolution", and "Log. Unit Count". The table content shows "No rows returned...".
- Create new Batch:** This section has a navigation bar with "Create new Batch", "Edit Batch", and "Prepare Batch".
- Receive Batch:** This section has a navigation bar with "Receive Batch", "Check In", "Vault In", "Vault Out", "ConservationIn", "ConservationOut", "Check Out", and "Return Batch". It features a "Batch Name:" input field with "EPHEME-2011-0004" and a "Receive Batch" button.

A modal window titled "Add Physical Units to Batch EPHEME-2011-0004" is overlaid on the "Receive Batch" section. It contains a "Batch" dropdown set to "EPHEME-2011-0004" and several configuration options:

- Get Bib. Data:** A button.
- Try all sources:** A checkbox.
- Scanning Mode:** A dropdown menu.
- Color Mode:** A dropdown menu set to "0".
- Resolution:** A dropdown menu set to "300".
- Origin:** A dropdown menu set to "NLF legal deposit copy".

At the bottom of the modal window, there are buttons for "Previous & Next" and "Close".

NATIONAL LIBRARY NETWORK SERVICES

The screenshot displays the Ephemera web application interface, which is used for managing digital collections. It features a navigation menu on the left and a main content area with several panels.

**Project Status Panel:** This panel allows users to view and filter project information. It includes a search bar for "Project Name" and "Remarks", and a list of filters such as "Where 'Current Sta...", "Where 'Current Loc...", "Where 'Created By...", and "Show hidden(delet...". An "Apply Filter" button is located below the filters.

**Current Statistic Panel:** This panel provides a summary of the current project's statistics. It includes a "Workflow Source" dropdown and tabs for "Current Statistic", "Recent Statistic", and "Scanning Statistic".

**Recent Statistic Panel:** This panel allows users to view and filter recent project statistics. It includes a "Timeframe" section with options for "last week", "last month", "total", "individual timeframe", "day", "week", and "month". It also includes a "Filter" section with a "Filter by batch..." dropdown and a "Refresh" button.

**Statistics Table:** The following table displays the recent statistics data:

Date / Timeframe	Items created	Items shipped	Items scanned	Items processed	Items returned
Wed 31.08.2011				1	
Tue 30.08.2011			1	1	
Mon 29.08.2011					
Sun 28.08.2011					
Sat 27.08.2011					
Fri 26.08.2011			1		
Thu 25.08.2011			1		
Wed 24.08.2011			1		
Tue 23.08.2011				1	
Mon 22.08.2011					
<b>Sum:</b>	<b>0</b>	<b>0</b>	<b>4</b>	<b>3</b>	<b>0</b>

The interface also shows a list of batches in the Project Status panel, including EPHEME-2011-0003, EPHEME-2011-0002, EPHEME-2011-0001, and EPHEME-2011-0004.

# 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



**SCORE: 12345**      **SAVED: 4/10**

LOG IN TO PLAY!

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



**LEVEL: 2**      **SCORE: 12500**

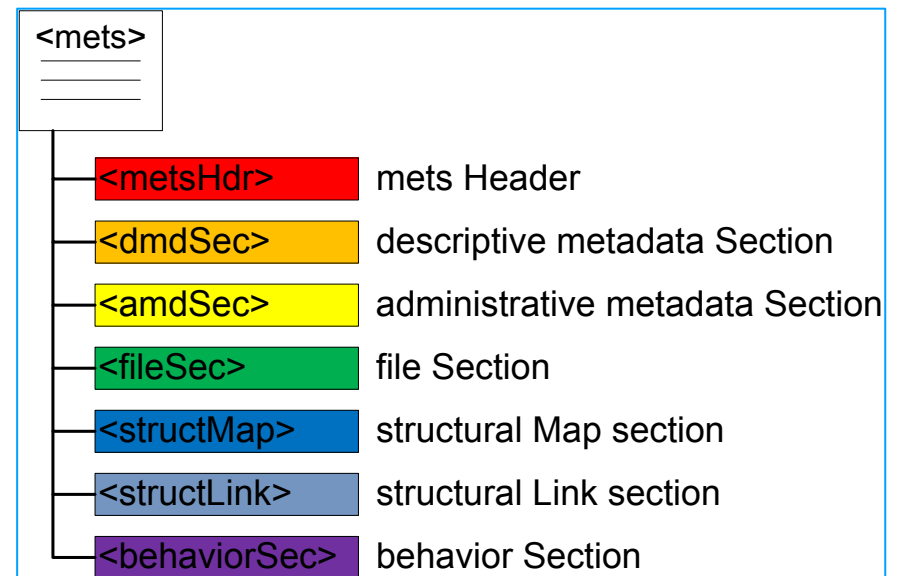
LOG IN TO PLAY!

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/>
- 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

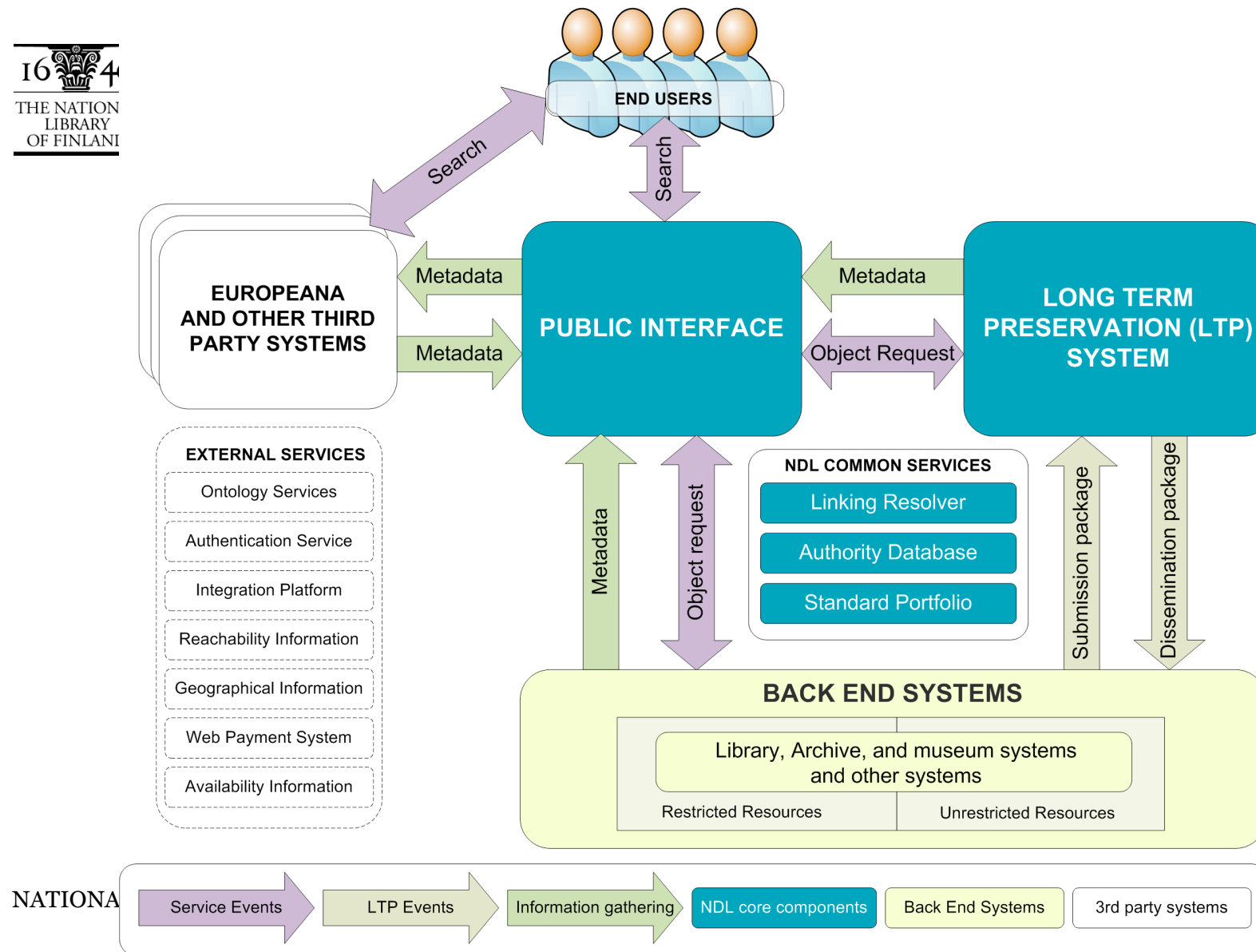
- METS with MARC, MODS, PREMIS and MIX
- For each page:
  - master image in JPEG2000 (lossless)
  - access image in JPEG
  - thumbnail
  - OCRd text in ALTO XML
- Single PDF with hidden text layer
- Whole SIP compressed to a zip file

## 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](http://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>



THE NATIONAL  
LIBRARY  
OF FINLAND