FIN-CLARIAH

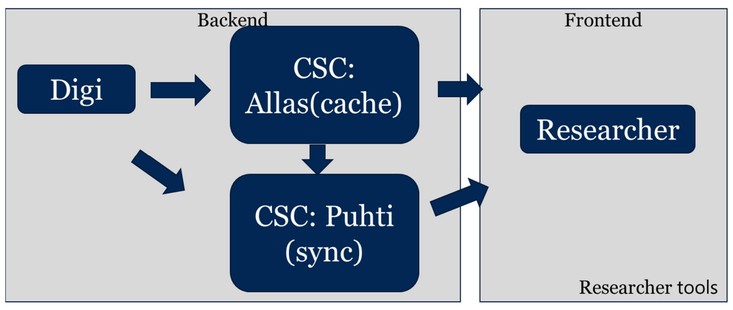
WP 3.1 INCREASINGLY AUTOMATED INGESTION OF MATERIAL

Tuula Pääkkönen
National Library of Finland, Research Library
Johanna Lilja
National Library of Finland, Research Library
Erno Liukkonen
National Library of Finland, Research Library
Martin Matthiesen
CSC – IT Center for Science
Anni Järvenpää
CSC – IT Center for Science

DELIVERABLES

D:3.1.1 Initial NLF data, D:3.1.2 Ingestion Framework , D:3.1.3 Versioning support, D:3.1.4 Incremental update process

We aim to deliver the copyright-free NLF material to the CSC Allas, for researcher use. This is either done via data dumps and/or by utilizing the digi.nationallibrary.fi OAI-PMH interface [3] with some customization.

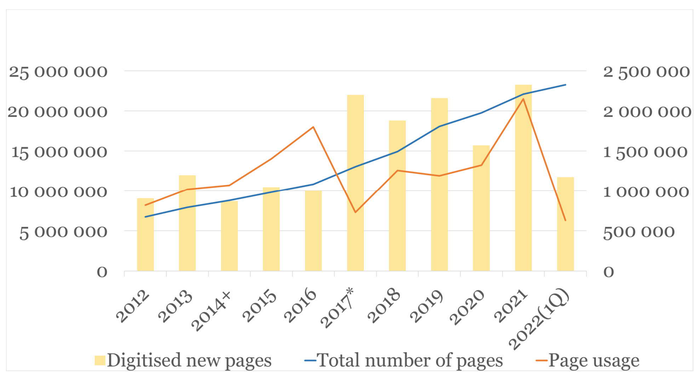


The aim is to be able to answer questions like:

- "I want Aamulehti from the year 1881".
- "I'm searching how many times person called 'Nikolai Bobrikoff' is visible in the materials"

ABOUT DIGITISATION

The digitising in the National Library of Finland (NLF) is based on the digitisation programme 2021-2024 [1]. Currently annually over 2 million pages is being digitised totalling a bit over 23 million pages.



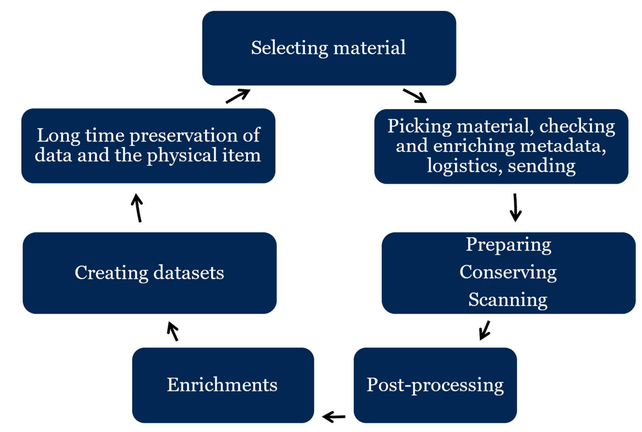
Digitisation statistics 2012-2022

The amount of copyright free data for Dariah-FI is around 11 terabytes (Tb) for access images and ALTO XMLs – for full export zip files it would be 65 Tb.

	ALTO	JPG	Zip files
Newspapers	1,466	7,567	56,190
Journals	0,101	0,761	3,928
Books	0,048	0,699	4,491
Total (Tb)	1,6	9,0	64,6

THE DIGITAL CHAIN

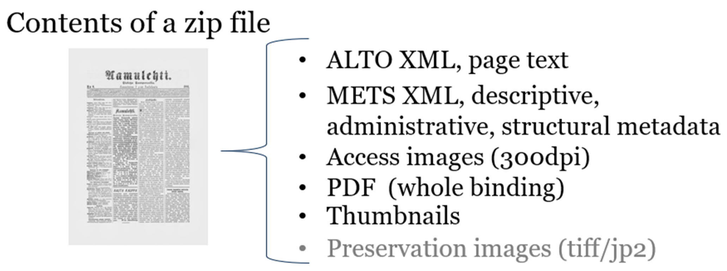
The digitisation process can be presented as a digital chain, with these major phases within the process:



The digital chain of digitisation

CURRENT STATUS

The digi.nationallibary.fi has digitised materials in ALTO XML and METS XML formats with access and preservation images within a zip file. The metadata of these bindings can be obtained via the OAI-PMH interface [2] or via the background search interface. Via interfaces you can download copyright-free materials.



Contents of a zip file of a binding

The NLF delivers materials to the FIN-CLARIN Language Bank annually. The Digi has the latest binding ids and is the main location of digitised materials.

OPEN QUESTIONS

Question 1. Can the material already in the Language Bank be utilized in the Dariah-Fi project?

Question 2. Is using the OAI-PMH interface such a solution, which would also work for other cultural heritage organisations or material providers to CSC Allas?

Question 3. How to handle the versioning of the materials? Who does the versioning? Is it enough to download the newest materials from the NLF's system? Are the old versions stored? How the versions are described, so that the researcher would know which version to use?

Question 4. How could restricted materials be handled in the future? (For example, would it be possible to agree with copyright holders to open an IP address from which it is allowed to download all materials?

PARTICIPATING ORGANISATIONS

The National Library of Finland is a key cultural heritage organisation, an infrastructure of knowledge and scholarship, and the developer of a digital operating culture and open science. Our services enable equal access to knowledge, promote the availability of the Finnish cultural heritage and create the preconditions for high-quality research. [4]

CSC, IT Center for Science, is a non-profit state enterprise with special tasks. As part of the national research system, it develops, integrates and provides high-quality information technology services and ensure that Finland remains at the forefront of development.

[1] https://bit.ly/nlfohjelma
[2] https://bit.ly/nlfoaipmh
[3] https://bit.ly/nlfinterface
[4] https://bit.ly/nlfstra