NATIONAL LIBRARY OF FINLAND

Fin-Clariah WP 3.1 Increasingly automated ingestion of material

Johanna Lilja, Tuula Pääkkönen, Martin Matthiesen

DARIAH.FI workshop November 9th 2022

Working group

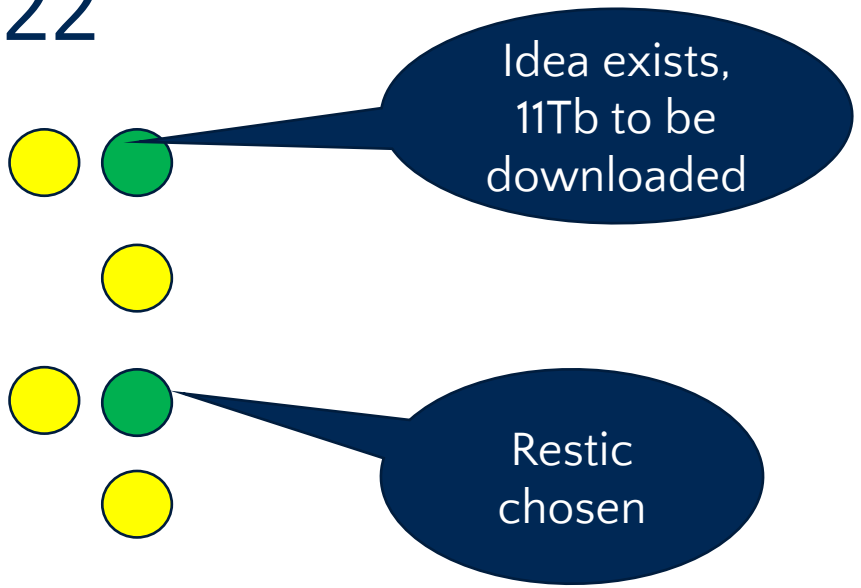
- Johanna Lilja (NLF) – leader
- Tuula Pääkkönen (NLF) – project manager
- Erno Liukkonen (NLF)
- Liisa Näpärä (NLF)
- Martin Matthiesen (CSC)
- Anni Järvenpää (CSC)
- Mikko Ojanen (UH)
- Eetu Mäkeli (UH)
- Risto Turunen (UH–FIN–CLARIAH)

WP 3.1 aiming at

- Development of a **data ingestion pipeline** where data providers such as cultural heritage institutions and companies can easily deposit their data for research use will be piloted.
- Initially, data will be accepted as data dumps, but work will be done to **increase automation in the transfer, versioning and incremental updates of the data.**
- **Starting with the out of-copyright digitized collections of the National Library of Finland** (digi.kansalliskirjasto.fi) containing over 24 million pages of newspapers, journals and books available as pictures and data – 50 terabytes
- The **solution will be scaled for other data providers** after the two-year piloting phase.

Overview of deliverables 10/2022

- D3.1.1. Initial NLF Data
- D3.1.2 Ingestion framework
- D3.1.3 Versioning support
- D3.1.4 Incremental update process



- Tool/way for end-user utilization of data

How researchers can easily use the data?

Done so far (10/2022)

- Technology defined (Apache airflow for workflow management)
  - Script created for downloading mets, and then alto files via Airflow
- CSC Project created
  - 20 Tb space for experiment (of 50 requested in Puhti)
  - 10 Tb Allas space, of requested 100Tb
- Restic (allas-backup) used for versioning
  - Potentially solution for in-copyright material (further project)

Initial NLF data

- Contains of out-of-copyright digitized publications
- Newspapers, journals, books
- Page text in ALTO XML, Access images (JPG)

	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**

To be discussed: How to administrate various versions of material in the service?

New Versions born in the NLF, e.g.

- Amendements of missing items
- Other amendments, e.g. annually new material is freed from copyright
- ReOCRred material

New versions born in research

- Annotations
- Enrichments
- Other, what?

Question of preservation images

- Preservation images are the original images taken in the scanner
  - Either tiff or jp2
  - Uncropped, unmodified images
  - Large size: if access image is 3Mb then preservation image is 30Mb
- Access image is 200 DPI (in newspapers, occasinally 300dpi)
- What are the researcher use cases for using preservation images?
  - The space preservation images require has cost impact

Suggestion

- Until the end of Q5 small pilot of research use (versions, images)
- Planning the next project 2024-2025
  - Access control for in-copyright data (CSC-KK)
  - NLF: licenses for in-copyright data unless the new copyright act made ingestion possible (close reading)
  - NLF: Webarchive?
- Q5 Scalability of ingestion (National Archives material?)
- Planning of further scalability with other CHOs