SUOMEN AKATEMIA
FINLANDS AKADEMI
ACADEMY OF FINLAND

Digitisation and Digital Library Presentation System – A Resource-Conscientious Approach

Tuula Pääkkönen, Kimmo Kettunen, Jukka Kervinen,

National Library of Finland

DHN18, 7.-9.3.2018

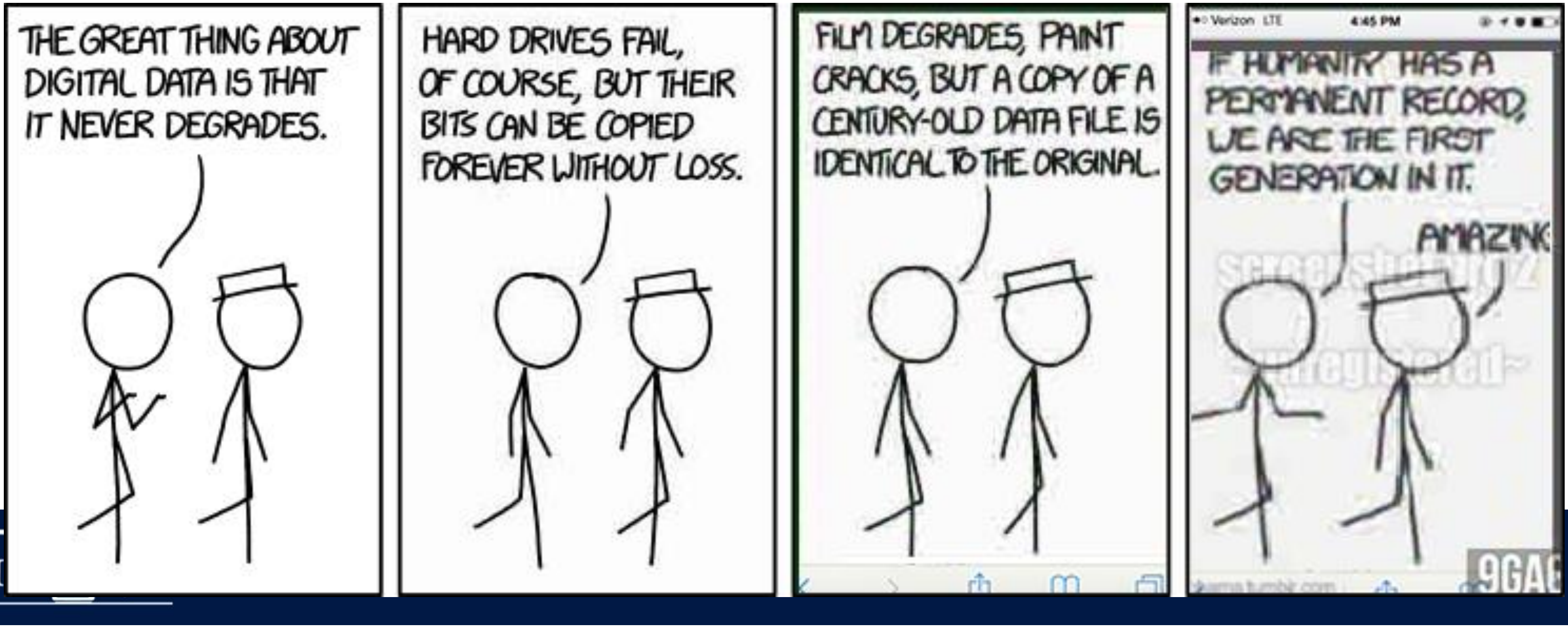
THE NATIONAL LIBRARY OF FINLAND

http://urn.fi/URN:NBN:fi-fe201803095826

Mission

- Most materials, most easily available for most users
  - Researchers, citizens, institutions

- Digitized?   Digitisation policy   &   Time, Quality, Cost


THE GREAT THING ABOUT DIGITAL DATA IS THAT IT NEVER DEGRADES.
HARD DRIVES FAIL, OF COURSE, BUT THEIR BITS CAN BE COPIED FOREVER WITHOUT LOSS.
FILM DEGRADES, PAINT CRACKS, BUT A COPY OF A CENTURY-OLD DATA FILE IS IDENTICAL TO THE ORIGINAL.
IF HUMANITY HAS A PERMANENT RECORD, WE ARE THE FIRST GENERATION IN IT.
AMAZIN

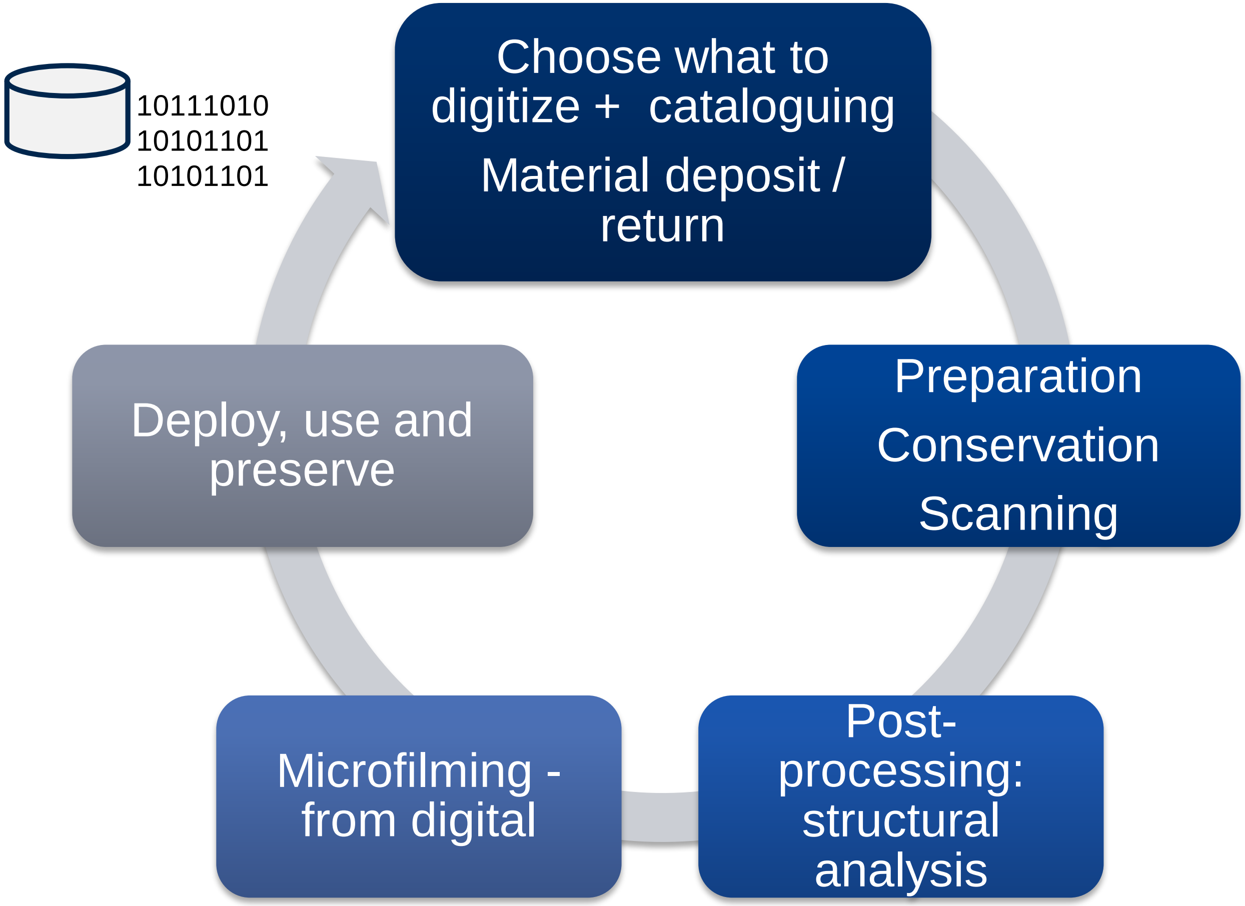
Source: https://xkcd.com/1683/

Glimpse to Nordic digitization

- Yesterday, we went past 13 million pages digitized (journals and newspapers combined)

Region	Pages (estimate, million pages)	Material Type	Reference
Denmark	33,3	Newspapers	http://www2.statsbiblioteket.dk/mediestream/avis
Finland	6,0	Newspapers	https://digi.kansalliskirjasto.fi/info?language=en
Iceland	5,4	Newspapers, periodicals	http://timarit.is/about_init.jsp?lang=en
Norway	22	Newspapers	R. Jøsevold, 'Digitization of copyright protected newspapers in Norway' (Liber 2015)
Sweden	17	Newspapers	http://feedback.tidningar.kb.se/viewtopic.php?id=89

Digital chain



10111010
10101101
10101101

Choose what to digitize + cataloguing

Material deposit / return

Preparation

Conservation

Scanning

Post-processing: structural analysis

Microfilming - from digital

Deploy, use and preserve

ALTO XML
METS XML

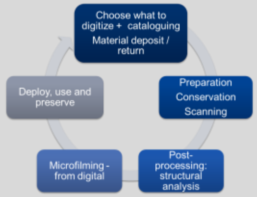
Digitisation & presentation system together

DIGI - NATIONAL LIBRARY'S DIGITAL COLLECTIONS
FRONTPAGE
NEWSPAPERS
JOURNALS
EPHEMERA
OTHER DIGITAL COLLECTIONS
DIGI.KANSALLISKIRJASTO.FI

Exporter

Importer

Digitisation

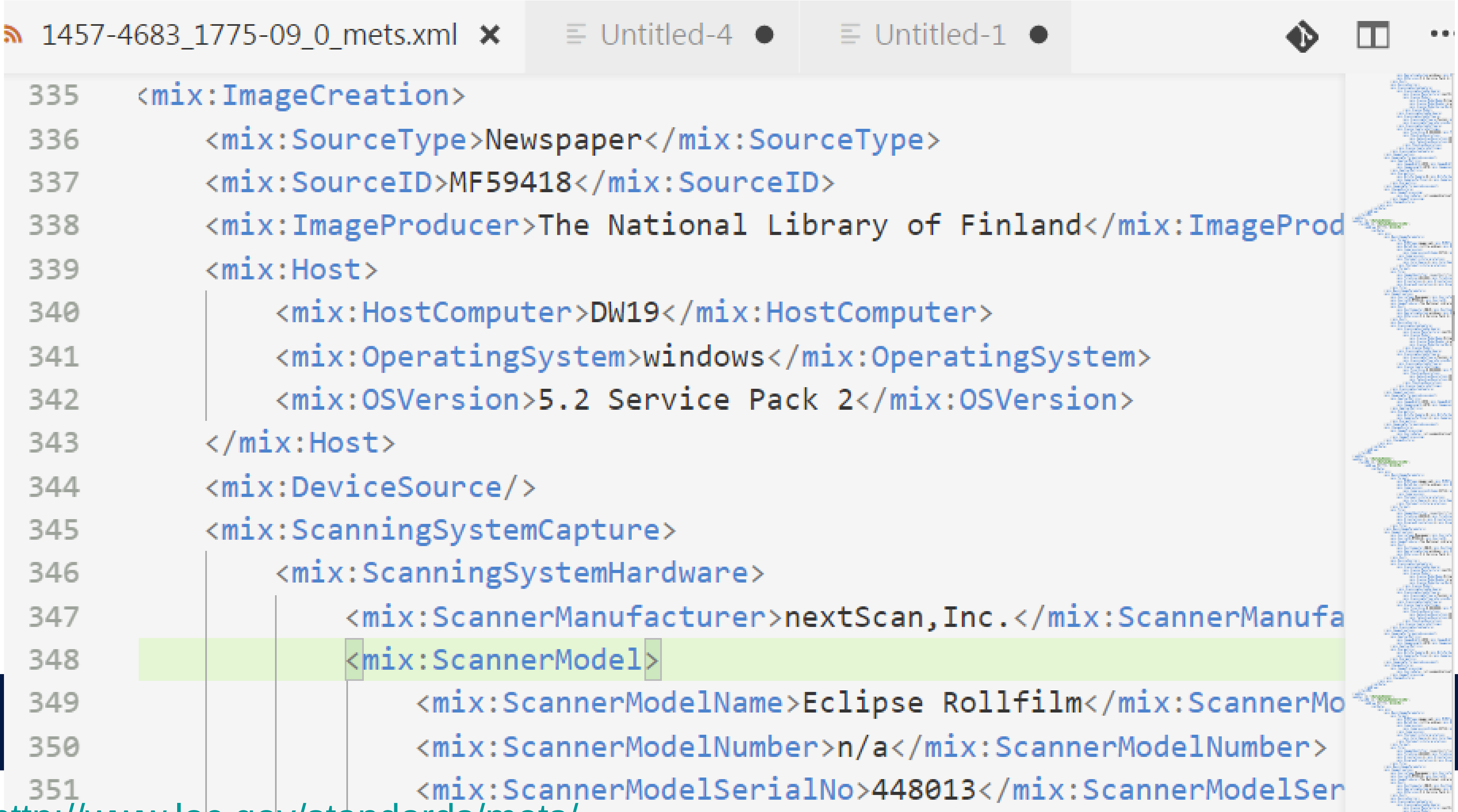

Choose what to digitize + cataloguing
Material deposit / return
Preparation
Conservation
Scanning
Post-processing: structural analysis
Microfilming - from digital
Deploy, use and preserve

Export package (zip)

- Preservation image (tiff)
- Access image (jpg, 300dpi)
- ALTO&METS XML
- Thumbnails
- Pdf

Audit trail of digitization

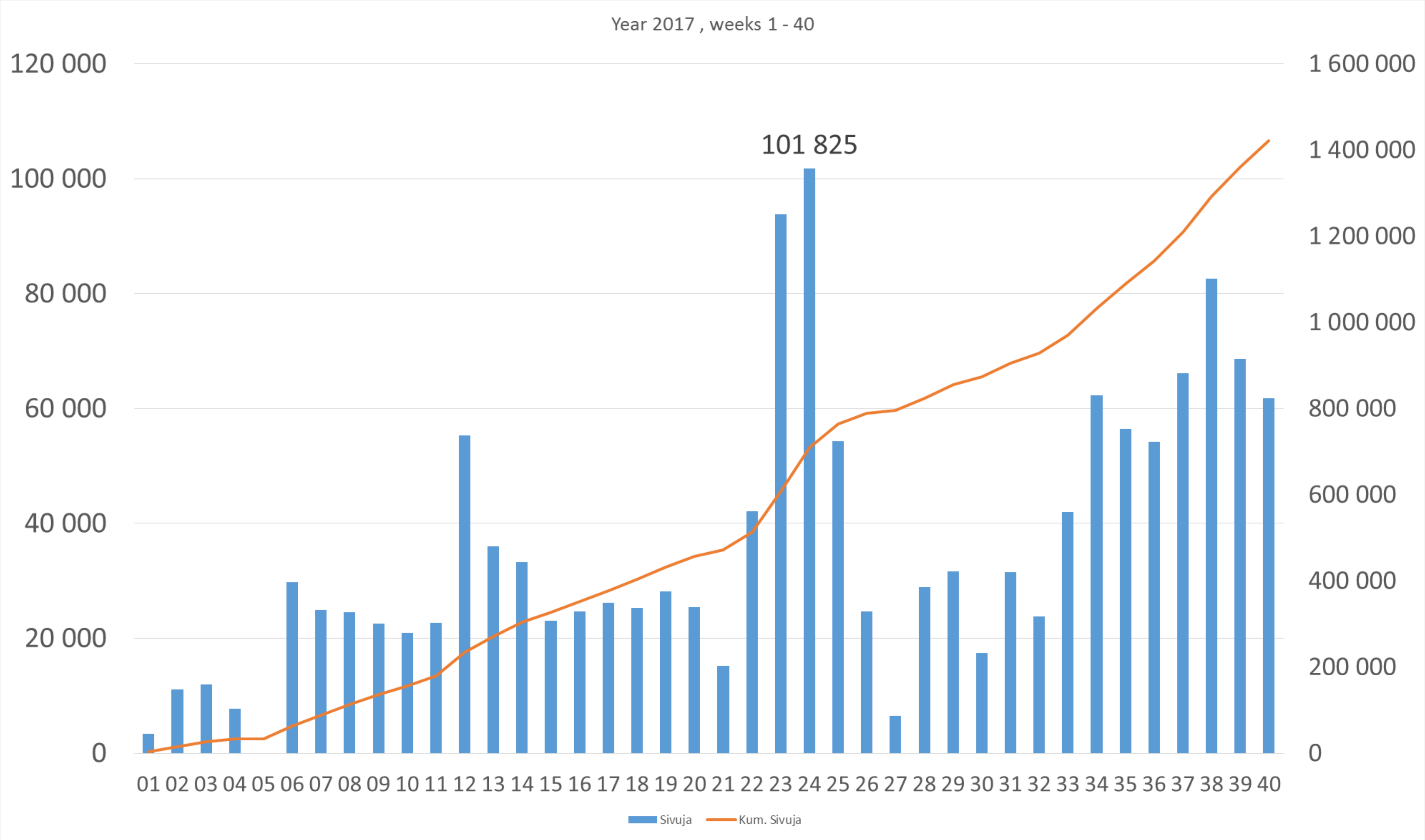
- Metadata of post-processing stored



```
<mix:ImageCreation>
    <mix:SourceType>Newspaper</mix:SourceType>
    <mix:SourceID>MF59418</mix:SourceID>
    <mix:ImageProducer>The National Library of Finland</mix:ImageProd
    <mix:Host>
        <mix:HostComputer>DW19</mix:HostComputer>
        <mix:OperatingSystem>windows</mix:OperatingSystem>
        <mix:OSVersion>5.2 Service Pack 2</mix:OSVersion>
    </mix:Host>
    <mix:DeviceSource/>
    <mix:ScanningSystemCapture>
        <mix:ScanningSystemHardware>
            <mix:ScannerManufacturer>nextScan,Inc.</mix:ScannerManufa
            <mix:ScannerModel>
                <mix:ScannerModelName>Eclipse Rollfilm</mix:ScannerMo
                <mix:ScannerModelNumber>n/a</mix:ScannerModelNumber>
                <mix:ScannerModelSerialNo>448013</mix:ScannerModelSer
```

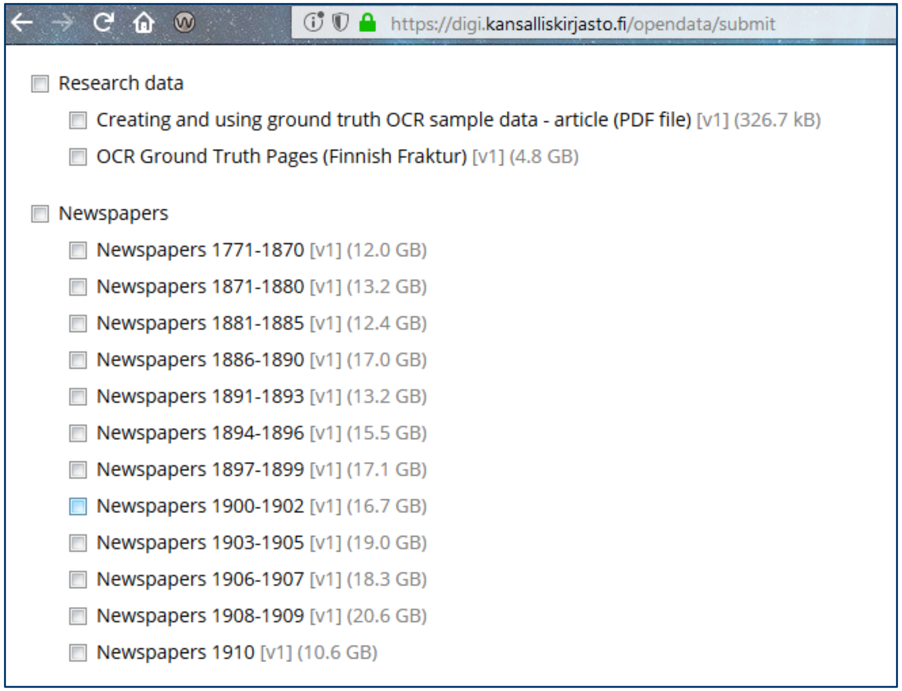
http://www.loc.gov/standards/mets/

Pages imported per week



Make room, make room!

- More and more material available as digital/digitized
- New formats and getting existing material digitized requires space somewhere
- Researcher data packages Enrichment work
  - Re-ocr, illustration content analysis, etc.



https://digi.kansalliskirjasto.fi/opendata/submit

□ Research data
- □ Creating and using ground truth OCR sample data - article (PDF file) [v1] (326.7 kB)
- □ OCR Ground Truth Pages (Finnish Fraktur) [v1] (4.8 GB)

□ Newspapers
- □ Newspapers 1771-1870 [v1] (12.0 GB)
- □ Newspapers 1871-1880 [v1] (13.2 GB)
- □ Newspapers 1881-1885 [v1] (12.4 GB)
- □ Newspapers 1886-1890 [v1] (17.0 GB)
- □ Newspapers 1891-1893 [v1] (13.2 GB)
- □ Newspapers 1894-1896 [v1] (15.5 GB)
- □ Newspapers 1897-1899 [v1] (17.1 GB)
- □ Newspapers 1900-1902 [v1] (16.7 GB)
- □ Newspapers 1903-1905 [v1] (19.0 GB)
- □ Newspapers 1906-1907 [v1] (18.3 GB)
- □ Newspapers 1908-1909 [v1] (20.6 GB)
- □ Newspapers 1910 [v1] (10.6 GB)

Final words

- More digitized pages available => more users

- Digitisation is a prioritization issue
  - National
  - Organisational
  - Between material types and projects

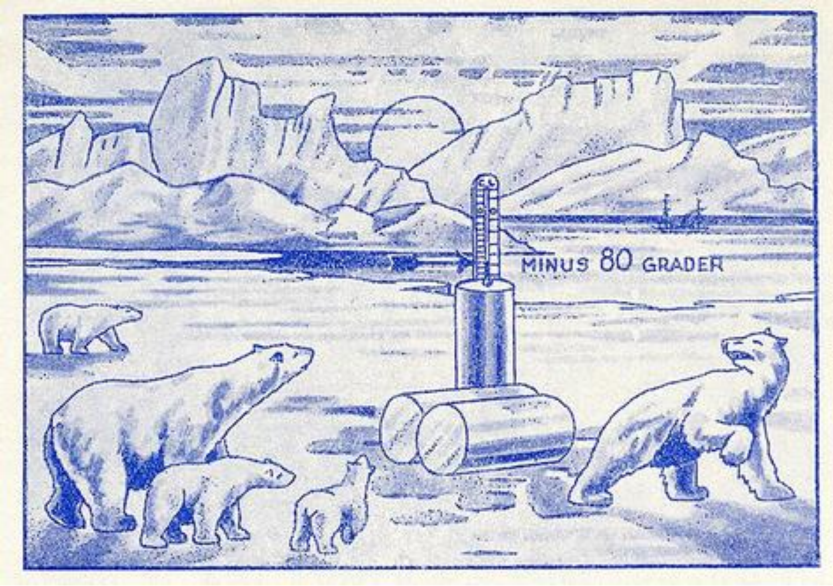
- Mass digitization & small-scale digitisation to live together?

Suomen
Julkiſia Sano

A. TILLANDER
Jewellers . Goldsmiths
Silversmiths

Almanacka
Eli
Ajan-Tieto,
Wuonna 1798.

Thank you!



Tuula.paakkonen@helsinki.fi

Image from: Kuiva-jää Jäähdytystekniikan voittosaavutus, , p. 2
https://digi.kansalliskirjasto.fi/pienpainate/binding/342428?page=2
National Library's Digital Collections

References

- K. Kettunen, E. Mäkelä, T. Ruokolainen, J. Kuokkala, and L. Löfberg, ‘Old Content and Modern Tools – Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910’, *Digit. Humanit. Q.*, vol. 11, no. 3, 2017.
- T. Pääkkönen, J. Kervinen, A. Nivala, K. Kettunen, and E. Mäkelä, ‘Exporting Finnish Digitized Historical Newspaper Contents for Offline Use’, *D-Lib Mag.*, vol. 22, no. 7/8, Jul. 2016.
- M. Koistinen, K. Kettunen, and T. Pääkkönen, ‘Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing’, in *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa*, Gothenburg, Sweden, 2017, pp. 277–283.

- https://wiki.helsinki.fi/display/Comhis/