

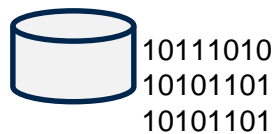
Digi.kansalliskirjasto.fi open data and interfaces & demo!

Tuula Pääkkönen,
Information Systems Specialist
HELDIG Forum 13.4.2018

On 2018: 1918-1929 open

- Note! Currently there is special agreement to have years 1918-1929 materials open for internet.
- Agreement done between Copyright organization Kopiosto and National Library **for this year!**

Digital chain



Choose what to digitize + cataloguing
Material deposit / return

Preparation
Conservation
Scanning

Deploy, use and preserve

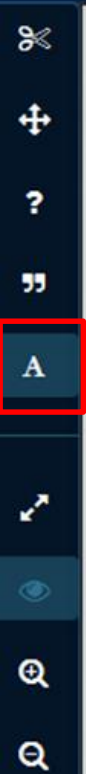
Microfilming -
from digital

Post-
processing:
structural
analysis

ALTO XML
METS XML



The Final Product



Otavan Joulukirjallisuus

Tässä näet kuvan, kun kissat pitivät kokouksen ja päättivät ajaa koirat pois koko Suomen maasta. Samassa kun he olivat tehneet päätöksen, niin —

saat „Joulukontista“ lukea ja nähdä miten
Se on hyvin naurettava juttu.

Tekstisisältö

Lataa sivun teksti: TXT **ALTO XML**

rivitetty

Tässä näet kuvan, kun kissat pitivät kokouksen ja päättivät ajaa koirat pois koko Suomen maasta. Samassa kun he olivat tehneet päätöksen, niin

niin, niin, sittenpäähän saat „ Joulukontista“ lukea ja nähdä miten hassusti kissoille kävi. Se on hyvin naurettava juttu.

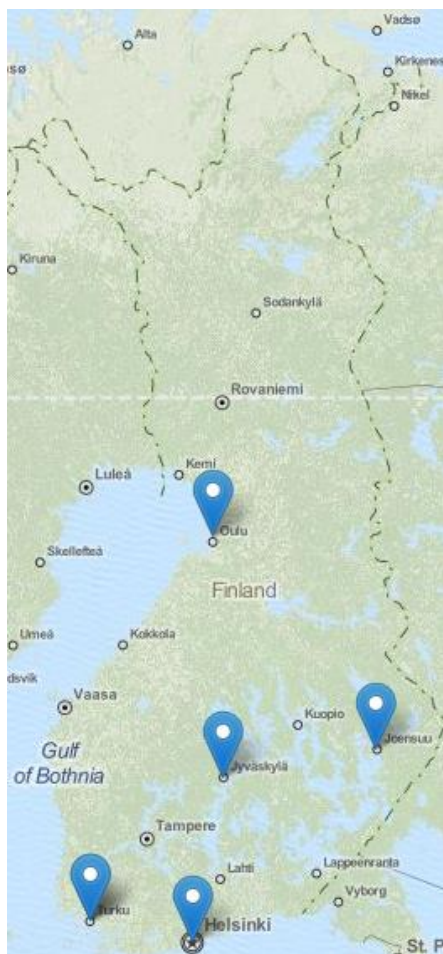
Joulukontissa on vakavia ja opettaviakin kertomuksia, neuvoja ja ajanvietettä.

Ei tietysti voi kertoa ja näyttää tässä kaikkea sitä hyvää, kaunista ja hupaista, jota „Joulukontissa“ on kukkuullaan, eikä tarvitsekaan, sillä tietysti sinä itsekkin saat jouluksi „Joulukontin“ sisareltasi, veljeltäsi, vanhemmiltasi! tai koulultasi.

Mutta se on tilattava heti, sillä voi käydä niin, ettei „Joulukontteja“ arvata varustaa tarpeeksi paljon ja loppuvat kesken. Silloin jää

Multiple places to access

- Legal deposit libraries for in-copyright material



- Various online services

All having bit different selection of material and tools of use.

Vs.



Data packages

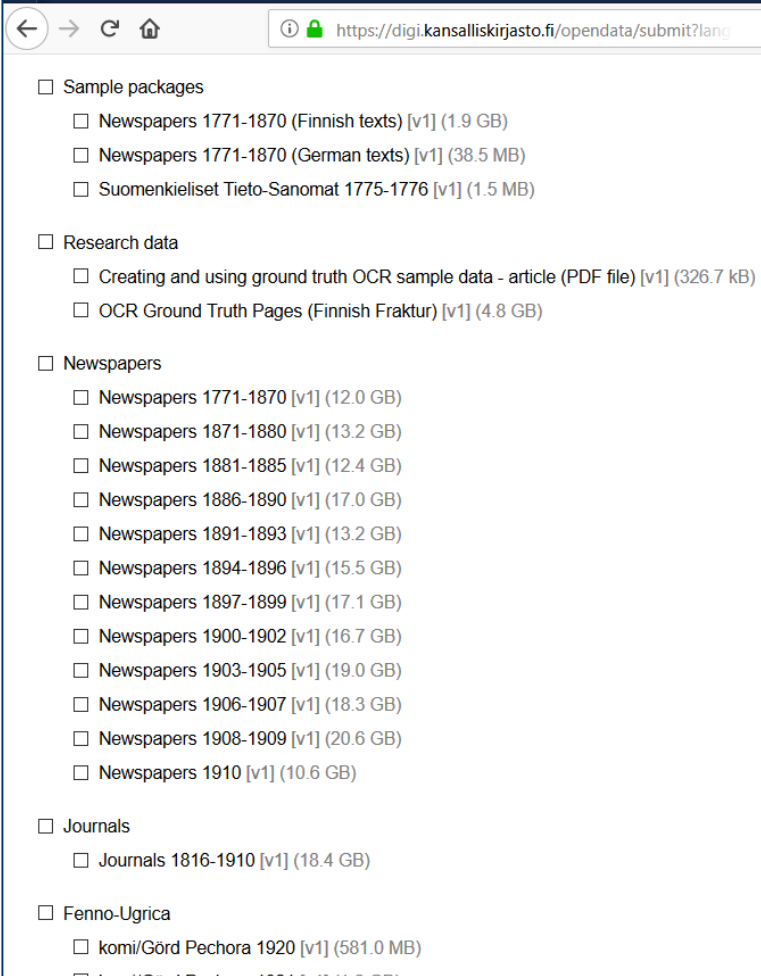
<https://digi.kansalliskirjasto.fi/opendata>

- You can download datapackages divided based on year ranges

- Material available until 1910

More info at

<https://wiki.helsinki.fi/display/Comhis/En+-+Digi.kansalliskirjasto.fi+Data>



The screenshot shows a web browser window with the URL <https://digi.kansalliskirjasto.fi/opendata/submit?lang>. The page displays a list of data packages for download, organized into several categories:

- Sample packages
 - Newspapers 1771-1870 (Finnish texts) [v1] (1.9 GB)
 - Newspapers 1771-1870 (German texts) [v1] (38.5 MB)
 - Suomenkieliset Tieto-Sanomat 1775-1776 [v1] (1.5 MB)
- 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)
- Journals
 - Journals 1816-1910 [v1] (18.4 GB)
- Fenno-Ugrica
 - komi/Görd Pechora 1920 [v1] (581.0 MB)

The contents of the data package

- Contains all the pages from newspapers and journals until **1910**.
- 1 XML file per page
- Custom XML containing 3 parts.
- Divided within package to years and languages.



Metadata

XML Alto (=text with layout info etc.)

CDATA (raw text)

Interfaces, too!

- **OAI-PMH** (traditional library interface, mainly used for harvesting, for example getting new records after specific date).

- **OpenURL** : handy way to refer to specific page e.g. by date

<http://digi.kansalliskirjasto.fi/openurl/query.html?genre=journal&date=1888-01-03&issn=0355-6913&spage=2>

- **JSON**: currently available for metadata of all titles

- More info at:

<https://wiki.helsinki.fi/display/Comhis/Interfaces+of+digi.kansalliskirjasto.fi>

JSON: Example Metadata of titles

- <https://digi.kansalliskirjasto.fi/api/newspaper/titles?language=fi>

```
JSON
Save Copy
0: {}
1:
  identification: "0355-6913"
  title: "Aamulehti"
  publishedTimes:
    0:
      from: 1881
      to: 2018
  digitizationCoverage: 0.41304347826086957
  type: "NEWSPAPER"
  languageCodes:
    0: "fin"
  digitizationTimes: [-]
  publishingPlaces:
    0:
      from: "1881-12-03"
      to: null
      countryCode: "fi"
  city: "Tampere"
  protectableTitleType: "ISSN"
  bindingCount: 15558
  pageCount: 127358
  latestBindingAdded: "2018-04-07"
  titleUrl: "https://digi.kansalliskirjasto.fi/api/newspaper/titles/0355-6913"
```

Also available as excel file


The screenshot shows the website interface for DIGI - NATIONAL LIBRARY'S DIGITAL COLLECTIONS. The navigation bar includes 'FRONTPAGE', 'NEWSPAPERS', 'JOURNALS', 'EPHEMERA', and 'OTHER DIGITAL COLLECTIONS'. The 'NEWSPAPERS' tab is active. Below the navigation bar, there are search filters: 'Name', 'Year', 'All langu' (language), and 'All publishing locat' (publishing location). There are also checkboxes for 'Only accessible material' and 'Only rec updated'. A large blue arrow points from the text 'Also available as excel file' to a download icon. The search results show 890 titles. Two titles are visible: 'ANK : kaupunkilehti Ankkuri' (2489-5288, 2017-2018, 2164 pages) and '★ Aamulehti' (0355-6913, 1881-1917, 1918-2018, 128574 pages).

Possibilities from data

Turvallinen | <https://voyant-tools.org/?corpus=7145af0cc4a59e4515dd04ec17992c28>


Voyant Tools

Cirrus Terms Links Reader TermsBerry Trends Document Terms

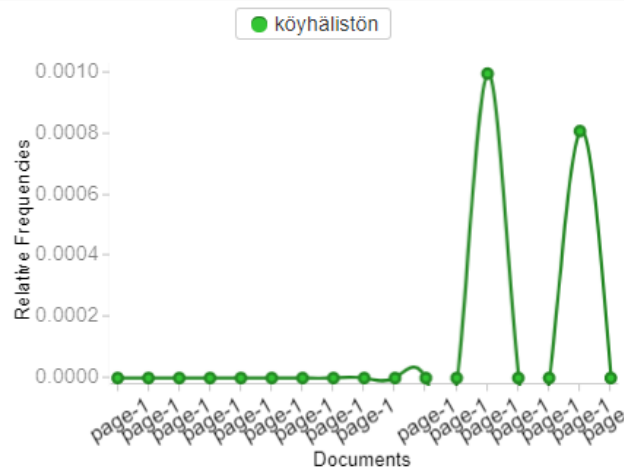


Scale Terms:

puhumattakaan tämän politikan lukuisista uhreista, jotka eivät yleensä tunne sitä yhtään. Katsomme senwuolfi enempi luin tarpeelliseksi ottaa siitä tällä het kellä pienen erittelyn. Maailman sofiademolraattisesia liikkeessä on jo pitemmän ajan ollut hnomattawisfa kaksi eri suun taa. Toinen suunta lähentelee radikaalista porwarillifuutta, sillä se pyrkii hankkimaan parannuksia köyhälistön oloihin fiweellistillä ja y> teensä sellaisilla keinoilla, mitkä lunkin aikaisissa olosuhteissa omat mahdollista. Tämän suunnan edustajia nimitetään usein sosialistisessa sanomalehdistössä rewisionisteiifi sen wnoksi, että he omat hyljänneet n. k. sosialistisen optimismin toteuttamisen mahdottomana



Relative Frequencies



Documents

Summary Documents Phrases Contexts Correlations Mandala

This corpus has 17 documents with 28,178 total words and 13,918 unique word forms. Created about 19 minutes ago.


Document Length:

- Longest: page-1 (3281); page-1 (3024); page-1 (2486); page-1 (2373); page-1 (2211)
- Shortest: (0); page-1 (58); page-1 (157); page-1 (1135); page-1 (1287)

Vocabulary Density:

- Highest: page-1 (0.914); page-1 (0.822); page-1 (0.793); page-1 (0.761); page-1 (0.707)

items:

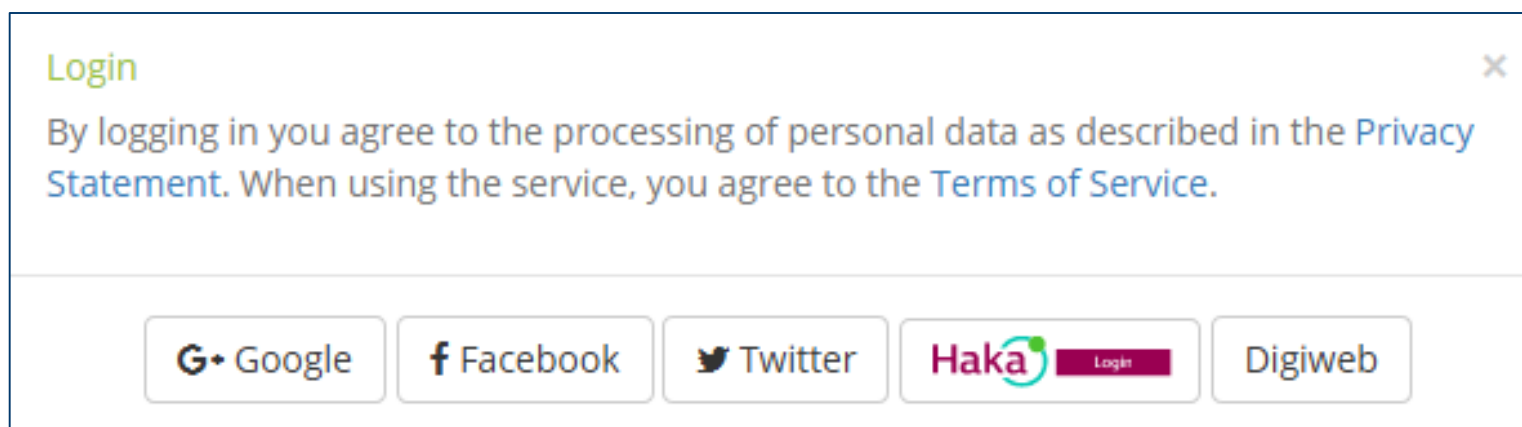


+ Add Clear labels

Voyant Tools, Stéfan Sinclair & Geoffrey Rockwell (© 2018) Privacy v. 2.4 (M5pr3)

Haka project 2017-2018+

- HAKA-identification at digi.kansalliskirjasto.fi for using in-copyright digital materials for **research and teaching** at universities
- Pilots with a few universities within Finland



Thank you!
-> Demo...



Suomen Kuvalehti, 03.10.1925, nro 40, s. 10
<https://digi.kansalliskirjasto.fi/aikakausi/binding/889602?page=10>