Table of Contents

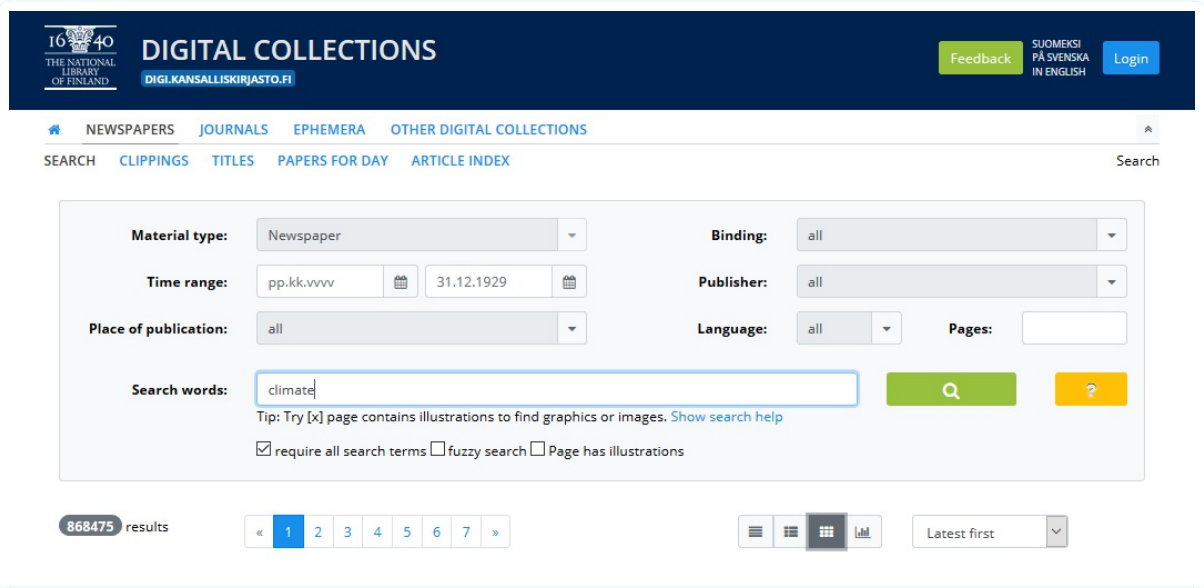
Introduction

digi.nationallibrary.fi (Digi) is the search- and presentation system for the digitized newspapers, journals, and ephemera (technical) of National Library of Finland.

This document tells you how to utilize Digi effectively in your research work.

Full Text Search

One of first ways to try out Digi is to utilize its search functions. You can do free text search or target the search to certain title, time period, publisher, publishing place or any combination of provided search fields. You can find the search under each material type.



Just enter the search words and all the hits are visible in the list below the search fields.

Downloadable Export Packages

The open data page of digi.nationallibrary.fi contains the downloadable export packages of all of the newspaper and journal pages until the year 1917. The pages have been divided by decades in order to keep package sizes at around 20 gigabytes.

You can download all of the packages and the download requires just filling in a short survey about your usage beforehand. The answers are used for statistical purposes.


DIGI - KANSALLISKIRJASTON DIGITOIDUT AINEISTOT
ETUSIVU
SANOMALEHDET
AIKAKAUSLEHDET
PIENPAINATTEET
MUUT AINEISTOT
Palaute
SUOMEKSI
PÅ SVENSKA
IN ENGLISH
Kirjaudu
Ur-Savo
KAUPPALEHTI
SUOMEN LIIKEMIESTEN ÄÄNENKANNATTAJA
1898
NOKIA
1936–37
PÄÄLLYS-
KENKIEN
KUVITETTU
HINNASTO
HANDLEDNING I BILMÅLNIN
DICCO
MIRANOL
MIRKO
DIGI.KANSALLISKIRJASTO.FI
10 923 634 SIVUA
HELSINGIN SANOMAT
SANOMALEHDET
Digitoitu yhteensä 4 425 731 sivua.
Vapaassa käytössä 2 953 798 sivua
(66%) (-1920).
Rajatussa käytössä 1 471 933 sivua
(34%) (1921-).
Vapaa
Rajattu
Tutustu Suomen historiaan ja menneeseen aikaan digitoitujen sanomalehtien kautta!
Kansalliskirjasto on digitoinut kaikki Suomessa vuosina 1771-1920 ilmestyneet sanomalehdet, ja ne ovat käytössä tämän palvelun kautta. Uudemmat digitoidut sanoma- ja aikakauslehdet ovat käytettävissä kaikissa vapaakappalekirjastoissa.
DIGI.KANSALLISKIRJASTO.FI/OPENDATA

Download content

If you want a specific title from Digi, then full export data packages might not be best option. Then it might be useful to download txt pages for just one title instead of everything.

Machine setup (Windows)

Install python

If you do not have python already installed, install python (either version 2.7 or 3.x depending your preference). Also Anaconda distribution can be useful as it includes many data science modules within its download package.

Choose the newspaper you want

Go to the titles view and take a note of the ISSN of the newspaper you want.

Setup environment

1) Open the command prompt to get to the command line:

Press R and type `cmd` .

2) Create an own directory to your PC.

```
mkdir C:\temp\datadownload

set MYWORKINGDIR=C:\temp\datadownload
```

3) Change to the directory where you downloaded the python file.

```
cd %MYWORKINGDIR%
```

4) Download the digi_downloadaltos.py and copy it to your working directory.

Note! The tool is not supported, use it at own risk.

Download pages

Run the provided helper script for downloading desired newspaper or journal.

```
python digi_downloadaltos.py -i 1458-851X
```

Expected output is:

```
python digi_downloadaltos.py -i 0018-2362
Digi.nationallibrary.fi - UNSUPPORTED downloader for ALTO XML or TXT from digi for given ISSN.
All years for issn 0018-2362: [1903, 1904, 1905, 1906, 1907, 1908, 1909, 1910, 1911, 1912, 1913, 1914, 1915, 1916, 1917, 1918, 19
19, 1920, 1921, 1922, 1923, 1924, 1925, 1926, 1927, 1928, 1929]
ISSN 0018-2362 - Year 1903. 0018-2362\1903\txt
Processing binding :499137
Proc.: http://digi-testi.kansalliskirjasto.fi/aikakausi/binding/499137/page-1.txt  to :  0018-2362\1903\txt
Proc.: http://digi-testi.kansalliskirjasto.fi/aikakausi/binding/499137/page-2.txt  to :  0018-2362\1903\txt
Proc.: http://digi-testi.kansalliskirjasto.fi/aikakausi/binding/499137/page-3.txt  to :  0018-2362\1903\txt
```

Supported command line parameters

You can see command line parameters via running the script by giving parameter -h

```
python digi_downloadaltos.py -h
```

```
Digi.nationallibrary.fi - UNSUPPORTED downloader for ALTO XML or TXT from digi for given ISSN.
usage: digi_downloadaltos.py [-h] -i ISSN [-f {alto,txt}]

Digi.nationallibrary.fi - UNSUPPORTED downloader for ALTO XML or TXT from digi
for given ISSN.

optional arguments:
  -h, --help            show this help message and exit
  -i ISSN, --issn ISSN  choose an issn and download altos for it
  -f {alto,txt}, --format {alto,txt}
                        choose data format ALTO (xml) or text (txt)
```

-f format

The default format to download is *text*. The text is as it has been originally obtained in the text recognization of the post-processing system of the digitization. It is good to realize that depending on the material, the OCR does contain errors.

If you choose 'alto', then you get the ALTO XML, which contains the layout information of the page and the words with their location on the page.

Analysing the data of digi.kansalliskirjasto.fi in a machine-readable format and creating enriched datasets

If you belong to the Haka-authorized user groups from specific universities of Finland and you need the data locally on your machine, follow these steps. This option is available for the duration of the Haka project.

Fill in the Haka-survey of Digi

1. Go to http://digi.kansalliskirjasto.fi
2. Click login from top right corner, and select 'Haka'
3. At first login (and at specific intervals) digi will show you a survey about how the materials will be used. Please fill the form, as that helps National Library of Finland in future negotiations in order to extend access to researchers.

Setup environment

If you have python installed, you are good to go, otherwise download and install python

Download the helper script digi_download_haka to specific directory.

Press Windows-key R and type `cmd` to start command prompt.

Download pages

Run the provided helper script:

```
python digi_download_haka.py -i 0356-0996
```

Expected output is:

```
# python digi_download_haka.py -i 0356-0996

Digi.nationallibrary.fi - UNSUPPORTED downloader via Haka authentication
Download ALTO XML or TXT from digi for given ISSN.
Approaching HAKA identification for helsinki, wait for bit...
Organisation identification page reached.
Organisation login phase: HAKA password for helsinki authorisation (asked at every run):
succesful.
Ready to go!
All years for Warkauden Lehti (issn 0356-0996): [1919, 1920, 1921, 1922, 1923, 1924, 1925, 1926, 1927, 1928, 1929]
ISSN 0356-0996 - Year 1929
Processing binding :1768795
https://digi.kansalliskirjasto.fi/sanomalehti/binding/1768795/page-1.txt -> ./0356-0996/1929/1768795_page-1.txt
https://digi.kansalliskirjasto.fi/sanomalehti/binding/1768795/page-2.txt -> ./0356-0996/1929/1768795_page-2.txt
https://digi.kansalliskirjasto.fi/sanomalehti/binding/1768795/page-3.txt -> ./0356-0996/1929/1768795_page-3.txt
https://digi.kansalliskirjasto.fi/sanomalehti/binding/1768795/page-4.txt -> ./0356-0996/1929/1768795_page-4.txt
Processing binding :1768785
https://digi.kansalliskirjasto.fi/sanomalehti/binding/1768785/page-1.txt -> ./0356-0996/1929/1768785_page-1.txt
https://digi.kansalliskirjasto.fi/sanomalehti/binding/1768785/page-2.txt -> ./0356-0996/1929/1768785_page-2.txt
https://digi.kansalliskirjasto.fi/sanomalehti/binding/1768785/page-3.txt -> ./0356-0996/1929/1768785_page-3.txt
https://digi.kansalliskirjasto.fi/sanomalehti/binding/1768785/page-4.txt -> ./0356-0996/1929/1768785_page-4.txt
Processing binding :1768776
https://digi.kansalliskirjasto.fi/sanomalehti/binding/1768776/page-1.txt -> ./0356-0996/1929/1768776_page-1.txt
https://digi.kansalliskirjasto.fi/sanomalehti/binding/1768776/page-2.txt -> ./0356-0996/1929/1768776_page-2.txt
https://digi.kansalliskirjasto.fi/sanomalehti/binding/1768776/page-3.txt -> ./0356-0996/1929/1768776_page-3.txt
https://digi.kansalliskirjasto.fi/sanomalehti/binding/1768776/page-4.txt -> ./0356-0996/1929/1768776_page-4.txt
12 files downloaded to 0356-0996 folder.
```

Supported command line parameters

```
Digi.nationallibrary.fi - UNSUPPORTED downloader via Haka authentication
Download ALTO XML or TXT from digi for given ISSN.
usage: digi_download_haka.py [-h] -i ISSN [-u {utu,uef,helsinki}]
                             [-f {alto,txt}]
```

-f format

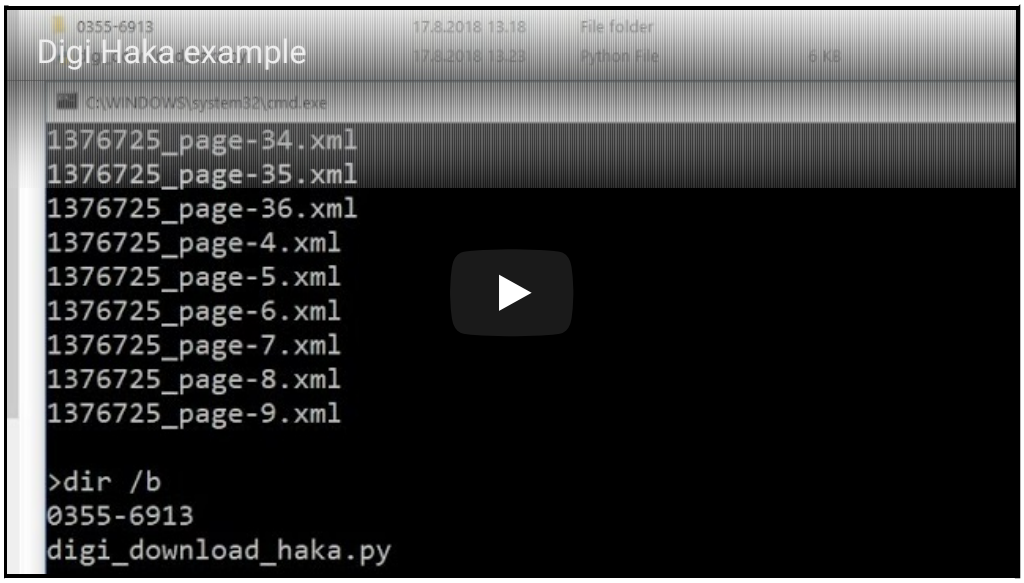
The default format to download is *text*. The text is as it has been originally received from the text recognization of the post-processing system of the digitization.

If you choose 'alto', then you get the ALTO XML, which contains the layout information of the page and the words with their location on the page.

-u university

With command line parameter -u there is an experimental login changer for different university. NB! requires customization based on the Haka login form of each university. Has been tested with University of Helsinki setup.

Demo



What next?

You can process the files you have downloaded in many ways.

- From XML files you can get back to text e.g. via tools provided by the Comhis-project
- It is also possible to utilize online tools, like Voyant Tools for text analysis, you can give data either via links, or upload your own data there directly.

Additional links

- COMHIS page, for seeing an application, which visualizes clusters of text in the Finnish newspapers.
- Data catalogue of the National Library of Finland
- The Language Bank's KORP-service enables search across many corpuses.
- Kopioston kopiointilupa yliopistoille (in Finnish)

Related international work

- Inspiration and additional resources: https://programminghistorian.org/
- Example of utilizing resources of National Library of Australia/ozglam-workbench
- NewsEye-project for creating new tools and methods for enhancing new views and perspectives from newspapers.

Good to know

- The provided helper scripts are experimental and utilize features of Digi, that can change over time. In future, the scripts might cease to work.
- If you need lots of materials, please contact us, so we can think whether a export package would be better option instead downloading every single page one-by-one.
- Some questions are also answered at Digi's FAQ https://digi.nationallibrary.fi/faq . Useful information exists also in terms and privacy page.