RaRa EESTI RAHVUS-RAAMATUKOGU

open digital libraries

Virtual Lab at the National Library of Estonia

Peeter Tinits, Oct 6, 2022

DH Estonia 2022

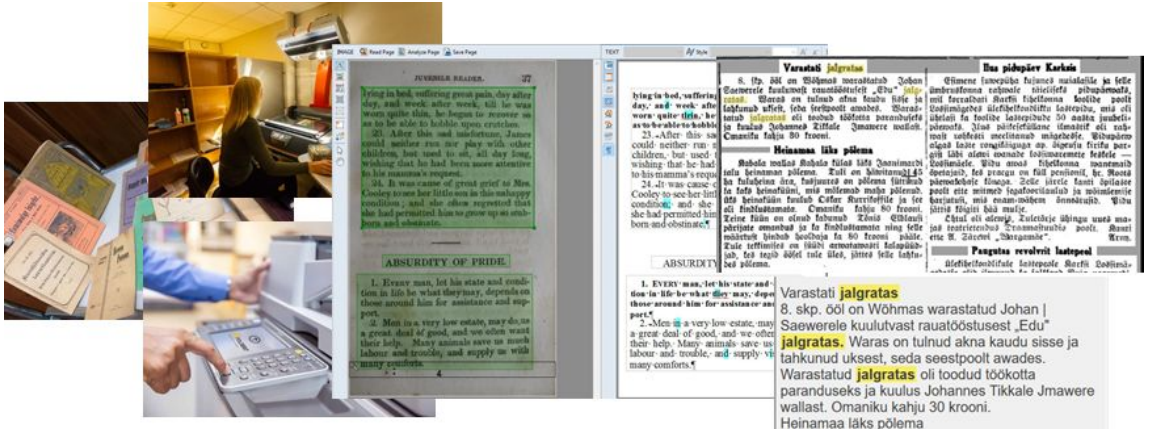
Libraries have large digital collections

Long time effort in digitization

- Focus on preservation
- Interfaces for reading

New frontiers in usage

- Collections as data
- Creative reuse



Example: Texts as data

RaRa newspapers
& periodicals

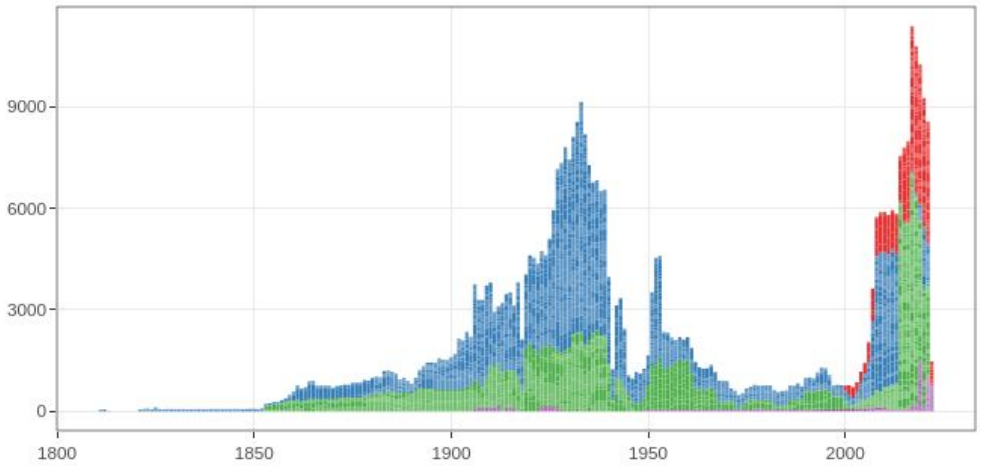
~9.1 M articles

~4.7 M pages

~465 K issues

~2601 publications

(+20-30% in last 2 years)



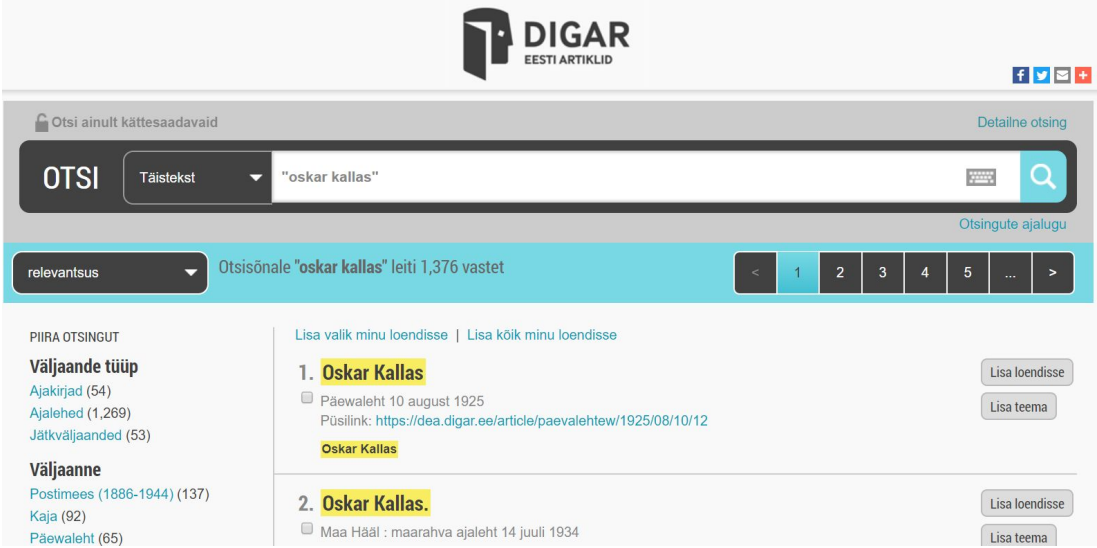
Researcher needs

Texts

Metadata

Tools

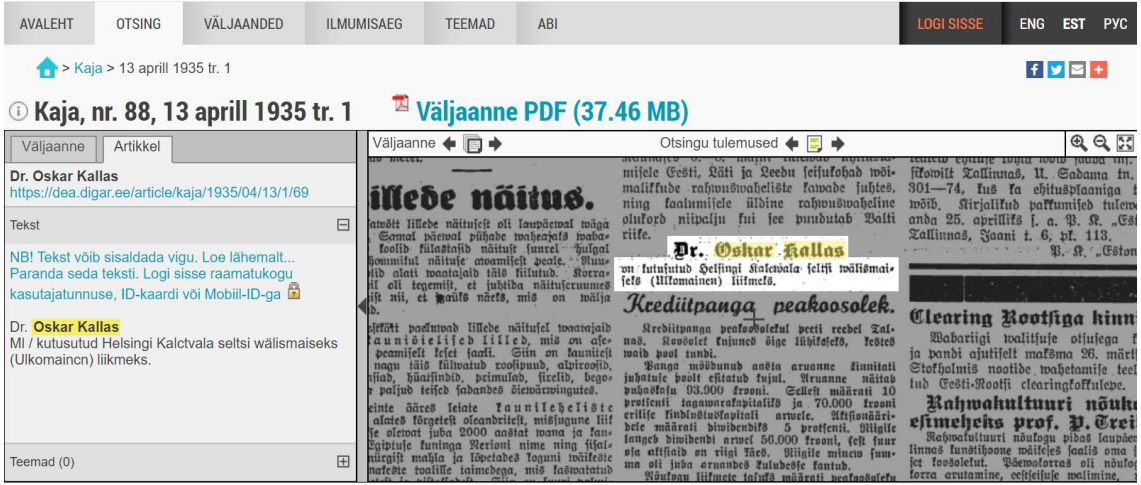
Approaches to collections: Search engine



RaRa EESTI RAHVUS-RAAMATUKOGU

https://dea.digar.ee

Approaches to collections: Search engine



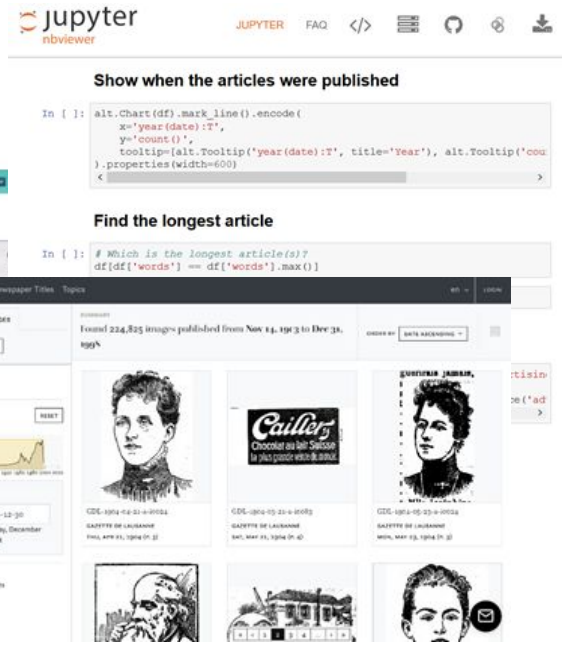
RaRa EESTI RAHVUS-RAAMATUKOGU

https://dea.digar.ee

More complex queries?

Interfaces & libraries

Delpher



Open Science movement

FAIR data

- Not just open, but findable+usable

Open Science Movement

- FAIR in science

Make analyses transparent, interoperable, reusable

(Heunis 2020)

Researcher needs

Texts

Metadata

Tools

GLAM Labs

GLAM Labs community
(galleries, libraries, archives, museums)

Creative uses of data.

Computational access to digital collections

Virtual Lab at RaRa

Working towards from data to use

- Access points
- Data enrichment
- Case studies

Learning from international examples:

Creative Europe: Open Digital Libraries (with Dutch Royal Library and Austrian National Library)

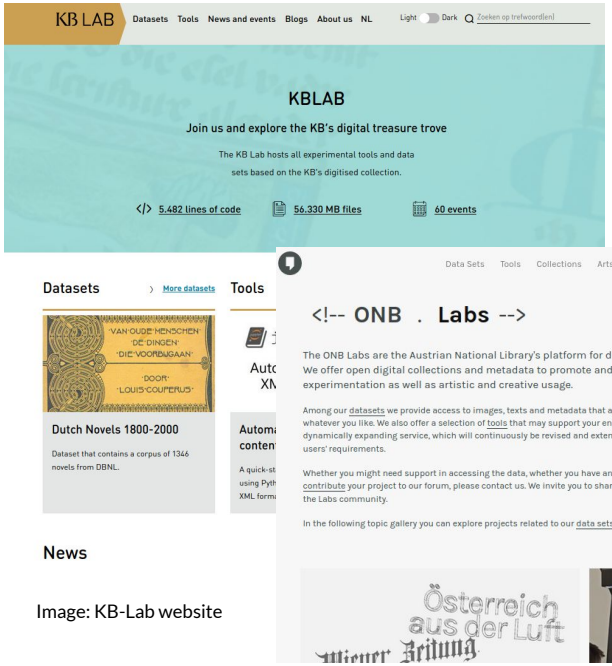


Image: KB-Lab website

Image: ÖNB-Lab website

Steps on the way

- Making the lab (last 2 years)
  - Service design (2021)
    - Mapping the needs - interviews with representative users and stakeholders
    - Reflecting and designing the plans on the basis of this
  - Legal analysis (2022)
    - What can we do in which limits
  - Platform (2022)
    - Updated website that caters for this (data, tools, case studies)
  - Migrating and making (2022 onwards)
    - Datasets and tools

Virtual lab as a platform

Tools

Datasets

User experience

Competences

Copyright and access to data

"My data"

Open access

Formats

Data quality

APIs

Enriched data

What to digitize?

Permanent IDs

Crowd-sourcing

Segmentation

Digitization quality

Integrating databases

Networks

Systematization and classification

Cooperation

Virtual lab as a service

Data available, data planned

Estonian National Bibliography (enriched publication metadata, people and organizations)

Digital archive text collection - metadata, fulltext access, ngrams (periodicals, books)

Thematic collections (e.g. images of postcards, parliamentary collections etc)

Goal: multiple formats where possible

Tools available, tools planned

Comfortable access to full texts and metadata (jupyter notebooks)

National Bibliography metadata explorer (point & click interface)

Ngram search on newspapers (like google ngrams)

APIs, SPARQL etc

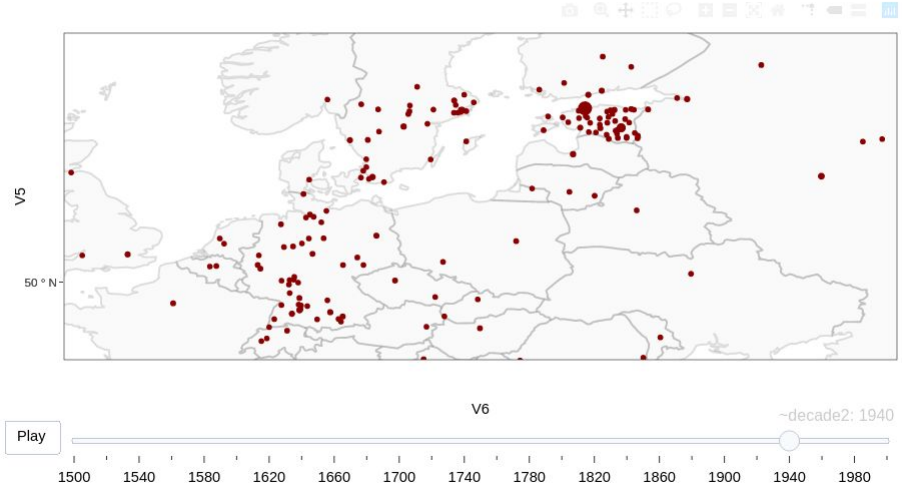
Bibliographic metadata explorer (work in progress)

Explore aspects of the bibliographic data

- Here, enriched with geoinfo
- But also just explore the contents

Get a better understanding of

- Dataset (gaps and biases)
- Cultural history



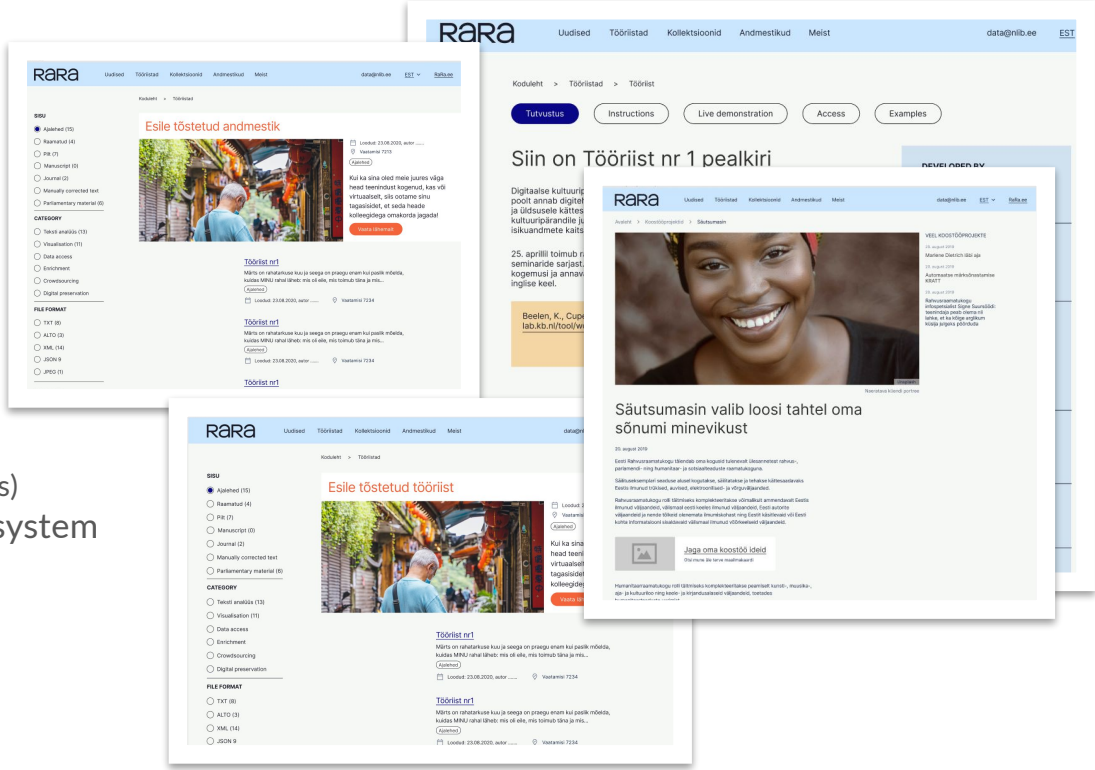
https://peetertinits.github.io/reports/nlib/all_works_geo.html

RaRa EESTI RAHVUS-RAAMATUKOGU

Virtual lab

Data, tools, case studies

- Website due to release in 2022
- Finding creative uses
  - Mapping what's being done
  - Encouraging use (scholarships, prizes)
- Getting the work done back into the system
  - Derived & enriched data
  - Algorithms and tools made



RaRa EESTI RAHVUS-RAAMATUKOGU

A call

If you want to help! If you see what you like or want to show how to do better.

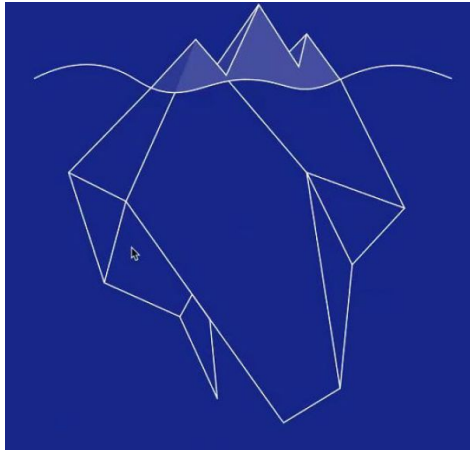
Talk to me after or e-mail at data@nlib.ee.

We may have a job for you. :)

Thank you

From the team at the National Library of Estonia

& only possible through the work of generations of librarians

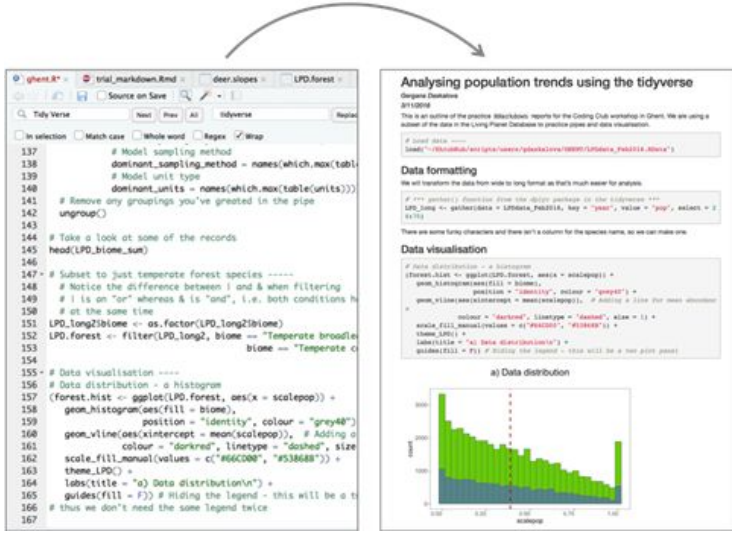


RaRa EESTI RAHVUS-RAAMATUKOGU

open digital libraries

Some extra slides

Open Science practice



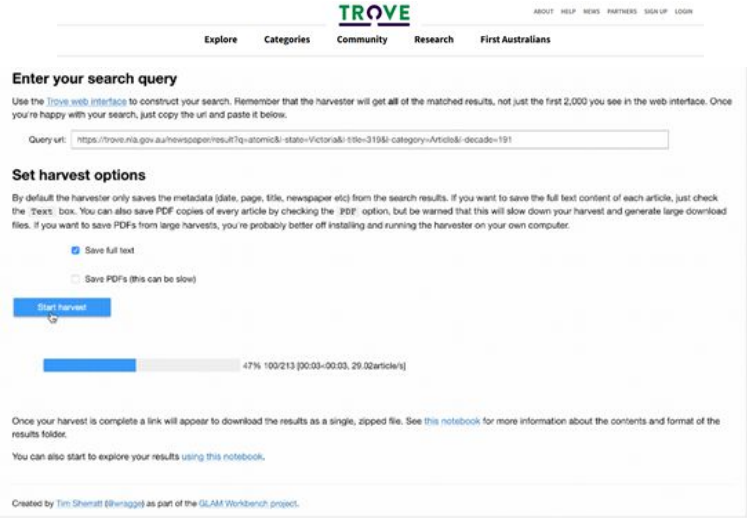
Data and code availability

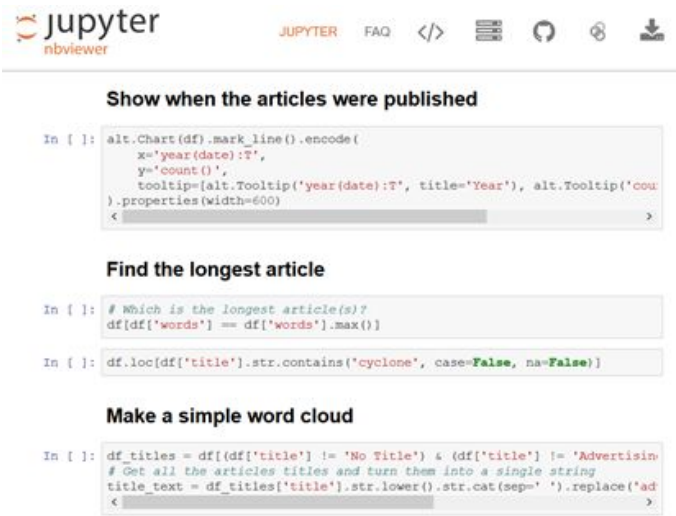
Data and code to reproduce the analysis and figures are available at https://osf.io/6ysda/

(Heunis 2020)

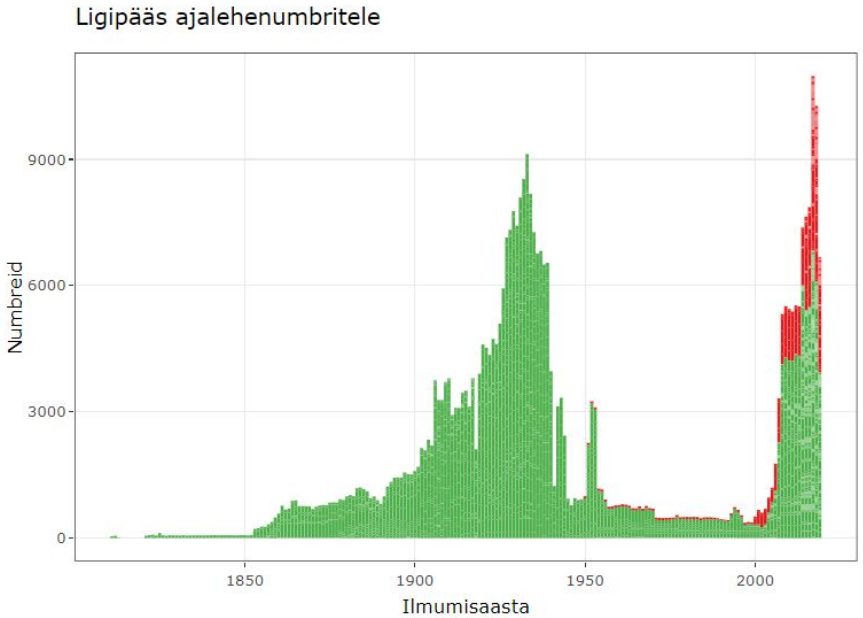
GLAMs with data

Access points to data via open code (e.g. GLAMworkbench)





Open materials at NLE



https://data.digar.ee

Interactive overviews

http://data.digar.ee/text/dea_info.html

http://data.digar.ee/text/dietrich_digar.html

Open code

Andmekogu

Andmekoguna kasutame Eesti Rahvusraamatukogu digiarhiivi Eesti artikleid, millele on olemas tekstikaeveligipääs.
Kollektsiooni materjalidest saab ülevaate siit http://data.digar.ee/text/dea_info.html. Ligipääs on hetkel ainult koodi läbi

```{r}
# Loe sisse metaandmete fail hpc serverilt.
all_issues <- fread("unzip -p /gpfs/hpc/projects/digar_txt/text/all_issues_access.zip",sep="\t")[access_now==T]

# Valime AJALEHED, 1920 ja 1940 vahel, kus on väljaande koodiks postimeesew
subset <- all_issues[str_detect(DocumentType,"NEWSPAPER")&year>1920&year<1940&keyid=="postimeesew"]

# Meile vajalike failide nimekiri
files <- subset[zippath_sections!="",unique(zippath_sections)]
collectionname <- "/gpfs/hpc/projects/digar_txt/text"
filelist <- paste0(collectionname,"/text_sections/", files)

```

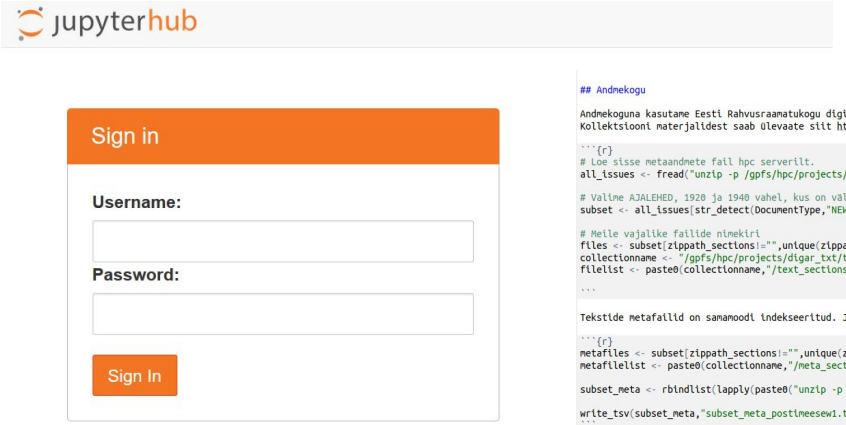
Tekstide metafailid on samamoodi indekseeritud. Järgmine koodijupp kogub kokku meie otsinguga seotud metainfo.

```{r}
metafiles <- subset[zippath_sections!="",unique(zippath_sections_meta)]
metafilelist <- paste0(collectionname,"/meta_sections/", metafiles)

subset_meta <- rbindlist(lapply(paste0("unzip -p ",metafilelist),fread,fill=T),idcol=T)

write_tsv(subset_meta,"subset_meta_postimeesew1.tsv")
```

https://data.digar.ee

Open data

Files at local computing cluster at the Information System of Estonian Science Agency (ETAIS)



Andmekogu

Andmekoguna kasutame Eesti Rahvusraamatukogu digiarhiivi Eesti artikleid, millele on olemas tekstikaeveligipääs.
Kollektsiooni materjalidest saab ülevaate siit http://data.digar.ee/text/dea_info.html. Ligipääs on hetkel ainult koodi läbi

```{r}
# Loe sisse metaandmete fail hpc serverilt.
all_issues <- fread("unzip -p /gpfs/hpc/projects/digar_txt/text/all_issues_access.zip",sep="\t")[access_now==T]

# Valime AJALEHED, 1920 ja 1940 vahel, kus on väljaande koodiks postimeesew
subset <- all_issues[str_detect(DocumentType,"NEWSPAPER")&year>1920&year<1940&keyid=="postimeesew"]

# Meile vajalike failide nimekiri
files <- subset[zippath_sections!="",unique(zippath_sections)]
collectionname <- "/gpfs/hpc/projects/digar_txt/text"
filelist <- paste0(collectionname,"/text_sections/", files)
```

Tekstide metafailid on samamoodi indekseeritud. Järgmine koodijupp kogub kokku mete otsinguga seotud metainfo.

```{r}
metafiles <- subset[zippath_sections!="",unique(zippath_sections_meta)]
metaflilelist <- paste0(collectionname,"/meta_sections/", metafiles)

subset_meta <- rbindlist(lapply(paste0("unzip -p ",metafilelist),fread,fill=T),idcol=T)

write_tsv(subset_meta,"subset_meta_postimeesew1.tsv")
```

Access points

RStudio, Jupyter