



# Virtual Lab at the National Library of Estonia

Peeter Tinitis, Oct 6, 2022

DH Estonia 2022

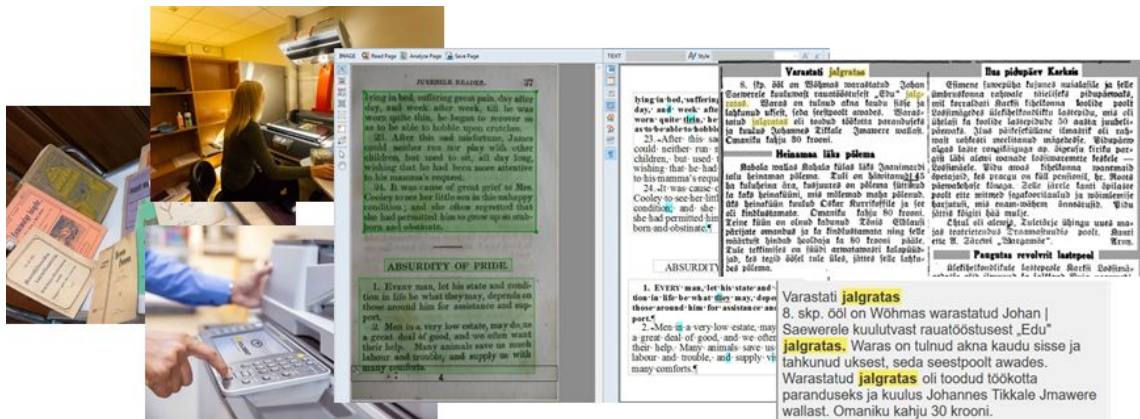
# Libraries have large digital collections

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

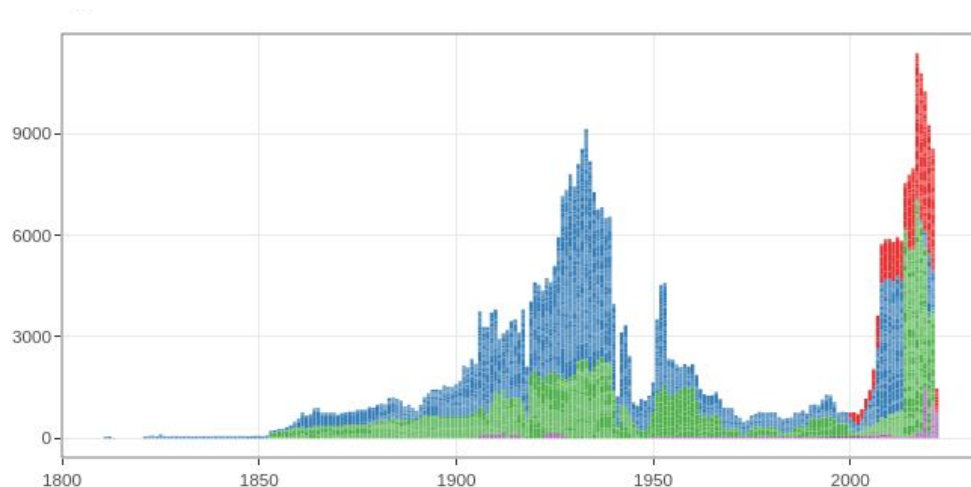
~9.1 M articles

~4.7 M pages

~465 K issues

~2601 publications

(+20-30% in last 2 years)





## Researcher needs

Texts



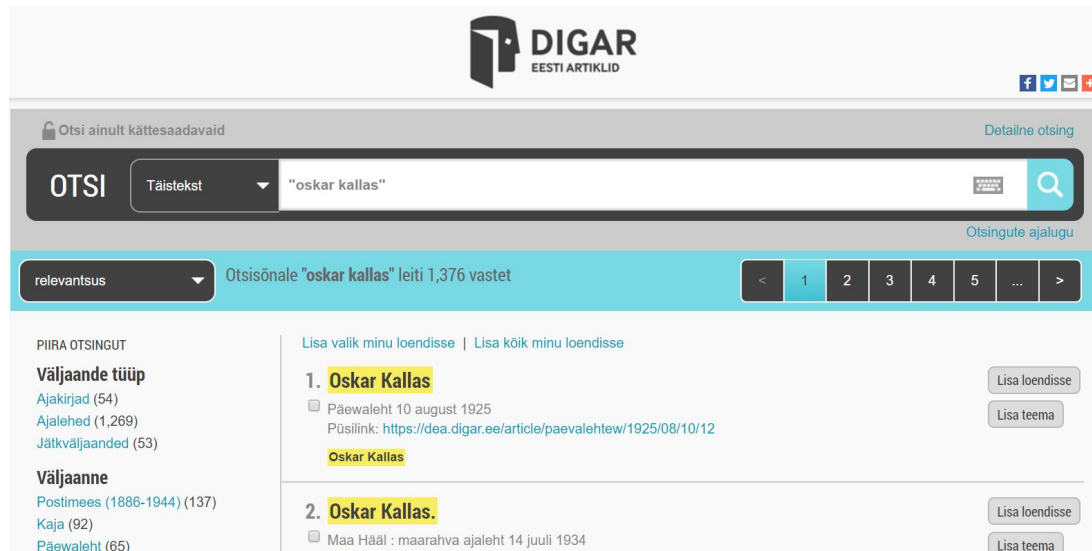
Metadata



Tools



# Approaches to collections: Search engine



The screenshot displays the DIGAR (Eesti Artiklid) search engine interface. At the top, the DIGAR logo is visible. Below it, a search bar contains the text "oskar kallas". To the left of the search bar is a dropdown menu labeled "OTSI" with "Täistekst" selected. To the right of the search bar is a magnifying glass icon. Below the search bar, a light blue bar indicates "Otsisõnale 'oskar kallas' leiti 1,376 vastet". Below this bar, there are two columns of results. The left column lists categories: "Väljaande tüüp" (Ajakirjad (54), Ajalehed (1,269), Jätkväljaanded (53)) and "Väljaanne" (Postimees (1886-1944) (137), Kaja (92), Päevaleht (65)). The right column shows search results: "1. Oskar Kallas" (Päevaleht 10 august 1925, Püsilink: https://dea.digar.ee/article/paevalehtew/1925/08/10/12) and "2. Oskar Kallas." (Maa Hääl : maarahva ajaleht 14 juuli 1934). Each result has a "Lisa loendisse" button. At the bottom right, there is a "Lisa teema" button.

OTSI Täistekst "oskar kallas"

Otsisõnale "oskar kallas" leiti 1,376 vastet

relevantsus

PIIRA OTSINGUT

**Väljaande tüüp**

Ajakirjad (54)

Ajalehed (1,269)

Jätkväljaanded (53)

**Väljaanne**

Postimees (1886-1944) (137)

Kaja (92)

Päevaleht (65)

Lisa valiik minu loendisse | Lisa kõik minu loendisse

1. **Oskar Kallas**

Päevaleht 10 august 1925

Püsilink: <https://dea.digar.ee/article/paevalehtew/1925/08/10/12>

**Oskar Kallas**

2. **Oskar Kallas.**

Maa Hääl : maarahva ajaleht 14 juuli 1934

Lisa loendisse

Lisa teema

Lisa loendisse

Lisa teema






## More complex queries?



The screenshot shows the DIGAR (Eesti Artiklid) search interface. At the top, the logo "DIGAR EESTI ARTIKLID" is visible. Below it, a search bar contains the text "names that cooccur with oskar kallas". To the left of the search bar is a dropdown menu labeled "Taistekst" with a downward arrow. To the right of the search bar is a magnifying glass icon. Above the search bar, there is a link "Otsi ainult kättesaadavaid" and a link "Detailne otsing". Below the search bar, there is a link "Otsingute ajalugu". At the bottom of the search bar, there is a message: "Otsisõnale names that cooccur with oskar kallas leiti 0 vastet".

# Interfaces & libraries

## Delpher



**jupyter nbviewer**

JUPYTER FAQ </> [Icons]

### Show when the articles were published

```
In [ ]: alt.Chart(df).mark_line().encode(
    x='year(date):T',
    y='count()',
    tooltip=[alt.Tooltip('year(date):T', title='Year'), alt.Tooltip('count()', title='Count')],
    properties(width=600)
```

### Find the longest article

```
In [ ]: # Which is the longest article(s)?
df[df['words'] == df['words'].max()]
```

Found 234,825 images published from Nov 14, 1923 to Dec 31, 1929.

1929

ORDER BY: DATE ASCENDING

Image	Publication Date	Publication Title
	1929-11-14	Gazette de Lausanne
	1929-11-14	Gazette de Lausanne
	1929-11-14	Gazette de Lausanne



# 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)

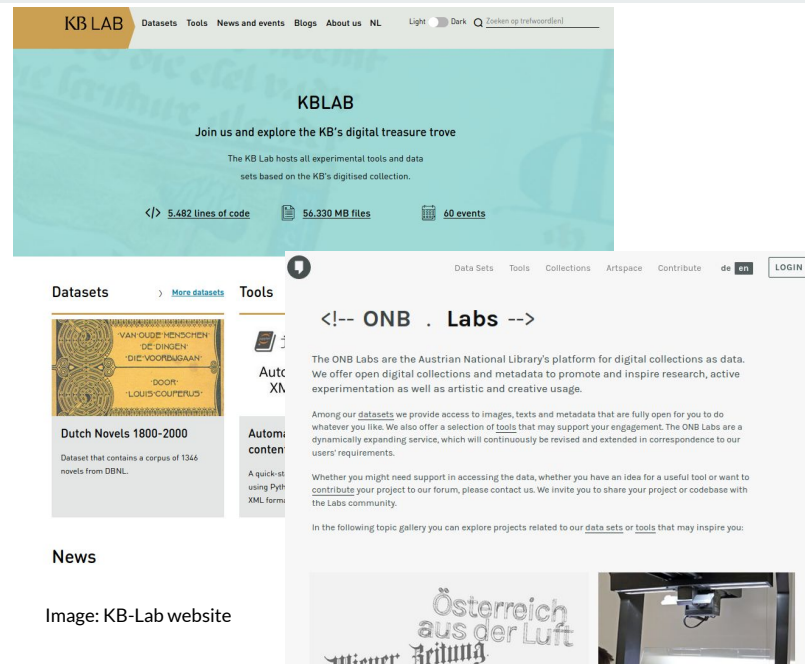


Image: KB-Lab website

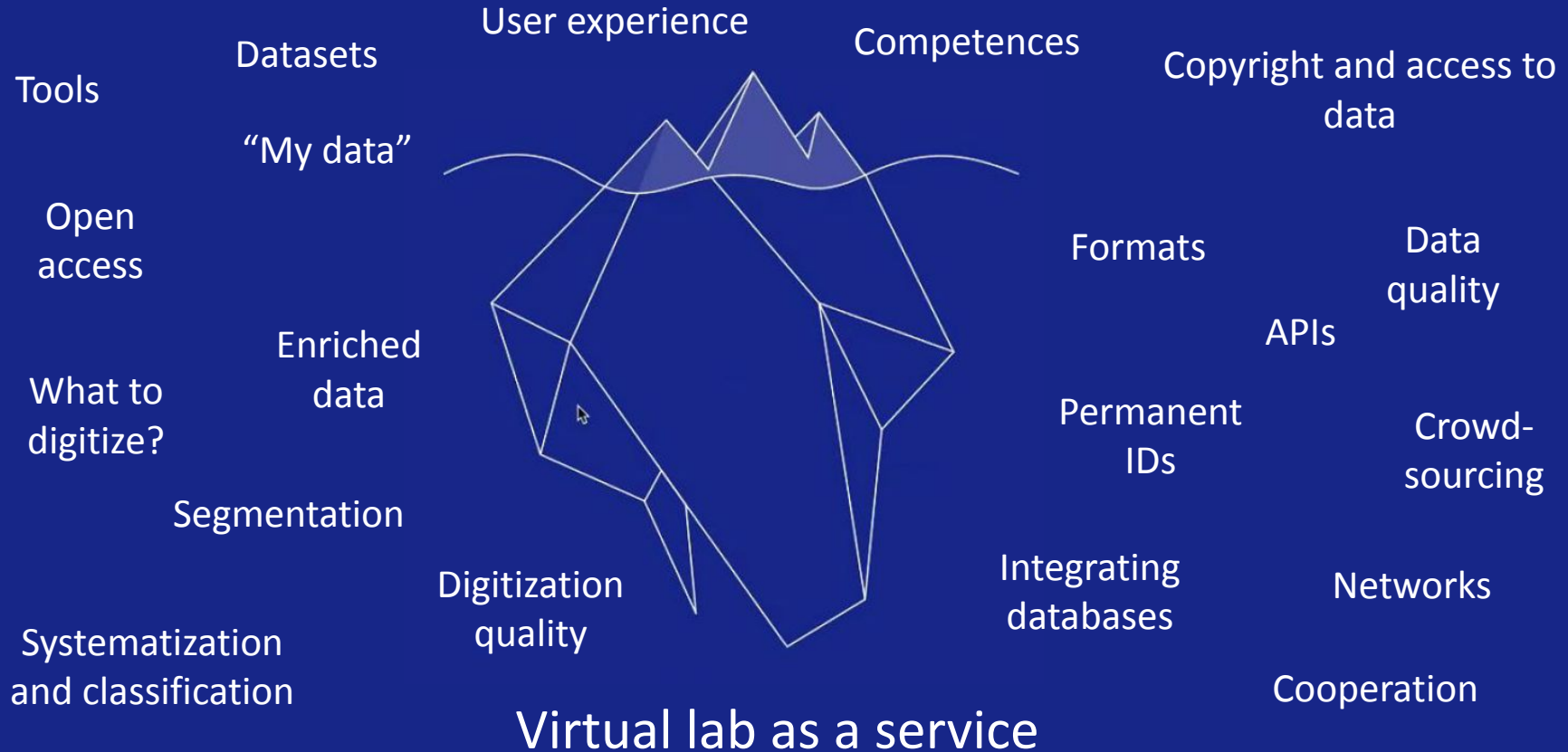
Image: ÖNB-Lab website



## 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





## 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



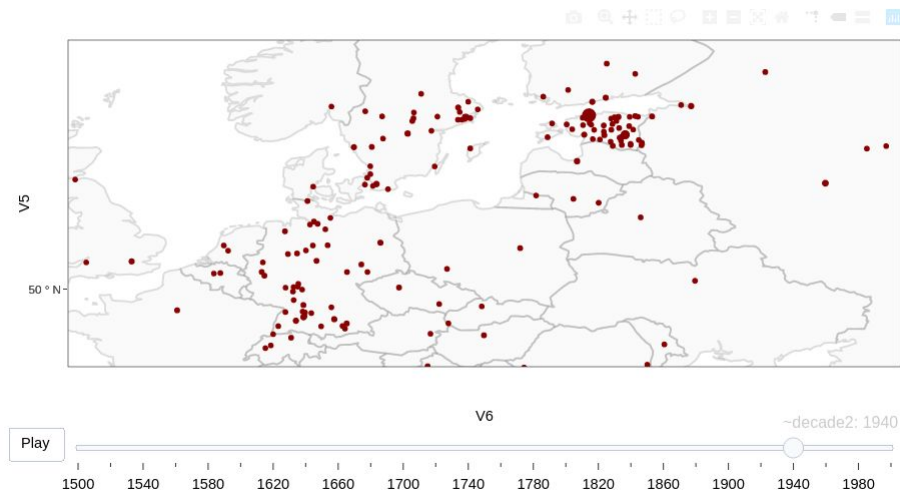
# Bibliographic metadata explorer (work in progress)

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

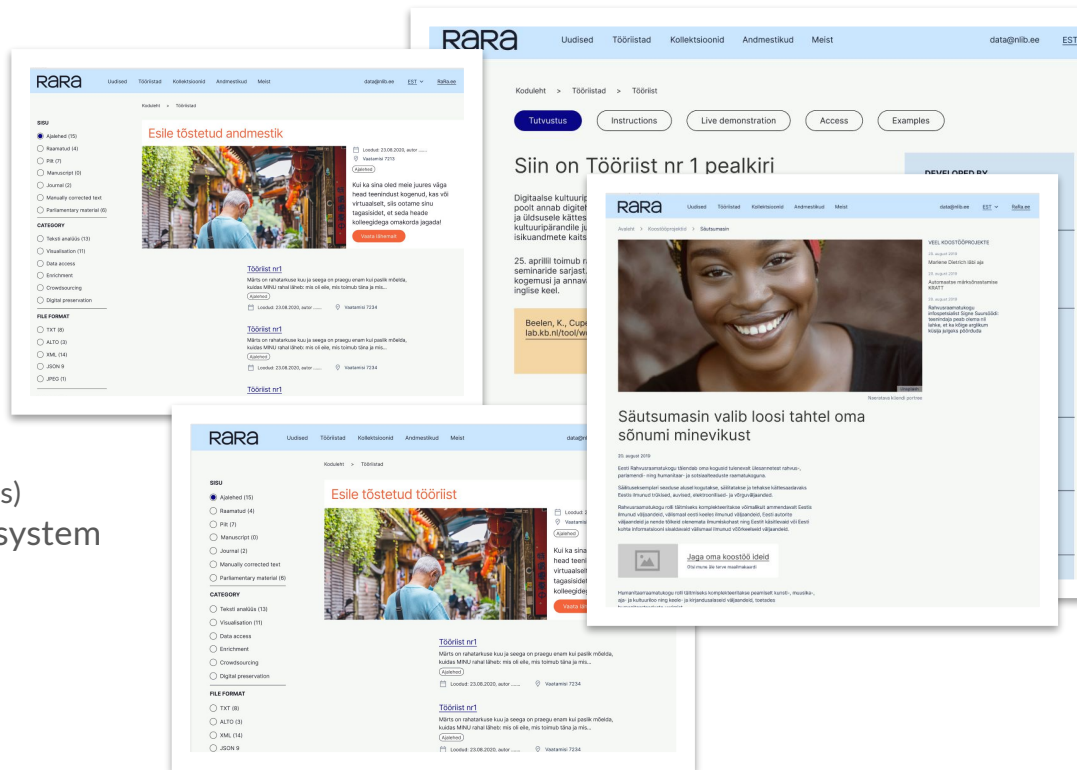


[https://peetertinits.github.io/reports/nlib/all\\_works\\_geo.html](https://peetertinits.github.io/reports/nlib/all_works_geo.html)

# 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





## 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](mailto:data@nlib.ee).

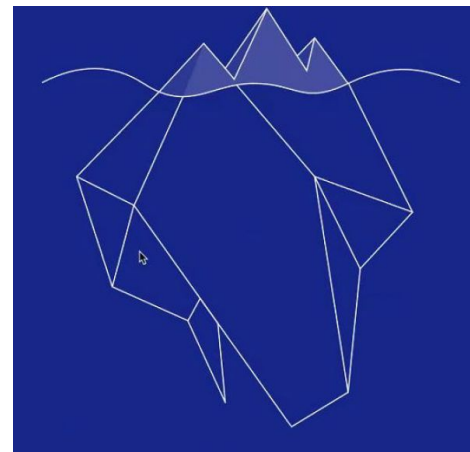
We may have a job for you. :)



# Thank you

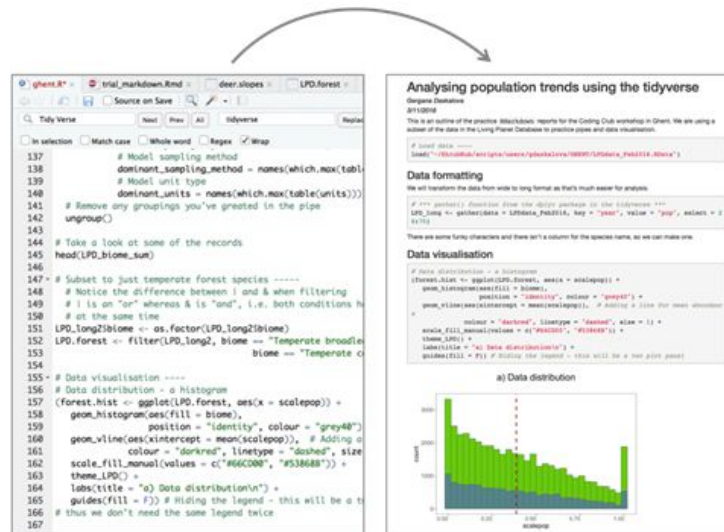
From the team at the National Library of Estonia

& only possible through the work of generations of librarians



Some extra slides

# Open Science practice



## Data and code availability

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

DATA AND CODE  
AVAILABLE ON REQUEST

SHARES CODE

SHARES  
DATA AND CODE

SHARES REPRODUCIBLE  
ANALYSIS ENVIRONMENT

SHARES OPEN & INTERACTIVE  
APPLICATION TO EXPLORE DATA

(Heunis 2020)

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

[Trove](#)
[ABOUT](#)
[HELP](#)
[NEWS](#)
[PARTNERS](#)
[SIGN UP](#)
[LOGIN](#)

[Explore](#)
[Categories](#)
[Community](#)
[Research](#)
[First Australians](#)

---

## Enter your search query

Use the [Trove web interface](#) to construct your search. Remember that the harvester will get all of the matched results, not just the first 2,000 you see in the web interface. Once you're happy with your search, just copy the url and paste it below.

Query url:

## Set harvest options

By default the harvester only saves the metadata (date, page, title, newspaper etc) from the search results. If you want to save the full text content of each article, just check the 'Text' box. You can also save PDF copies of every article by checking the 'PDF' option, but be warned that this will slow down your harvest and generate large download files. If you want to save PDFs from large harvests, you're probably better off installing and running the harvester on your own computer.

☒ Save full text  
☐ Save PDFs (this can be slow)

[Start Harvest](#)

47% 100/213 [00:03-00:03, 29.02article/s]

Once your harvest is complete a link will appear to download the results as a single, zipped file. See [this notebook](#) for more information about the contents and format of the results folder.

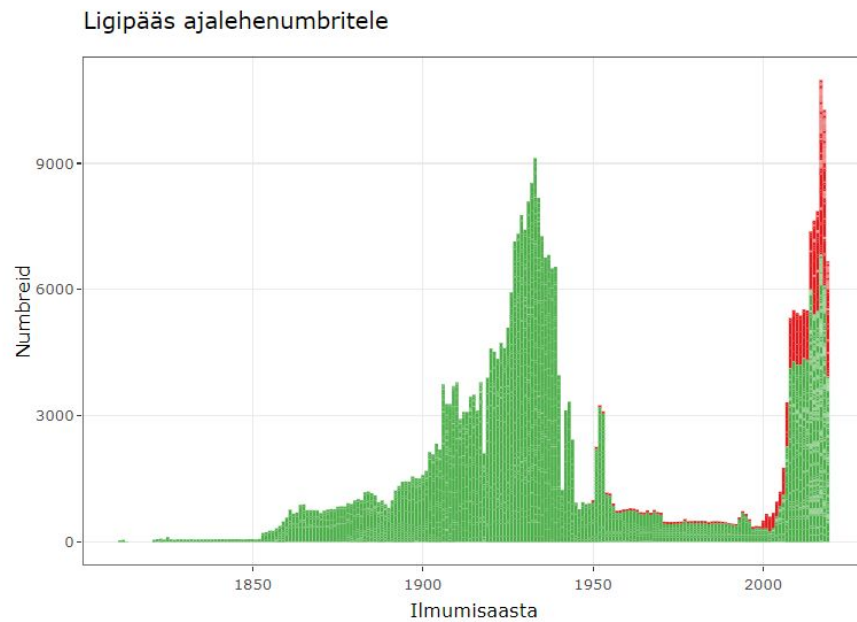
You can also start to explore your results [using this notebook](#).

Created by Tim Shennatt (@wags) as part of the GLAM Workbench project.

```

In [ ]: alt.Chart(df).mark_line().encode(
        x='year(date):T',
        y='count()',
        tooltip=[alt.Tooltip('year(date):T', title='Year'), alt.Tooltip('count', title='Count')],
        properties(width=600)
    )
```

# Open materials at NLE







# Interactive overviews

[http://data.digar.ee/text/dea\\_info.html](http://data.digar.ee/text/dea_info.html)

[http://data.digar.ee/text/dietrich\\_digar.html](http://data.digar.ee/text/dietrich_digar.html)

# Open code

## Andmekogu

Andmekoguna kasutame Eesti Rahvusraamatukogu digiarhiivi Eesti artikleid, millele on olemas tekstikaueveligipääs. Kollektiooni materjalidest saab ülevaate siit [http://data.digar.ee/text/dea\\_info.html](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)

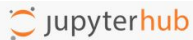
subset_meta <- rbindlist(lapply(paste0("unzip -p ", metafilelist), fread, fill=T, idcol=T))

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)



Sign in

Username:

Password:

Sign In

## Andmekogu

Andmekogu kasutane Eesti Rahvusraamatukogu diglarhiivi Eesti artikleid, millele on olemas tekstikaeveligipääs. Kollektiooni materjalidest saab ülevaate siit [http://data.digar.ee/text/dea\\_info.html](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/dlgar_txt/text/all_issues_access.zip",sep="\t")[access_now==1]

# Valime AJALEHD, 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 nimekirj
files <- subset(zippath_sections=="",unique(zippath_sections))
collectionname <- "/gpfs/hpc/projects/dlgar_txt/text"
filelist <- paste0(collectionname,"/text_sections/", files)

'''
```

Tekstide metafailid on sanamoodi indekseeritud. Järgmine koodijupp kogub kokku neile otsinguga seotud metainfo.

```
'''(r)
metafiles <- subset(zippath_sections=="",unique(zippath_sections_meta))
metafilelist <- paste0(collectionname,"/meta_sections/", metafiles)

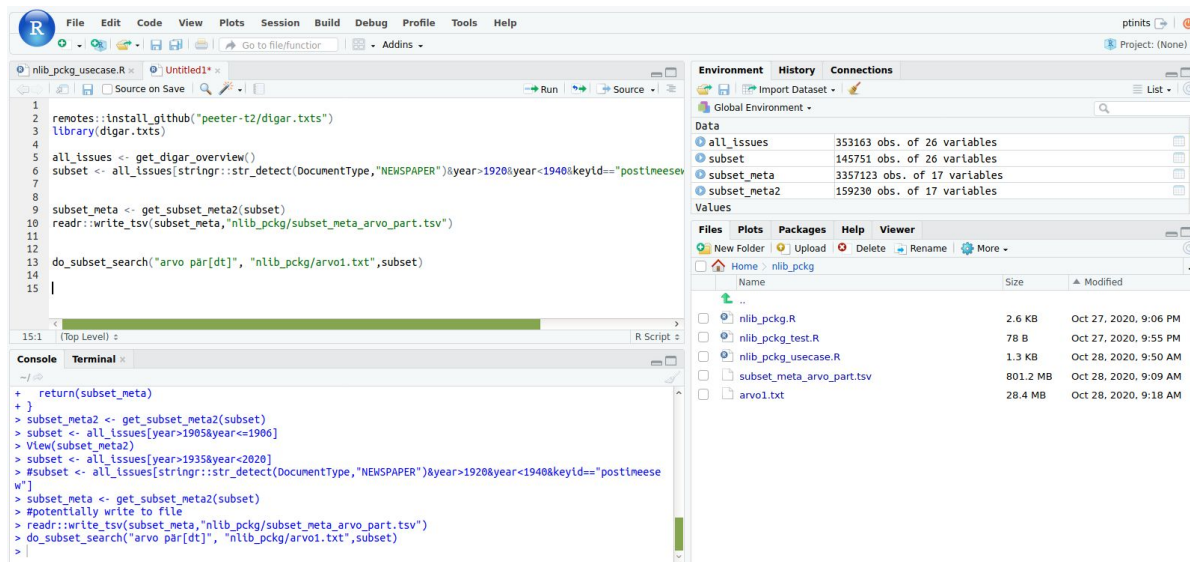
subset_meta <- rbindlist(lapply(paste0("unzip -p ",metafilelist),fread,fill=T),idcol=1)

write_tsv(subset_meta,"subset_meta_postimeesew1.tsv")

'''
```

# Access points

RStudio, Jupyter



The screenshot displays the RStudio interface with the following components:

- Source Editor:** Contains an R script for data processing. The code includes installing a GitHub package, loading it, filtering a dataset by year and key, and saving the results.
- Console:** Shows the execution output, including the return of the subsetted data and the successful writing of the TSV file.
- Environment Pane:** Lists the objects in the global environment, including the original dataset and the subsetted data.

```
1 remotes::install_github("peeter-tz/digar.txts")
2 library(digar.txts)
3
4 all_issues <- get_digar_overview()
5 subset <- all_issues[stringr::str_detect(DocumentType, "NEWSPAPER") & year > 1920 & year < 1940 & keyId == "postineese"]
6
7 subset_meta <- get_subset_meta2(subset)
8 readr::write_tsv(subset_meta, "nlib_pckg/subset_meta_arvo_part.tsv")
9
10 do_subset_search("arvo par[dt]", "nlib_pckg/arvo1.txt", subset)
```

**Console Output:**

```
+ return(subset_meta)
+ }
> subset_meta2 <- get_subset_meta2(subset)
> subset <- all_issues[year>1905&year<=1906]
> View(subset_meta2)
> subset <- all_issues[year>1935&year<2020]
> #subset <- all_issues[stringr::str_detect(DocumentType, "NEWSPAPER") & year>1920&year<1940&keyId=="postineese"]
> subset_meta <- get_subset_meta2(subset)
> #potentially write to file
> readr::write_tsv(subset_meta, "nlib_pckg/subset_meta_arvo_part.tsv")
> do_subset_search("arvo par[dt]", "nlib_pckg/arvo1.txt", subset)
> |
```

**Environment Pane:**

| Object       | Size                         | Modified |
|--------------|------------------------------|----------|
| all_issues   | 353163 obs. of 26 variables  |          |
| subset       | 145751 obs. of 26 variables  |          |
| subset_meta  | 3357123 obs. of 17 variables |          |
| subset_meta2 | 159230 obs. of 17 variables  |          |

**Files Pane:**

| Name                      | Size     | Modified              |
|---------------------------|----------|-----------------------|
| ..                        |          |                       |
| nlib_pckg.R               | 2.6 KB   | Oct 27, 2020, 9:06 PM |
| nlib_pckg_test.R          | 78 B     | Oct 27, 2020, 9:55 PM |
| nlib_pckg_usecase.R       | 1.3 KB   | Oct 28, 2020, 9:50 AM |
| subset_meta_arvo_part.tsv | 801.2 MB | Oct 28, 2020, 9:09 AM |
| arvo1.txt                 | 28.4 MB  | Oct 28, 2020, 9:18 AM |