



# Virtual Lab at the National Library of Estonia

Peeter Tinitis, Urmas Sinisalu

DHNB 2023

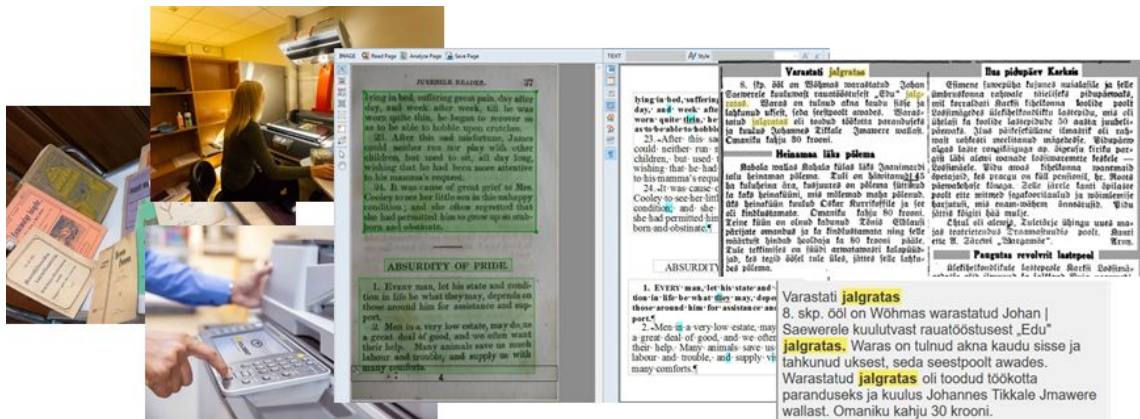
# Libraries have large digital collections

Long time effort in digitization

- Focus on preservation
- Interfaces for reading

New frontiers in usage

- Collections as data
- Creative reuse

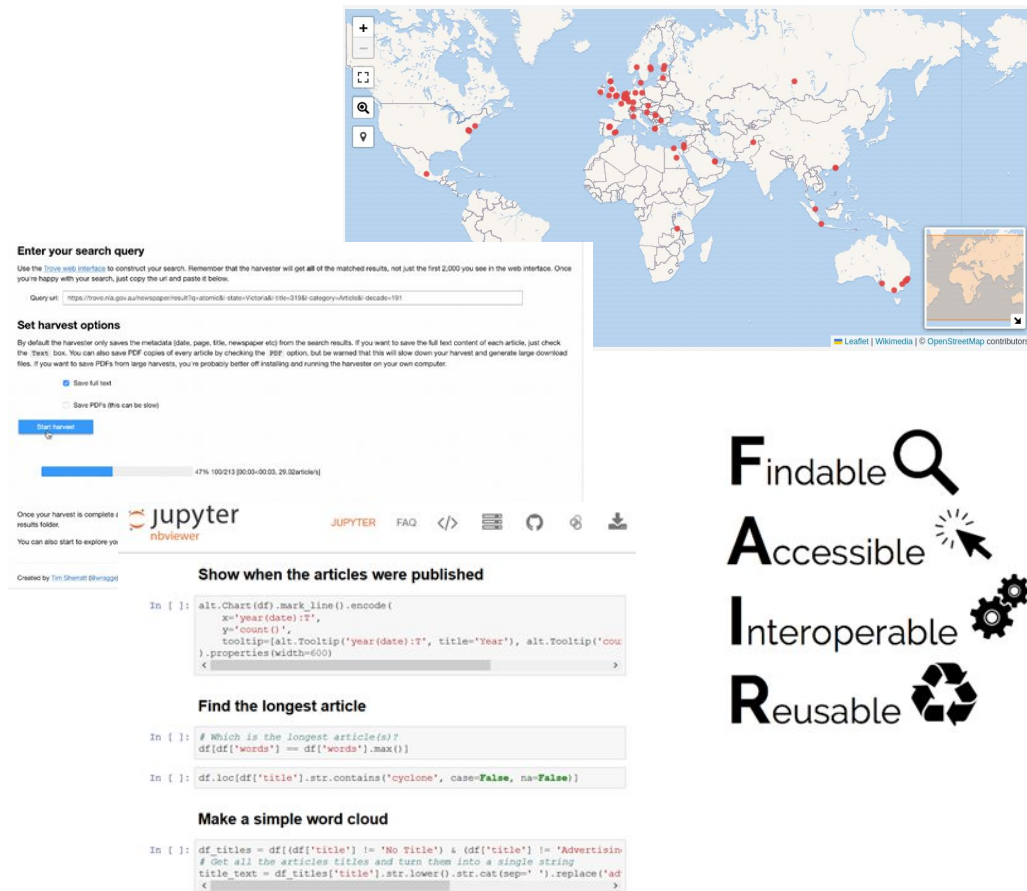


# GLAM Labs

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

Experimental projects in using data.

Computational access to digital collections



**F**indable   
**A**ccessible   
**I**nteroperable   
**R**eusable

# 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)

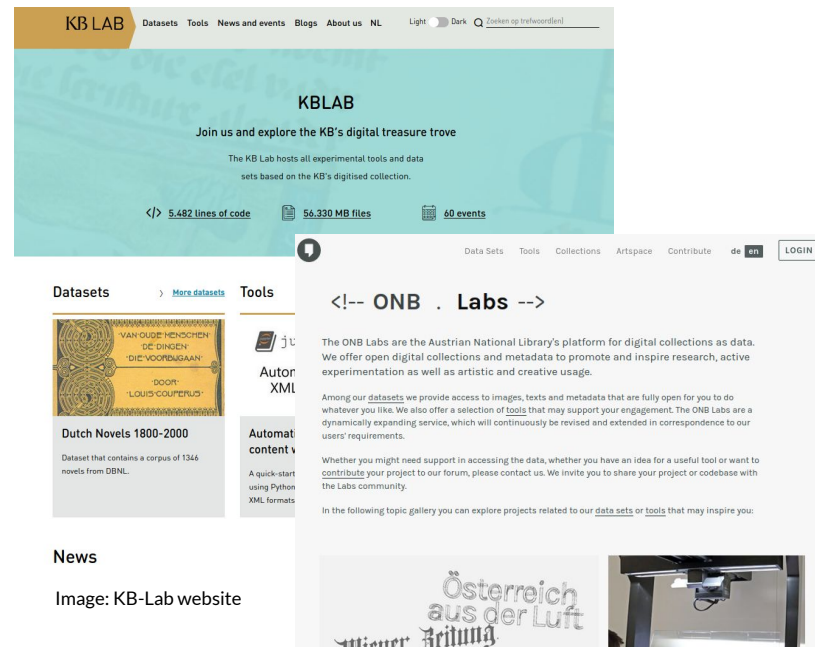


Image: ÖNB-Lab website



## Road map (2019-2023)



- 2019 Plans for the lab
- 2020 - 2023 Project ODL Creative Europe
  - Service design
  - Legal analysis
  - Building the web
- 2020 - ...
  - Preparing tools and datasets
  - Case studies in use
- ...
- 2023.03.30 - Launch





## Steps on the way

### Service design

- Who is it for?
- What do they need?
- What can we do.

Charted other labs, interviewed 13 representative local stakeholders, summarized and synthesized the results.

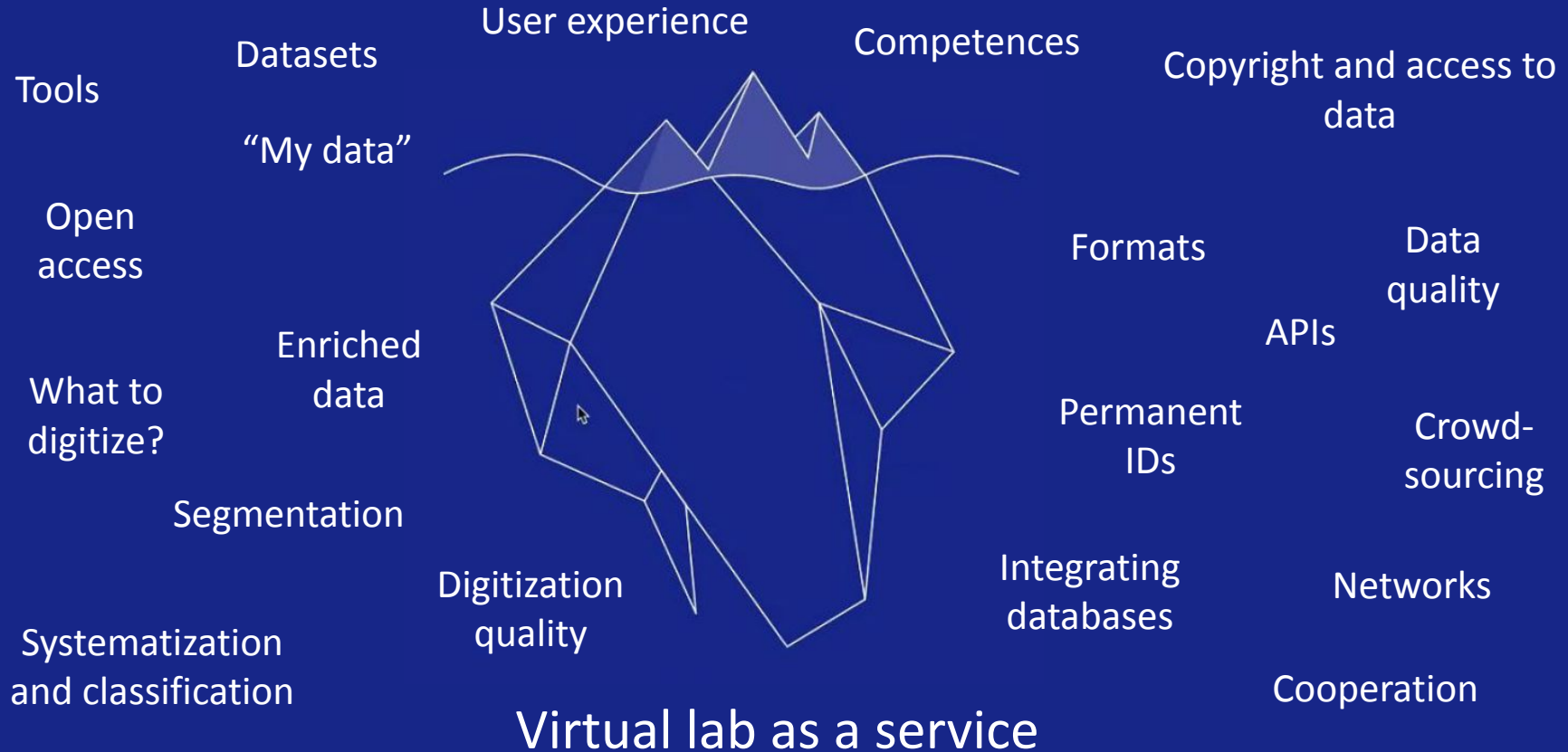
### Legal analysis

- What we want to do.
- What are the constraints?
- What options do we have?

A legal expert completed a report while consulting with us.



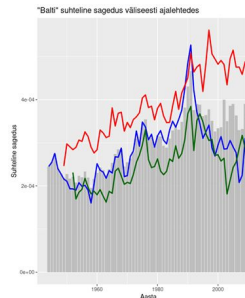
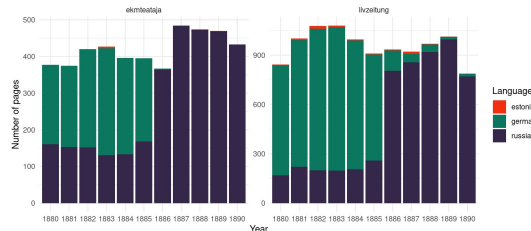
# Virtual lab as a platform



# Case studies

Project at University of Tartu (EKKD72, 2022)

- Case studies on newspaper corpora 1850-2020 at National Library
  - Small groups of students working with supervision
  - Diverse topics (environmental conflicts, foreign words, language of materials)
  - Learning to understand the data and its use
- Testing and feedback on use



## Kes on esinaine?

Milal leidakse organisatsiooni juhtiva naise kohta 'esinaine'?

03.02.2023 • DEa, Eesti keel, Juhtumiuuring



## Laensõnad eesti keeles

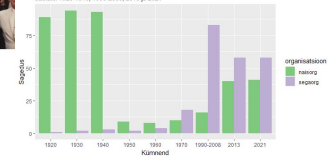
Kas ingliskeelsed sõnad tunguvad eesti keelde ja ohustavad see

02.02.2023 • DEa, Eesti keel, Juhtumiuuring



## Ajalooline märgenud Valla-Eesti aialgheides

akirjanduse

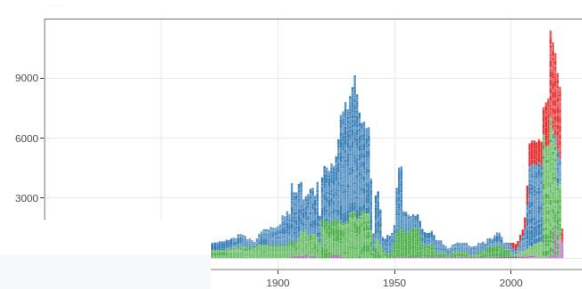




# Building up

## Contents

- Tools, data, guides and examples
- 9M newspaper articles, 120k printed works, 300k entries in the National Bibliography
- Thematic collections and subsets
- Data visualizations and case studies



2. Activate the package that was installed, use

```
library(digar.txts, lib.loc=~/.R/pkg/)
```

3. Use `get_digar_overview()` to get overview of the collections (issue-level).

```
all_issues <- get_digar_overview()
```

4. Build a custom subset through any tools in R. Here is

```
subset <- all_issues %>%  
  filter(DocumentType=="NEWSPAPER") %>%  
  filter(year>1888&year<1948) %>%  
  filter(keyid=="postimeesew")
```

**RARA** Andmestikud Tõrlistad Blogi Uudised Sündmused Meist

Avalaht > Andmestikud

**JÄRJELKORD**

☐ Tähestiku järjekorras

☐ Avalikustamise kuupäev

**TÜÜP**

☒ Kõik

☐ Metaandmed (12)

☐ Ülevaade (1)

**KATEGOORIA**

☒ Kõik

☐ Ajakirjad (3)

☐ Ajalehed (2)



☐ Bibliograafiad (2)

☐ Kaandid (2)

☐ Postkaardid (2)

Vaata veel

**Andmestike ülevaade**



[Jätkväljaanded](#)

DIGAR-is leiduvate jätkväljaanne

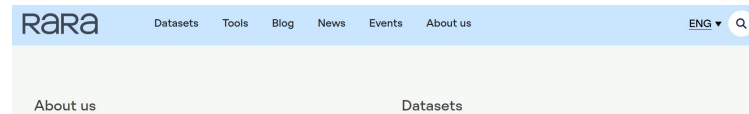
22.02.2023 x



## Community - Goals

What we strive for

- A place to easily find and use library data
- A platform to showcase case studies
- A workflow to integrate used data
- A way to find value for library and creators



Launch 03.30.2023

Ask for more about process and results!

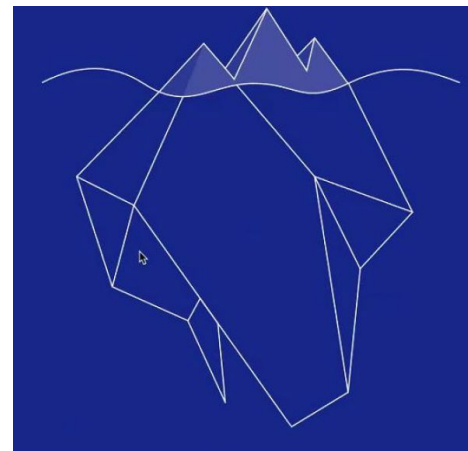
[digilab@nlib.ee](mailto:digilab@nlib.ee)



# Thank you

From the team at the National Library of Estonia

& only possible through the work of generations of librarians







## Researcher needs

Texts



Metadata



Tools



# 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, 1924 to Dec 31, 1924.

ORDER BY: DATE ASCENDING

Image	Publication Date	Source
	1924-11-14	Gazette de Lausanne
	1924-11-20	Gazette de Lausanne
	1924-11-20	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







## 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



## 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

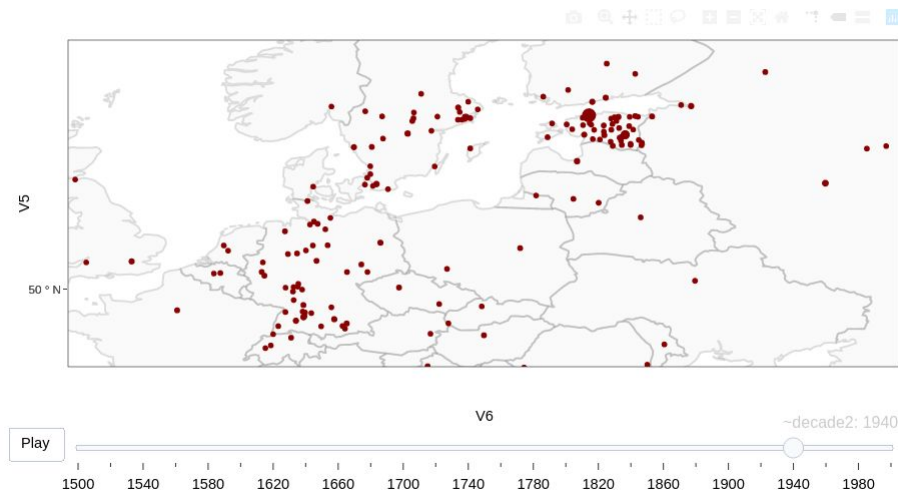
# Bibliographic metadata explorer (work in progress)

Explore aspects of the bibliographic data

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

Get a better understanding of

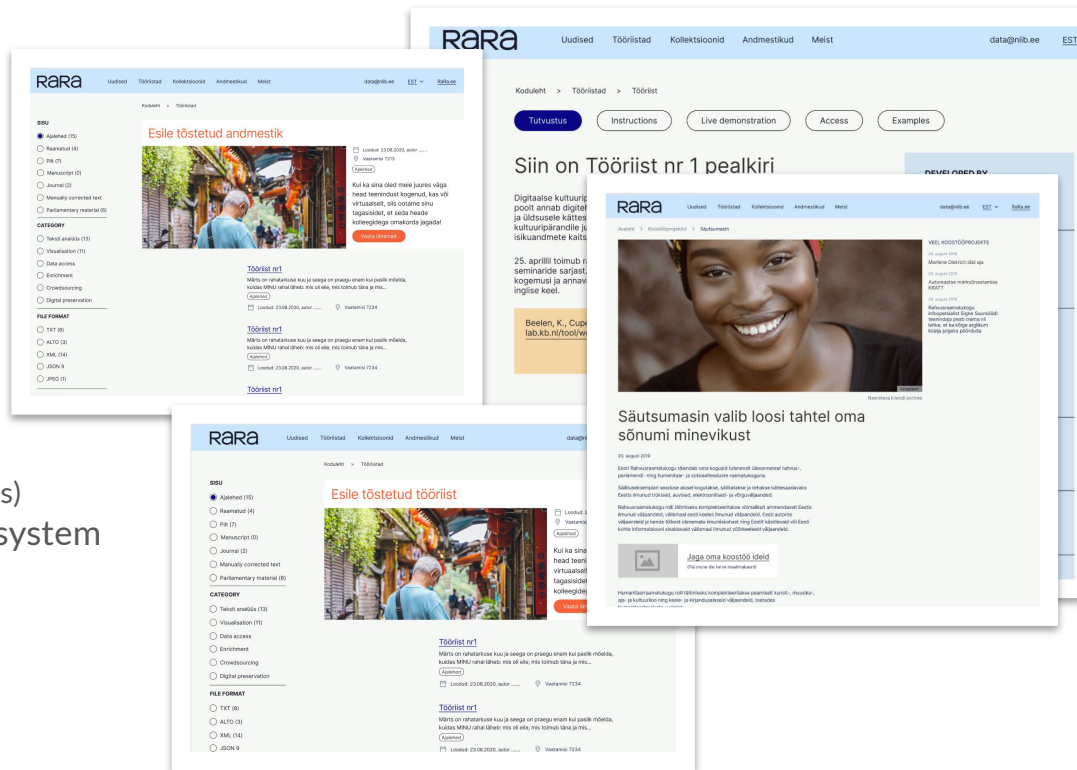
- Dataset (gaps and biases)
- Cultural history



# 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).

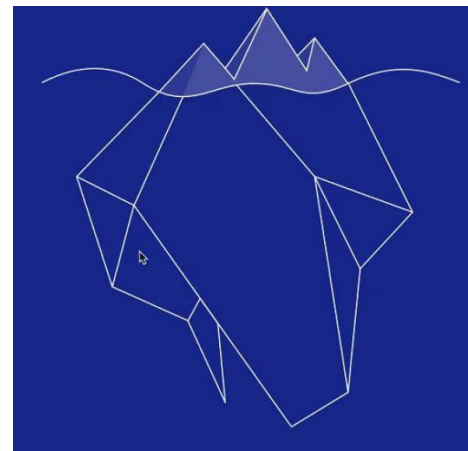
We may have a job for you. :)



# Thank you

From the team at the National Library of Estonia

& only possible through the work of generations of librarians

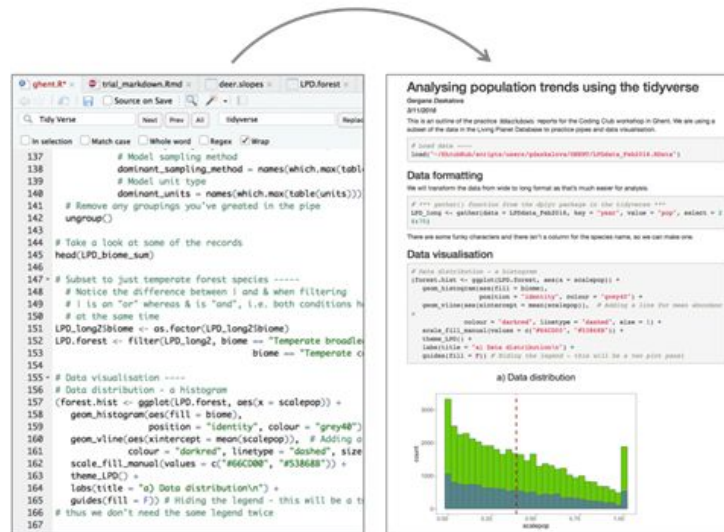


o p e n  
digital  
libraries

Some extra slides



# Open Science practice



## Data and code availability

Data and code to reproduce the analysis and figures are available at <https://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)

## Example: Texts as data

RaRa newspapers  
& periodicals

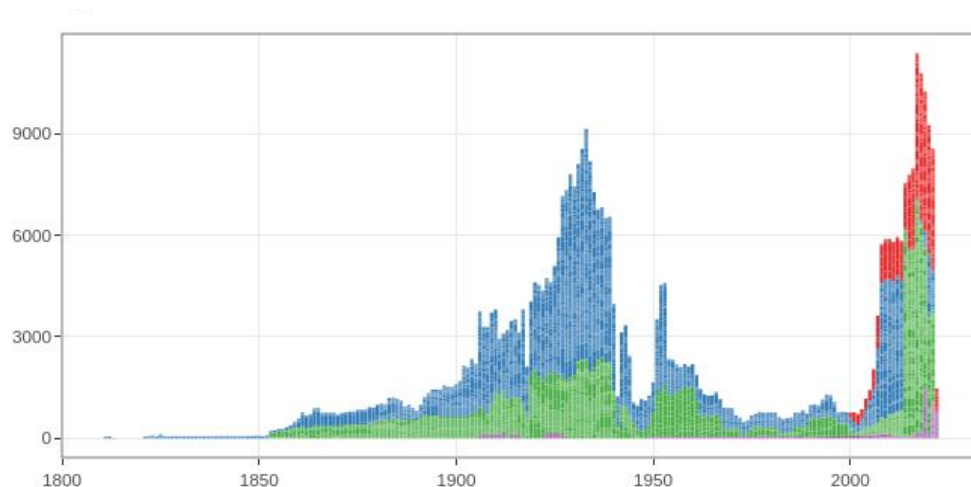
~9.1 M articles

~4.7 M pages

~465 K issues

~2601 publications

(+20-30% in last 2 years)



# GLAMs with data

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

ExploreCategoriesCommunityResearchFirst Australians

trove

ABOUTHELPNEWSPARTNERSIGN-UPLOGIN

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, 29.92article/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 Sherratt \(@tsherratt\)](#) as part of the [GLAM Workbench project](#).

jupyter

nbviewer

JUPYTERFAQ</>⋮⌂⌛⬇

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')])

Find the longest article

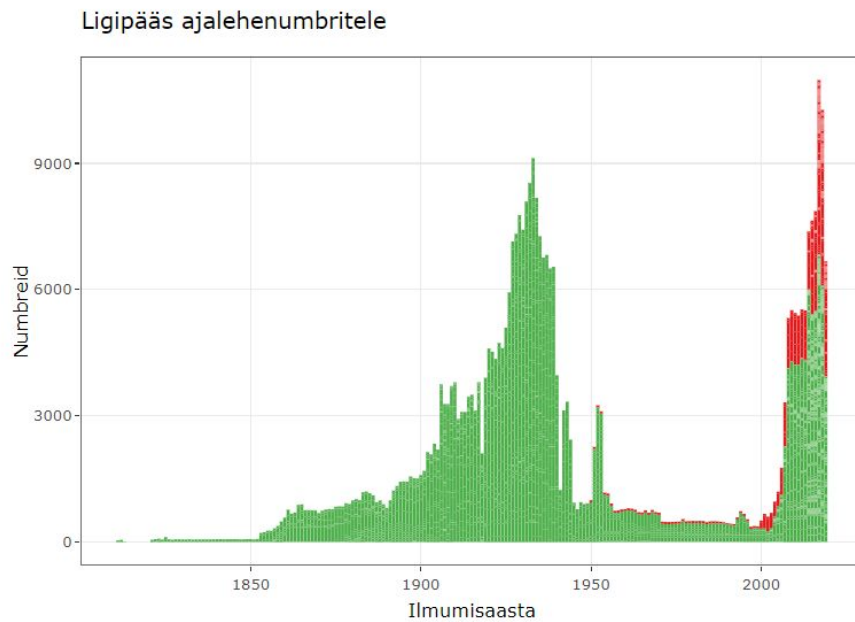
In [ ]: # Which is the longest article(s)?  
df[df['words'] == df['words'].max()]

In [ ]: df.loc[df['title'].str.contains('cyclone', case=False, na=False)]

Make a simple word cloud

In [ ]: df\_titles = df[(df['title'] != 'No Title') & (df['title'] != 'Advertisin...')]  
# Get all the articles titles and turn them into a single string  
title\_text = df\_titles['title'].str.lower().str.cat(sep=' ').replace('ad...')

# 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 tekstikauevõlgipäas. Kollektiooni materjalidest saab ülevaate siit [http://data.digar.ee/text/dea\\_info.html](http://data.digar.ee/text/dea_info.html). Ligipäas on hetkel ainult koodi labi

```
'''(r)
# Loe sisse metaandmete faili 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)

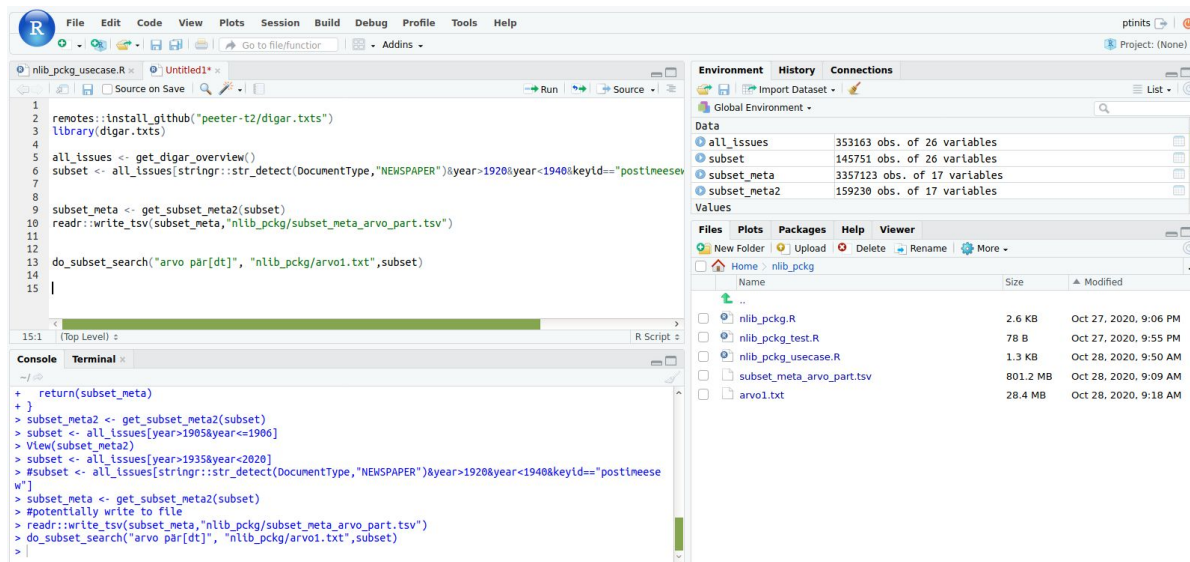
subset_meta <- rbindlist(lapply(paste0("unzip -p ",metafilelist),fread,fill=T),idcol=1)

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, showing the number of observations and variables for each.

```
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)
```

**Environment Pane Data:**

| Object       | Observations | Variables |
|--------------|--------------|-----------|
| all_issues   | 353163       | 26        |
| subset       | 145751       | 26        |
| subset_meta  | 3357123      | 17        |
| subset_meta2 | 159230       | 17        |

**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 |