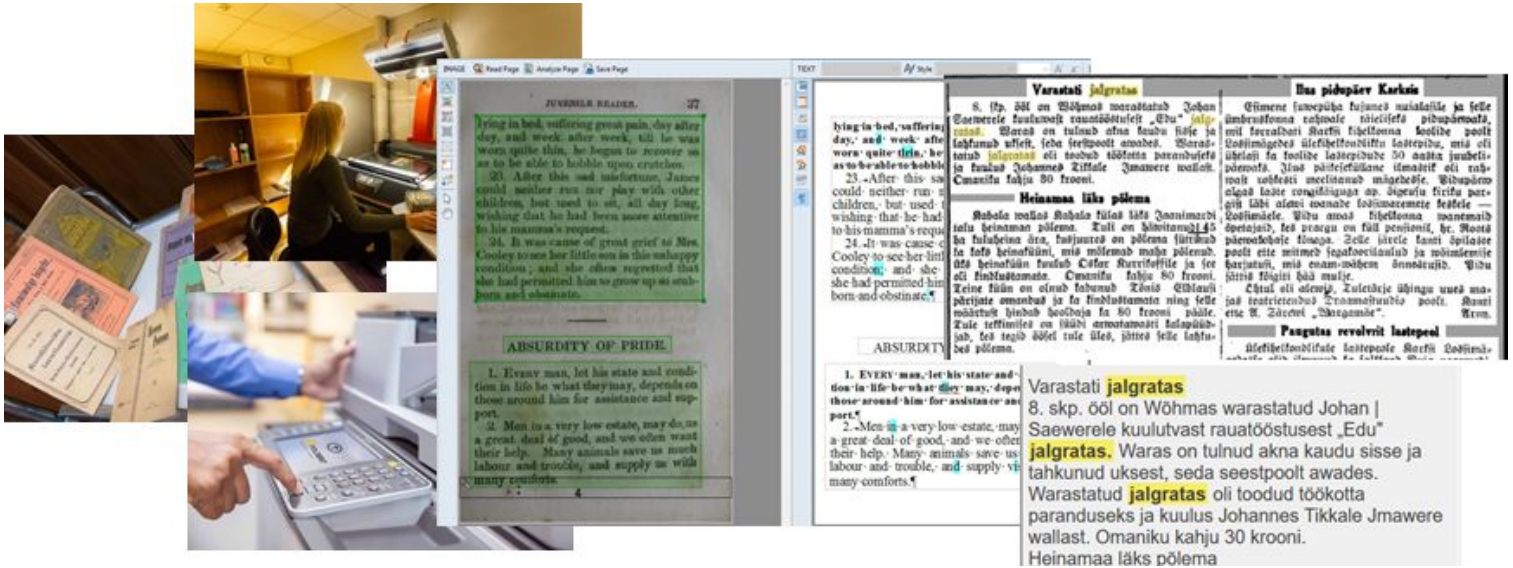
Using digital texts at the National Library of Estonia

Peeter Tinits, Oct 28, 2020
Nordplus workshop

NATIONAL LIBRARY OF ESTONIA

Large digital collections



NATIONAL LIBRARY OF ESTONIA

Texts as data

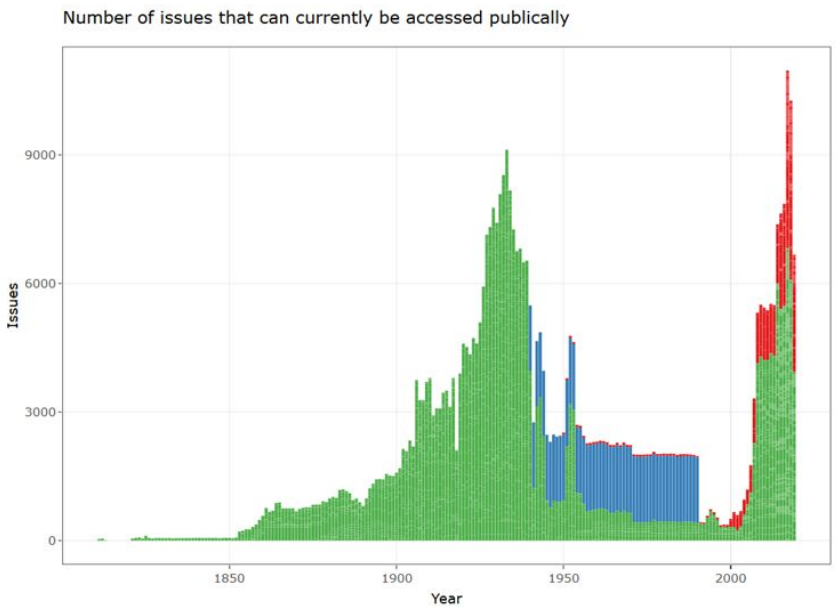
Newspapers
& periodicals

~6.3 M articles

~3.6 M pages

~390 K issues

~2200 publications



NATIONAL LIBRARY OF ESTONIA

https://data.digar.ee

Researcher needs

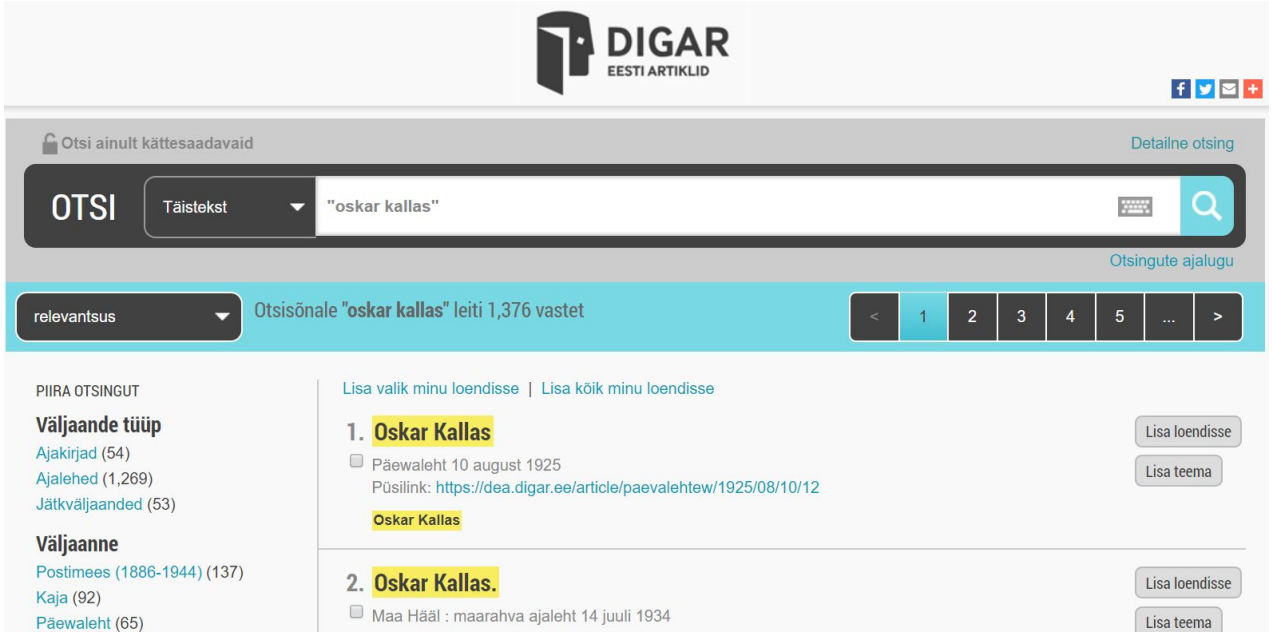
Texts

Metadata

Tools

NATIONAL LIBRARY OF ESTONIA

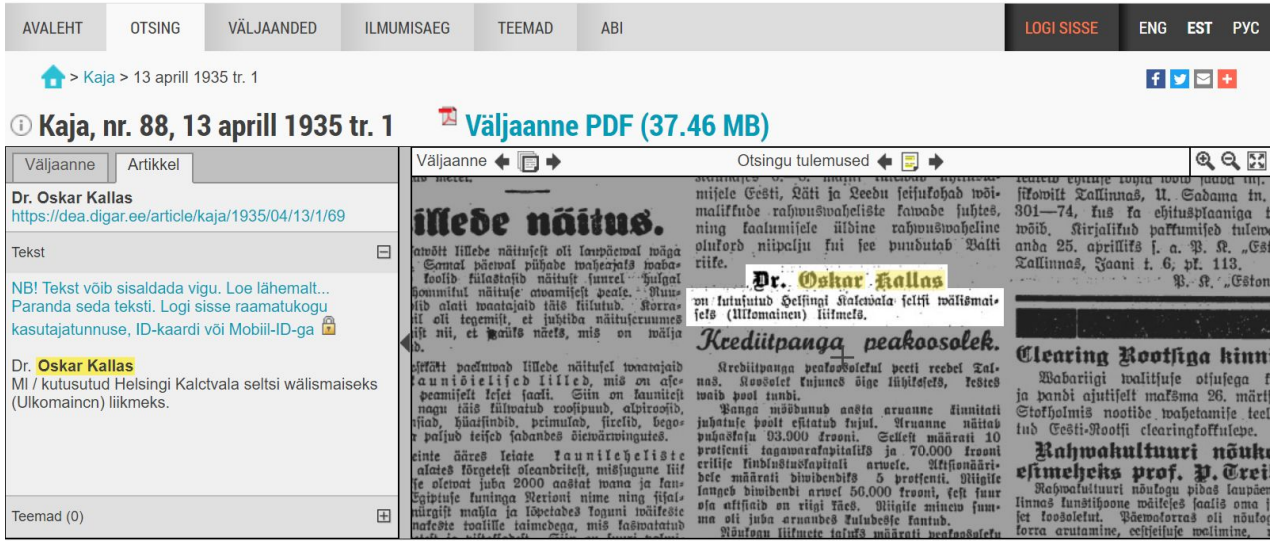
Search engine



NATIONAL LIBRARY OF ESTONIA

https://dea.digar.ee

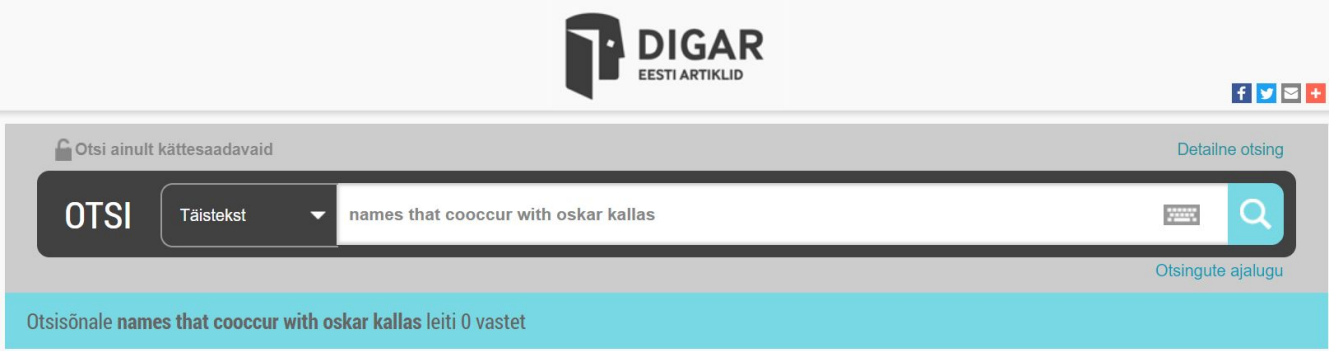
Search engine



NATIONAL LIBRARY OF ESTONIA

https://dea.digar.ee

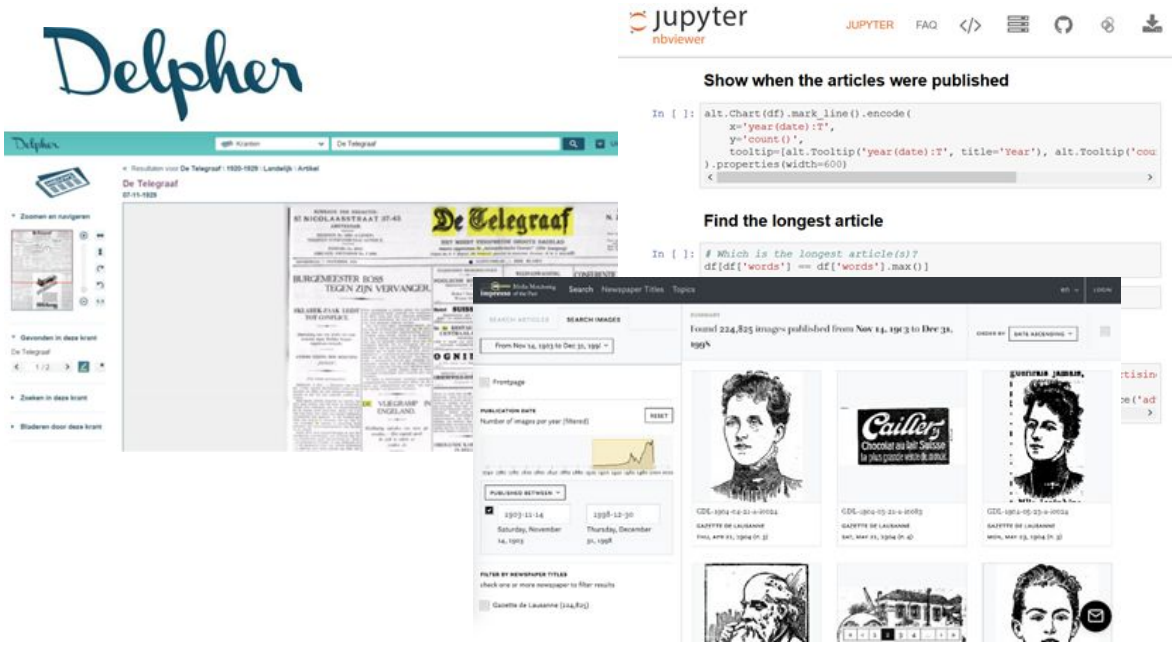
More complex queries?



NATIONAL LIBRARY OF ESTONIA

https://dea.digar.ee

Interfaces & libraries



NATIONAL LIBRARY OF ESTONIA

Open Science movement

FAIR data

- Not just open, but findable+usable

Open Science Movement

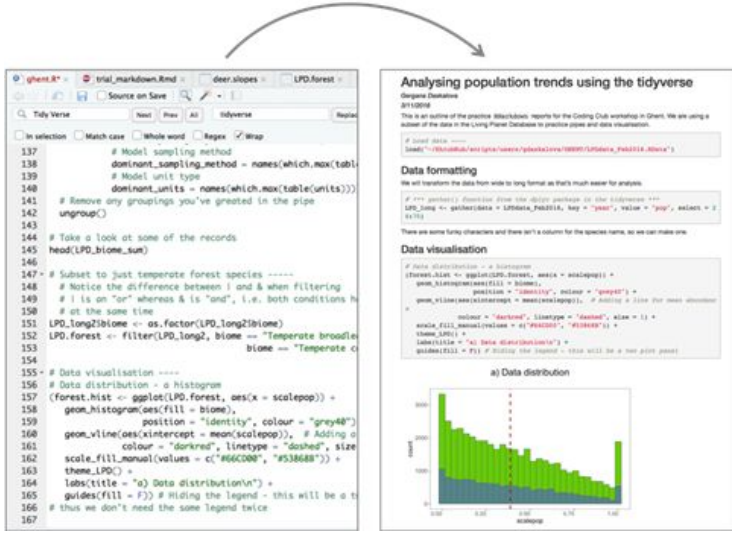
- FAIR in science

Make analyses transparent, interoperable, reusable

CENTER FOR OPEN SCIENCE

NATIONAL LIBRARY OF ESTONIA

Open Science practice



Data and code availability

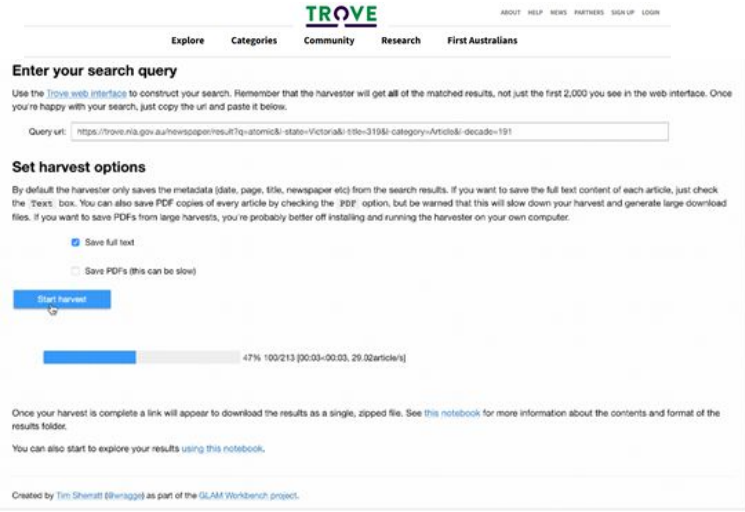
Data and code to reproduce the analysis and figures are available at https://osf.io/6ysda/

(Heunis 2020)

NATIONAL LIBRARY OF ESTONIA

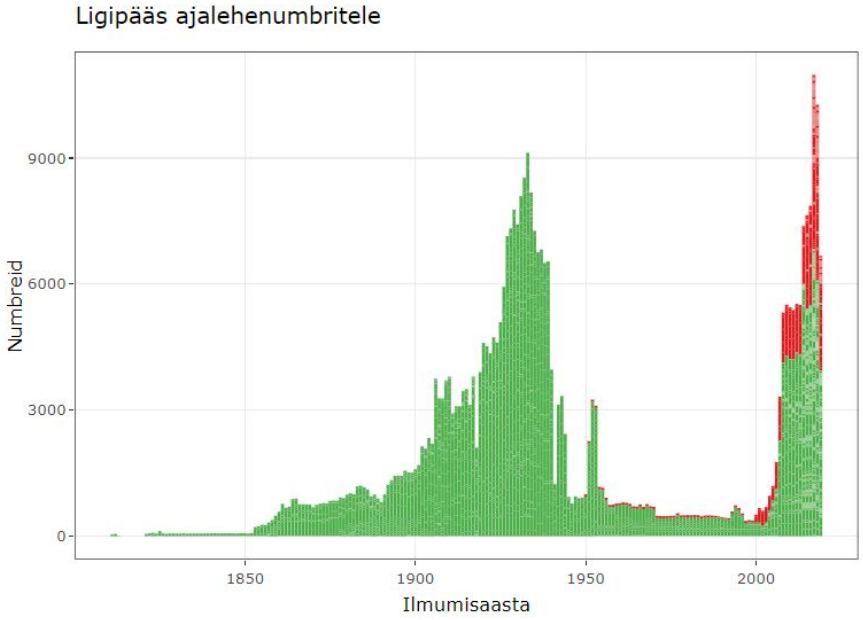
GLAMs with data

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



NATIONAL LIBRARY OF ESTONIA

Open materials at NLE



NATIONAL LIBRARY OF ESTONIA

https://data.digar.ee

Interactive overviews

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

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

NATIONAL LIBRARY OF ESTONIA

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)

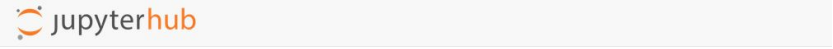
write_tsv(subset_meta,"subset_meta_postimeesew1.tsv")
```

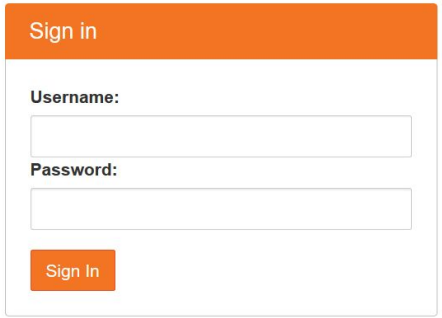
NATIONAL LIBRARY OF ESTONIA

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 meie otsinguga seotud metainfo.

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

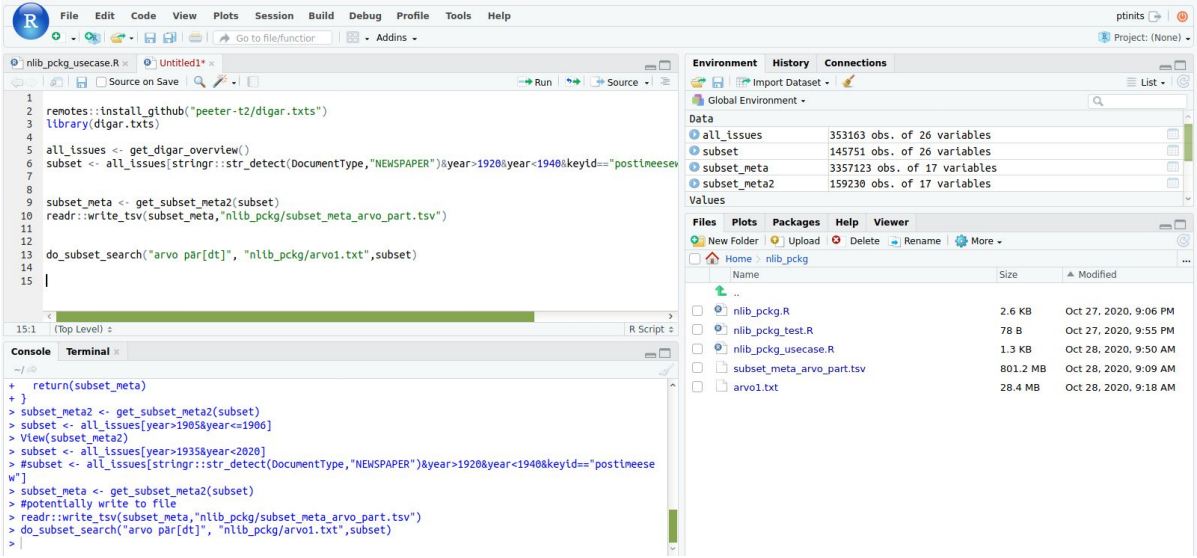
subset_meta <- rbindlist(lapply(paste0("unzip -p ",metafilelist),fread,fill=T),idcol=T)

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

NATIONAL LIBRARY OF ESTONIA

Access points

RStudio, Jupyter



NATIONAL LIBRARY OF ESTONIA