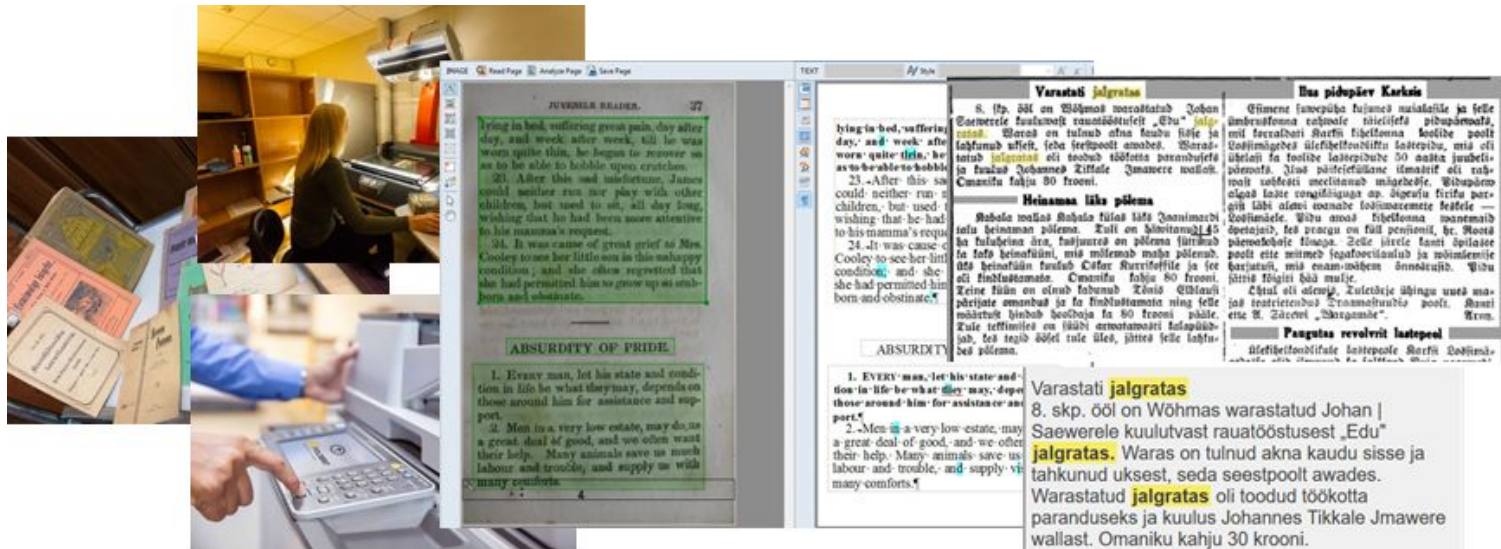

Using digital texts at the National Library of Estonia

Peeter Tinitis, Oct 28, 2020
Nordplus workshop



Large digital collections



Texts as data

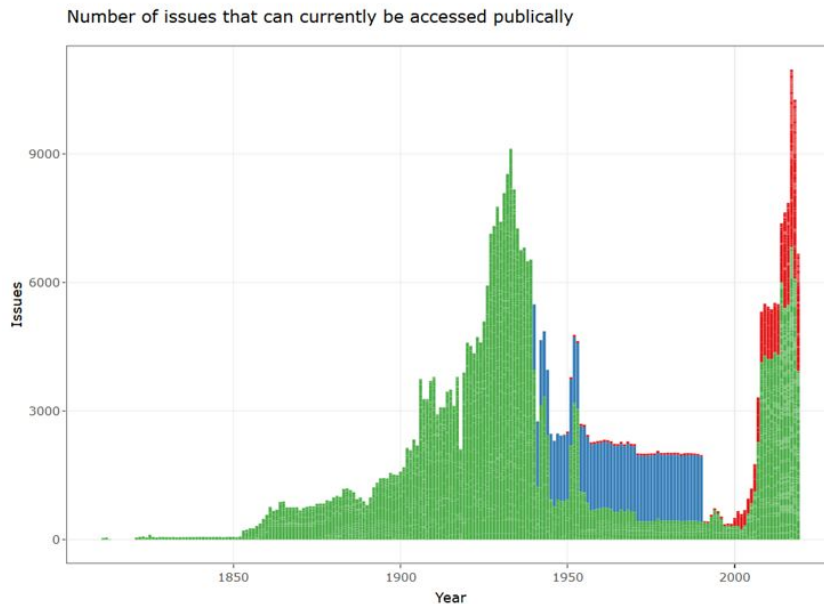
Newspapers
& periodicals

~6.3 M articles

~3.6 M pages

~390 K issues

~2200 publications



Researcher needs

Texts




Metadata







Tools



Search engine

 DIGAR
EESTI ARTIKLID




Otsi ainult kättesaadavaid

Detailne otsing

OTSI

Täistekst

"oskar kallas"



Otsingute ajalugu

relevantsus

Otsisõnale "oskar kallas" leiti 1,376 vastet

< 1 2 3 4 5 ... >

PIIRA OTSINGUT

Väljaande tüüp

[Ajakirjad](#) (54)

[Ajalehed](#) (1,269)

[Jätkväljaanded](#) (53)

Väljaanne

[Postimees \(1886-1944\)](#) (137)

[Kaja](#) (92)

[Päewaleht](#) (65)

Lisa valik minu loendisse | Lisa kõik minu loendisse

1. Oskar Kallas

☐ Päewaleht 10 august 1925

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

Oskar Kallas

Lisa loendisse

Lisa teema

2. Oskar Kallas.

☐ Maa Hääl : maarahva ajaleht 14 juuli 1934

Lisa loendisse

Lisa teema

Search engine

AVALEHTOTSINGVÄLJAANDEDILMUMISAEGTEEMADABI

LOGI SISSEENGESTPYC

> Kaja > 13 aprill 1935 tr. 1

Väljaanne

Artikkel

Dr. Oskar Kallas

<https://dea.digar.ee/article/kaja/1935/04/13/1/69>

Tekst

NB! Tekst võib sisaldada vigu. Loe lähemalt...
Paranda seda teksti. Logi sisse raamatukogu kasutajatunnuse, ID-kaardi või Mobiil-ID-ga

Dr. Oskar Kallas

MI / kutsusutud Helsingi Kalctvala seltsi välismaiseks (Ulkomainen) liikmeks.

Teemad (0)

Väljaanne

Otsingu tulemused

Illede näitus.

amäti illede näituselt oli laupäeval möga...
Samaal püüdnud pühade vahetaval...
...koosil külastasid näitusel...
...nõuad näitusel...
...id olid näitusel...
...il oli tegemisi, et juhtida näitusruumes...
...ist nii, et näitus näeks, mis on mõlgi...

Dr. Oskar Kallas

on kutusutud Helsingi Kalevala seltsi välismaiseks (Ulkomainen) liikmeks.

Krediidipanga peakoosolek.

Krediidipanga peakoosolekul peeti reedel Tallinas...
...Krediidipanga peakoosolekul...
...Krediidipanga peakoosolekul...
...Krediidipanga peakoosolekul...

Clearing Rootsiiga hinn

Wabariigi walituse otsusega...
...ja pandi ajutiselt maksma 26. märtsi...
...Stofholmis nootide vahetamise leel...
...lud Eesti-Rootsi clearingfoffulepe.

Rahwahakultuuri nõu

Reformakultuuri nõu...
...Reformakultuuri nõu...
...Reformakultuuri nõu...

More complex queries?



The screenshot shows the DIGAR (Eesti Artiklid) search interface. At the top, the logo "DIGAR EESTI ARTIKLID" is displayed. 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 "OTSI" with "Täistekst" selected. 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".

Otsi ainult kättesaadavaid

OTSI Täistekst names that cooccur with oskar kallas

Detailne otsing

Otsingute ajalugu

Otsisõnale names that cooccur with oskar kallas leiti 0 vastet

Interfaces & libraries

Delpher

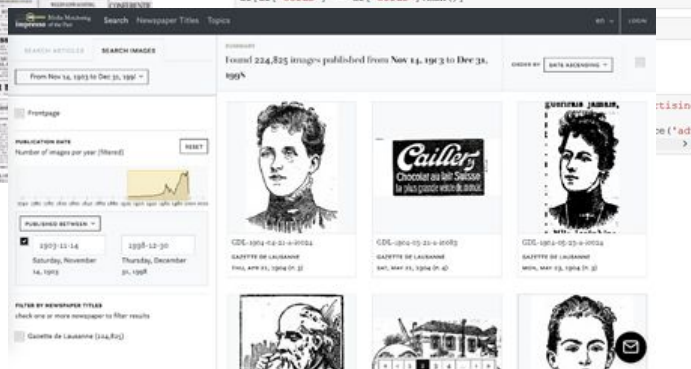


Show when the articles were published

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

Find the longest article

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



Open Science movement

FAIR data


- Not just open, but findable+usable

Open Science Movement

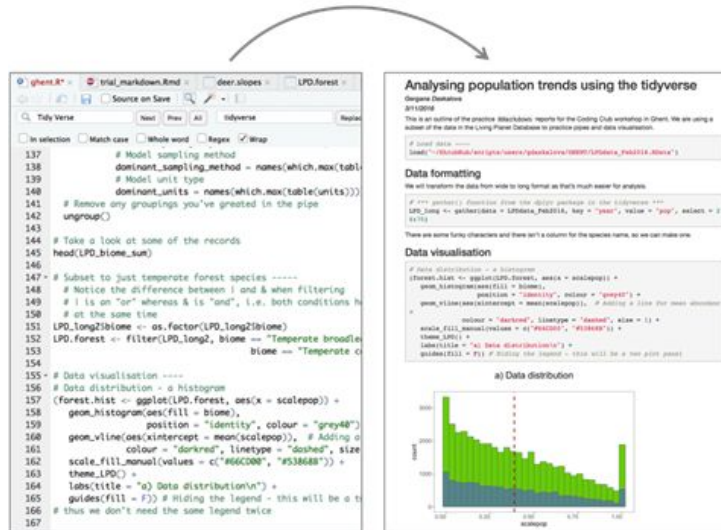
- FAIR in science

Make analyses transparent,
interoperable, reusable

Findable 
Accessible 
Interoperable 
Reusable 


— CENTER FOR —
OPEN SCIENCE

Open Science practice



Data and code availability

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

DATA AND CODE
AVAILABLE ON REQUEST

SHARES CODE

SHARES
DATA AND CODE

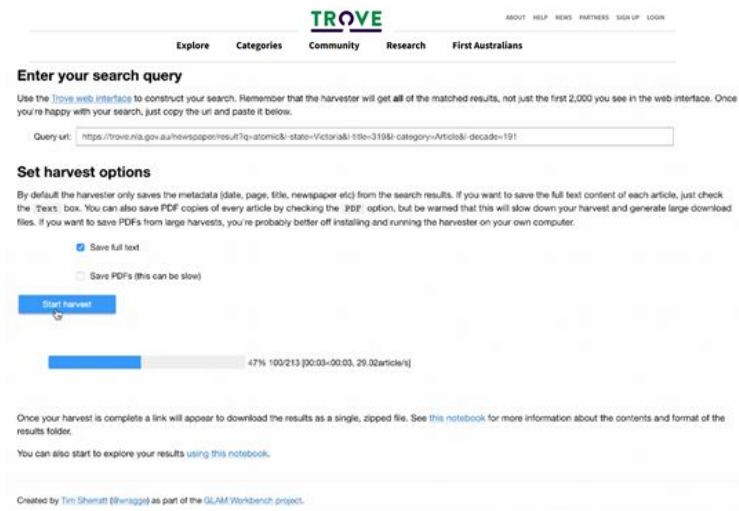
SHARES REPRODUCIBLE
ANALYSIS ENVIRONMENT

SHARES OPEN & INTERACTIVE
APPLICATION TO EXPLORE DATA



GLAMs with data

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



The screenshot shows the TROVE website with a search bar and navigation links. Below the search bar, there's a section for "Enter your search query" with a text input field containing a URL. Below that, there's a section for "Set harvest options" with checkboxes for "Save full text" and "Save PDF's (this can be slow)". A "Start harvest" button is visible. A progress bar shows 47% completion. At the bottom, there's a note about downloading results and a link to a notebook.

TROVE

Explore Categories Community Research First Australians

ABOUT HELP NEWS PARTNERS SIGN UP LOGIN

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 PDF's (this can be slow)

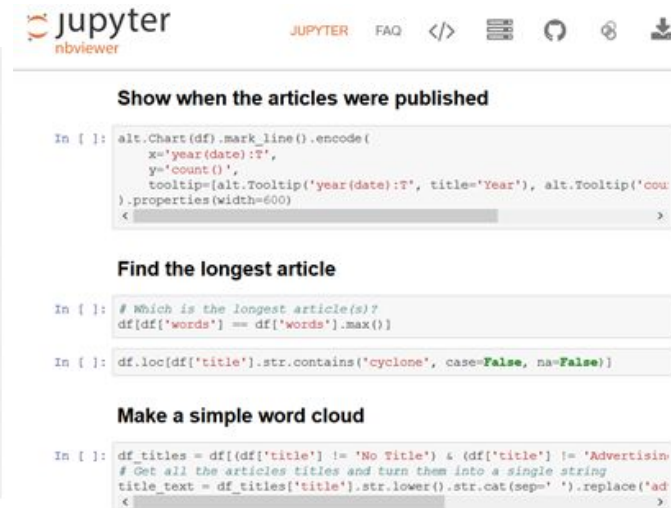
[Start harvest](#)

47% 100/213 [00:03-00:03, 26.02article/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 Sherratt \(@sherratt\)](#) as part of the [GLAMworkbench](#) project.



The screenshot shows a Jupyter Notebook interface with three code cells. The first cell is titled "Show when the articles were published" and contains a line plot. The second cell is titled "Find the longest article" and contains a line of code to find the longest article. The third cell is titled "Make a simple word cloud" and contains a line of code to create a word cloud.

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

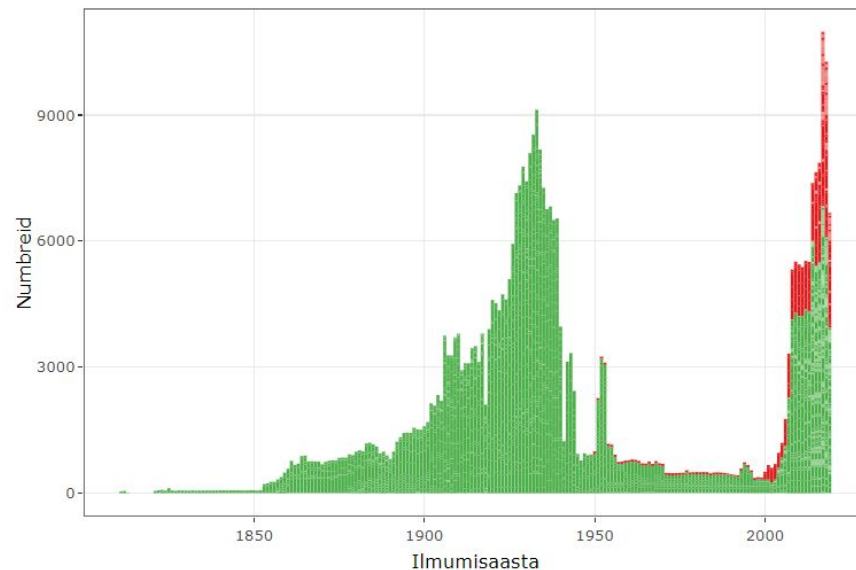
```
In [ ]: # Which is the longest article(s)?
df[df['words'] == df['words'].max()]
```

Make a simple word cloud

```
In [ ]: df_titles = df[(df['title'] != 'No Title') & (df['title'] != 'Advertisement')]
# 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

Ligipääs ajalehenumbritele



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. Kollektsooni 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 /gpf/hpc/projects/digar_txt/text/all_issues_access.zip", sep="\t")[access_now==T]

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 <- "/gpf/hpc/projects/digar_txt/text"
filelist <- paste0(collectionname, "/text_sections/", files)

```
```

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

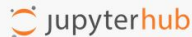
```
```{r}
metafiles <- subset[zippath_sections!="", unique(zippath_sections_meta)]
metafilelist <- paste0(collectionname, "/meta_sections/", metafiles)

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

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

Open data

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



Sign in

Username:

Password:

Sign In

```
## Andmekogu

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

'''(r)
# Loe sisse metaandmete faili hpc serverilt.
all_issues <- fread("unzip -p /gpps/hpc/projects/digar_txt/text/all_issues_access.zip", sep="\t")[access_now==T]

# Valime AJALEHD, 1928 ja 1940 vahel, kus on väljaande koodiks postineesew
subset <- all_issues[str_detect(DocumentType, "NEWSPAPER")&year=1928&year=1940&keyId="postineesew"]

# Meile vajalikke failide nimekirj
files <- subset(zippath_sections!="", unique(zippath_sections))
collectionname <- "/gpps/hpc/projects/digar_txt/text"
filelist <- paste0(collectionname, "/text_sections/", files)

...

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

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

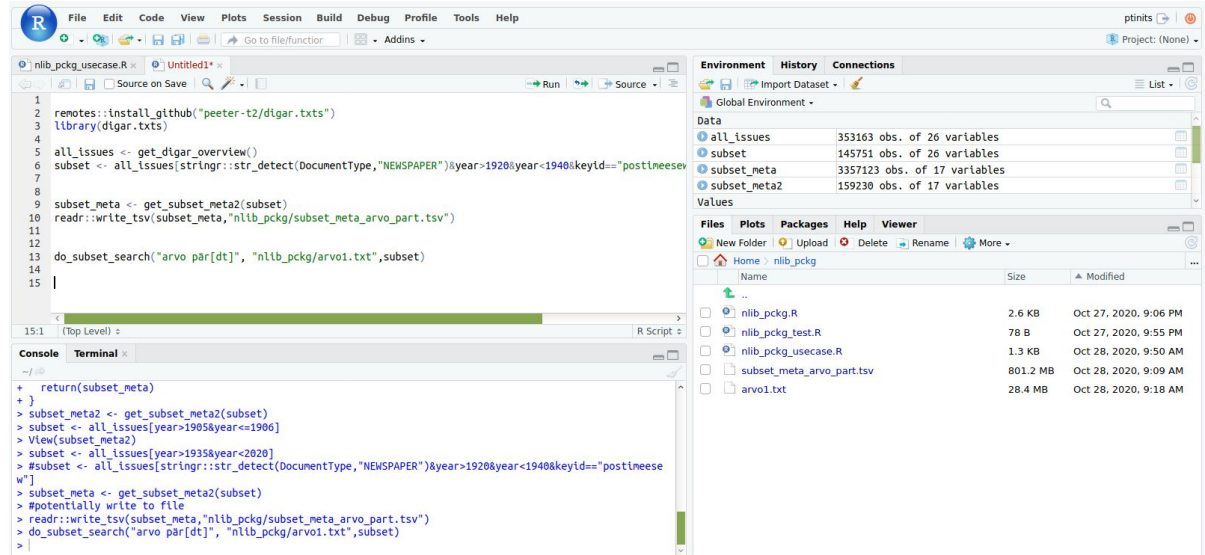
subset_meta <- rbindlist(lapply(paste0("unzip -p ", metafilelist), fread, fill=T), idcol=T)
write_tsv(subset_meta, "subset_meta_postineesew1.tsv")

...

```

Access points

RStudio, Jupyter



The screenshot displays the RStudio IDE interface. The main editor window shows an R script with the following code:

```
1 renotes::install_github("peeter-t2/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=="post/neesee
6
7
8 subset_meta <- get_subset_meta2(subset)
9 readr::write_tsv(subset_meta,"nlib_pckg/subset_meta_arvo_part.tsv")
10
11 do_subset_search("arvo par[dt]", "nlib_pckg/arvo1.txt",subset)
12
13
14
15
```

The Environment pane on the right shows the following objects:

| 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 | |

The Files pane on the right shows the following files:

| File | 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 |

The Console pane at the bottom shows the following 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=="post/neesee
w")
> 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)
>
```