Mining Texts at the National Library of Estonia

Potential for Learning and Discovery

Peeter Tinits

12.03.2020

Tallinn, Estonia

Digitizing texts

Large digitization programmes:

Estonian written word to the world (2014-2018)
Accomplished: 9% of books, 26% of periodicals

Action Plan for Digitization of Cultural heritage (2018-2023)
Goal: 28% of printed publications (+700,000 newspaper pages, + 800,000 book pages)

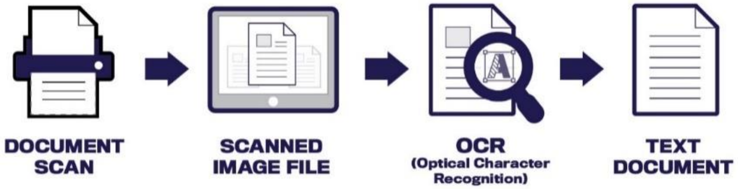
Legal Deposit Copy Act – since 2017 digital copies of estonian Publications

Current state: 13% of books, 36% of periodicals in the registry

Digital print archives (DIGAR):
61,042 prints

Digitized periodicals (DEA.DIGAR):
390,186 issues,
3,725,232 pages
6,441,167 articles
2285 titles

Digitization

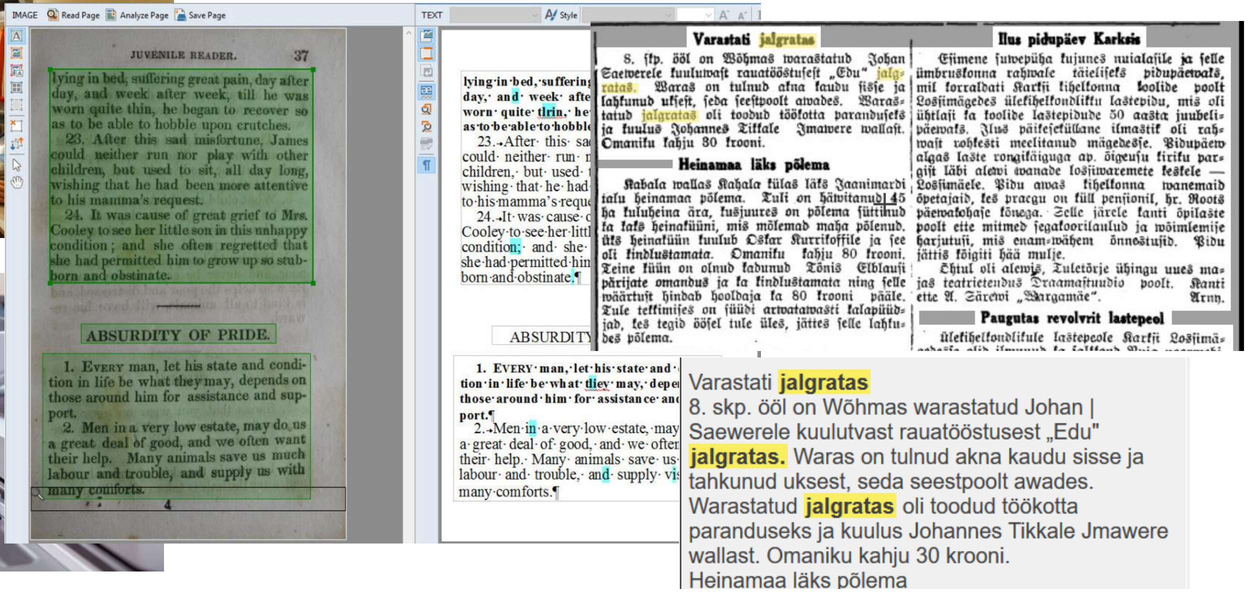


DOCUMENT SCAN

SCANNED IMAGE FILE

OCR (Optical Character Recognition)

TEXT DOCUMENT



JUVENILE READER. 37

lying in bed, suffering great pain, day after day, and week after week, till he was worn quite thin, he began to recover so as to be able to hobble upon crutches.

23. After this sad misfortune, James could neither run nor play with other children, but used to sit, all day long, wishing that he had been more attentive to his mamma's request.

24. It was cause of great grief to Mrs. Cooley to see her little son in this unhappy condition; and she often regretted that she had permitted him to grow up so stubborn and obstinate.

ABSURDITY OF PRIDE.

1. EVERY man, let his state and condition in life be what they may, depends on those around him for assistance and support.

2. Men in a very low estate, may do us a great deal of good, and we often want their help. Many animals save us much labour and trouble, and supply us with many comforts.

4

lying·in·bed,·suffering
day,· and· week· afte
worn· quite· tlrin,· he
as·to·be·able·to·hobble
23.→After· this· sa
could· neither· run· n
children,· but· used·
wishing· that· he· had
to·his·mamma's·reque
24.→It· was· cause· c
Cooley·to·see·her·littl
condition;· and· she·
she·had·permitted·hin
born·and·obstinate.¶

ABSURDITY

1. EVERY·man,·let·his·state·and·
tion·in·life·be·what·tliey·may,·deper
those·around·him·for·assistance·and
port.¶
2.→Men·in·a·very·low·estate,·may
a·great·deal·of·good,·and·we·often
their·help.·Many·animals·save·us·
labour·and·trouble,·and·supply·vi
many·comforts.¶

Varastati jalgratas

8. ſkp. ööl on Wõhmas warastatud Johan Saewerele kuuluwaſt rauatööstuſeſt „Edu" jalgratas. Waras on tulnud akna kaudu ſiſſe ja lahkunud ukſeſt, ſeda ſeeſtpoolt awades. Warastatud jalgratas oli toodud töökotta paranduſeks ja kuulus Johannes Tikkale Jmawere wallaſt. Omaniku kahju 30 krooni.

Heinamaa läks põlema

Kabala wallas Kahala külas läks Jaanimardi talu heinamaa põlema. Tuli on häwitanud 45 ha kuluheina ära, kuſjuures on põlema ſüttinud ka kaks heinaküüni, mis mõlemad maha põlenud. Üks heinaküün kuulub Oskar Kurrikoffile ja ſee oli kindlustamata. Omaniku kahju 80 krooni. Teine küün on olnud kadunud Tõnis Elblauſi pärijate omandus ja ka kindlustamata ning ſelle wäärtuſt hindab hooldaja ka 80 krooni pääle. Tule tekkimiſes on ſüüdi arwatawasti kalapüüdjad, kes tegid ööſel tule üles, jättes ſelle lahkudes põlema.

Ilus pidupäev Karksis

Esimene ſuwepüha kujunes nuialaſile ja ſelle ümbruskonna rahwale täieliſeks pidupäewaks, mil korraldati Karkſi kihelkonna koolide poolt Loſſimägedes ülekihelkondlikku lastepidu, mis oli ühtlaſi ka koolide lastepidude 50 aasta juubelipäewaks. Ilus päikeſeküllane ilmastik oli rahwaſt rohkesti meelitanud mägedesſe. Pidupäew algas laste rongikäiguga ap. õigeuſu kiriku pargiſt läbi alewi wanade loſſiwaremete keskele — Loſſimäele. Pidu awas kihelkonna wanemaid õpetajaid, kes praegu on küll penſionil, hr. Roots päewakohaſe kõnega. Selle järele kanti õpilaste poolt ette mitmed ſegakoorilaulud ja wõimlemiſe harjutuſi, mis enam-wähem õnnestuſid. Pidu jättis kõigiti hää mulje.

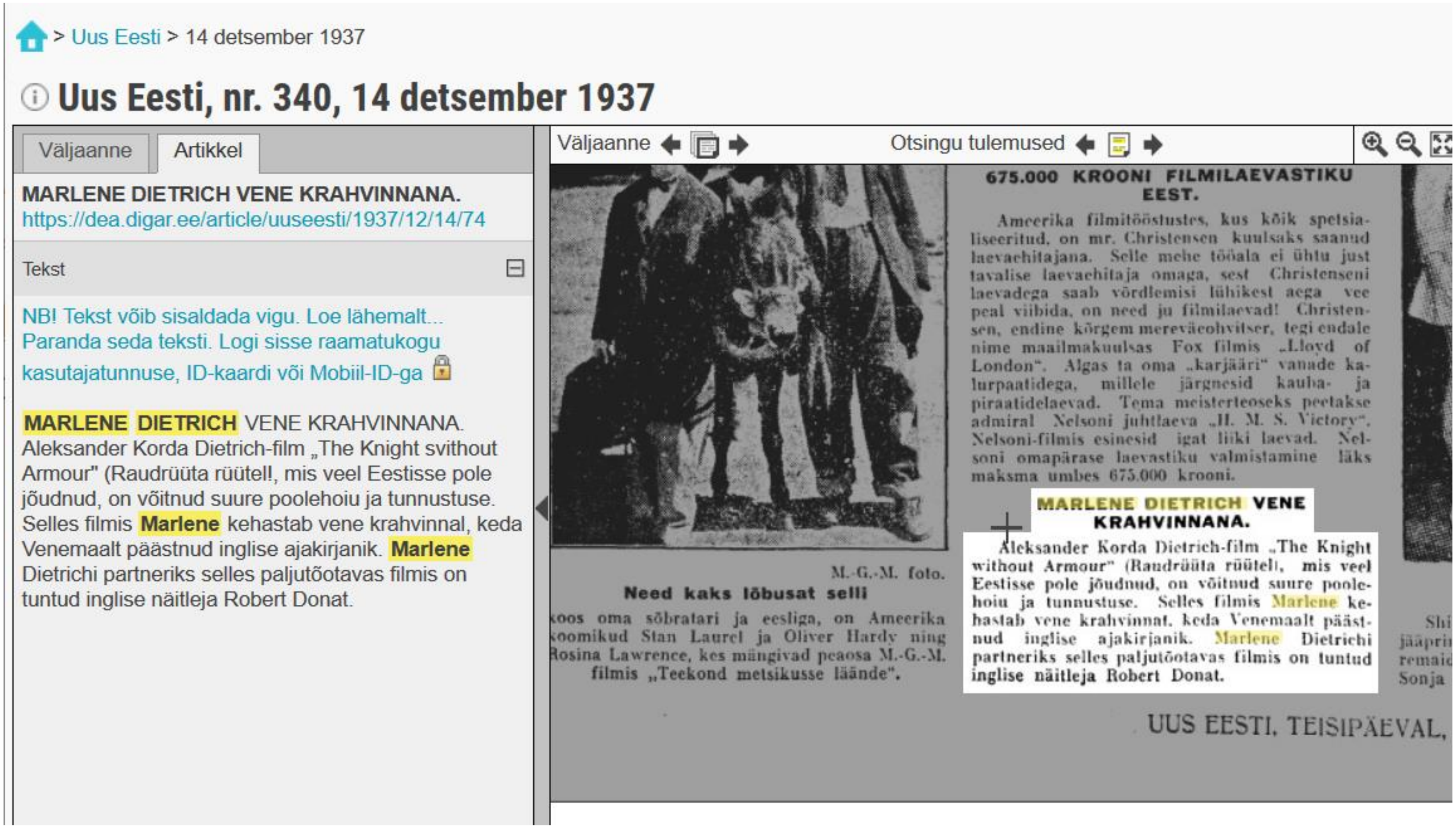
Õhtul oli alewis, Tuletõrje ühingu uues majas teatrietendus Draamaſtuudio poolt. Kanti ette A. Särewi „Wargamäe". Arny.

Paugutas revolvrit lastepeol

Ülekihelkondlikule lastepeole Karkſi Loſſimä-

Varastati **jalgratas**
8. skp. ööl on Wõhmas warastatud Johan |
Saewerele kuuluvast rauatööstusest „Edu"
jalgratas. Waras on tulnud akna kaudu sisse ja
tahkunud uksest, seda seestpoolt awades.
Warastatud **jalgratas** oli toodud töökotta
paranduseks ja kuulus Johannes Tikkale Jmawere
wallast. Omaniku kahju 30 krooni.
Heinamaa läks põlema

DIGAR: search possibilities



> Uus Eesti > 14 detsember 1937

Uus Eesti, nr. 340, 14 detsember 1937

MARLENE DIETRICH VENE KRAHVINNANA.
https://dea.digar.ee/article/uuseesti/1937/12/14/74

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

MARLENE DIETRICH VENE KRAHVINNANA. Aleksander Korda Dietrich-film „The Knight svithout Armour" (Raudrüüta rüütel!, mis veel Eestisse pole jõudnud, on võitnud suure poolehoiu ja tunnustuse. Selles filmis Marlene kehastab vene krahvinnal, keda Venemaalt päästnud inglise ajakirjanik. Marlene Dietrichi partneriks selles paljutõotavas filmis on tuntud inglise näitleja Robert Donat.

M.-G.-M. foto.

Need kaks lõbusat selli

koos oma sõbratari ja eesliga, on Ameerika koomikud Stan Laurel ja Oliver Hardy ning Rosina Lawrence, kes mängivad peaosa M.-G.-M. filmis „Teekond metsikusse läände".

675.000 KROONI FILMILAEVASTIKU EEST.

Ameerika filmitööstustes, kus kõik spetsialiseeritud, on mr. Christensen kuulsaks saanud laevaehitajana. Selle mehe tööala ei ühtu just tavalise laevaehitaja omaga, sest Christenseni laevadega saab võrdlemisi lühikest aega vee peal viibida, on need ju filmilaevad! Christensen, endine kõrgem mereväeohvitser, tegi endale nime maailmakuulsas Fox filmis „Lloyd of London". Algas ta oma „karjääri" vanade kalurpaatidega, millele järgnesid kauba- ja piraatidelaevad. Tema meisterteoseks peetakse admiral Nelsoni juhtlaeva „H. M. S. Victory". Nelsoni-filmis esinesid igat liiki laevad. Nelsoni omapärase laevastiku valmistamine läks maksma umbes 675.000 krooni.

MARLENE DIETRICH VENE KRAHVINNANA.

Aleksander Korda Dietrich-film „The Knight without Armour" (Raudrüüta rüütel), mis veel Eestisse pole jõudnud, on võitnud suure poolehoiu ja tunnustuse. Selles filmis Marlene kehastab vene krahvinnat, keda Venemaalt päästnud inglise ajakirjanik. Marlene Dietrichi partneriks selles paljutõotavas filmis on tuntud inglise näitleja Robert Donat.

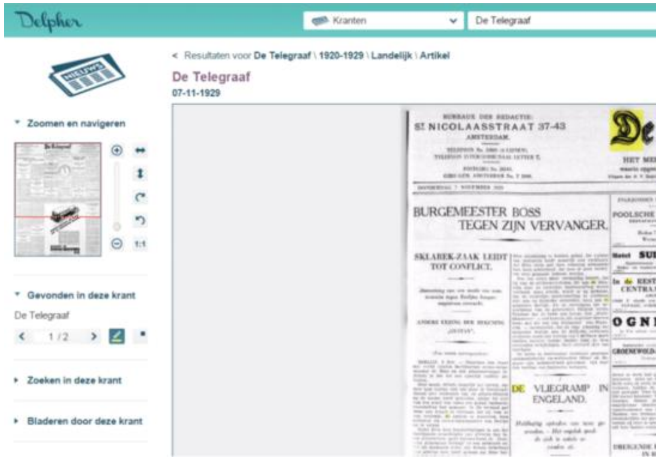
Shi
jääpri
remaid
Sonja

UUS EESTI, TEISIPÄEVAL,

Text and data mining approach

Developed by a # of institutions and researchers

Delpher

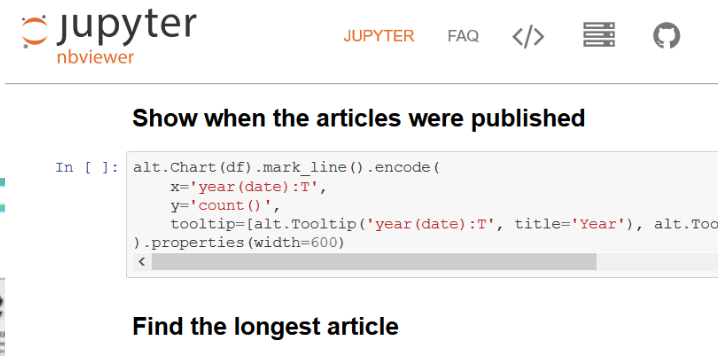


Delpher

< Resultaten voor **De Telegraaf** \ **1920-1929** \ **Landelijk** \ **Artikel**

De Telegraaf

07-11-1929



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.Too
        ).properties(width=600)
```

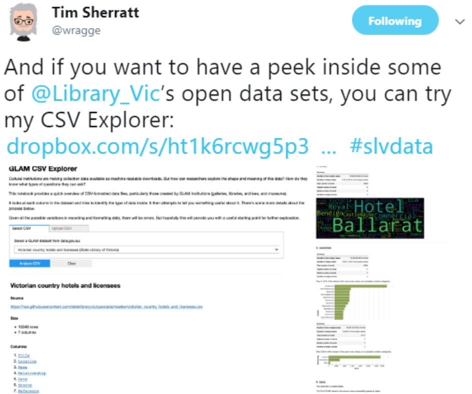
Find the longest article

```
In [ ]: # Wh
        df[d

In [ ]: df.l
```

Mal

```
In [ ]: df_t
        # Ge
        titl
```



Tim Sherratt
@wragge

And if you want to have a peek inside some of @Library_Vic's open data sets, you can try my CSV Explorer: dropbox.com/s/ht1k6rcwg5p3 … #slvdata

GLAM CSV Explorer

Cultural institutions are making collection data available as machine readable downloads. But how can researchers explore the shape and meaning of this data? How do they know what types of questions they can ask?

This notebook provides a quick overview of CSV-formatted data files, particularly those created by GLAM institutions (galleries, libraries, archives, and museums).

It looks at each column in the dataset and tries to identify the type of data inside. It then attempts to tell you something useful about it. There's some more details about the process below.

Given all the possible variations in recording and formatting data, there will be errors. But hopefully this will provide you with a useful starting point for further exploration.

Select a GLAM dataset from data.gov.au: Victorian country hotels and licensees (State Library of Victoria)

Victorian country hotels and licensees

Source

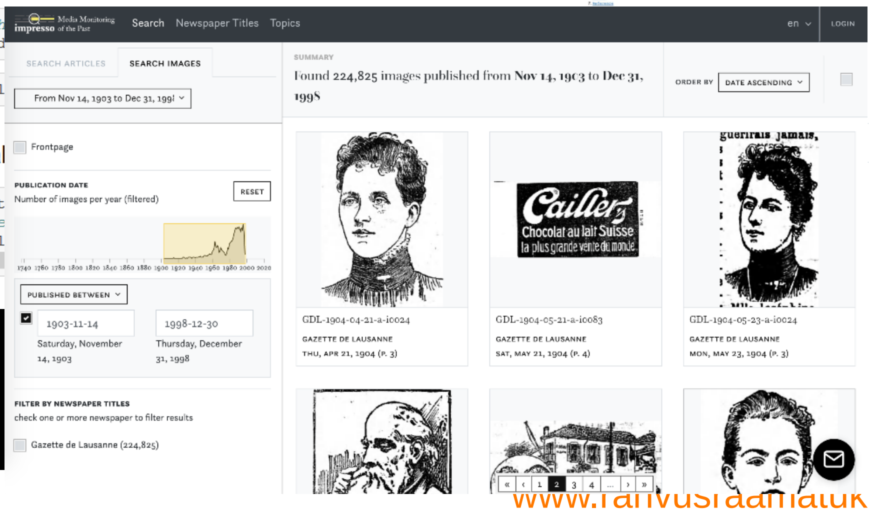
https://raw.githubusercontent.com/statelibraryvic/opendata/master/victorian_country_hotels_and_licensees.csv

Size

- 16348 rows
- 7 columns

Columns

1. Title
2. Location
3. Name
4. Relationship
5. Date
6. Source
7. Reference



From Nov 14, 1903 to Dec 31, 1998

□ Frontpage

PUBLICATION DATE

Number of images per year (filtered)

☑ 1903-11-14 Saturday, November 14, 1903 — 1998-12-30 Thursday, December 31, 1998

FILTER BY NEWSPAPER TITLES

check one or more newspaper to filter results

□ Gazette de Lausanne (224,825)

SUMMARY

Found 224,825 images published from Nov 14, 1903 to Dec 31, 1998

GDL-1904-04-21-a-i0024
GAZETTE DE LAUSANNE
THU, APR 21, 1904 (P. 3)

GDL-1904-05-21-a-i0083
GAZETTE DE LAUSANNE
SAT, MAY 21, 1904 (P. 4)

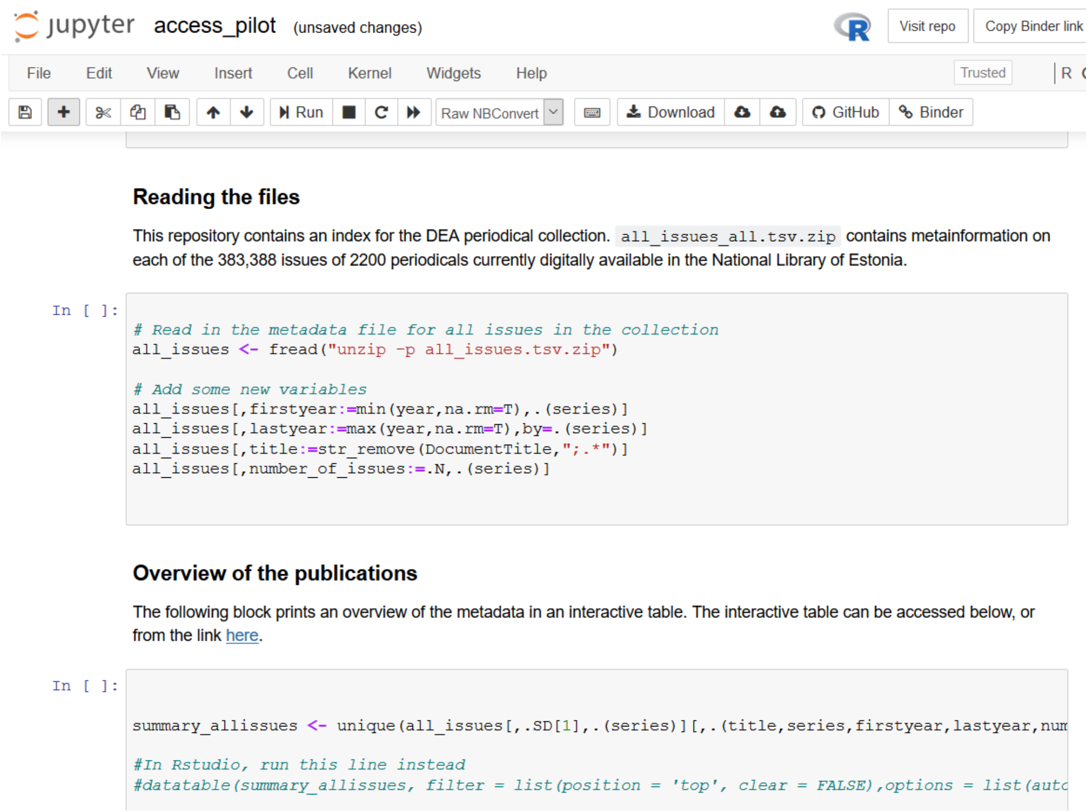
GDL-1904-05-23-a-i0024
GAZETTE DE LAUSANNE
MON, MAY 23, 1904 (P. 3)

impresso Media Monitoring of the Past

EESTI RAHVUS-RAAMATUKOGU

Text and data mining in NLE

Improving access for text and data mining



Reading the files

This repository contains an index for the DEA periodical collection. `all_issues_all.tsv.zip` contains metainformation on each of the 383,388 issues of 2200 periodicals currently digitally available in the National Library of Estonia.

```
# Read in the metadata file for all issues in the collection
all_issues <- fread("unzip -p all_issues.tsv.zip")

# Add some new variables
all_issues[,firstyear:=min(year,na.rm=T),.(series)]
all_issues[,lastyear:=max(year,na.rm=T),by=.(series)]
all_issues[,title:=str_remove(DocumentTitle,";.*")]
all_issues[,number_of_issues:=.N,.(series)]
```

Overview of the publications

The following block prints an overview of the metadata in an interactive table. The interactive table can be accessed below, or from the link here.

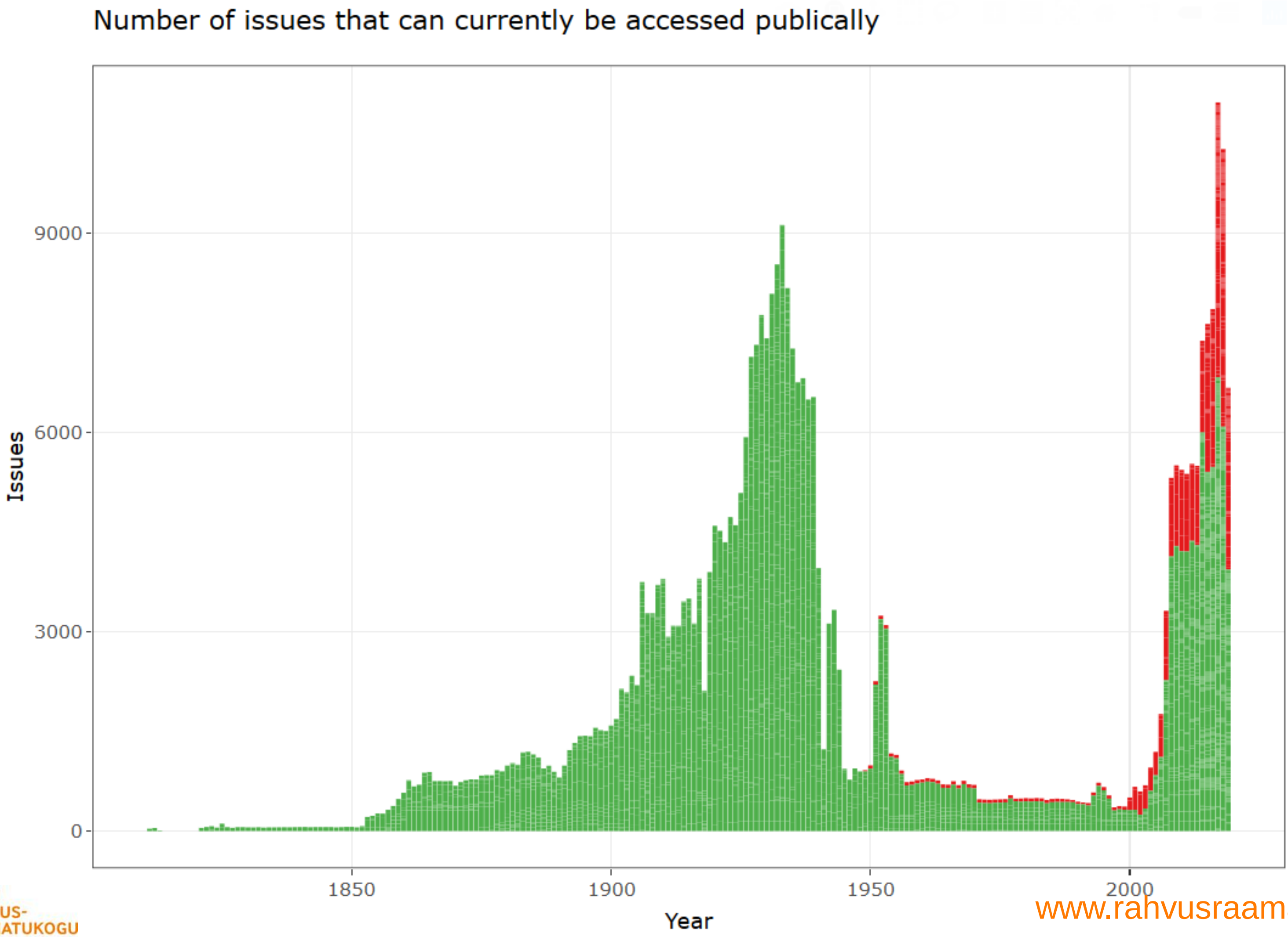
```
summary_allissues <- unique(all_issues[,.SD[1],.(series)][,.(title,series,firstyear,lastyear,num

#In Rstudio, run this line instead
#datatable(summary_allissues, filter = list(position = 'top', clear = FALSE),options = list(autc
```

Pilot app for data access

Overviews

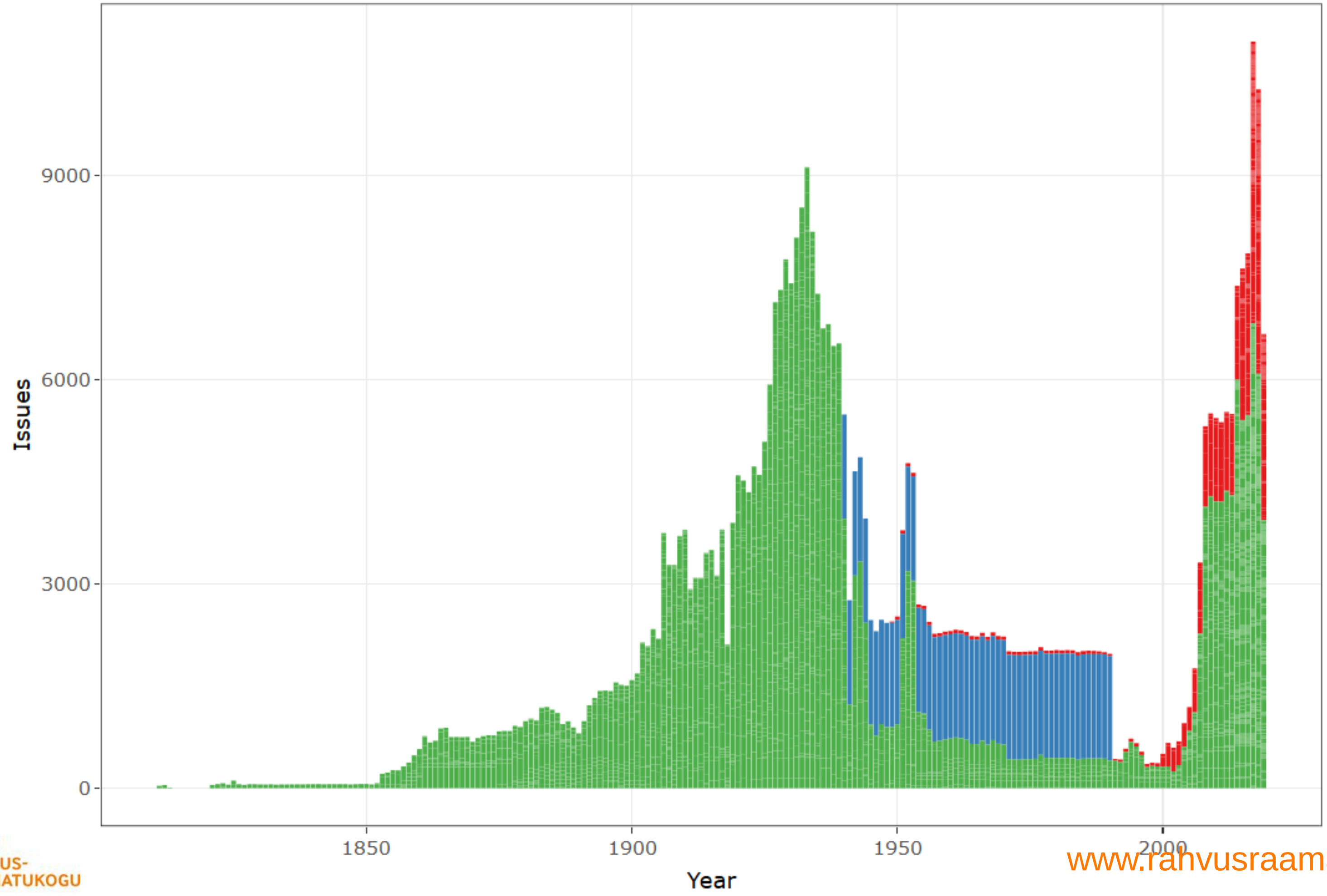
Visualization link



Overviews

Visualization link

Number of issues that can currently be accessed publically

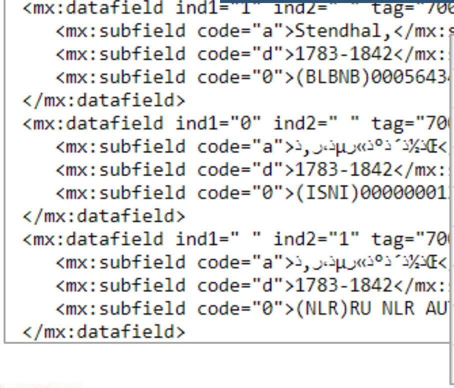


Bibliography as data

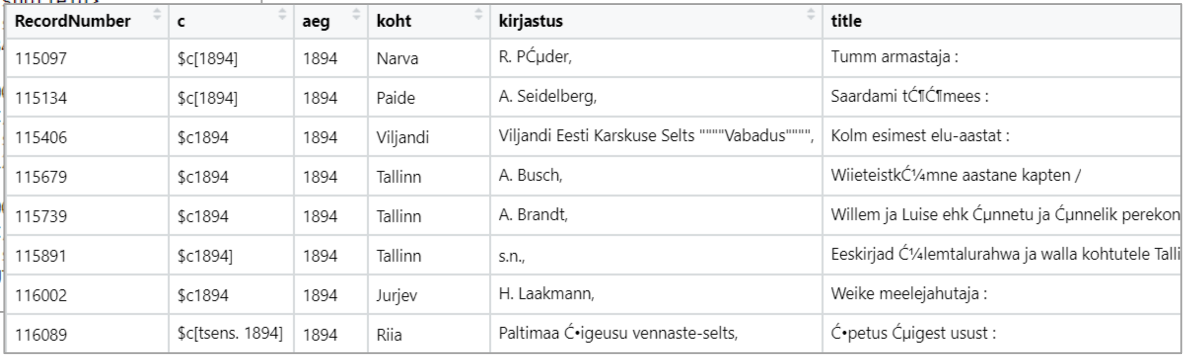
Data on 351,301 printed publications

Aims for complete coverage (quite close)

- Published in Estonia, or abroad but related to Estonia



```
<mx:datafield ind1="1" ind2=" " tag="700">
   <mx:subfield code="a">Stendhal,</mx:subfield>
   <mx:subfield code="d">1783-1842</mx:
   <mx:subfield code="0">(BLBNB)0005643
</mx:datafield>
<mx:datafield ind1="0" ind2=" " tag="70
   <mx:subfield code="a">ذ,رذµرذ«ذ°ذ´ذ½ذŒ<
   <mx:subfield code="d">1783-1842</mx:
   <mx:subfield code="0">(ISNI)00000001
</mx:datafield>
<mx:datafield ind1=" " ind2="1" tag="70
   <mx:subfield code="a">ذ,رذµرذ«ذ°ذ´ذ½ذŒ<
   <mx:subfield code="d">1783-1842</mx:
   <mx:subfield code="0">(NLR)RU NLR AU
</mx:datafield>
```



RecordNumber	c	aeg	koht	kirjastus	title
115097	$c[1894]	1894	Narva	R. PĆµder,	Tumm armastaja :
115134	$c[1894]	1894	Paide	A. Seidelberg,	Saardami tĆ¶Ć¶mees :
115406	$c1894	1894	Viljandi	Viljandi Eesti Karskuse Selts """"Vabadus"""",	Kolm esimest elu-aastat :
115679	$c1894	1894	Tallinn	A. Busch,	WiieteistkĆ¼mne aastane kapten /
115739	$c1894	1894	Tallinn	A. Brandt,	Willem ja Luise ehk Ćµnnetu ja Ćµnnelik perekon
115891	$c1894]	1894	Tallinn	s.n.,	Eeskirjad Ć¼lemtalurahwa ja walla kohtutele Talli
116002	$c1894	1894	Jurjev	H. Laakmann,	Weike meelejahutaja :
116089	$c[tsens. 1894]	1894	Riia	Paltimaa Ć•igeusu vennaste-selts,	Ć•petus Ćµigest usust :

EESTI RAHVUS-RAAMATUKOGU

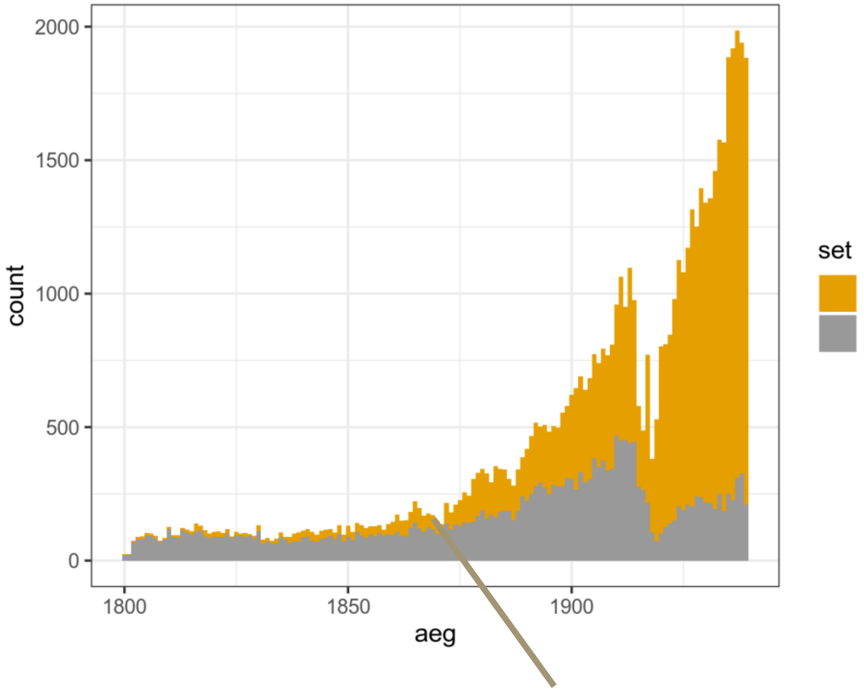
www.rahvusraamatukogu.ee

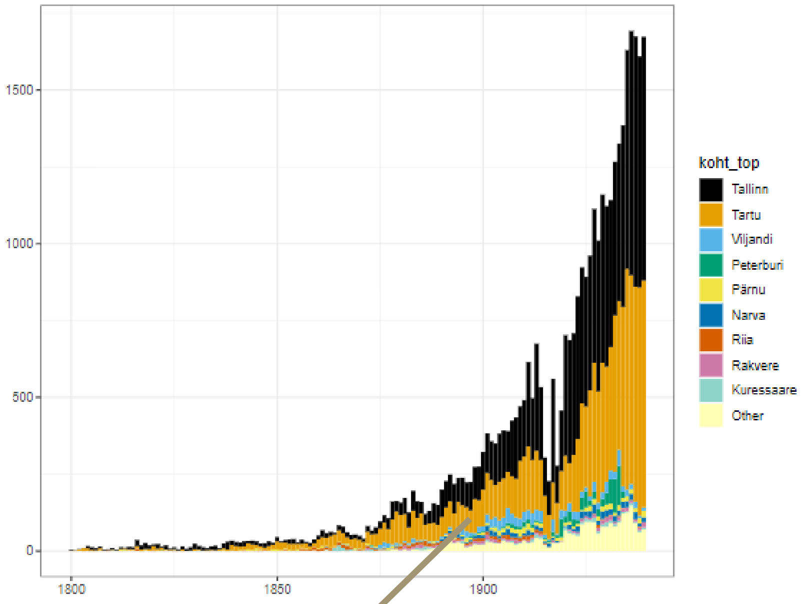
Uses

Some examples in collections as data

For learning

Number of publications by language, by place





„National Awakening“

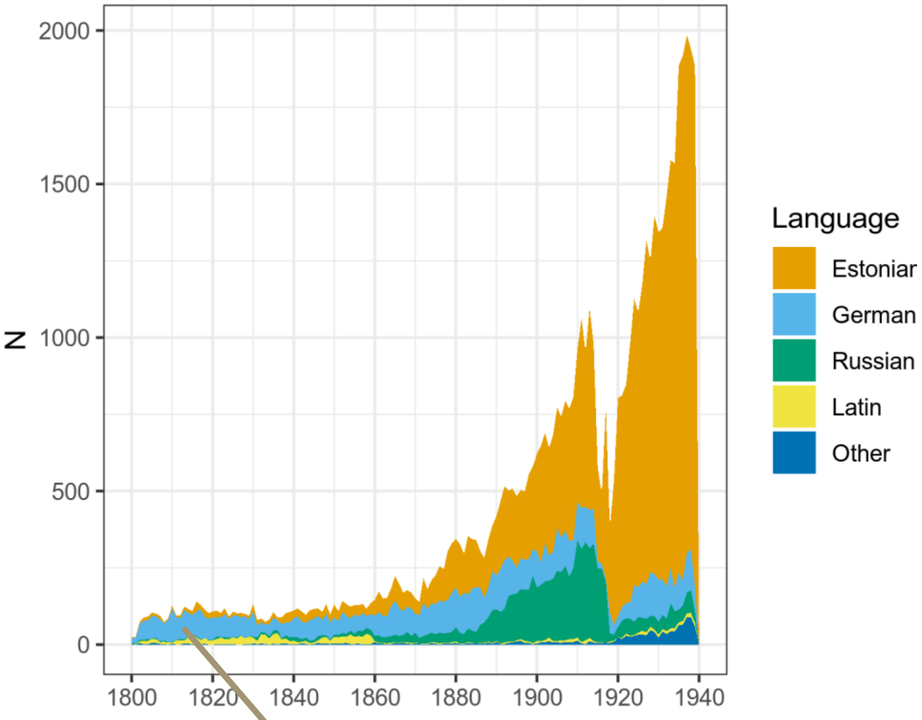
Tartu leads publishing

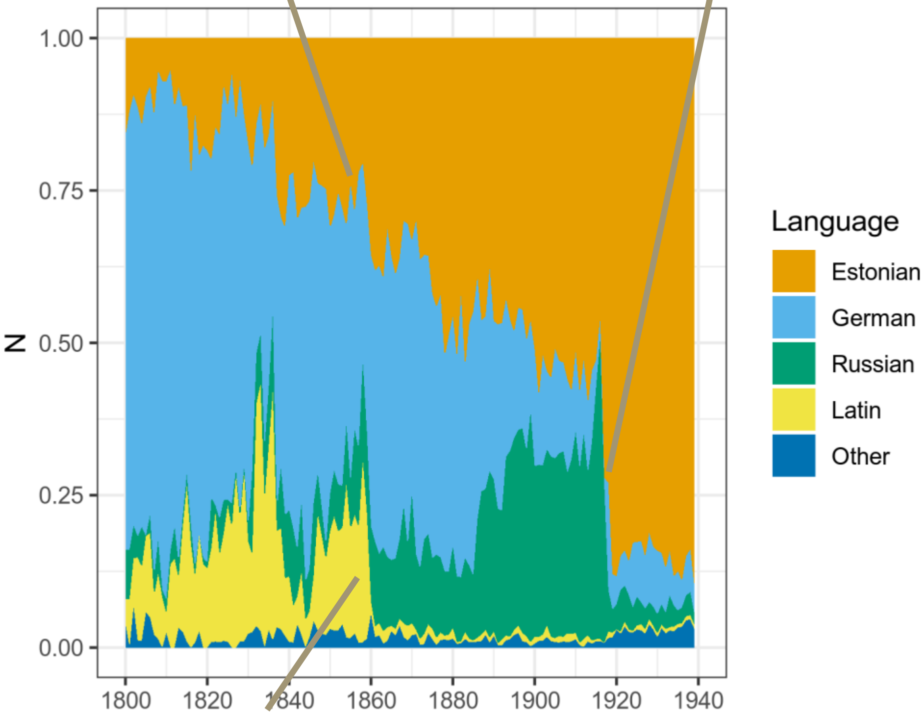
For learning

Languages of prints 1800-1940

The use of Estonian grows gradually and surely

Independent Estonia





Loss of Latin as the language of science

Baltic German Publications throughout

EESTI RAHVUS-RAAMATUKOGU

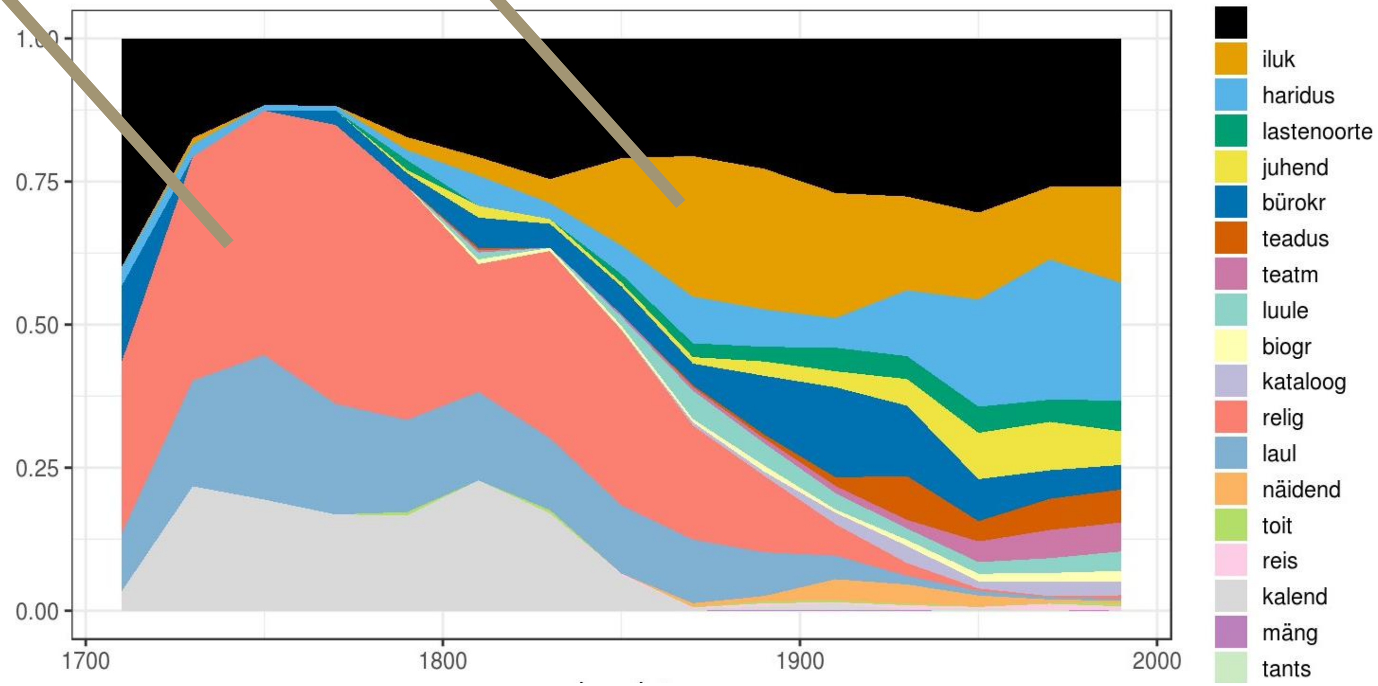
www.rahvusraamatukogu.ee

For learning

Genres of prints 1700-2000

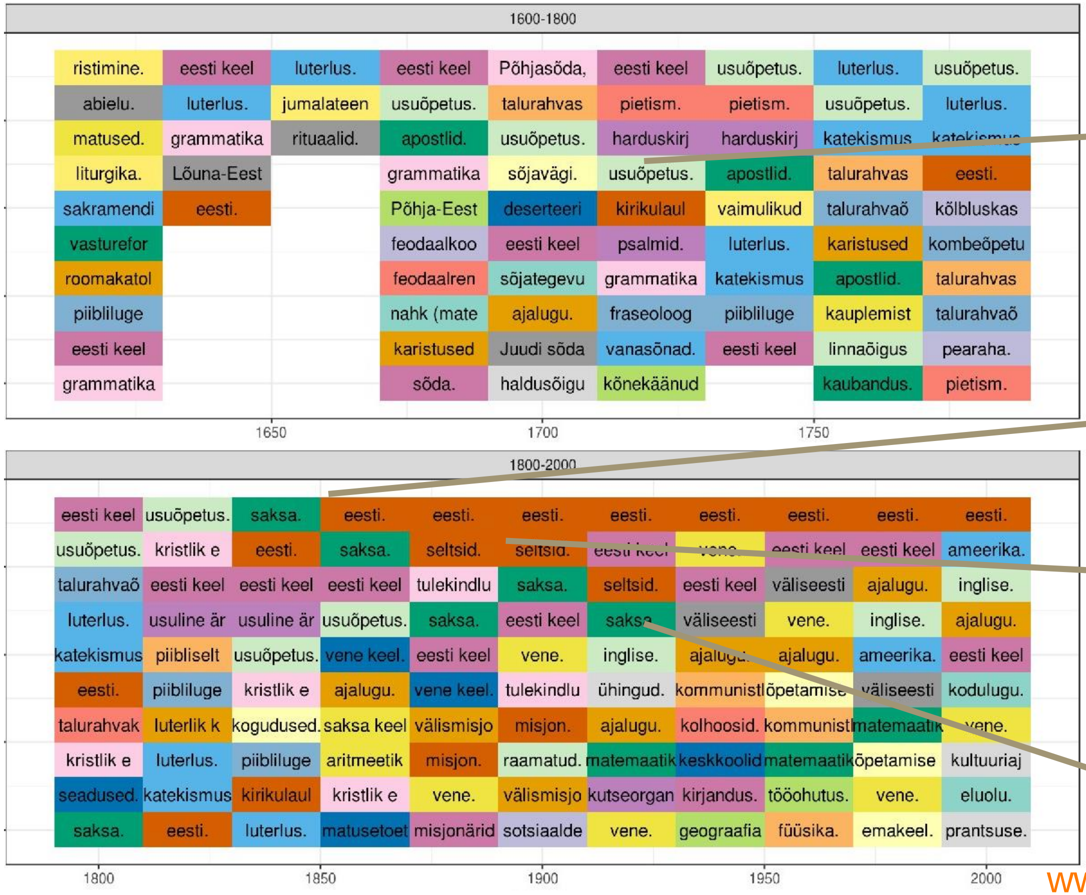
Religious texts

Fiction



For discovery

Top keywords 1600-2000 (20-year chunks)



Religious texts

Estonia(n) as keywo

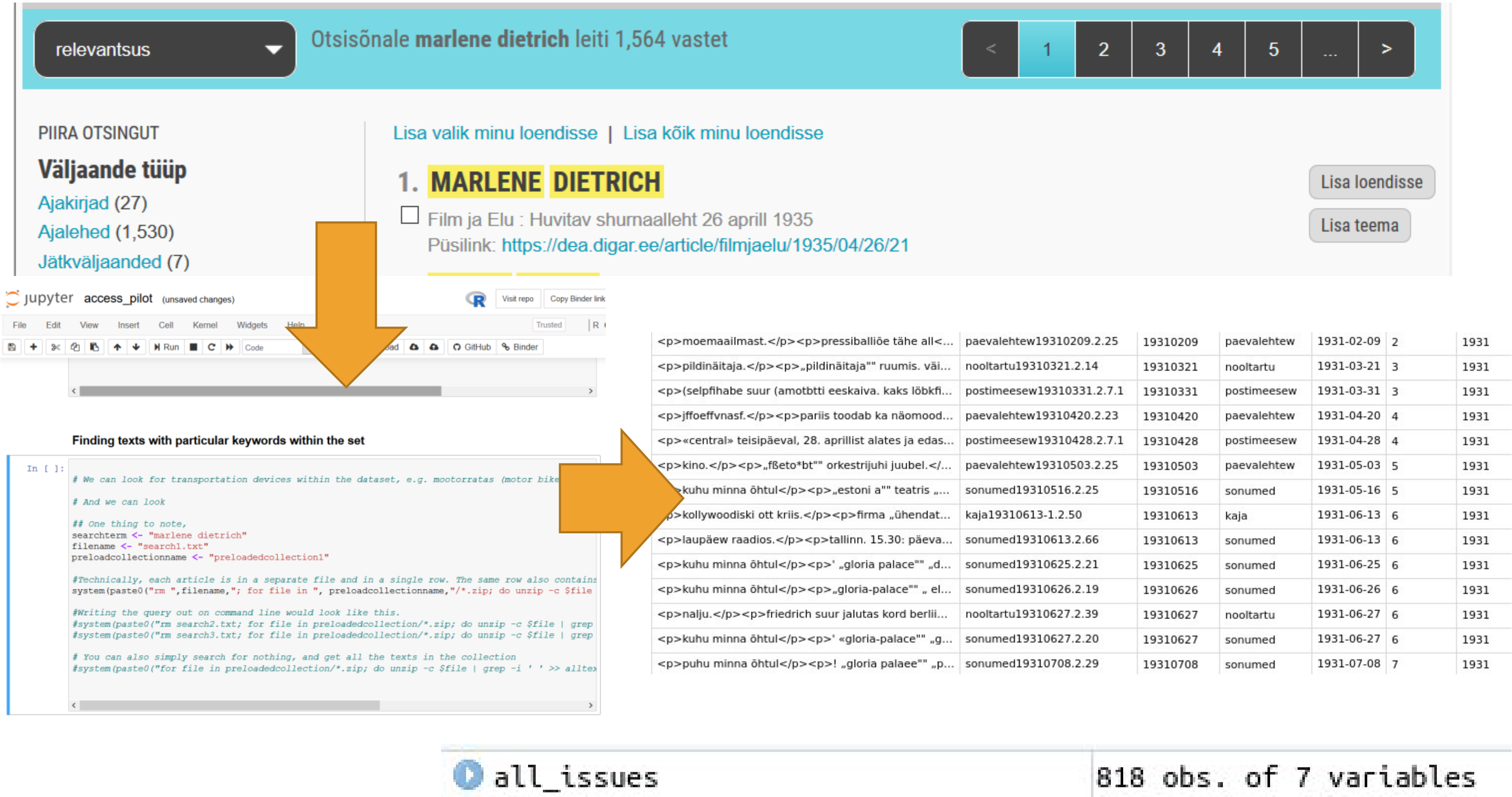
Civic societies

Disappearance of
German keyword

EESTI
RAHVUS-
RAAMATUKOGU

www.rahvusraamatukogu.ee

For discovery



1. MARLENE DIETRICH

□ Film ja Elu : Huvitav shurnaalleht 26 aprill 1935

Püsilink: https://dea.digar.ee/article/filmjaelu/1935/04/26/21

Finding texts with particular keywords within the set

```
In [ ]:
# We can look for transportation devices within the dataset, e.g. mootorratas (motor bike

# And we can look

## One thing to note,
searchterm <- "marlene dietrich"
filename <- "search1.txt"
preloadcollectionname <- "preloadedcollection1"

#Technically, each article is in a separate file and in a single row. The same row also contains
system(paste0("rm ",filename,"; for file in ", preloadcollectionname,"/*.zip; do unzip -c $file

#Writing the query out on command line would look like this.
#system(paste0("rm search2.txt; for file in preloadedcollection/*.zip; do unzip -c $file | grep
#system(paste0("rm search3.txt; for file in preloadedcollection/*.zip; do unzip -c $file | grep

# You can also simply search for nothing, and get all the texts in the collection
#system(paste0("for file in preloadedcollection/*.zip; do unzip -c $file | grep -i ' ' >> alltex
```

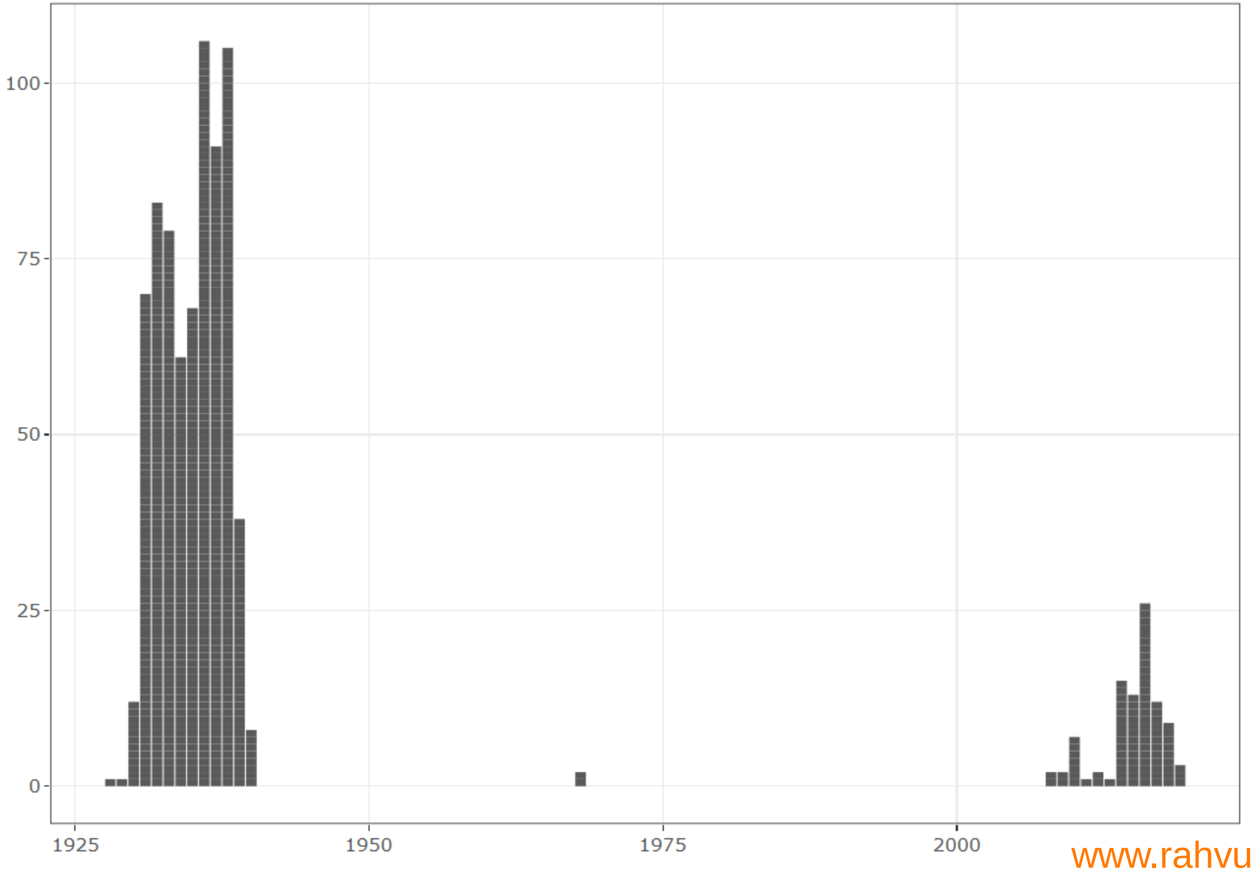
<p>moemaailmast.</p><p>pressiballiõe tähe all<...	paevalehtew19310209.2.25	19310209	paevalehtew	1931-02-09	2	1931
<p>pildinäitaja.</p><p>„pildinäitaja"" ruumis. väi...	nooltartu19310321.2.14	19310321	nooltartu	1931-03-21	3	1931
<p>(selpfihabe suur (amotbtti eeskaiva. kaks löbkfi...	postimeesew19310331.2.7.1	19310331	postimeesew	1931-03-31	3	1931
<p>jffoeffvnasf.</p><p>pariis toodab ka näomood...	paevalehtew19310420.2.23	19310420	paevalehtew	1931-04-20	4	1931
<p>«central» teisipäeval, 28. aprillist alates ja edas...	postimeesew19310428.2.7.1	19310428	postimeesew	1931-04-28	4	1931
<p>kino.</p><p>„fßeto*bt"" orkestrijuhi juubel.</...	paevalehtew19310503.2.25	19310503	paevalehtew	1931-05-03	5	1931
<p>kuhu minna õhtul</p><p>„estoni a"" teatris „...	sonumed19310516.2.25	19310516	sonumed	1931-05-16	5	1931
<p>kollywoodiski ott kriis.</p><p>firma „ühendat...	kaja19310613-1.2.50	19310613	kaja	1931-06-13	6	1931
<p>laupäew raadios.</p><p>tallinn. 15.30: päeva...	sonumed19310613.2.66	19310613	sonumed	1931-06-13	6	1931
<p>kuhu minna õhtul</p><p>' „gloria palace"" „d...	sonumed19310625.2.21	19310625	sonumed	1931-06-25	6	1931
<p>kuhu minna õhtul</p><p>„gloria-palace"" „ el...	sonumed19310626.2.19	19310626	sonumed	1931-06-26	6	1931
<p>nalju.</p><p>friedrich suur jalutas kord berlii...	nooltartu19310627.2.39	19310627	nooltartu	1931-06-27	6	1931
<p>kuhu minna õhtul</p><p>' «gloria-palace"" „g...	sonumed19310627.2.20	19310627	sonumed	1931-06-27	6	1931
<p>puhu minna õhtul</p><p>! „gloria palaee"" „p...	sonumed19310708.2.29	19310708	sonumed	1931-07-08	7	1931

```
all_issues          818 obs. of 7 variables
```

For discovery

Visualization link

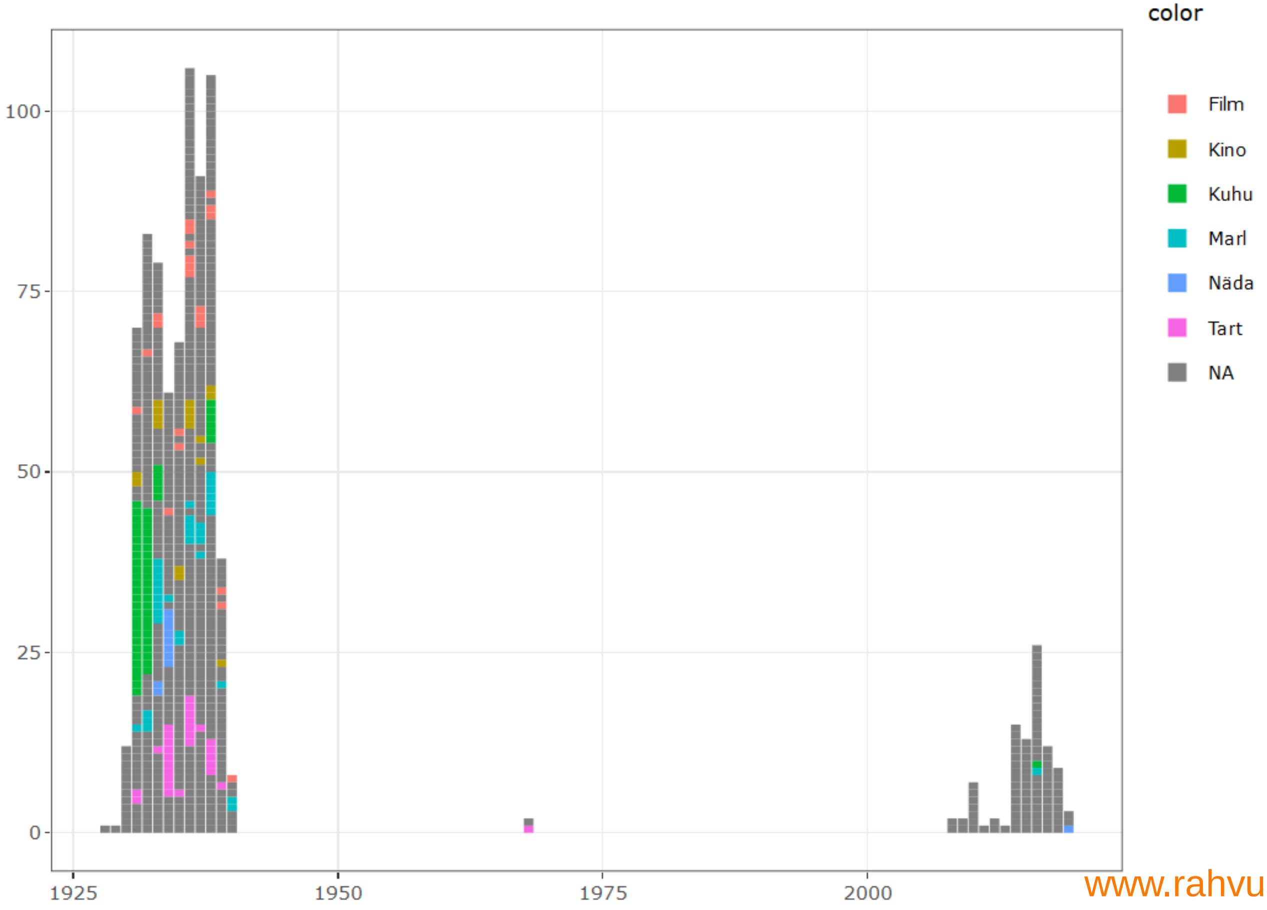
Texts featuring „marlene“ and „dietrich"



For discovery

Visualization link

Texts featuring „marlene“ and „dietrich"

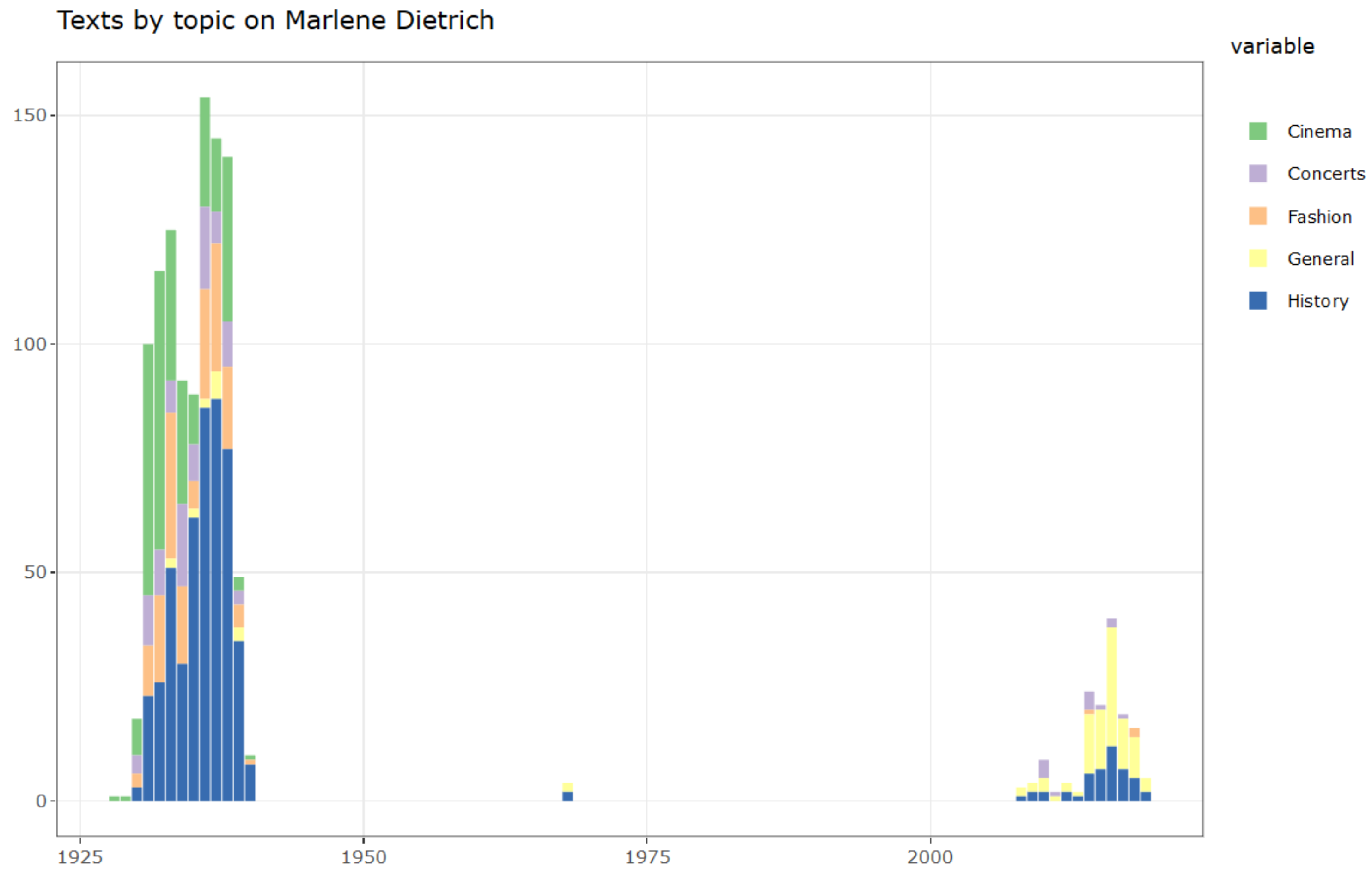


EESTI RAHVUS- RAAMATUKOGU

www.rahvusraamatukogu.ee

For discovery

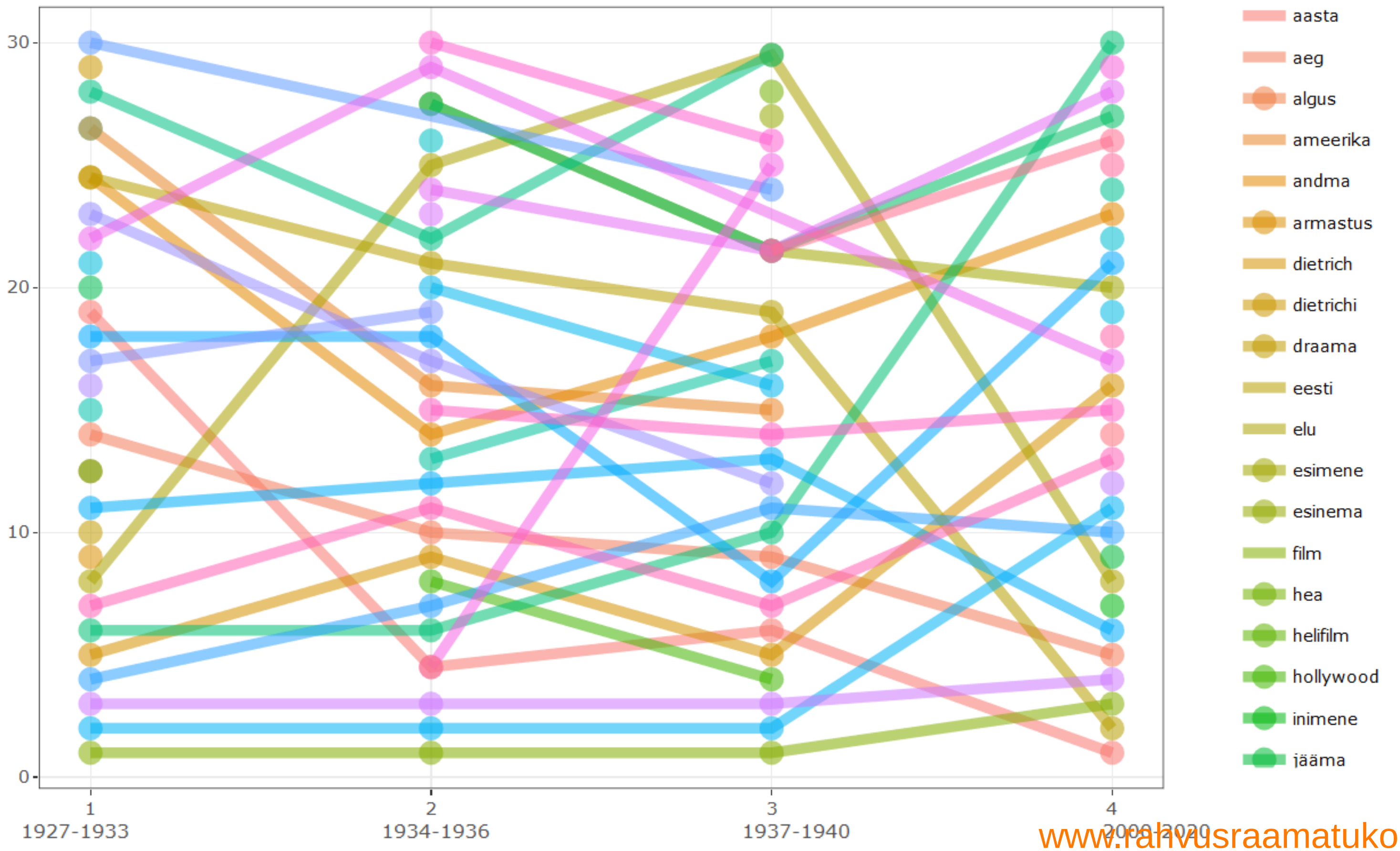
Visualization link



For discovery

Visualization link

Words in texts over time

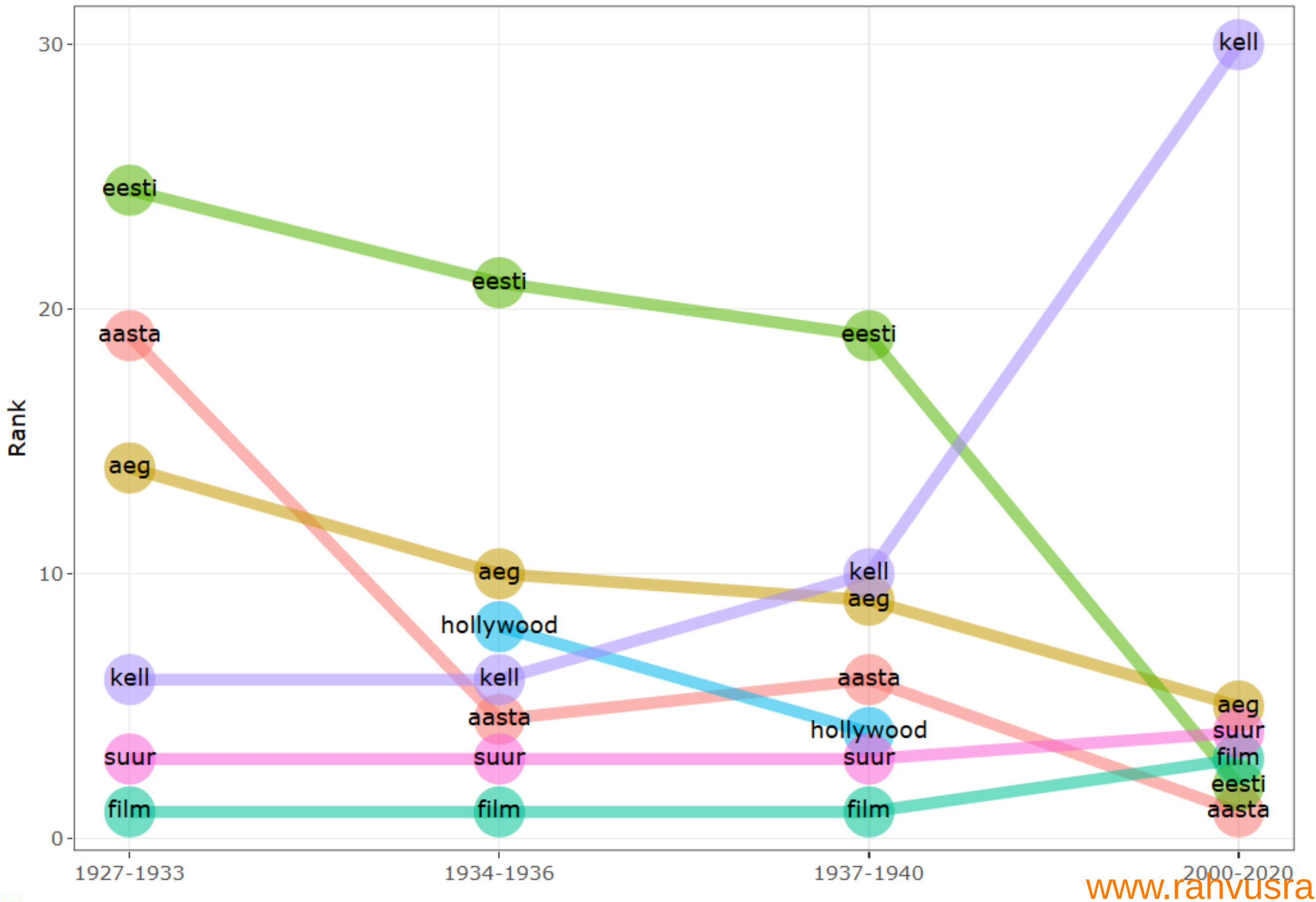


EESTI
RAHVUS-
RAAMATUKOGU

For discovery

Visualization link

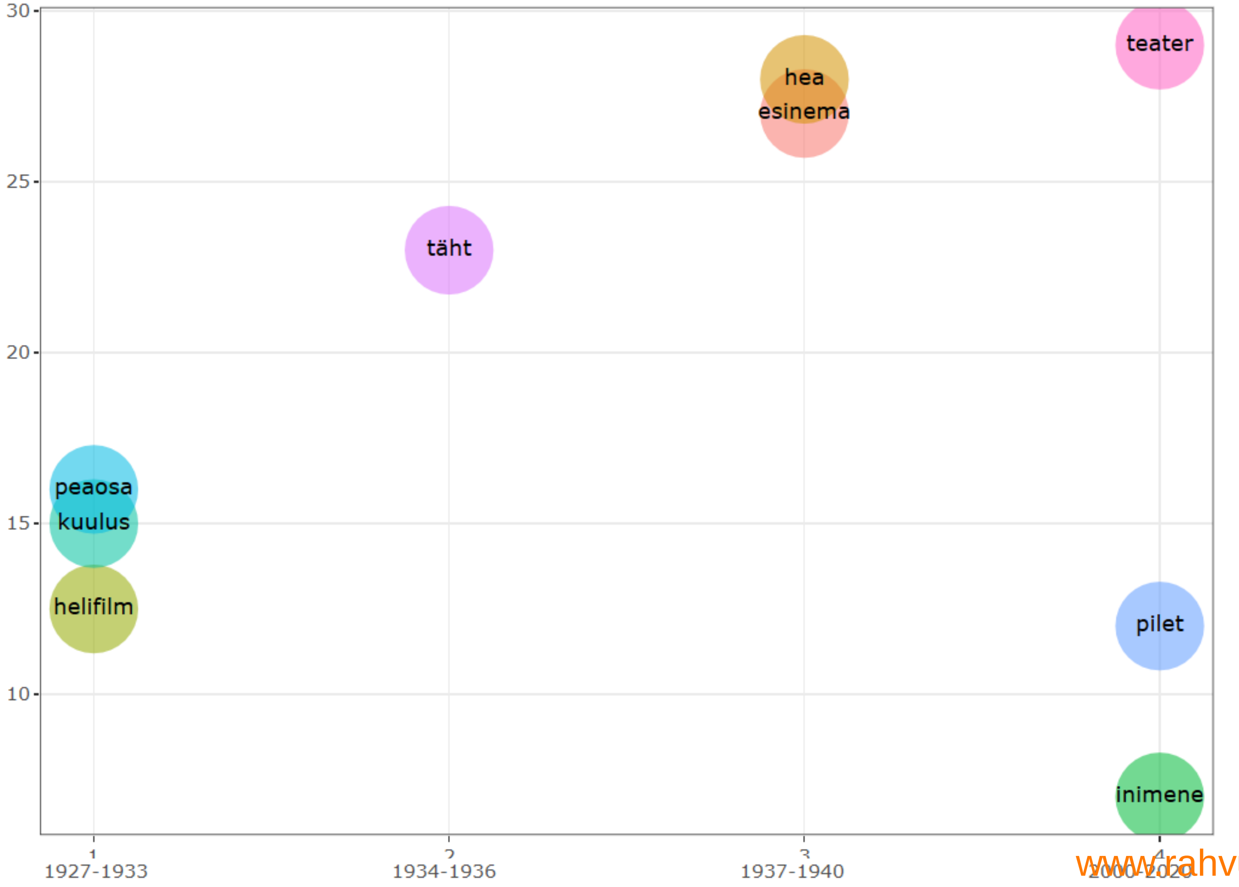
Words by frequency in texts on Marlene Dietrich



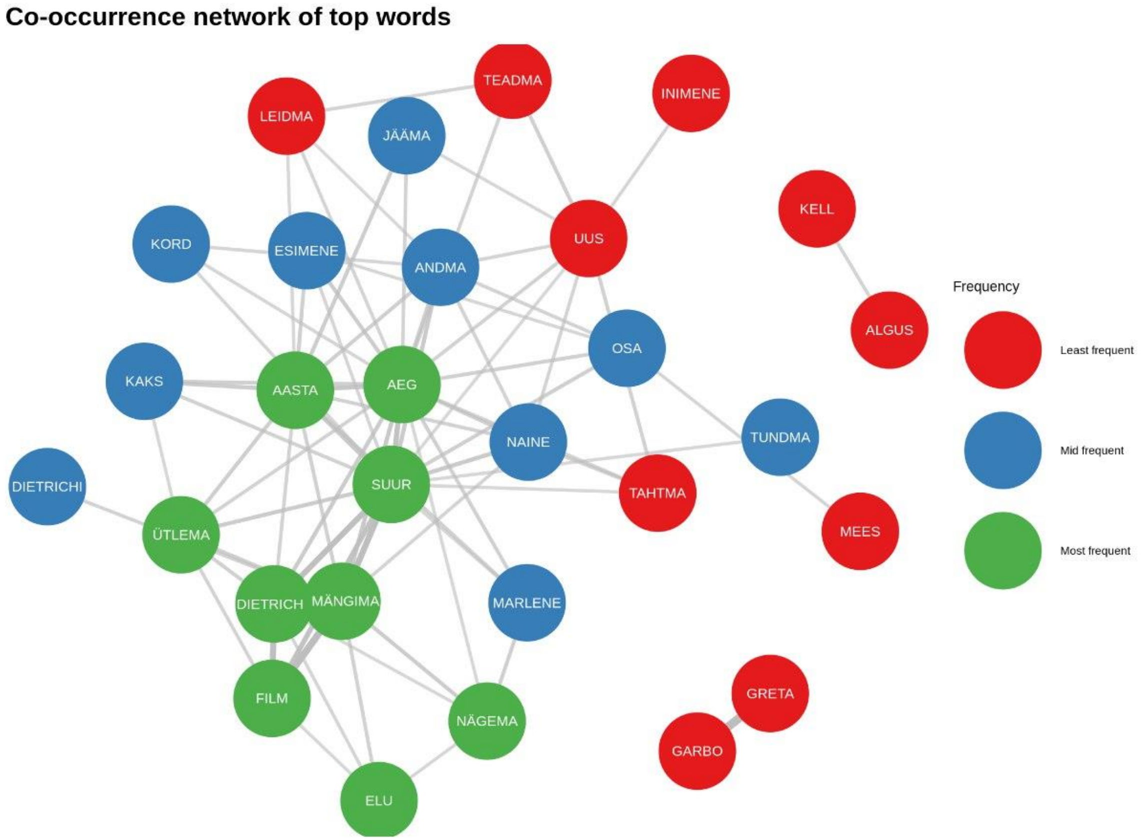
For discovery

Visualization link

Words that were common only for a short time



For discovery



Thank you!

All links

https://peetertinits.github.io/slides/plots/RR2020/current_state_color.html

https://peetertinits.github.io/slides/plots/RR2020/future_state_color.html

https://peetertinits.github.io/slides/plots/RR2020/dietrich_articlecount.html

https://peetertinits.github.io/slides/plots/RR2020/dietrich_articlecount_wtypes.html

https://peetertinits.github.io/slides/plots/RR2020/dietrich1.html

https://peetertinits.github.io/slides/plots/RR2020/dietrich_all_lines.html

https://peetertinits.github.io/slides/plots/RR2020/dietrich_continued.html

https://peetertinits.github.io/slides/plots/RR2020/dietrich_singles.html