

Adapting newspaper and bibliographic data for historical research: tools and case studies in Estonia

Peeter Tinitis
08.02.2024

Helsinki Digital Humanities Research Seminar

Outline

0. Thinking about researchers and archives

1. Digitized newspapers

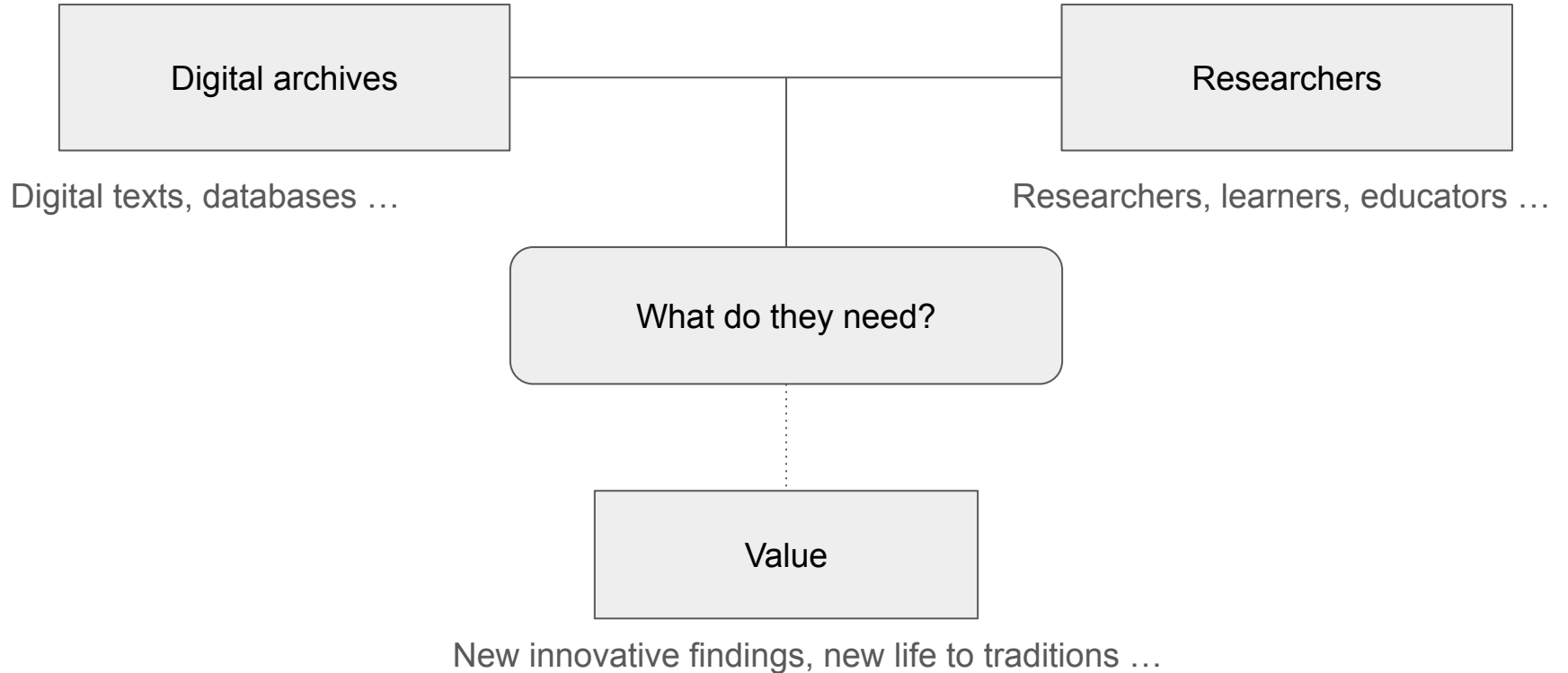
- a. What is in
- b. Data explorers
- c. Example cases

2. Bibliographic data

- a. What is in it
- b. Longer case study: emergence of Estonian written language community 1800-1940

3. Discussion

The use of digital archives



Thinking about the use

Perspective from data holders --- Perspective from researchers

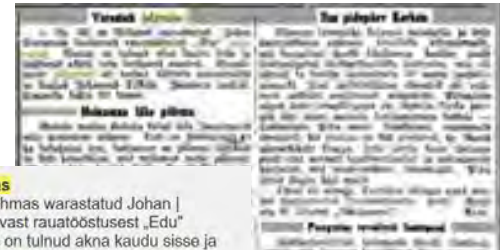


UNIVERSITY OF TARTU

Centre for Digital Humanities
and Information Society



1. Digitized newspapers



Brief historical overview



- 1993 microfilms, 2003 digitization.
 - First major newspapers Perno Postimees (1857-1885), Postimees (1886-1920), Päewaleht (1905-1912)
- 2004 online indexed reading portal
- 2014 text search portal (85 publications, 100 000 pages)
- 2024 text search portal (total 2787 publications, 5.8M pages, 15.6M articles)
 - Full text access to open data
 - New portal in 2026 includes data mining access in the system

Digitization done by a number of institutions over 20 years following current priorities with technology of the time. Workflow standards have varied over time and across institutions.

Biggest collection is in the National Library, but data is in a few different digital archives

Example case

You want to use newspapers to understand attitudes to nature in the 1920s.

Full text access:


0 m., nahk- või ülem kalingorköites 1000 m. Eesti **Loodus**. 8 iseseisvat artiklit Eesti ranniku, mere, geoloogi värvikiillaset, meeleolurikkamat ja sügavamalt **Loodus**ilku näinud. Ka kaasreisijate näod kõnelewad sedasama. See üsna magus lause on aga wabandataw: on nähtud **Loodus**t Antoni kaudu, ja tema psühholoogiale lause wastab oma office'ist või büroost wälja sõidab wabasse **Loodus**esse uut jõudu koguma. Kes kaugemale sõita ei jõua ole suutnud kõigutada isegi ilmasõda. Siin wabas **Loodus**es andub inglane täieliku innuga oma armsale lõbus rkk. Füüsilise maateaduse ülesannete kogu, I wihk «**Loodus**e maatlusd". Hind 85 markm Maateaduse ülesannete

Data balance
(e.g. corpus size)

Data quality
(e.g. OCR)

Representativeness
(do we have enough?)

Materials to answer the question




Study design

A user needs to know

A researcher would need to know

- What data is available (many archives)
- The quality of the data (varies a lot)
- Representativeness of the data (archive vs published)

Good to know this in the stage of planning & project-writing



Study design

Learning by doing

Case studies to understand the researcher perspective

E.g. 6 case studies made by 12 students 2 instructors in U. Tartu

Feedback into building the interfaces & updating data

E.g. full text access toolkit, ngram reader

What is in the dataset?

General overview

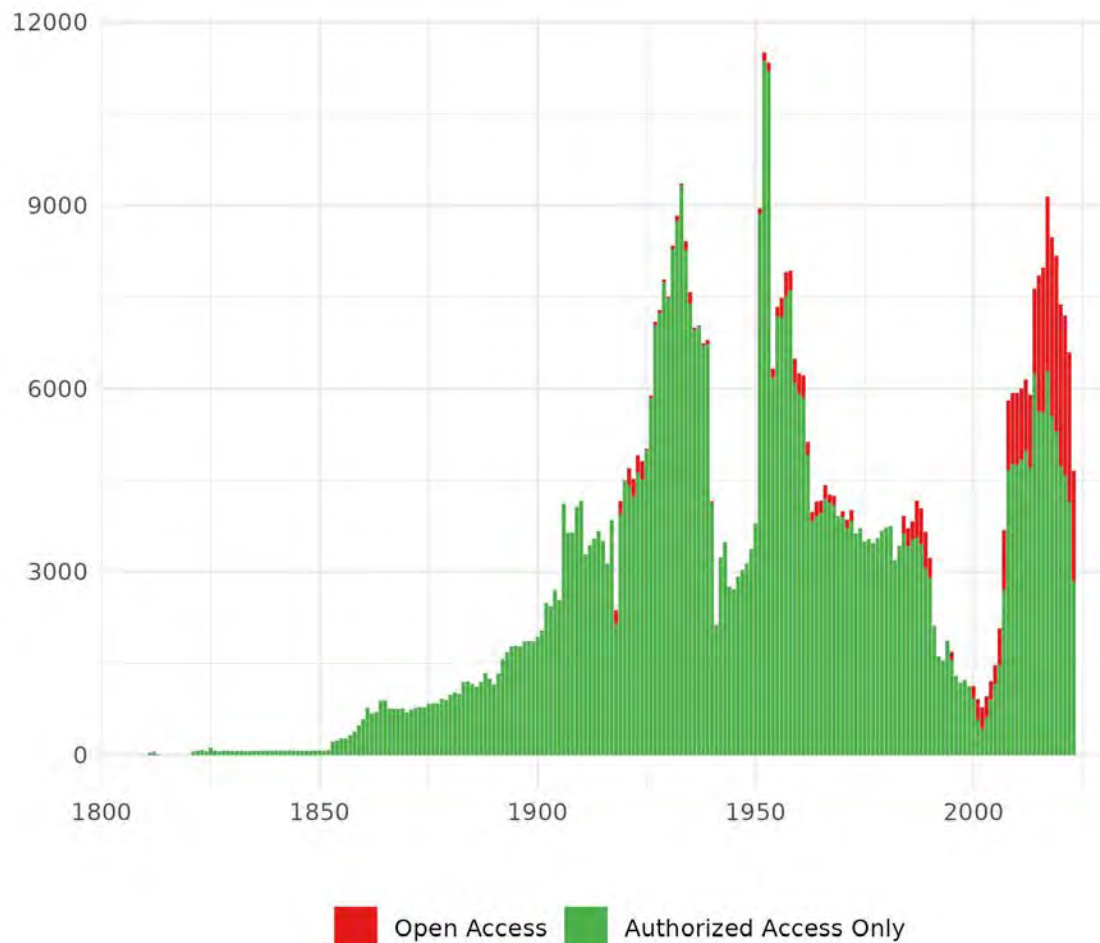
Distribution of issues over time
(n = 622,061)

Green = open access (92.8%)

Red = authorized access (7.2%)

Interactive:

https://peetertinits.github.io/reports/nlib/dea_info.html



Which newspapers have been digitized?

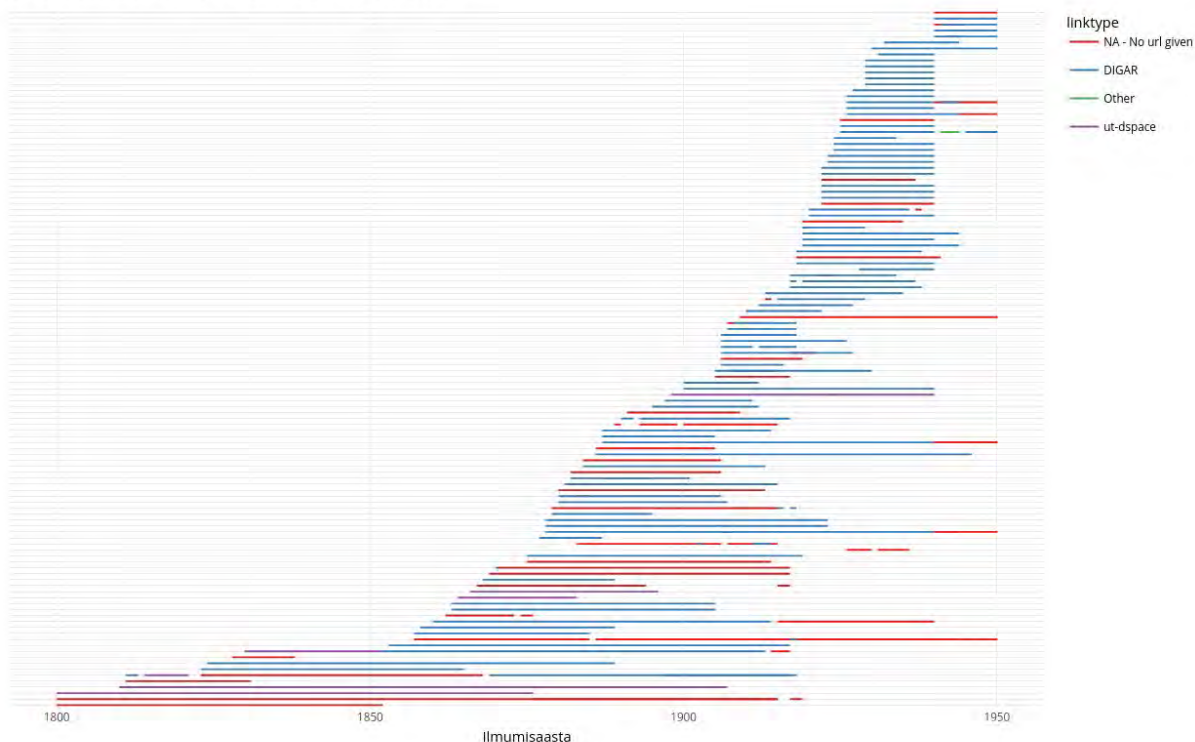
Data in historical context

Data in National Bibliography that is linked to digital archives (with some links added here)

Tool to understand the coverage, select sources when designing a study

https://digilab.shinyapps.io/digitized_newspapers2/

Ajalehed kollektsoonis (Üle 10 aasta ilmumist, n = 117)

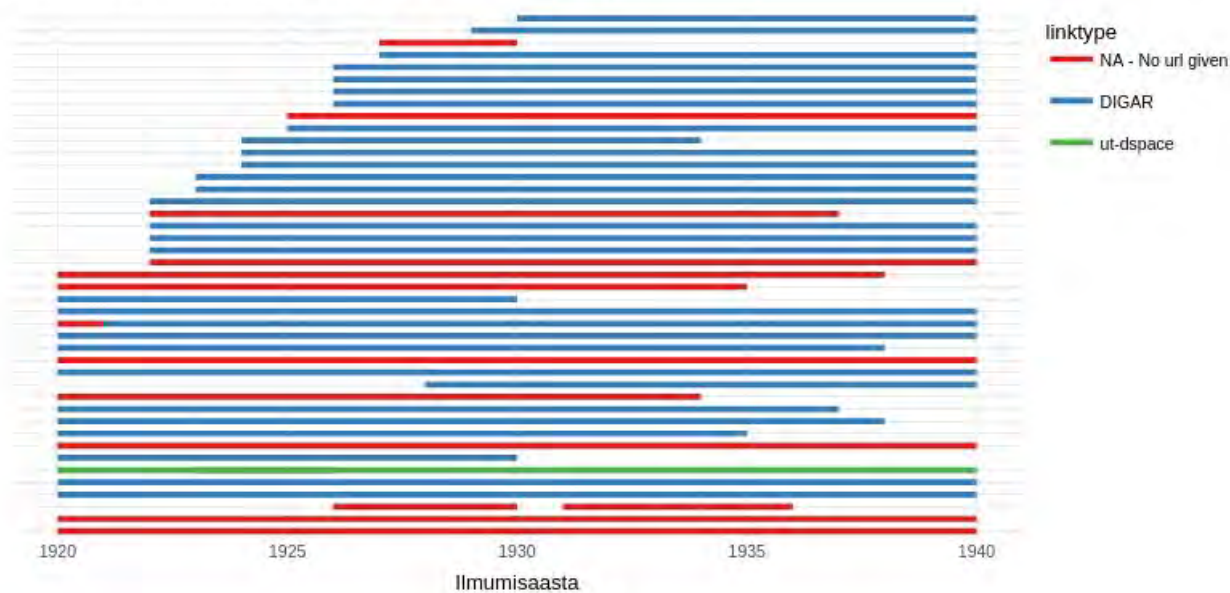


Which newspapers have been digitized?

An example - newspapers from 1920-1940 published for at least 10 years



Ajalehed kolleksioonis (Üle 10 aasta ilmumist, n = 43)



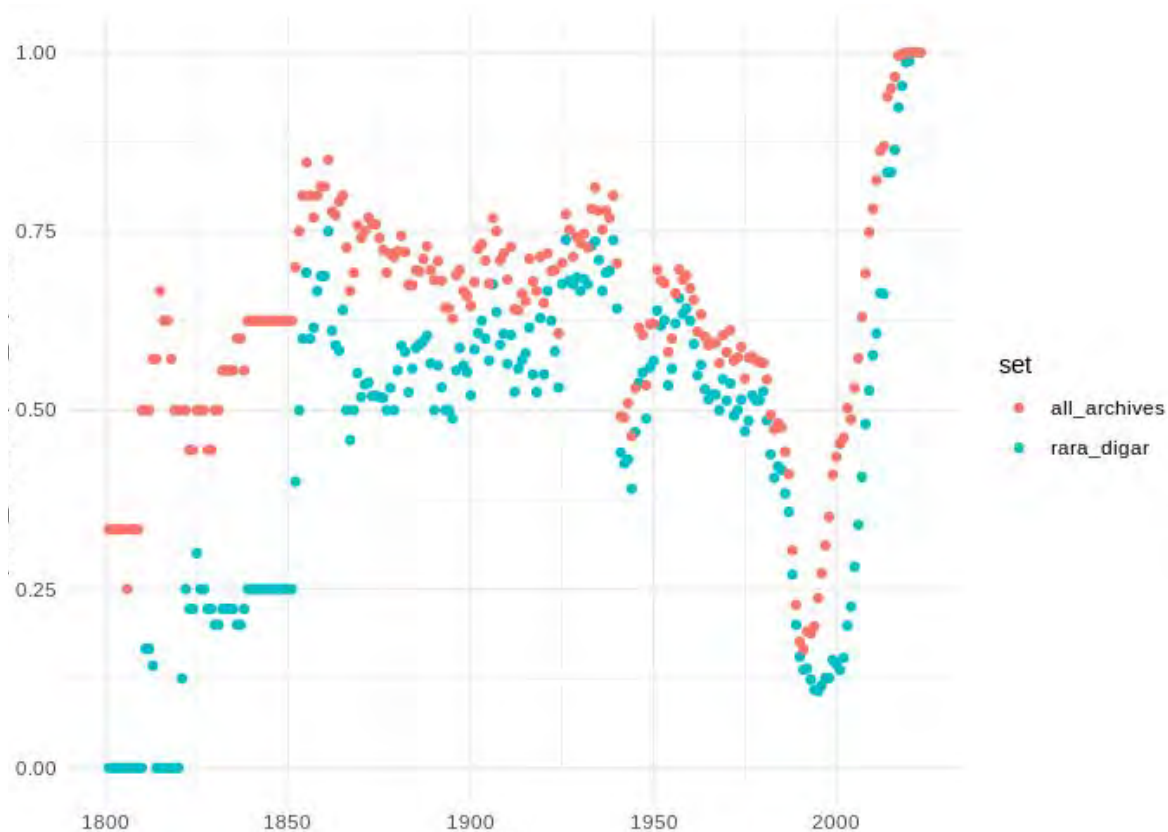
How much has been digitized?

Coverage of digitization

For each year, the share of newspapers in digital collections.

E.g. in 1900, 43 newspapers are known, 27 of which are digitized = 62.8%

1800-2020, 58.7% newspapers digitally available in some form



What is in the dataset?

Metadata explorer

Explore dataset with your own parameters in mind.

Interactive tool

<https://digilab.rara.ee/en/tools/newspapers-metadata-browser/>

Chosen sidebar options applied to all tabs except for the Timeline

Year of issue

1800 1918 1940 2020

1800 1822 1844 1866 1888 1910 1932 1954 1976 1998 2020

Language

German Estonian Russian Finnish English Swedish Dutch Latin French

Country of issue

Estonia Latvia Germany Russia United States Finland Australia Norway Sweden Poland Canada United Kingdom Netherlands

Newspaper type

Weekly newspaper Other newspaper Local newspaper Daily newspaper Supplement Exile newspaper Newspaper before 1944 Institutional newspaper

RESET SELECTION

SHARE LINK TO THIS SELECTION

Explore DIGAR collection

HOME METADATA NEWSPAPER SEARCH TIMELINE LANGUAGES COUNTRIES TYPES

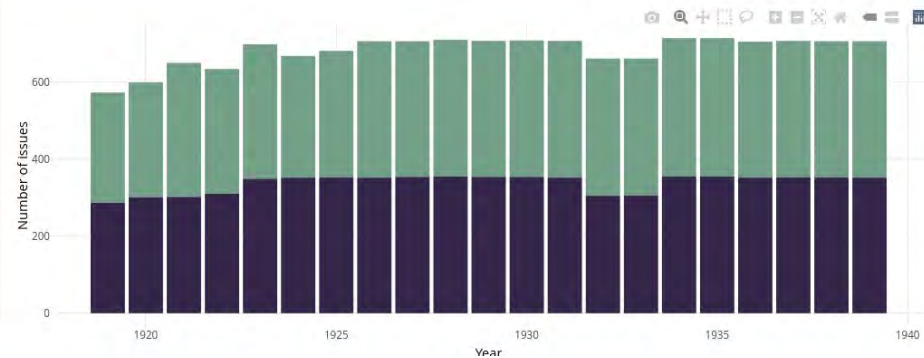
Search by title

Here one or multiple newspapers can be searched by title and filtered using sidebar. The output displays the issues distribution on a timeline and newspaper's metadata in the table

All titles are included in the search, but if there is no output in the plot and the table, the sidebar options such as year range and language should be loosened. The table below show summarised information for each newspaper.

Choose newspaper(s):

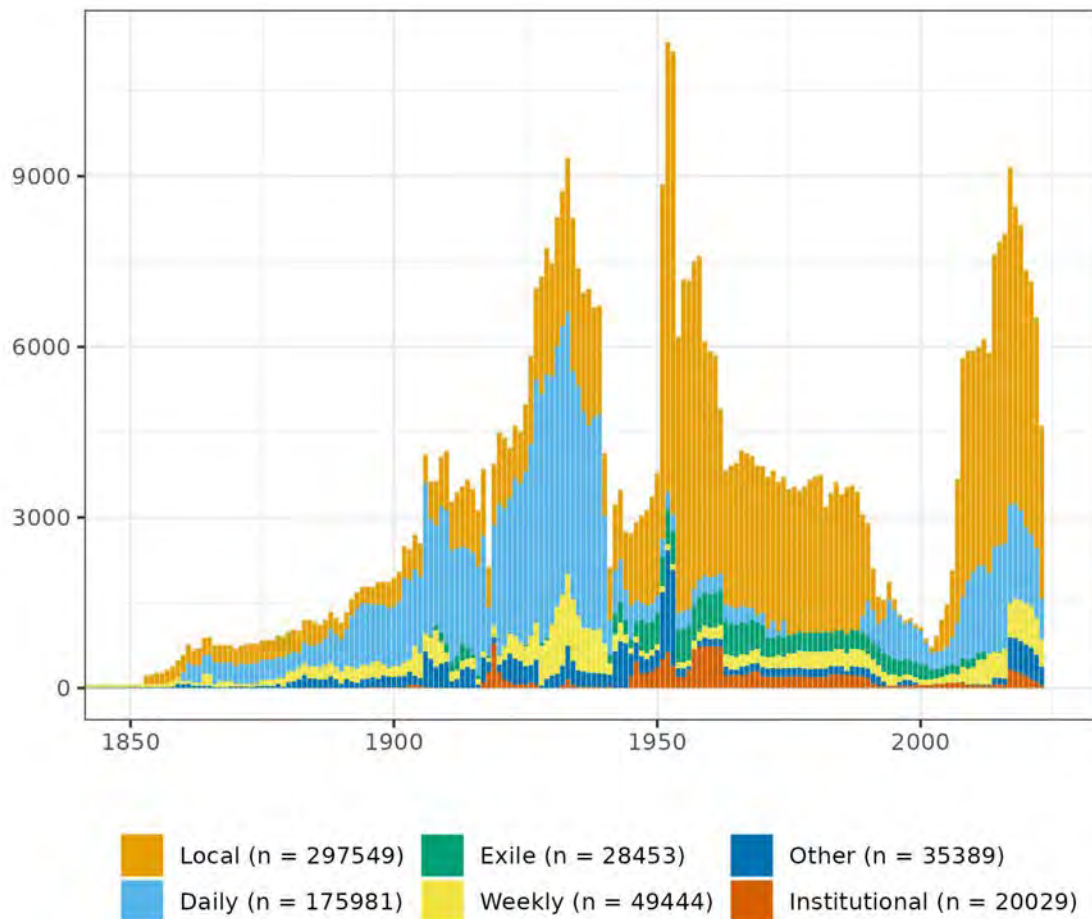
Päevaleht Postimees



What's in the dataset?

Overview by publication type

- Daily newspapers
- Weekly newspapers
- Local newspapers
- Exile newspapers
- Institutional newspapers



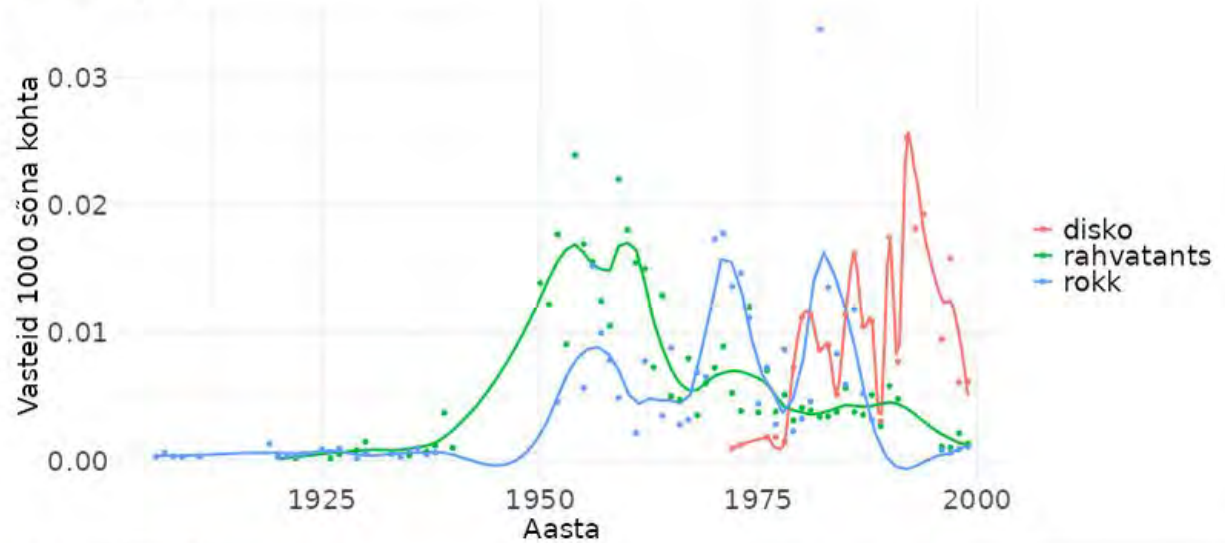
What's in the dataset

Ngram search

Quick peek at the text

Terms per 1000 words

Selected newspapers
over the period to
balance the dataset.



Rahvusraamatukogu sõnamitmike loendaja

Otsing

Töötlus

- Sõnad (muutmata)
 Lemmad (sõnade algvormid)

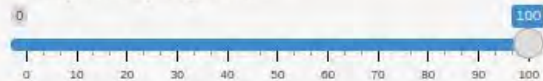
Aastad vaatluse all



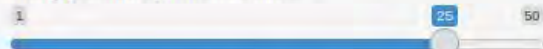
Joone kõverus



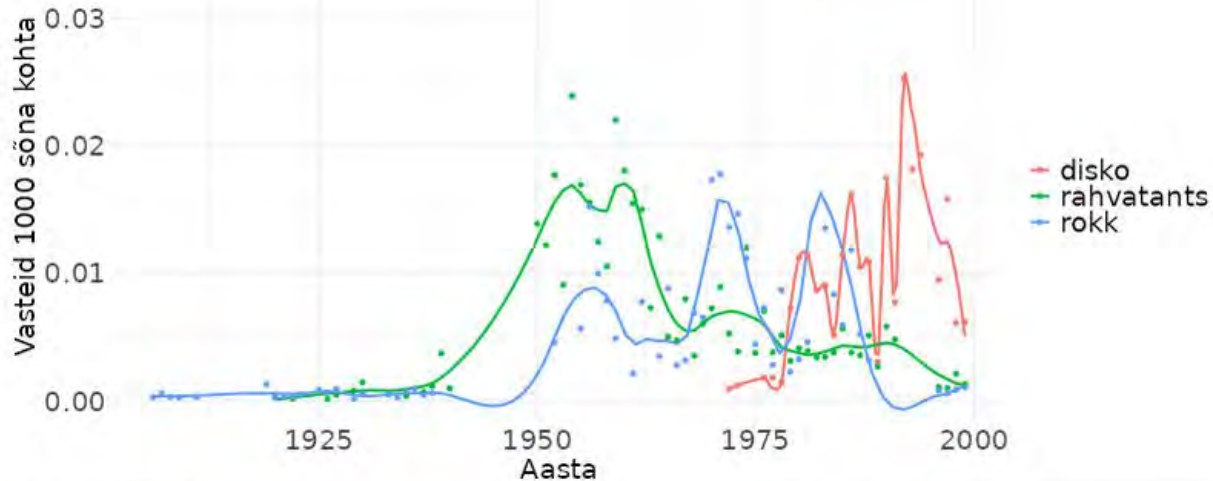
Lõika joonisel y-telg ära siin



Märgi joonisel iga mitmes aasta



RaRa sõnamitmike loendaja koondab sõnade ja sõnamitmike sagedusi Rahvusraamatukogu ajalehekollektsioonides. Sisestades ülal otsingusse mõne sõna, näeb paremal joonisel selle sagedust kollektsioonides 1857-2023. Terminite eristamiseks kasuta koma, tab-i või enterit. Otsides sõnamitmikku eraldi sõnad selle sees tühikuga. Kui teie otsitud sõna ei kuvata joonisel, ei ületanud ta vajalikku künnist, et



Näita kirjeid kaup

Otsi (regex lubatud)

Sõna/Sõnamitik	Tüüp	Mitu korda kokku	Mitmes dokumendis	Mitu sõna	
<input type="text" value="disko"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	
55904	disko	lemma	3176	2519	1
85010	disko	word	1991	1676	1
99895	diskor	lemma	1652	1127	1
134944	retrodisko	lemma	1159	593	1
135147	retrodisko	word	1157	608	1

Full text access

JupyterHub access to freely work with data & download if needed.



Sign in

Username:

Password:

Sign in

Server Options

Select a job profile:

Default: 1 CPU core, 8 GB memory, 6h timelimit

Start

Elektriuringu kood

Me kasutame uuringus Digiteeritud Eesti Artiklite andmekogu Eesti Rahvusraamatukogus. Ligipääsu vahendid on laiemalt lahti kirjeldatud siin.

Kõigepealt loeme sisse vajalikud paketid `tidyverse` ja `tidytext` ning DEA tekstidega töötamiseks loodud `digar.txts`. Neil pakettidel on

```
suppressPackageStartupMessages(library(tidyverse, lib.loc="/gpfs/space/projects/digar_txt/R/4.3/"))
suppressPackageStartupMessages(library(tidytext, lib.loc="/gpfs/space/projects/digar_txt/R/4.3/"))
suppressPackageStartupMessages(library(digar.txts, lib.loc="/gpfs/space/projects/digar_txt/R/4.3/"))
suppressWarnings(dir.create("plots"))
```

Andmestikust olevatest materjalidest saab ülevaate käsuga `get_digar_overview()` `digar.txts` pakettist. Tulemuseks on tabel, kus on kõi

Andmed

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

```
[1] "Issue metadata read"
```

Praeguses uurimuses huvitavad meid Postimehe numbrid vahemikus 1880-1940. Teeme töötamiseks vastava alamhulga

```
subset <- all_issues %>%
  filter(DocumentType=="NEWSPAPER") %>%
  filter(year>1880&year<1940) %>%
  filter(keyid=="postimeesew")
#fwrite(subset, "data/subset_postimeesew1.tsv", sep="\t")
```

Sellele alamhulgale võtame eraldi välja ka metaandmestiku iga artikli kohta, mis valimi ajalehenumbrites on.

```
subset_meta <- get_subset_meta(subset)
#fwrite(subset_meta, "data/subset_meta_postimeesew1.tsv", sep="\t")
```

Valimist ülevaate saamiseks vaatame kui palju on igal aastal valimis 1) artikleid ja 2) sõnu.

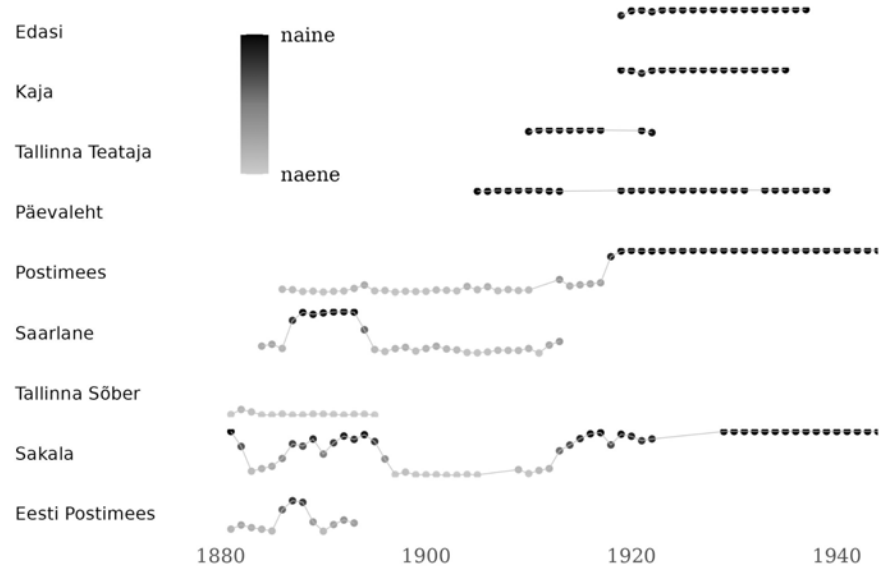
Full text access

This allows more deliberate studies

E.g. the use of linguistic terms of interest in different newspapers.

Can look for relative frequencies of ngrams within newspapers.

~ Impact of prescription and/or editors on language.



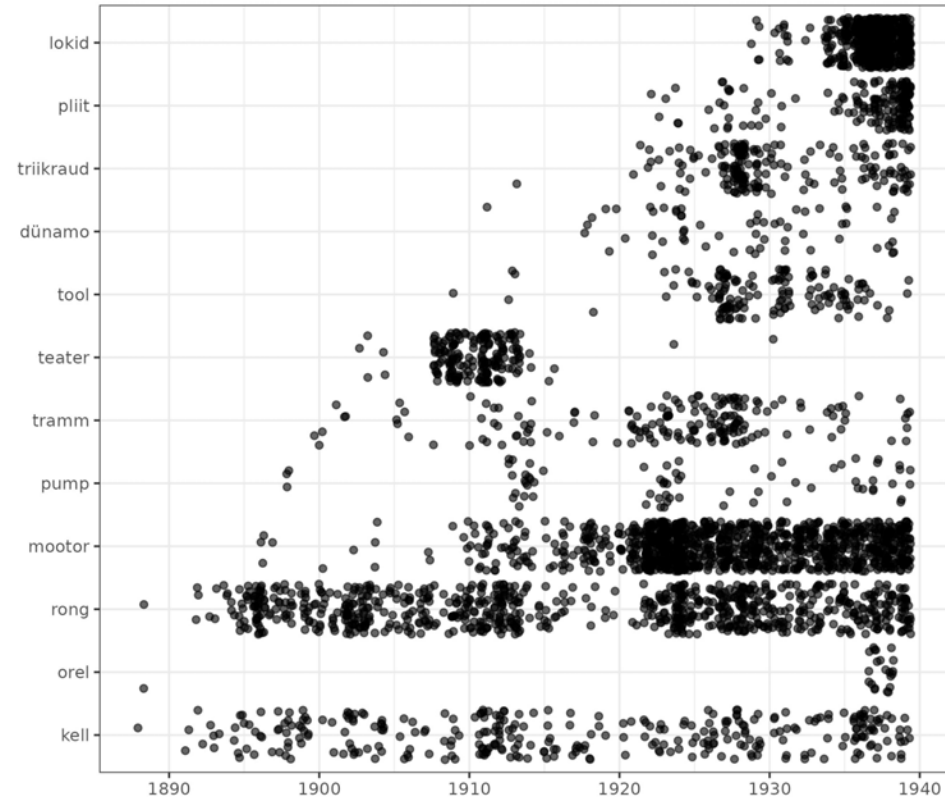
Full text access

This allows more deliberate studies

E.g. the distribution of electricity related products in early 20th century.

Can look for electricity related terms, classify them into types, and select the meaningful ones.

~ History of technology and advertisement or popular discourse



Quality of materials

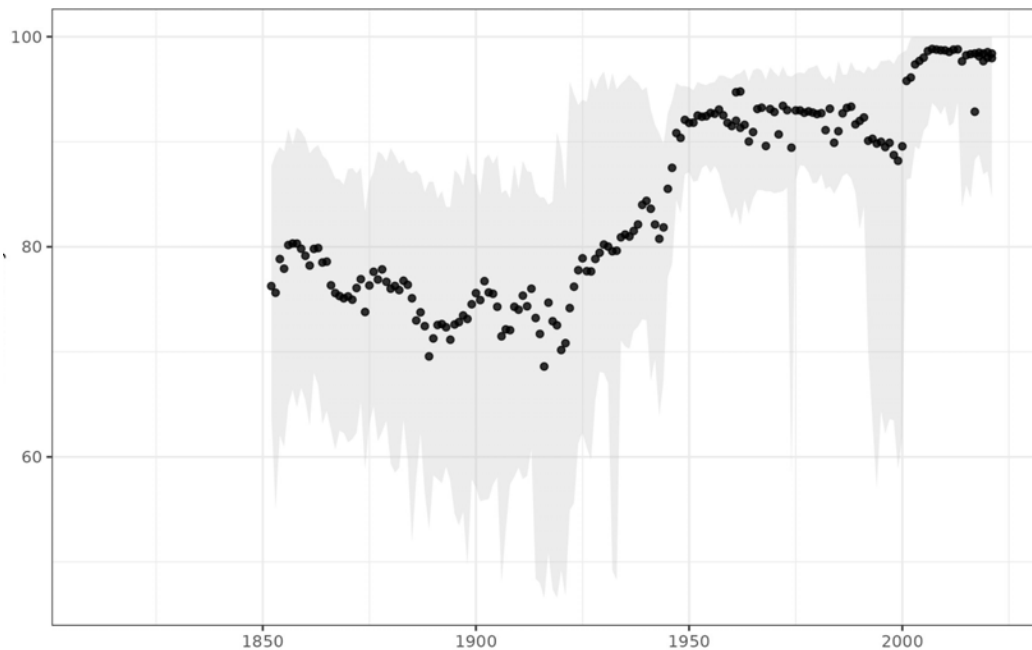
Data quality differs.

- Earlier digitized works have poorer quality.
- Quality not evaluated during digitization.
- Since ~2014, OCR Confidence stored for articles.

Working to evaluate now.

- Possibilities to auto-correct?
- May need to redigitize either from image or from scratch.
- Problems for some tasks, not others.

Mean OCR Confidence + spread of 95% of articles



Digitized newspapers

Keeping the researcher in focus:

- Overviews
- Access
- Data explorers
- Data in context

Current topics:

- Data quality

2. Bibliographic data

* Luggemise kirjandus, v. D. W. Masing, 1824.
 * Dppetus loida neid luggemise lehti laste
 ga prindis, v. D. W. Masing, 1821.
 * A. B. D., Duch, v. D. H. 1823.
 * Tšieline A. B. D. - Namat it. v. D. W.
 Masing, 1823.
 * Tšieline Nisistaja dpetus it. v. D. W.
 Masing, 1825.
 * Patsuse Namat, Bönna Nihelkonna Pd
 Teeli laste kullase. v. E. H. Kõder. 1827-
 2) Manuscripte.
 * Mures Jutti E. Andres Schreibeubch,
 1814.
 * Tamme Andres E. Jafob's Schreibeubch,
 1814.
 * Andres u. Jafob's Entlühungen. 3 Bde.
 1815 u. 16.
 * Bönne A. B. D. - ja luggemise-ramat, van
 D. H. 1815.
 * Andres u. Jafob's Geographie, nach Was
 paris ritsem Kurfus. 1816.
 * Abraham Holtera's Entlühungen. 2 Bde. 1816
 und 17.
 * Mures Abraham E. Karle's u. Bilenida

X. Rechnenkunst.
 1) Druckchriften.
 * Argoiddamise eht Arwamisse. Kuu
 D. H. v. J. J. 1808.
 * Arwamisse Namat, v. D. W. Masing.
 * Arwo - eht Nehtendamise etsemplid.
 Mr. Holter. 1823.
 2) Manuscripte.
 * Arwamisse - ramat it. von D. H. 1814
 XI. Pädagogik.
 1) Druckchriften.
 * Ma. rahwa laste-koewotamisse it. v.
 digten.
 2) Manuscripte.
 * Zeitschrift, dem esthnischen Schul-
 bungs- und Bildungsweisen gemäss
 d. H. Erles Heft. 1817.
 * Plan zu einem Schulmeister. Semis
 Eifen, nebst einer Geschichte desselb.
 D. W. Masing. 1827.

title_responsibility_statement	publication_place	publisher	publication_date
Anton Hansen Tammeaare ; illustreeritud Harald Eskma	Tallinn	Eesti Raamat	1978
Mari Laar	Stockholm	Välis-Eesti & EMP	1997
Mihhail Makarov ; (Kare. E. Tall)	Tallinn	Eesti Raamat	1886
Neel Bark ; inglise keeles Mõinud A. Hanker ; kirjandus- Eesti teadusajakirja Seltsi Eesti Teaduste Akadeemia j.	Tallinn	Eesti Raamat	1994
Eesti Raamat	Valgus		1993
Eesti Raamat	Tallinn	Eesti Raamat	1995
Alexandre Dumas (junior) ; kirjandus keeles Mõinud j.	Tallinn	Eesti Raamat	1992
Linda Petti	Tallinn	Valgus	1993
Hennrich Bill ; [ilukta keeles Mõinud M. Sirkel ; kirjandus- Eesti teadusajakirja Seltsi Eesti Teaduste Akadeemia j.]	Tallinn	Eesti Raamat	1993
Marina Tsvetajeva ; [koostanud, eestindanud ja järelel- toimetajad: Baloo Kõnn, Sven Kiviküll, Ants Noormetsa]	Tallinn	Eesti Raamat	1998
L. Elkon	Uppsala	Eesti Meeslaidu Selts Melbourne'is	1994
Linda Petti	Tallinn	Valgus	1992
Juhan Jää ; [ilustratsioonid H. Mugaiste]	Tartu	Noor-Eesti	1938
A. H. Tammeaare	Tartu	Noor-Eesti	1917
Peter Brook ; inglise keeles Mõinud M. Pall ; Eesti Teadus- Eesti teadusajakirja Seltsi Eesti Teaduste Akadeemia j.]	Tallinn	Eesti Raamat	1993
Mari Laar	Stockholm	Välis-Eesti & EMP	1994
Isakman by Anur Vatarna ; toimetatud by Paul Hagns- Harald Sialava ; [ilustratsioonid R. Baitson]	Kaepo	Õhellaan-usthoosti	1995
Harald Sialava ; [ilustratsioonid R. Baitson]	Tallinn	Eesti Raamat	1984

```
ind2=" " tag="700">
a">Stendhal,</mx:subfield>
d">1783-1842</mx:subfield>
0">(BLBNB)000564341</mx:subfield>
```

```
ind2=" " tag="700">
a">برنارد شو,</mx:subfield>
d">1783-1842</mx:subfield>
0">(ISNI)0000000120981326</mx:subfield>
```

```
ind2="1" tag="700">
a">برنارد شو,</mx:subfield>
d">1783-1842</mx:subfield>
0">(NLR)RU NLR AUTH 775430</mx:subfield>
```

```
=LDR 00376nz a2200133n 4500
+001 a11124799
+003 ERRR
+008 9904051n|adnnaabn|||||||a|||||
+040 ||$aERRR$best$eERR
+043 ||$cee
+100 1|$aAdojaan, Julius.$d1912-1994.
+400 1|$aАдоян, Юлийс.$d1912-1994.
+670 1|$aEesti elulood. Tallinn, 2000.
+680 1|$iEesti õigusteadlane, Aleksander Adojaan
```

```
=LDR 00589nz a2200181n 4500
+001 a11124805
+003 ERRR
+008 981223n|adnnaabn|||||||a|||||
+040 ||$aERRR$best$eERR
+043 ||$cee
+100 1|$aAdson, Artur.$d1889-1977.
+400 1|$aAdson, Karl Artur.
+400 1|$aAdson, Arthur.
+400 1|$aAncon, Karl Arthur.
+400 1|$aAncon, A. Pryp.$d1889-1977.
+500 1|$aBollmann, Peeter.$cPseud.$d1889-19
+670 1|$aEesti kirjanduse leksikon. 2000.
+680 1|$iEesti luuletaja, näitekirjanik ja kriitik. M.
```

Research use

Thinking again in terms of research use <-> data

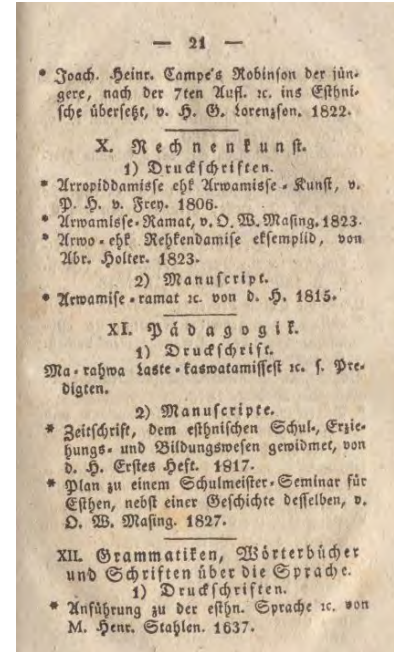
The researcher needs

- Access and usability
- Overviews of the data
- Connections with other datasets
- Representativeness of the data



National bibliographies

- Register publications connected to a language and/or region
- Collect systematic data (titles, authors, formats, etc.) on printed publications
- Usually maintained by libraries and accessed within the library catalogue search



Excerpt from Rosenplänter's article
"Übersicht der estnischen Literatur" (1832)

Estonian National Bibliography

Consists of publications that are either:

- a. published in Estonia
- b. in Estonian
- c. by Estonians
- d. about Estonia or Estonians



<https://digilab.rara.ee/en/>

Estonian National Bibliography

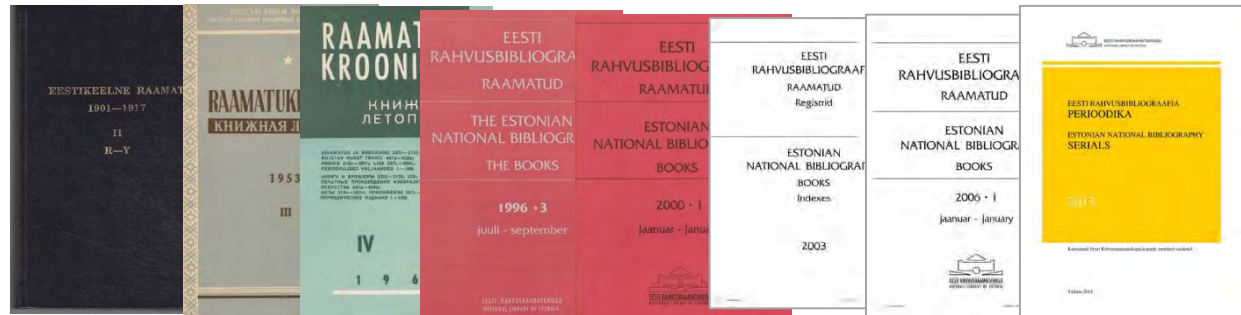
Origins in 1832: H. J. Rosenplänter aimed to collect an overview of all published in Estonian or about Estonia.

Then: 384 books listed in Estonian 1505-1832. Now known: 919 books 1505-1832.

Followed by a number of comprehensive bibliographies throughout the period, since 2004 an electronic database.



Excerpt from Rosenplänter's article "Übersicht der estnischen Literatur" (1832)



- LDR 00376nz a2200133n 4500
 - 001 a11124799
 - 003 ErRR
 - 008 990405[n]adnnaabn|||||||/|a|/|||
 - 040 ||\$aErRR\$bestscErRR
 - 043 ||\$cee
 - 100 1|\$aAdojaan, Julius.\$d1912-1994.
 - 400 1|\$aAdoan, Юлий.\$d1912-1994.
 - 670 ||\$aEesti elulood. Tallinn, 2000.
 - 680 ||\$iEesti õigusteadlane, Aleksander Adojaani vend.
-
- LDR 00589nz a2200181n 4500
 - 001 a11124805
 - 003 ErRR
 - 008 981223[n]adnnaabn|||||||/|a|/|||
 - 040 ||\$aErRR\$bestscErRR
 - 043 ||\$cee
 - 100 1|\$aAdson, Artur.\$d1889-1977.
 - 400 1|\$aAdson, Karl Artur.
 - 400 1|\$aAdson, Arthur.
 - 400 1|\$aAdson, Karl Arthur.
 - 400 1|\$aAdson, Artur.\$d1889-1977.
 - 500 1|\$aBollmann, Peeter.Scpsseud.\$d1889-1977.
 - 670 ||\$aEesti kirjanike leksikon, 2000
 - 680 ||\$iEesti luuletaja, näitekirjanik ja kriitik. M. Uuden abi

The database format

- Has been maintained in Marc21, following library standards.
- Converted it into more compact tables for research purposes.
- Now as a .tsv in Zenodo
- We are working on a curated, simplified version of it.

```
-LDR 00376nz a2200133n 4500
-001 a1112495
-003 EtRR
-008 990405[n]adnnaabn|||||Vaf|||||
-040 \${aEtRR$best$EtRR
-043 \${cSee
-100 1\${aAdojaan, Julius.$d1912-1994
-400 1\${aAdon, Kõrv,.$d1912-1994
-670 \${eEesti elulood. Tallinn, 2000
-680 \${eEesti õigusteadlane, Aleksander Adojaani vend.
```

```
-LDR 00589nz a2200181n 4500
-001 a11124905
-003 EtRR
-008 981223[n]adnnaabn|||||Vaf|||||
-040 \${aEtRR$best$EtRR
-043 \${cSee
-100 1\${aAdson, Arhur.$d1889-1977
-400 1\${aAdson, Karl Arhur.
-400 1\${aAdson, Arhur.
-400 1\${aAdson, Karl Arhur.
-400 1\${aAnzon, Ap'ryp.$d1889-1977
-500 1\${aBollmann, Peeter.$cpcpsuud.$d1889-1977
-670 \${eEesti kirjanduse lokaalikon. 2000
-680 \${eEesti kujutaja, nätekinriinik ja kriitik M. Unden abikaasa. Kasutanud pseudonüümi Peeter Bollmann.
```

title_responsibility_statement	publication_place	publisher	publication_date
Anton Hansen-Tammäär ; illustreeritud Herald Esbma	Tallinn	Eesti Raamat	1974
Mart Lauer	Stockholm	Välö-Eesti & ISAP	1997
Mihail Makarov ; (Kaz. E. Tall)	Tallinn	Eesti Raamat	1986
Pjotr Burk ; inglise keelest tõlkinud A. Händel ; (Tajenda-)	Tallinn	Eesti Raamat	1994
Eesti Lõõtsaõuajajate Seltsi Eesti Teaduste Akademia ja	Tallinn	Valgus	1993
Krist Hännus ; (Tõlkinud Henrik Sepemaa ; koostanud.)	Tallinn	Eesti Raamat	1995
Aleksander Duma (Janin) ; (prantsuse keelest tõlkinud)	Tallinn	Eesti Raamat	1992
Linda Petti	Tallinn	Valgus	1993
Heinrich Bill ; (saksa keelest tõlkinud M. Sirkel ; kujund.)	Tallinn	Eesti Raamat	1993
Marina Tsvetajeva ; (koostanud, eestindanud ja järeltõl.)	Tallinn	Eesti Raamat	1994
toimetajad: Rakvi Elmel, Sveth Kivvül, Ants Noormets...	Melbourne	Eesti Meeslaulu Selts Melbourne's	1994
J. Ekman	Uppsala	Tu Sõna	1993
Linda Petti	Tallinn	Valgus	1992
James Herriot ; (inglise keelest tõlkinud Jaagup Ahari ; ka-	Tallinn	Valgus	1994
Juhan Jaak ; (ilustreeritud H. Muugisti)	Tartu	Noor-Eesti	1934
A. H. Tammäär	Tartu	Noor-Eesti	1927
Peter Bronk ; inglise keelest tõlkinud M. Päll ; Eesti Tek.	Tallinn	Eesti Raamat	1993
Mart Laan	Stockholm	Välö-Eesti & ISAP	1994
ladinust by Anne Valerna ; toimitanud by Paul Hagns...	Kaopio	Insellman-instituut	1995
Harald Soidsepp ; (koostanud R. Pangaaga)	Tallinn	Eesti Raamat	1994

```
<mx:datafield ind1="1" ind2=" " tag="700">
  <mx:subfield code="a">Stendhal,</mx:subfield>
  <mx:subfield code="d">1783-1842</mx:subfield>
  <mx:subfield code="0">(BLBNB)000564341</mx:subfield>
</mx:datafield>
<mx:datafield ind1="0" ind2=" " tag="700">
  <mx:subfield code="a">مذكرات</mx:subfield>
  <mx:subfield code="d">1783-1842</mx:subfield>
  <mx:subfield code="0">(ISNI)0000000120981326</mx:subfield>
</mx:datafield>
<mx:datafield ind1=" " ind2="1" tag="700">
  <mx:subfield code="a">مذكرات</mx:subfield>
  <mx:subfield code="d">1783-1842</mx:subfield>
  <mx:subfield code="0">(NLR)RU NLR AUTH 775430</mx:subfield>
</mx:datafield>
```

The data at the moment

- Comprehensively described with good coverage
- Non-harmonised
- A lot of pre-processing and cleaning to be useable

```
-LDR 00376nz a2200133n 4500
-001 a11124799
-003 ErRR
-008 990405[n]adnnaabn|||||||/a|||||
-040 ||$aErRR$best$cErRR
-043 |$c$e
+100 1|$aAdojaan, Julius.$d1912-1994.
-400 1|$aАдоян, Юлиус.$d1912-1994.
+670 1|$aEesti elulood. Tallinn, 2000.
-680 1|$iEesti õigusteadlane, Aleksander Adojaani vend.
```

```
-LDR 00589nz a2200181n 4500
-001 a11124805
-003 ErRR
-008 981223[n]adnnaabn|||||||/a|||||
-040 ||$aErRR$best$cErRR
-043 |$c$e
+100 1|$aAdson, Artur.$d1889-1977.
-400 1|$aAdson, Karl Artur.
-400 1|$aAdson, Arthur.
+400 1|$aAdson, Karl Arthur.
-400 1|$aAdson, Artur.$d1889-1977.
+500 1|$aBollmann, Peeter.Scps$eud.$d1889-1977.
-670 1|$aEesti kirjanike leksikon, 2000.
-680 1|$iEesti luuletaja, näitekirjanik ja kriitik. M. Unden abikaasa. Kasutanud pseudonüümi Peeter Bollmann.
```

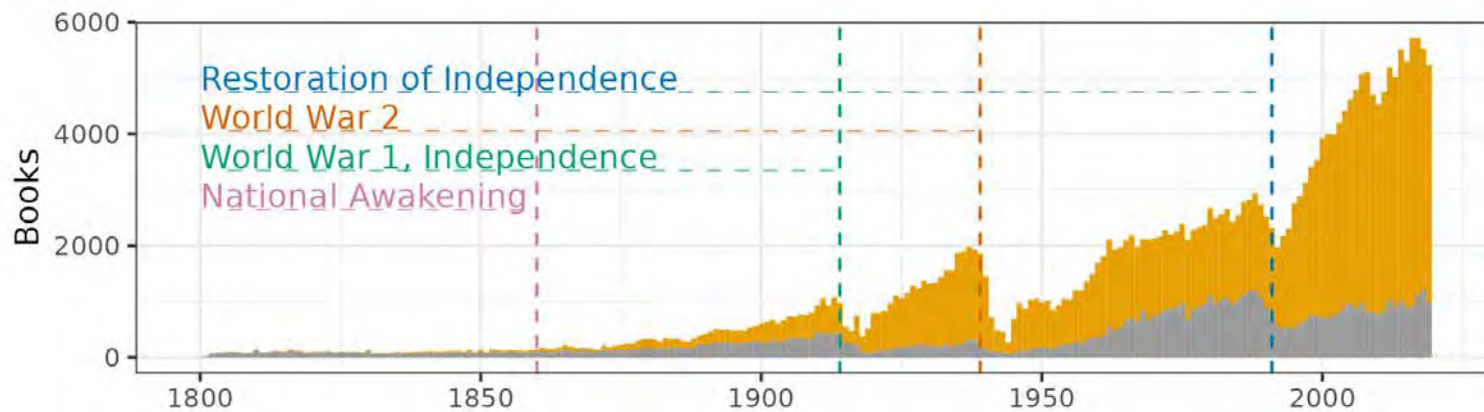
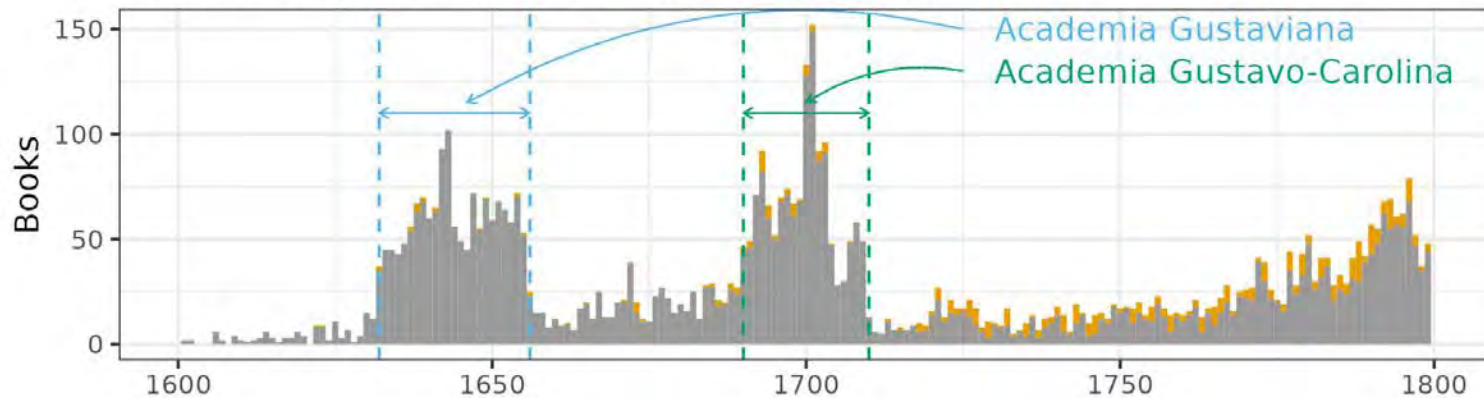
title_responsibility_statement	publication_place	publisher	publication_date
Antoni Hansen.Tammearst ; illustreerinud Harald Esimo	Tallinn	Eesti Raamat	1974
Mart Laar	Stockholm	Välje-Eesti & EMP	1977
Mihhail Makarev ; (kirj. E. Tall)	Tallinn	Eesti Raamat	1966
Paavo Burk ; inglise keelest tõlkinud A. Hänke ; [koostaja...]	Tallinn	Eesti Raamat	1994
Eesti kirjandusajajate Seltsi Eesti Teaduste Akadeemia j...	Tallinn	Vilgus	1993
Eevi Hansson ; (tõlkinud Henrik Sepemaa ; kujundanud...	Tallinn	Eesti Raamat	1995
Alexandre Dumaa (junior) ; (prantsuse keelest tõlkinud j...	Tallinn	Eesti Raamat	1992
Linda Petti	Tallinn	Vilgus	1993
Heinrich Bill ; [saksa keelest tõlkinud M. Sirks] ; kujund...	Tallinn	Eesti Raamat	1993
Marlene Teetajava ; [koostanud, eestindanud ja järeltõ...	Tallinn	Eesti Raamat	1994
Tõimetaja: Ralfo Eilini, Sven Rivilüü, Ants Naarmets...	Melbourne	Eesti Meeslaulu Selts Melbourne'is	1994
J. Elman	Lappeala	Etu Sõna	1993
Linda Petti	Tallinn	Vilgus	1992
Jarmo Hirsut ; (inglise keelest tõlki Jaagup Alevis, ka...	Tallinn	Vilgus	1994
Juhan Jää ; [ilustreerinud H. Mugasti]	Tartu	Noor-Eesti	1934
A. H. Tammeäär	Tartu	Noor-Eesti	1917
Peter Bronk ; inglise keelest tõlkinud M. Päll ; Eesti Pev...	Tallinn	Eesti Raamat	1993
Mart Laar	Stockholm	Välje-Eesti & EMP	1994
laulnud by Anne Valarna ; toimittanud by Paul Niiga...	Kaepio	(Sveitsi muusika)	1995
Harald Sielava ; (kujundanud R. Pangaste)	Tallinn	Eesti Raamat	1994

```
<mx:datafield ind1="1" ind2=" " tag="700">
<mx:subfield code="a">Stendhal,</mx:subfield>
<mx:subfield code="d">1783-1842</mx:subfield>
<mx:subfield code="0">(BLBNB)000564341</mx:subfield>
</mx:datafield>
<mx:datafield ind1="0" ind2=" " tag="700">
<mx:subfield code="a">دور، محمد</mx:subfield>
<mx:subfield code="d">1783-1842</mx:subfield>
<mx:subfield code="0">(ISNI)0000000120981326</mx:subfield>
</mx:datafield>
<mx:datafield ind1=" " ind2="1" tag="700">
<mx:subfield code="a">دور، محمد</mx:subfield>
<mx:subfield code="d">1783-1842</mx:subfield>
<mx:subfield code="0">(NLR)RU NLR AUTH 775430</mx:subfield>
</mx:datafield>
```

A look at the data

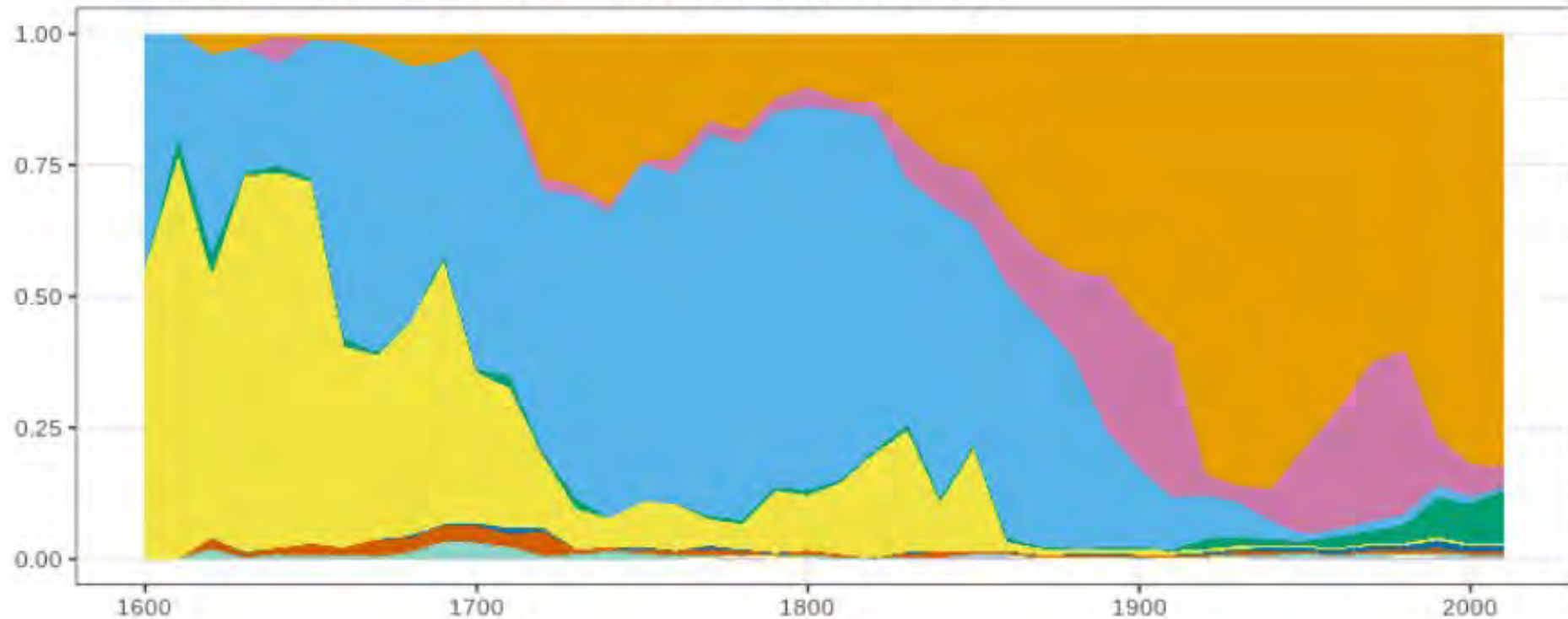
Books in ENB (1600-2020)

Some historical periods and events added for comparison



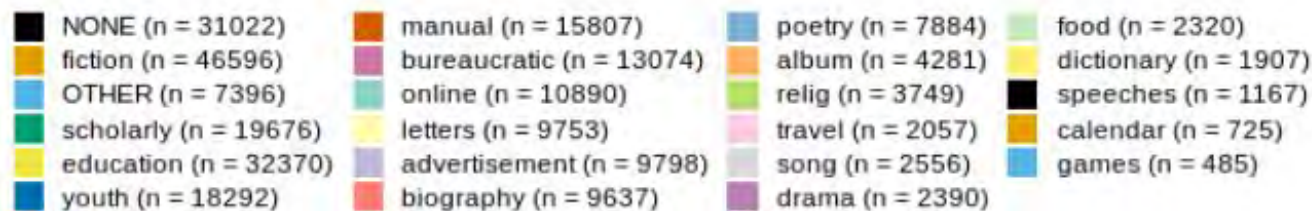
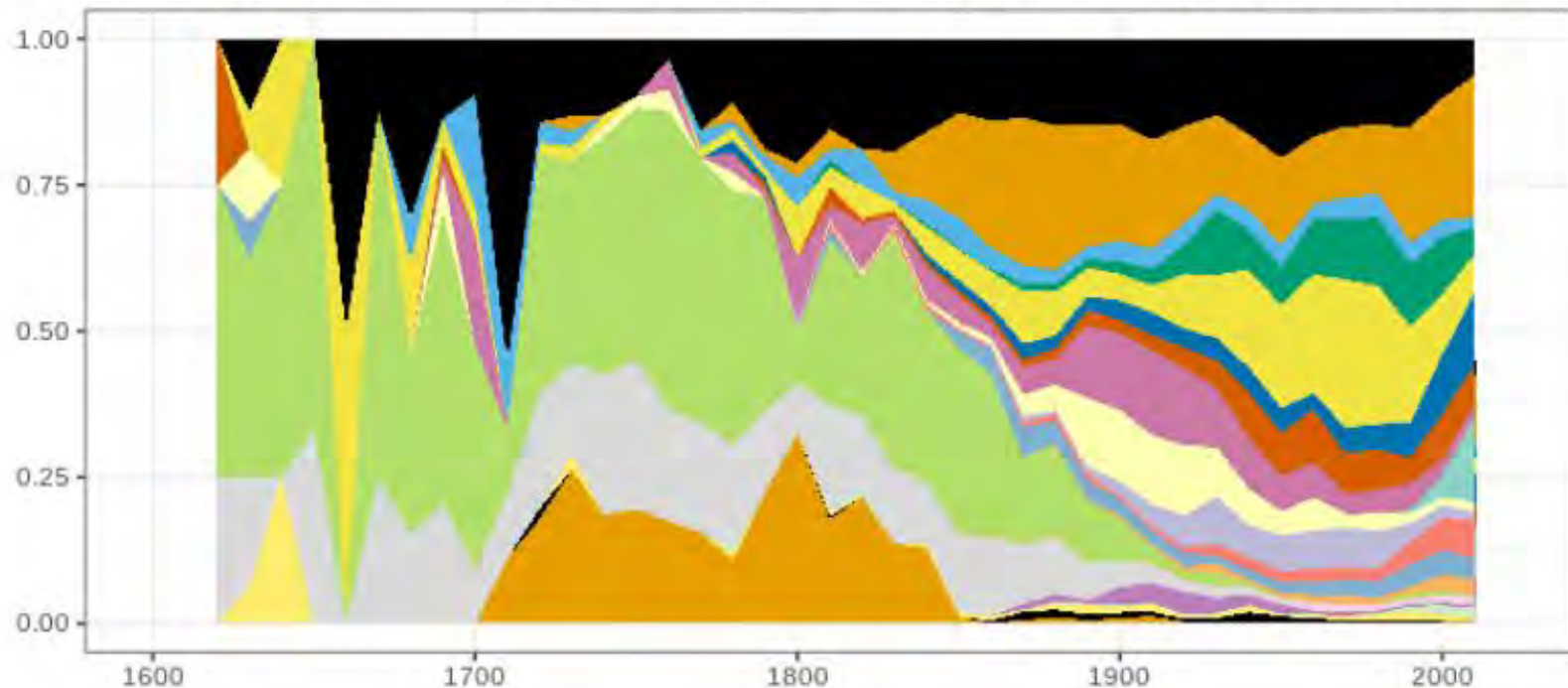
estonian other languages

Languages in Estonian National Bibliography (1600-2020)



estonian (n = 199677) german (n = 17602) latin (n = 4675) danish (n = 1662) polish (n = 553)
russian (n = 37866) english (n = 13453) finnish (n = 2303) swedish (n = 1291) latvian (n = 489)

Genres of Estonian language texts in Estonian National Bibliography (1600-2020)



Case study: the emergence of the Estonian
written language community 1800-1940.

Crash course: social history of the Estonian language

- Very long (recorded) oral tradition
- Relatively new as a literary language
- Tied to social hierarchy:
 - From the Baltic crusades (13th century), German was established as the main language of commerce and culture in the region
 - Increasing overlap between linguistic and social divisions, which in turn reinforced each other
 - Using German was the prerequisite of almost any kind of social mobility
- (Slow) changes in the 19th century:
 - Liberation of peasants (≈ Estonians) from serfdom in 1816/19; additional social reforms during the 19th century, urbanization, industrial development
 - Enlighteners and estophiles among the Baltic German elite
 - National Awakening period (1850s-1870s)
 - Russification (1890s)
- “Coming of age”:
 - Modernist movements in the early 20th century
 - Official language of the Estonian Republic, est. 1918



The “Baltic provinces” of the Russian Empire, 1860
(David Rumsey Map Collections)

Changes to the language community 1800-1940

- From mostly oral and low prestige to a first written language community.
 - Literati focussed on Estonian community
- Mass literacy
 - 1800 - 66% could read, but it was mostly used for religious schooling
 - 1840s-1850s better reading comprehension and learning to write
 - 1897 - 91% read, 77% write
- Urbanization, newspapers, legal documents - written communication becomes common



Estonia & Livonia in the Russian Empire 1800

European context

European written traditions are long, but mostly for the elite.

19c, mass literacy in some places. Huge educational reforms.

Estonian was in Scandinavian-German sphere of influence.

* Measure not fully comparable

Estonians*

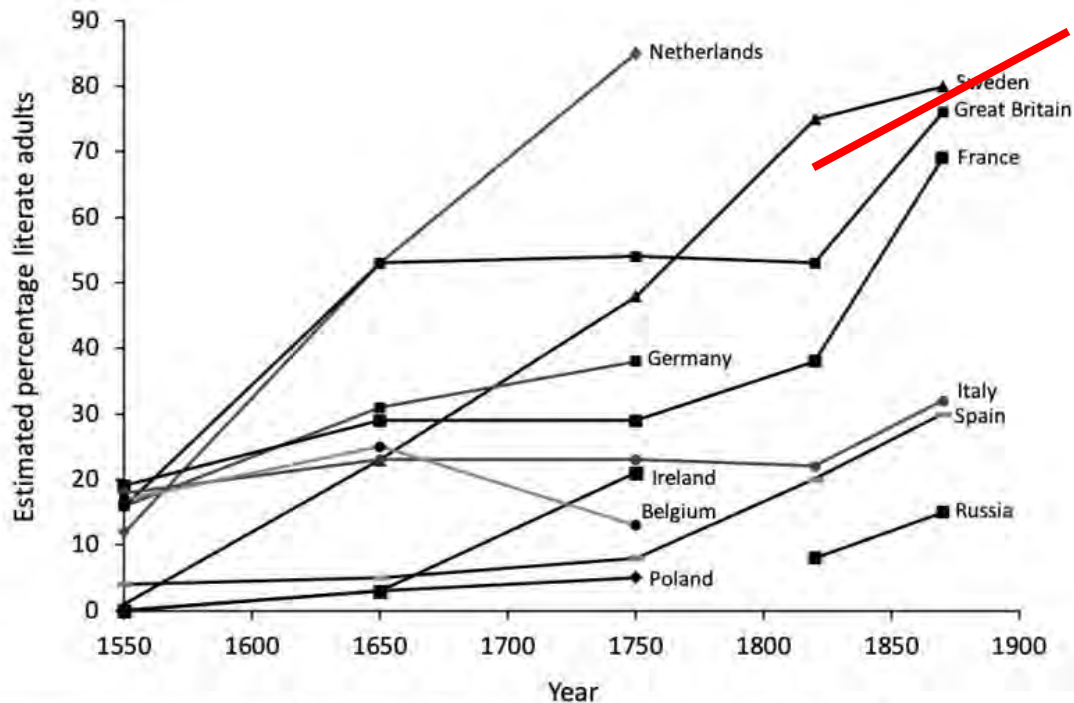


FIGURE P.1. Literacy rates for various European countries from 1550 to 1900. These estimates are based on book publishing data calibrated using more direct measures of literacy.⁸

Figure from Henrich 2020, data from Burigh & Zanden 2009

The emergence of the community

How did the Estonian language community emerge from a German-dominated cultural area?

- Languages of publication and political events
- Influx of new authors
- Language choice for individuals
- Language choice in communities
- The role of schools for authors

Estonian National Bibliography as a data source

Books published is a window to the language community

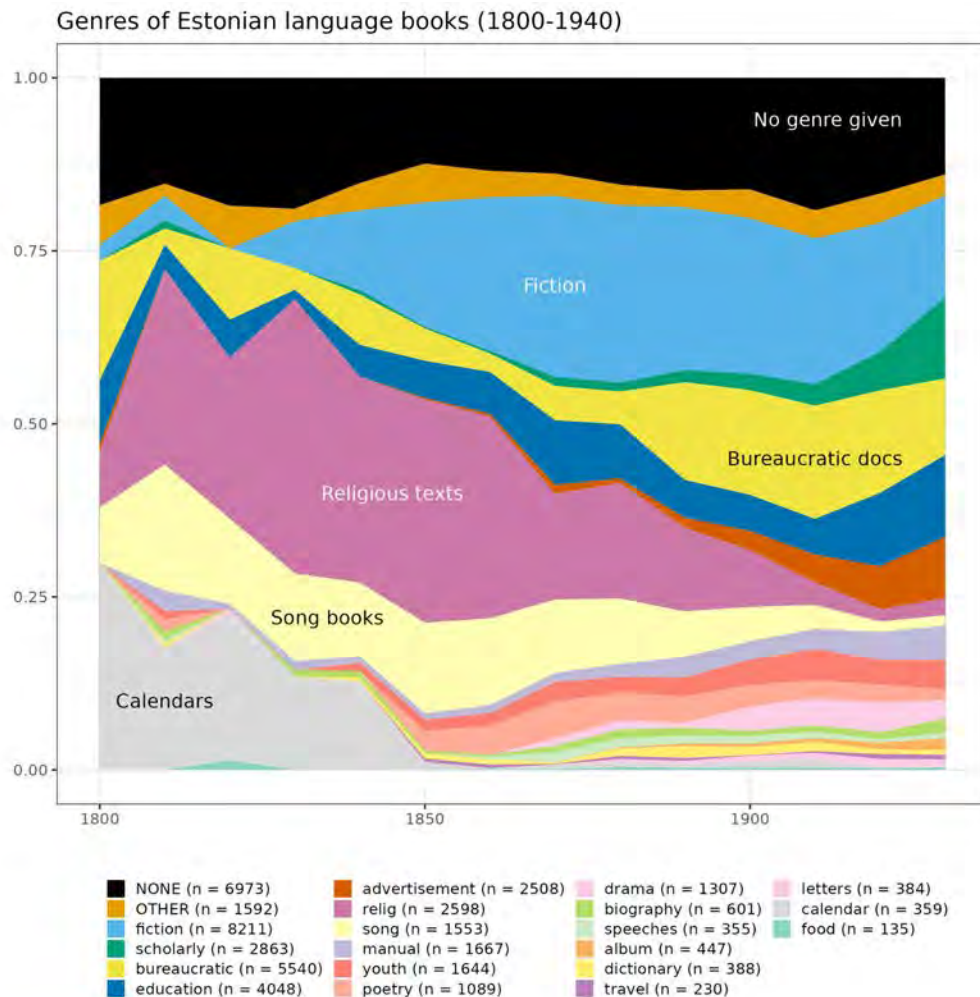
- Though not all of it. There's also newspapers, letters etc.

We can take books published in 1800-1940

- Includes books in Estonian and other languages
- Expected to have fairly good coverage of known books published
- Has data on authors and contributors to each book

Data

- 43,570 works
 - 25,986 Estonian (60%)
 - 17,588 other languages (40%)
- 9768 people
 - Authors, publishers, illustrators etc.
 - 5571 with birth dates (57%)
 - 711 multilingual authors (16%)
- Harmonized and cleaned author names and book metadata.
- Augmented metadata by detecting language from book titles.

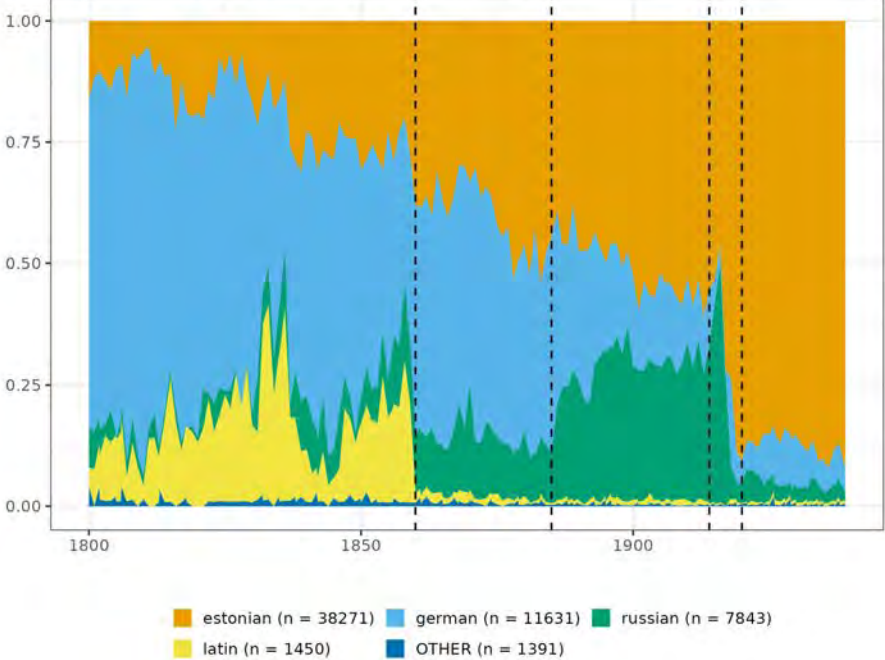


Emergence of the Estonian literary community

Number of books in Estonian and other languages (1800-1940)



Share of language in books (1800-1940)



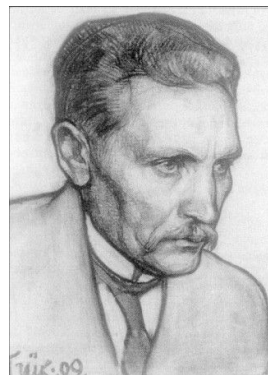
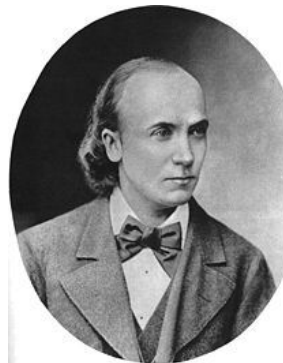
Contributors on the books

Each book has a list of contributors

- Writers, illustrators, editors, publishers
- They are given as full name with years

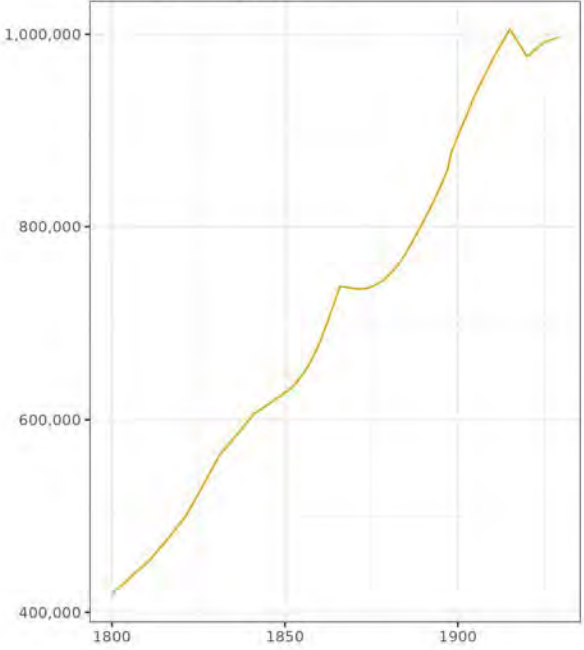
We can look at who was involved and when

- E.g.
 - Friedrich Brandt (1830-1890), publisher (n = 206)
 - Matthias J. Eisen (1857-1934), folklorist (n = 327)
 - Johannes Käis (1885-1950), teacher (n = 242)
 - Richard Kivit (1888-1981), illustrator (n = 450)

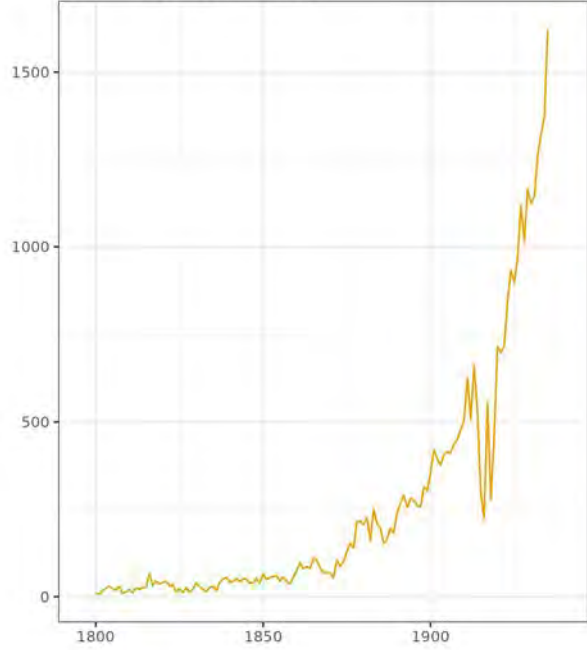


Population growth and publishing trends

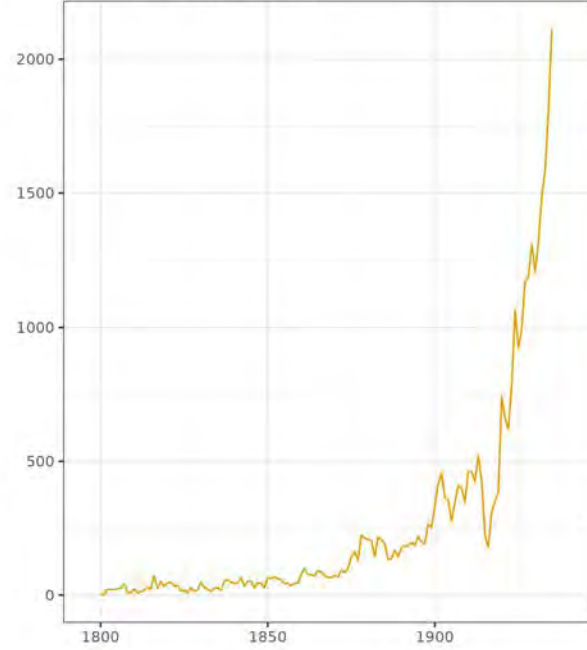
Estonian native speakers (1800-1930)
Residing in modern Estonia



Estonian language books
Per million native speakers

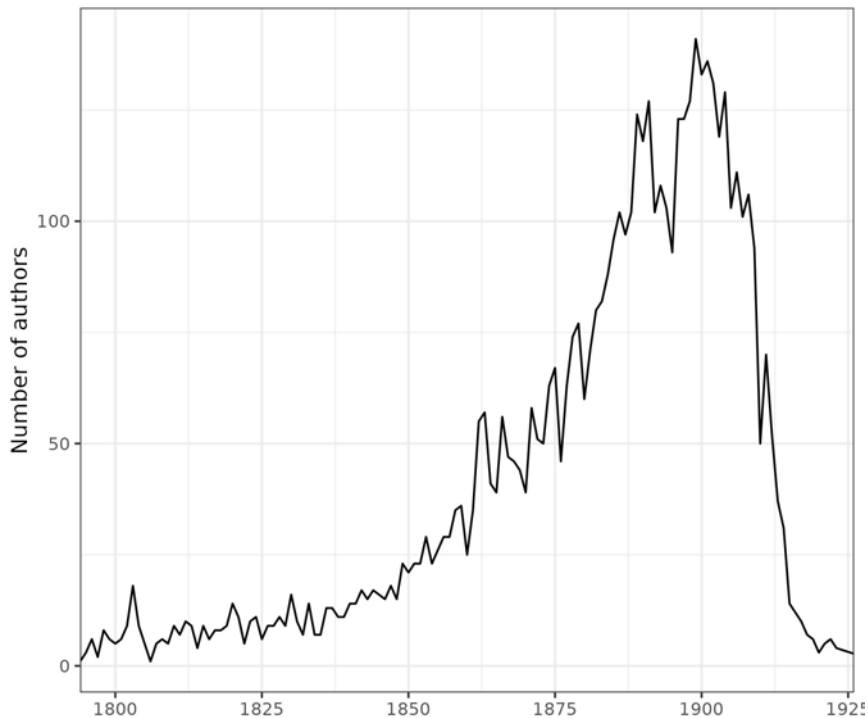


Estonian literary actors
Per million native speakers

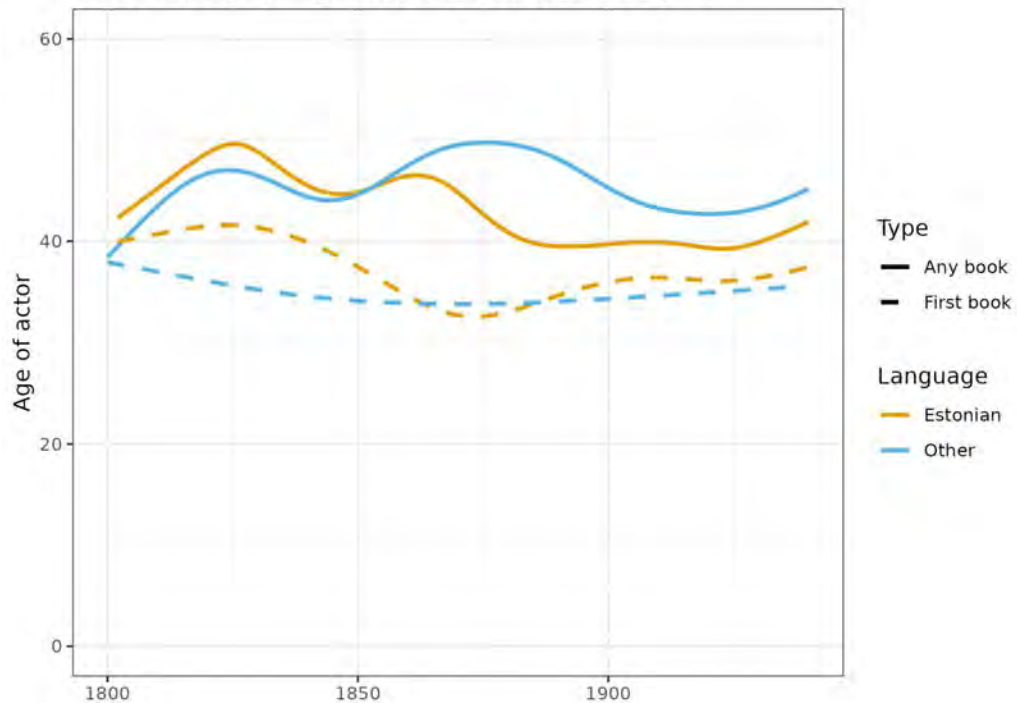


New people to written language community

Birthyears of actors active in 1800-1940

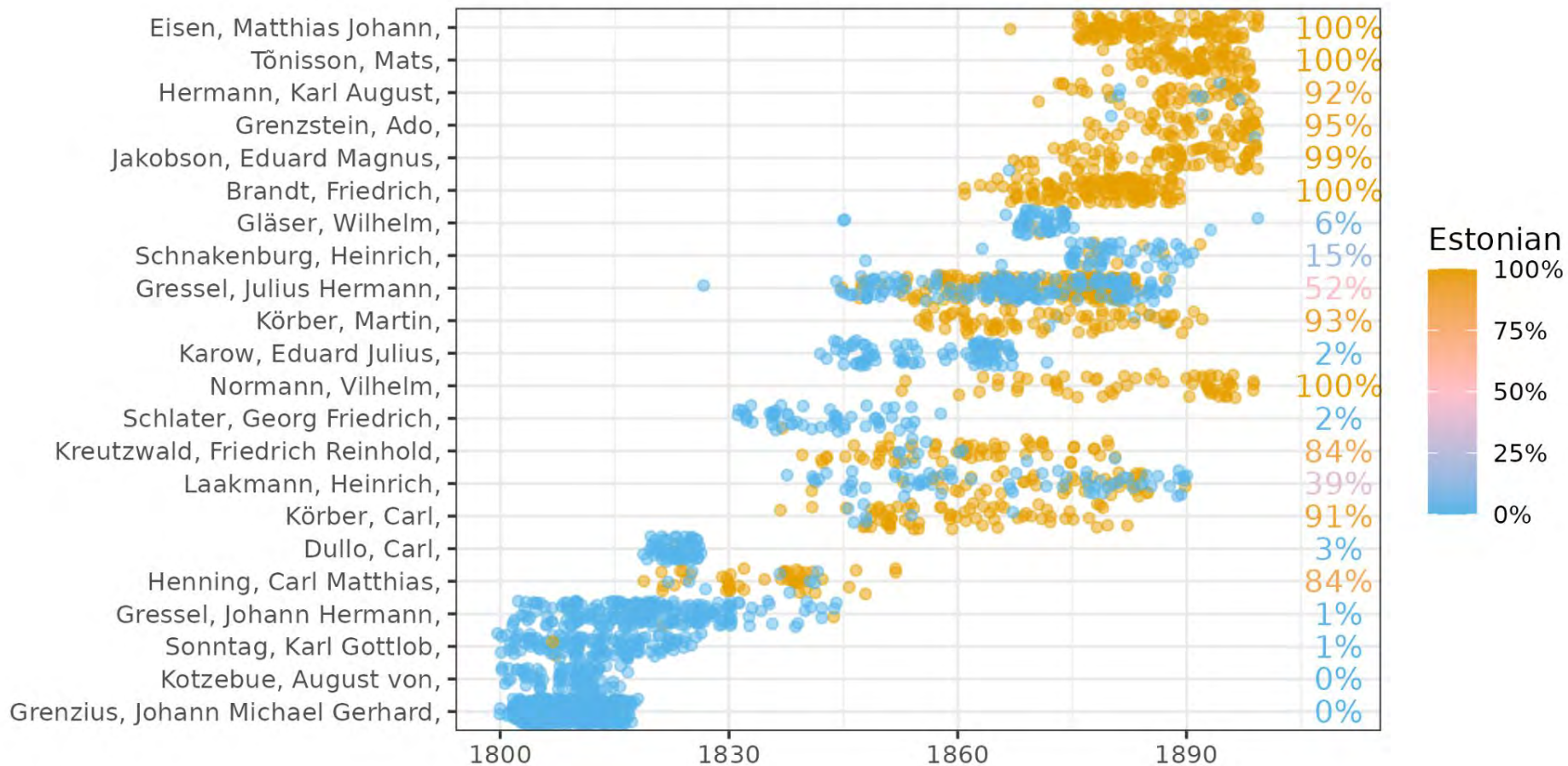


Age of average active living actor by language (1800-1940)
Across all books (—) and for first books published (- - -)



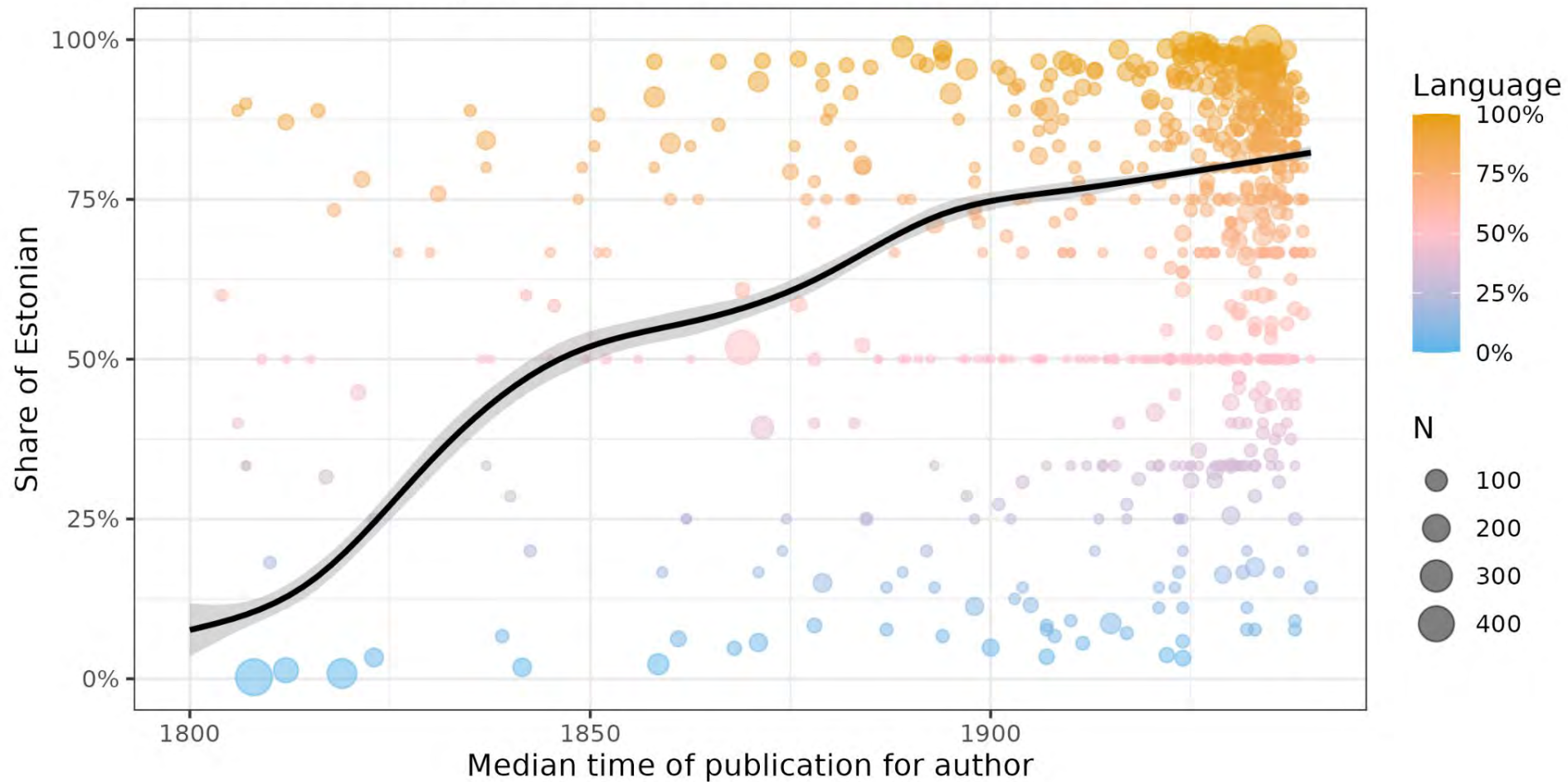
Language of books during actor's life

Actors with > 50 works, lifetime share on the right



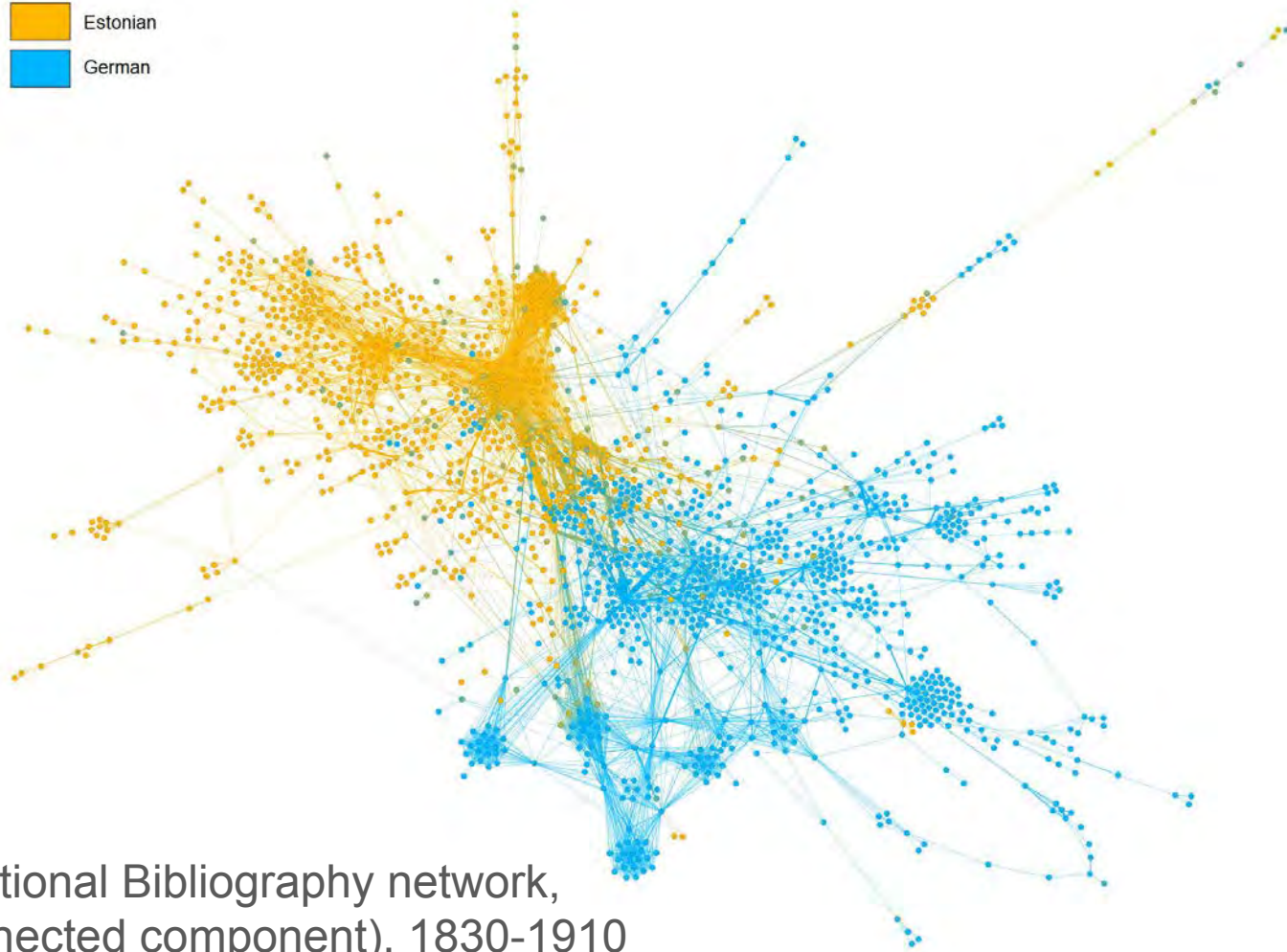
Language choice for multilingual actors (n = 711)

Lifetime share of language for actor vs median time of publication



Networks

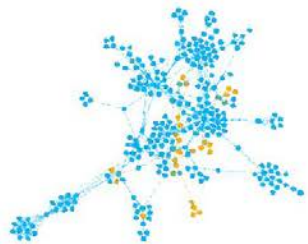
- The National Bibliography data can also be viewed as a literary network, where nodes are actors and edges are works (e.g. Hill *et al* 2019)
- Shows the relationships between authors, publishers, illustrators, editors and others
- Can be constructed diachronically and be coded to represent attributes like language



Estonian National Bibliography network,
(largest connected component), 1830-1910

Largest connected components

1830-1849



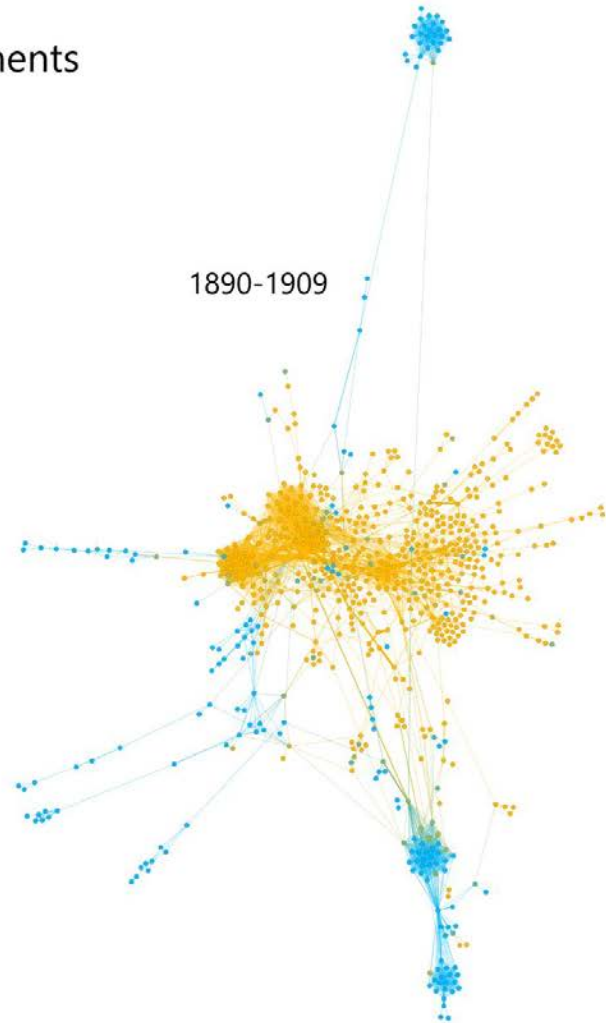
1850-1869



1870-1889

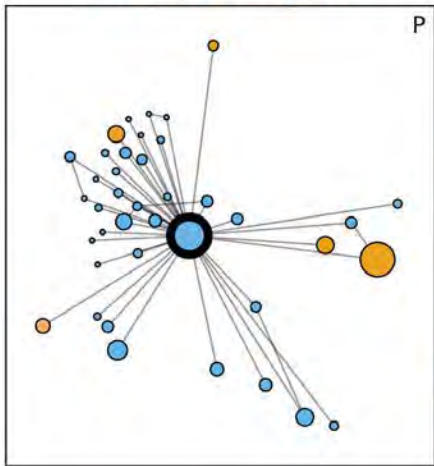


1890-1909

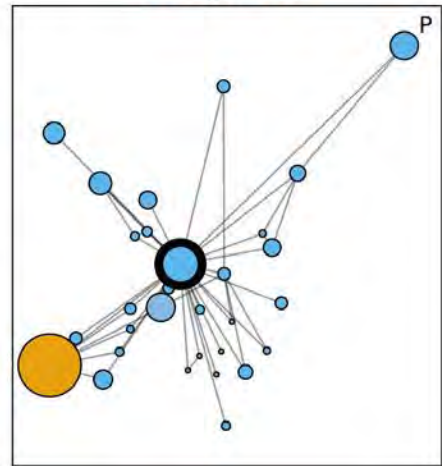


Ego networks (radius=1) of most central nodes by betweenness, 1850-1869

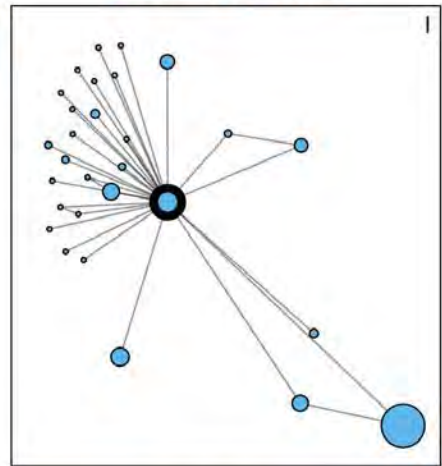
1. Karow, Eduard Julius



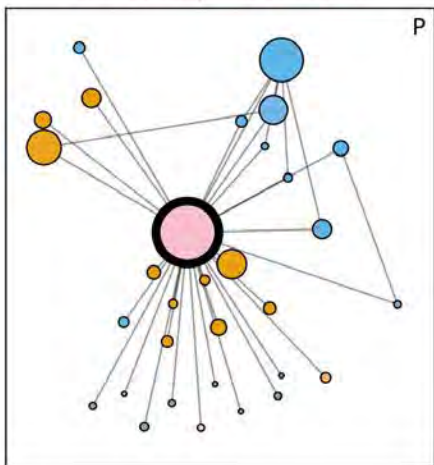
2. Kluge, Franz



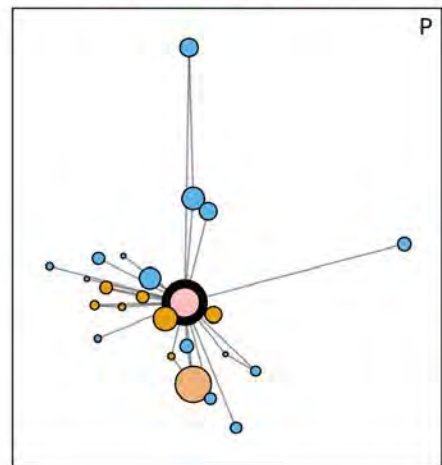
3. Schulz, Carl Anton



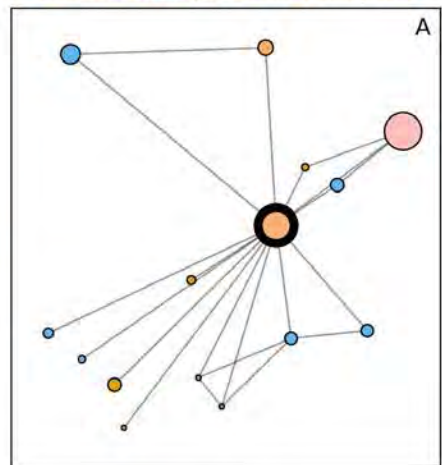
4. Gressel, Julius Hermann



5. Laakmann, Heinrich

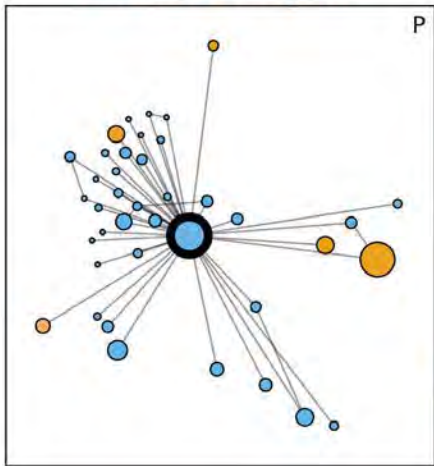


6. Kreutzwald, Friedrich Reinhold

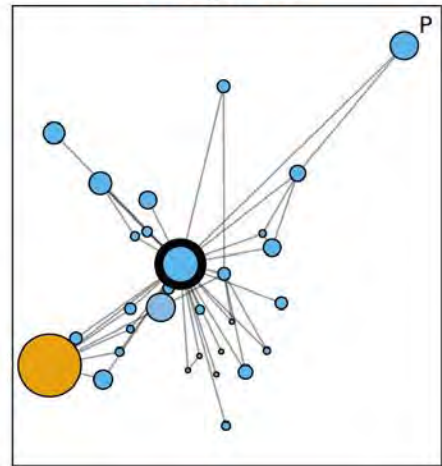


Ego networks (radius=1) of most central nodes by betweenness, 1850-1869

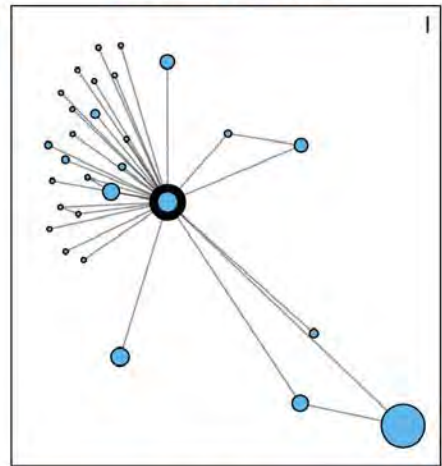
1. Karow, Eduard Julius



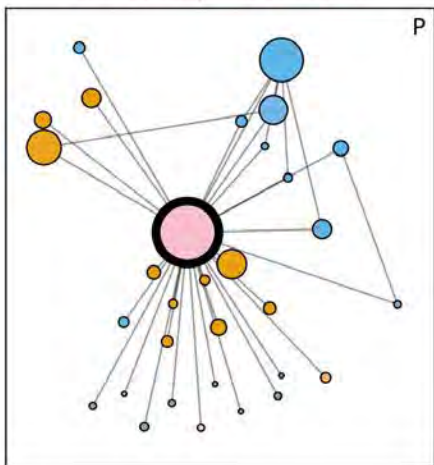
2. Kluge, Franz



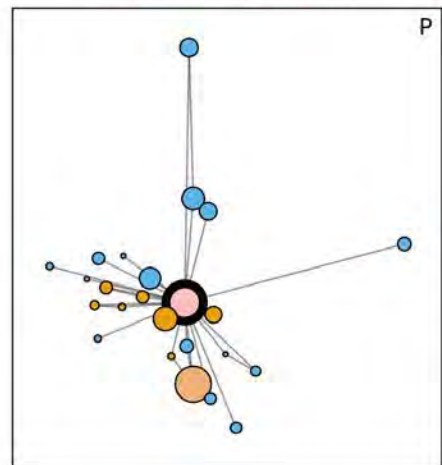
3. Schulz, Carl Anton



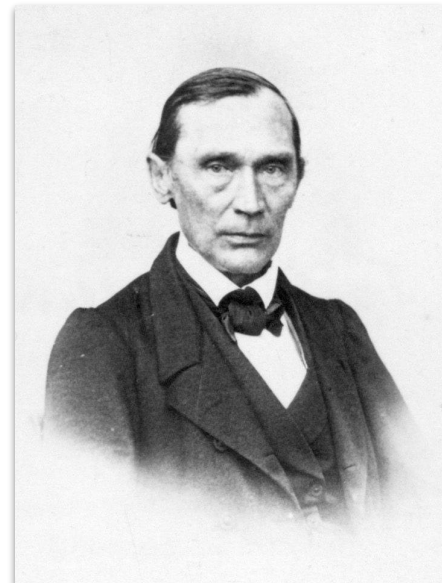
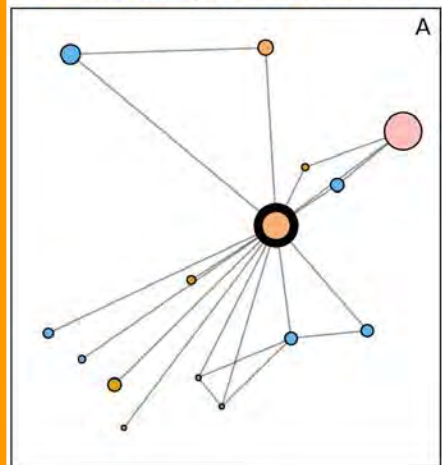
4. Gressel, Julius Hermann



5. Laakmann, Heinrich



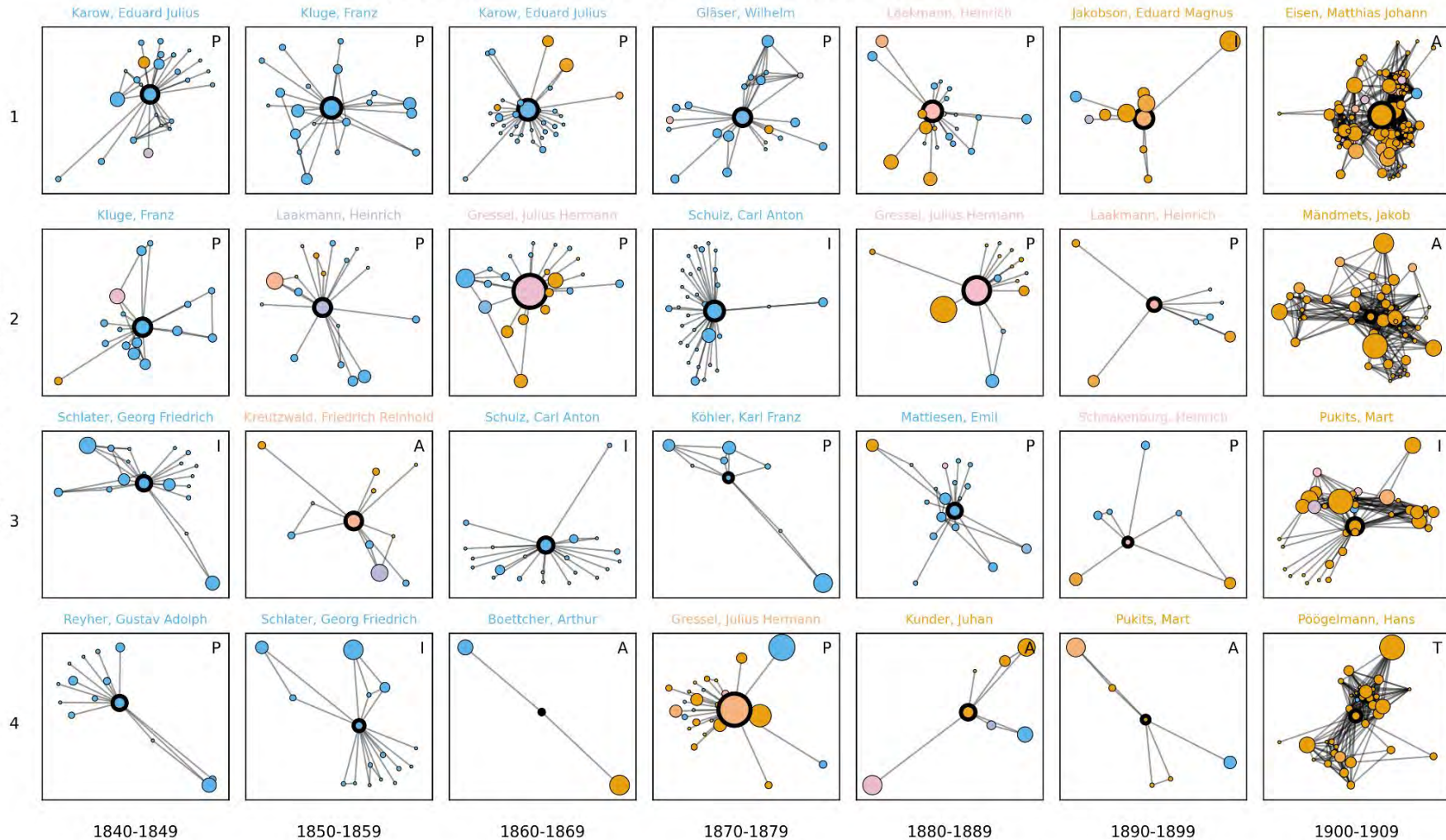
5. Kreutzwald, Friedrich Reinhold



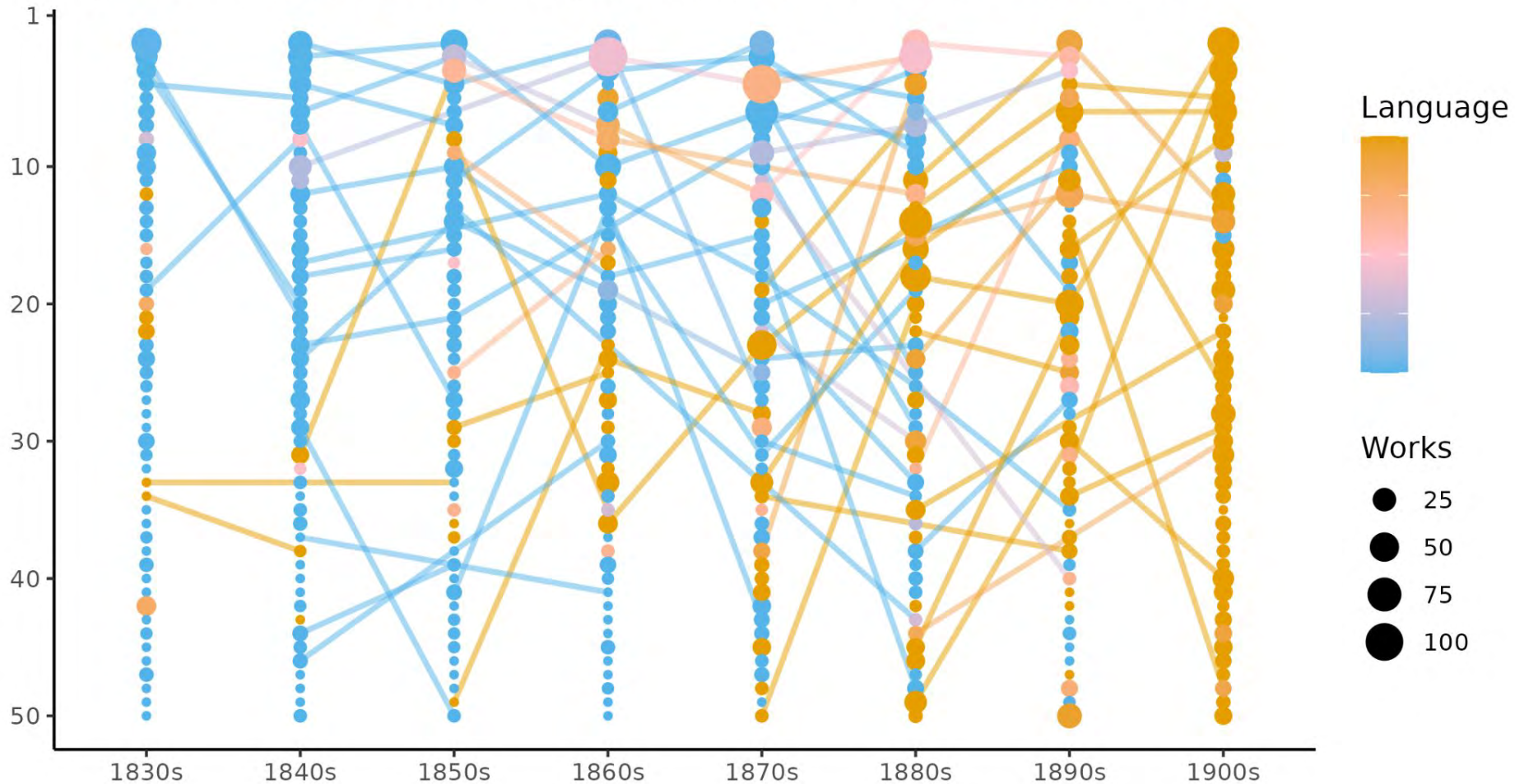
- Leading figure in the early National Awakening
- Author of the Estonian national epic "Kalevipoeg"
- Published mainly in Estonian, but also German and other languages

Ego networks (radius=1) of most central nodes, 1840-1910

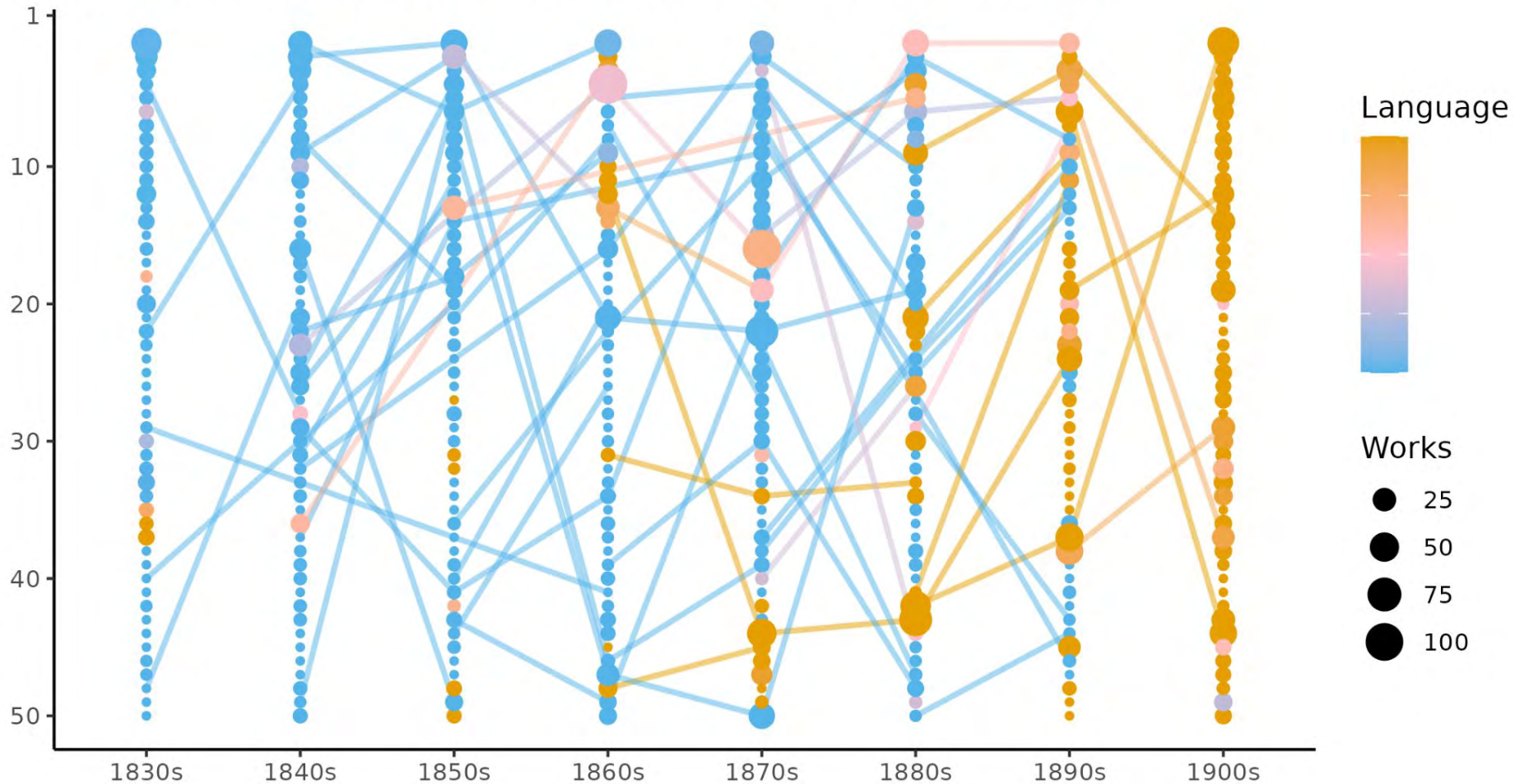
betweenness centrality rank



Top 50 most central nodes by decade, ranked by betweenness



Top 50 most central nodes by decade, ranked by closeness



Linked biographies

Many contributors have a link to VIAF

Some more data in other datasets



Julius Aamisepp (Q1295523)

in more languages [edit](#)

Language	Label	Descriptor	Also known as
English	Julius Aamisepp	(S) www.aamisepp.ee	
Estonian	Julius Aamisepp	(S) www.aamisepp.ee	Julius Aamisepp
Russian	Аамисепп, Юлий Цезаревич	(S) www.aamisepp.ee	
Finnish	Juulius Aamisepp	(S) www.aamisepp.ee	

in entered languages

Statements

instance of	human	edit
-------------	-------	----------------------



Linked biographies

- Gathered birth places from various datasets
- Geo-coded them for coordinates



001	ERRRja11124591	ÜMEL:cutted
003	ERRR	
006	960322	pehnenasteen ja
006	ja	(ERRRja11124591
040	ja	ERRR to est za: ERRR to EITUR
043	za	err
100	1	za Aamisepp, Julius, ja 1883-1960
400	1	za Aamisepp, Julius
600	1	za Aamisaam, Kiney, ja 1883-1960
620	1	za EE, 1 ka, 1960; EE, 14 ka, 2000
680	1	za Eesti põllumajandusteadlase, sordistaja, põllumajandus (1947)



Arhiiv: Arhiiv | Vastu | Muuda | Muuda lähtekoti | Täida ajalugu | Osta | Vajadustel

Julius Aamisepp

Julius Aamisepp (1. september (vkj 20. august) 1883 Harjumaal, Kloostri vakt, Karlepa, Karlepa-Tõnu talu – 10. jaanuar 1950 Jõgeva) oli eesti sordistaja.

Eesti biograafiline andmebaas ISIK

avalehekülg · hindekirj ja otsing · ametikohad [110] · asutused [470] · isikud [4486] · kaastöö välistajatele [1133] · nime eeskujud [927] · pöördnumbrid [9179] · sugulased [309] · sünnikohad [625] · surmakohad [166] · tegevusala [599] · lühendid ja allikad · admin

AAMISEPP, JULIUS

Nimi	Aamisepp, Julius
Nime normitud kuju	Aamisepp, Julius
Eluajaatumid	1883-09-01
Sünniaeg	1883-09-01
Sünnikoht	Karlepa-Tõnu v. Kloostri v. Harjumaal.
Surmaaeg	1950-01-19
Surmakohat	Jõgeva
Märkus	maetud Tartu Raadi kalmistule
Üldandmed	Sordistaja
Tegevusala	1883-1960 Harjumaal Kloostri valla Karlepa-Tõnu, Taluperemees Siim A., Liisa Vrajer. Aa 1917 Anna Maria Volmer Ants (1918-19), Vilho Zasius (1920-43), Ilmar (1927-92), v.a. 1883-1960 Harjumaal, Kloostri v. Karlepa-Tõnu talu Jõgeva, tegevdalad, vangistatud, põllumajandusteadlane, sordistaja, põllumajandus (1947)
Biograafilised andmed	1883-1960 Harjumaal Kloostri valla Karlepa-Tõnu, Taluperemees Siim A., Liisa Vrajer. Aa 1917 Anna Maria Volmer Ants (1918-19), Vilho Zasius (1920-43), Ilmar (1927-92), v.a. 1883-1960 Harjumaal, Kloostri v. Karlepa-Tõnu talu Jõgeva, tegevdalad, vangistatud, põllumajandusteadlane, sordistaja, põllumajandus (1947) Keskbüroo juhatusel liige, Harju maakonnakõukogu liige ja sekretär 1917-20, 1918 Eesti diviisi staabi mobilisatsiooniosakonna asjaajaja, Vabariikliku ülemjuhataja staabi mõjandusülem, Eesti Sordiparenduse Seltsi Jõgeva sordikasvatuse osakonnajuhataja 1920-50, "Väikeeloma-kasvataja" toimetaja 1919-21, ENSV IA korrespondentile 1946, ÜN 1947-50 (II ko); ENSV teenelise teadlane 1945, NE preemia 1947, Stalini preemia 1948. Surnud 19. I 1950 Jõgeval, maetud Tartu Raadi kalmistule. EAT, EE, EBLI, EAT2, ENEL, ENES, EABL, VJv, ENES(14), ETeadbl. – ISOTAMM 2

Julius Aamisepp (Eesti keeles)

→ Vaata ka lingitud lingid

Language	Label	Description	Also known as
English	Julius Aamisepp	Swedish-speakin	
Estonian	Julius Aamisepp	Swedish-speakin	Julius Aamisepp
Finnish	Aamisaam, Kiney, Oskari	Swedish-speakin	
Finnish	Aamisaam, Kiney, Oskari	Swedish-speakin	

All entries require login

Statements

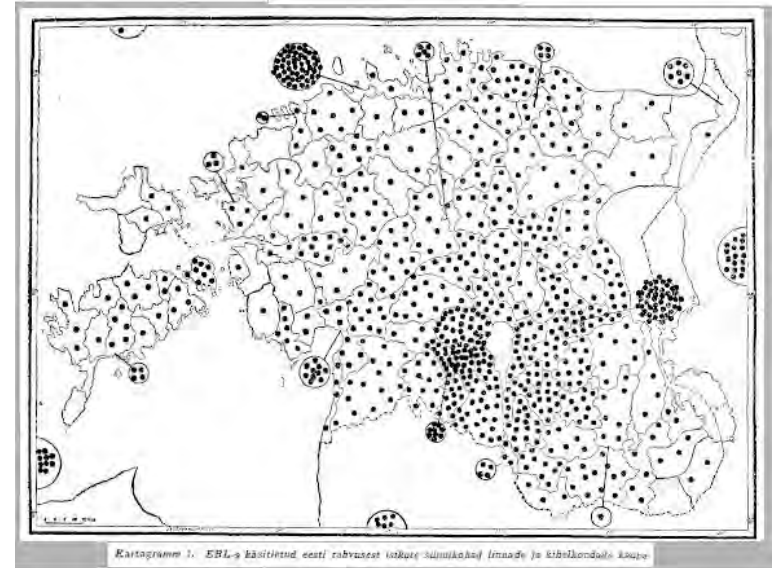
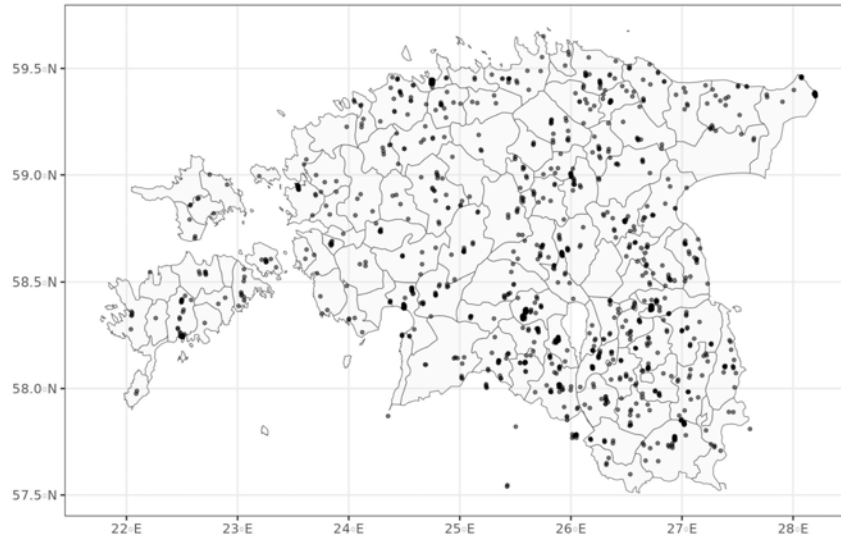
Instance of

Number

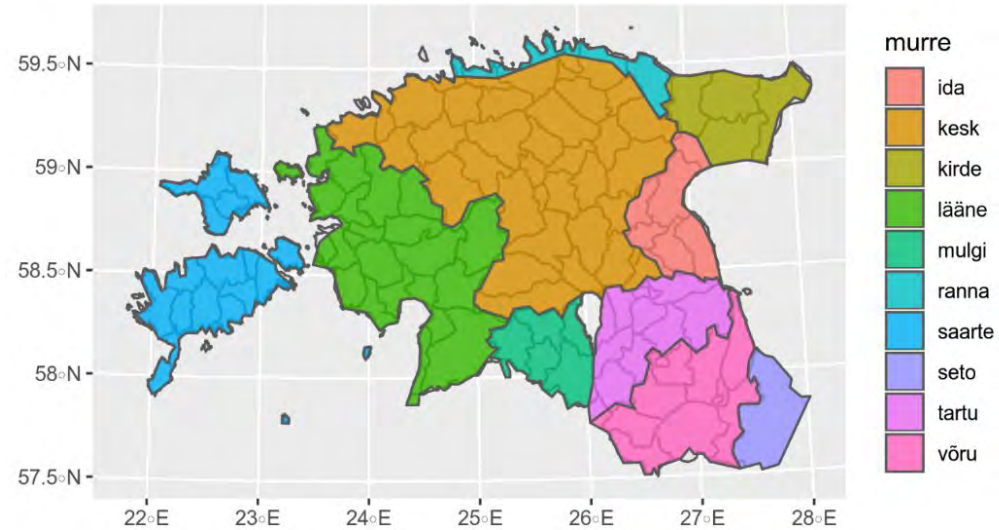
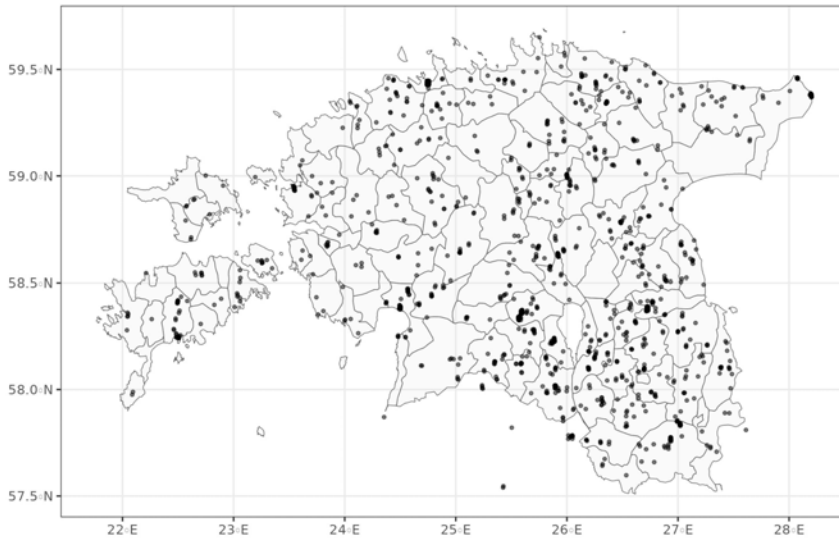
m	Aamisepp, Julius
slecht	männlich
e	Pressearchiv des Herder-Instituts Marburg Wikipedia (Stand: 25.09.2018): https://en.wikipedia.org/wiki/Julius_Aamisepp
	Lebensdaten: 1883-1950
	Estland (XA-EE)
rafischer	Geburtsort: Karlepa (Gem. Lääne-Harju)
bezug	Sterbeort: Jõgeva
Beruf(e)	Agrarwissenschaftler

100	1	za Aamisepp, Julius, ja 1883-1960
376	1	za 1 ja 200211
500	1	za (DE-1911)096201100 10 (DE-686)1066201100 10 http://is-nb.info/standards/elementset/igndb:all/Relationship zw / D Beziehung familiaer zw / V Sohn
548	1	za 1883-1960 14 dtd 14 http://is-nb.info/standards/elementset/igndb:all/AndDeath zw / D Lebensdaten
548	1	za 01 05 1883-19 01 1960 24 dtd 24 http://is-nb.info/standards/elementset/igndb:all/AndDeath zw / D Exakte Lebensdaten
550	1	za (DE-101)04200164X 10 (DE-686)4200164-1 10 http://is-nb.info/standards/elementset/igndb:all/AndDeath zw / D Charakteristischer Beruf
650	1	za (DE-101)041779436 20 (DE-686)177943-6 20 http://is-nb.info/standards/elementset/igndb:all/AndDeath zw / D Beruf
661	1	za Karlepa 24 ortg 24 http://is-nb.info/standards/elementset/igndb:all/AndDeath zw / D Geburtsort zw / V Gem. Lääne-Harju

Mapped birthplaces of the contributors

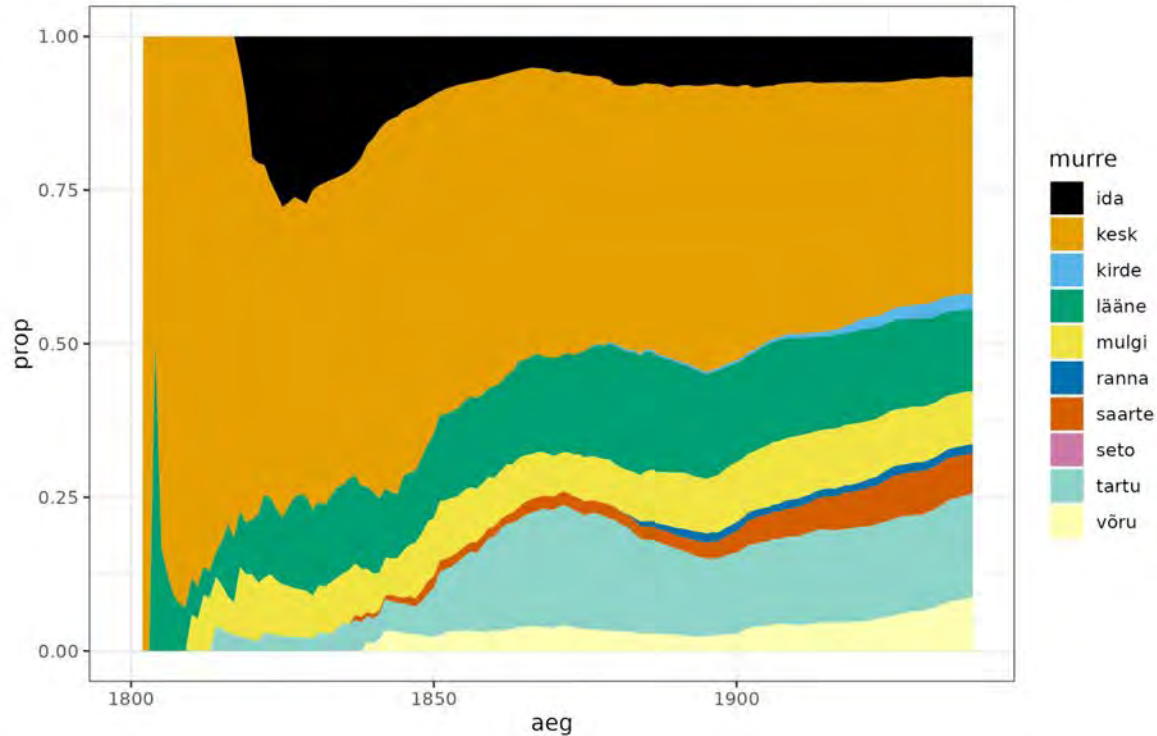


Mapped birthplaces and dialects



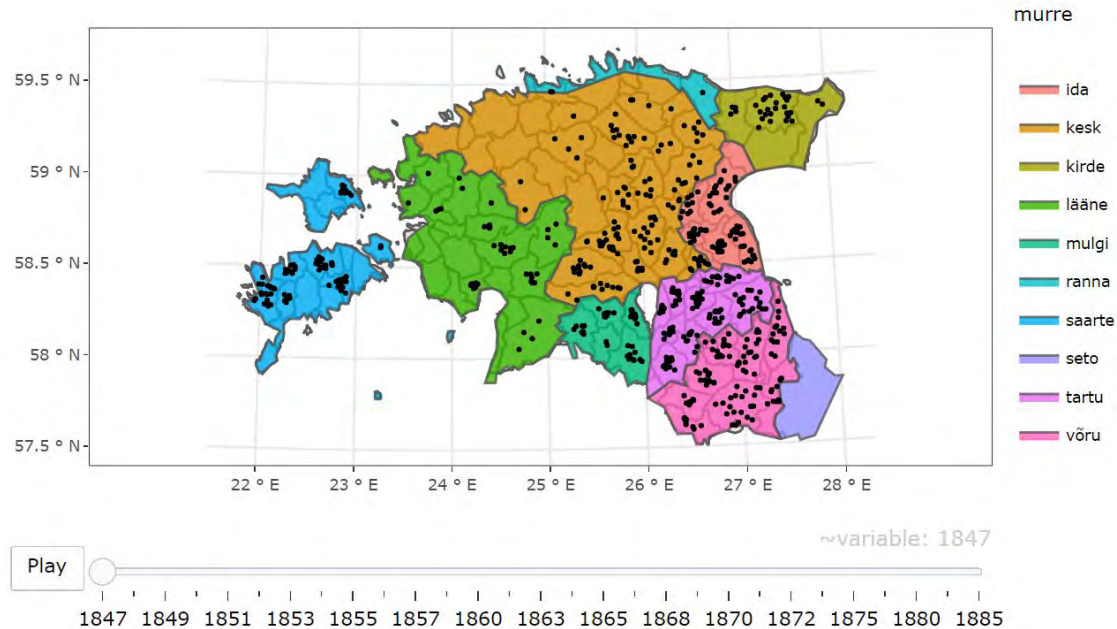
Share of dialects in the dataset

Cumulative share of different dialects of Estonian 1800-1940



Public schools in 19c

- Are founded from 1830s to 1880s



Estonia & Livonia

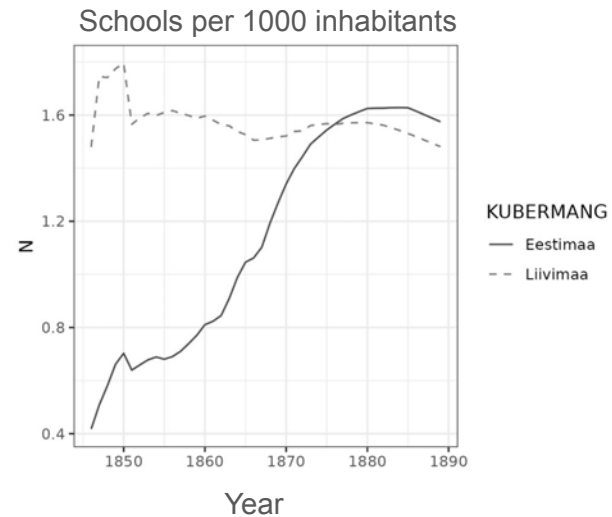
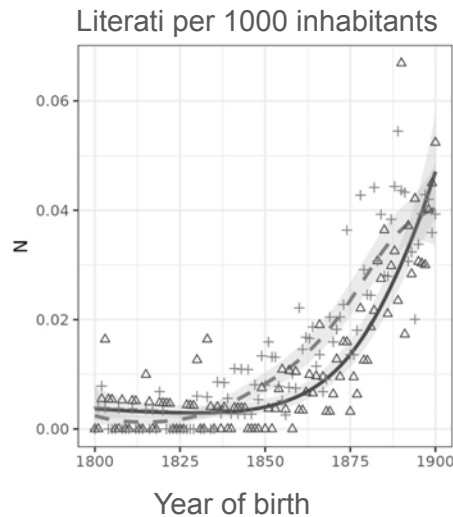
- Administrative rules differ
- Livonia was founding schools quicker, and peasants grew quicker in wealth
- Influence on written language



Origins of the contributors

Odds of a child born to join the literati in their lifetime (~3.3 kids per 1000)

- 1850
 - 1:800 Livonia
 - 1:1600 Estonia
- 1875
 - 1:400 Livonia
 - 1:800 Estonia
- 1900
 - 1:200 Livonia
 - 1:200 Estonia



Results

Research questions

- We can see shifts in language use as response to major historical events
- The 19th century Estonian literary community is shaped by young new actors
- The Estonian community initially emerges *from within* the existing Baltic German dominated network - not alongside it
- Schools played an important role in supporting it

Adapting bibliographic data for research use

Has a strong research potential

- Needs harmonization, processing
- Benefits hugely from linking to other datasets
- Needs overviews of what is in the data, and what is the coverage
- We're building a curated dataset with this in mind

Adapting digital archives for research use

Keep the researcher in mind

- Access
- Overviews and contents
- Tools to explore
- Case studies, case studies, case studies
-
- ... repeat

Thank you!

Team at the University of Tartu



UNIVERSITY OF TARTU

Centre for Digital Humanities
and Information Society

Team in the National Library of Estonia



La
la

Literature

- Leo Lahti, Jani Marjanen, Hege Roivainen & Mikko Tolonen (2019) Bibliographic Data Science and the History of the Book (c. 1500–1800), *Cataloging & Classification Quarterly*, 57:1, 5-23, DOI: [10.1080/01639374.2018.1543747](https://doi.org/10.1080/01639374.2018.1543747)
- Mark J Hill, Ville Vaara, Tanja Säily, Leo Lahti & Mikko Tolonen (2019) Reconstructing Intellectual Networks: From the ESTC's bibliographic metadata to historical material. In: *Proceedings of the Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 4th Conference.*