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

Peeter Tinits
08.02.2024
Helsinki Digital Humanities Research Seminar

Outline

0. Thinking about researchers and archives
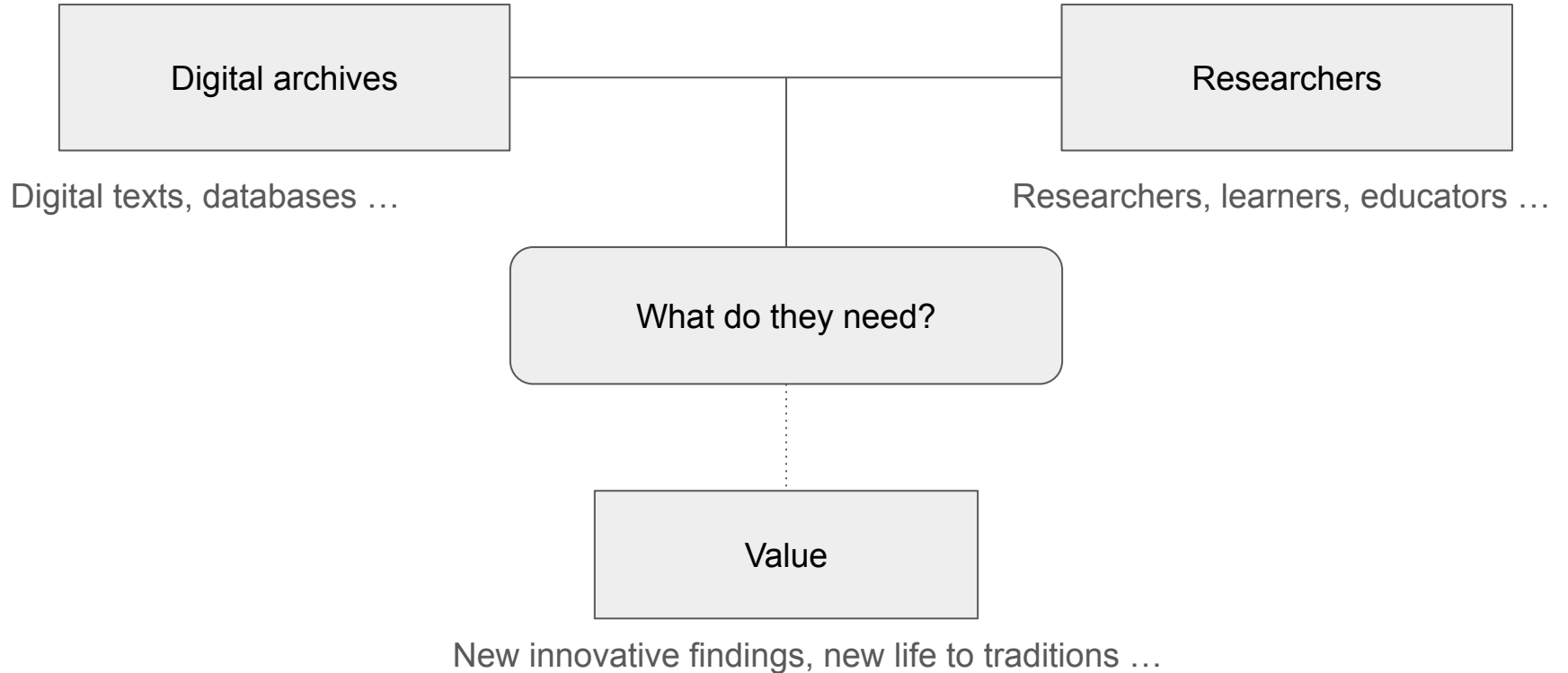
1. Digitized newspapers
   a. What is in
   b. Data explorers
   c. Example cases
2. Bibliographic data
   a. What is in it
   b. Longer case study: emergence of Estonian written language community 1800-1940
3. Discussion

The use of digital archives

Digital archives

Digital texts, databases …

Researchers

Researchers, learners, educators …

What do they need?

Value

New innovative findings, new life to traditions …

Thinking about the use

Perspective from data holders --- Perspective from researchers

1. Digitized newspapers
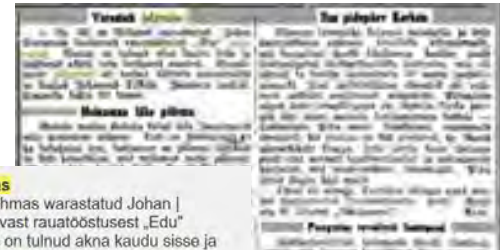
Brief historical overview

- 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- wöi ülem kalingorköites 1000 m. Eesti loodus. 8 iseseiswat artiklit Eesti ranniku, mere, geolo
i wärwikiillasemat, meeleolurikkamat ja sügawamat loodusilu näinud. Ka kaasreisijate näod kõnelewad sedasa
See üsna magus lause on aga wabandataw: on nähtud loodust Antoni kaudu, ja tema psüholoogiale lause wastab
a oma office'ist wõi büroost wälja sõidab wabasse loodusesse uut jõudu koguma. Kes kaugemale sõita ei jõua
 ole suutnud kõigutada isegi ilmasõda. Siin wabas looduses andub inglane täieliku innuga oma armsale lõbus
rk. Füüsilise maateaduse ülesannete kogu, I wihk «Looduse maatlussd"". Hind 85 markm Maateaduse ülesannete
```

Study design

Data balance
(e.g. corpus size)

Data quality
(e.g. OCR)

Representativeness
(do we have enough?)

Materials to answer the question

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
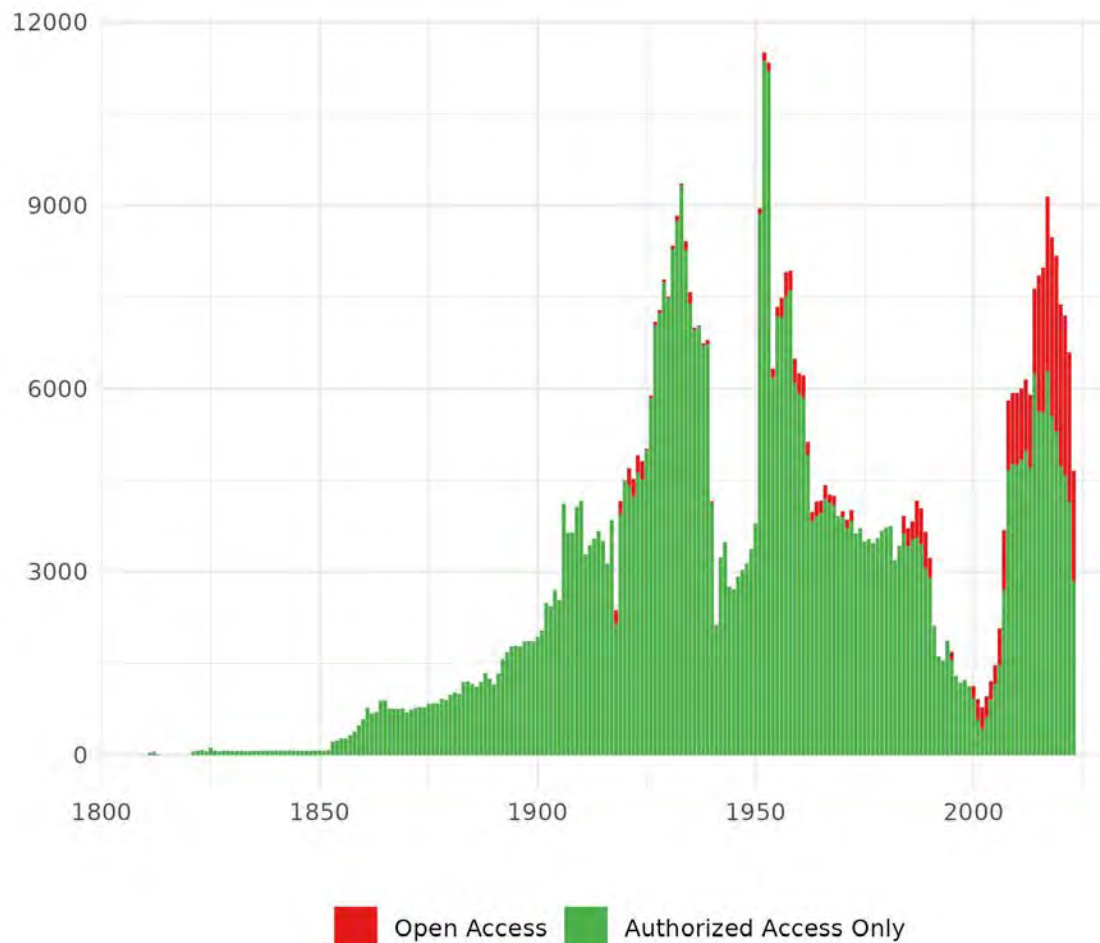
Which newspapers have been digitized?

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 kollektsioonis (Üle 10 aasta ilmumist, n = 117)

linktype
- NA - No url given
- DIGAR
- Other
- ut-dspace

Ilmumisaasta

1800   1850   1900   1950
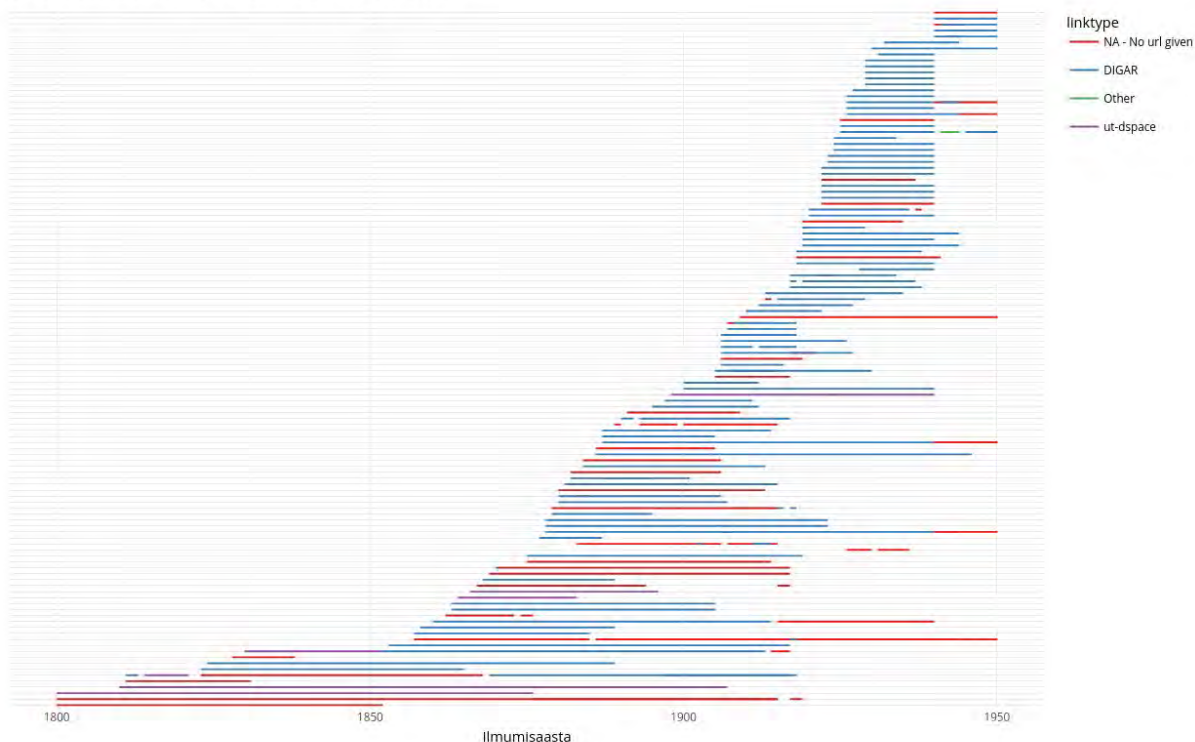
Which newspapers have been digitized?

An example - newspapers from 1920-1940 published for at least 10 years
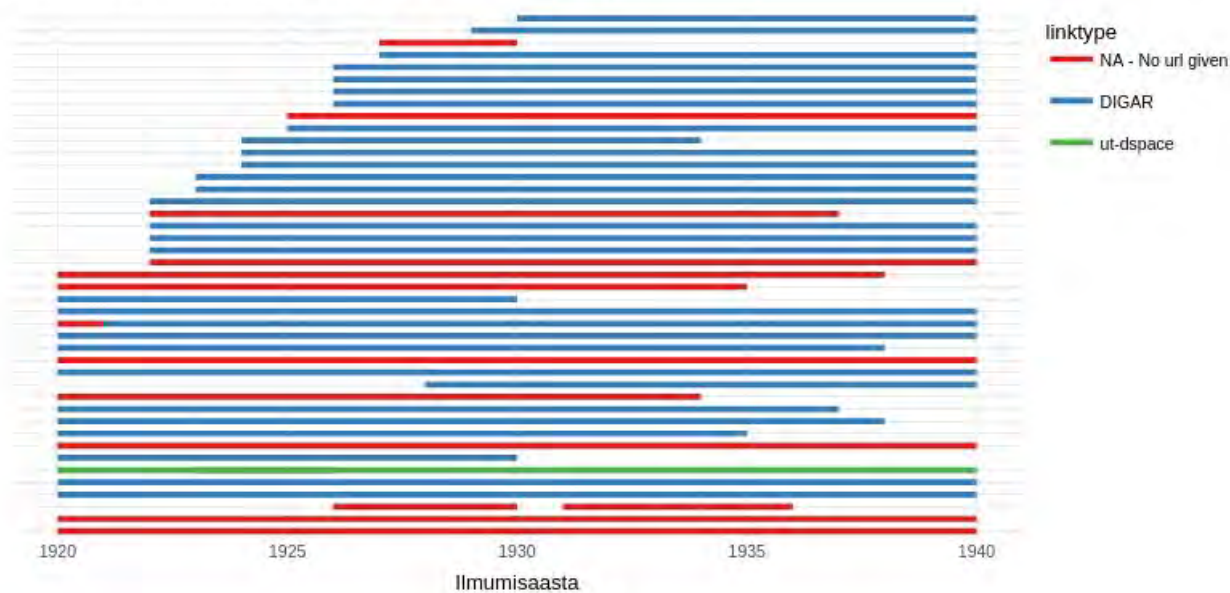
How much has been digitized?

<u>Coverage of digitization</u>

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
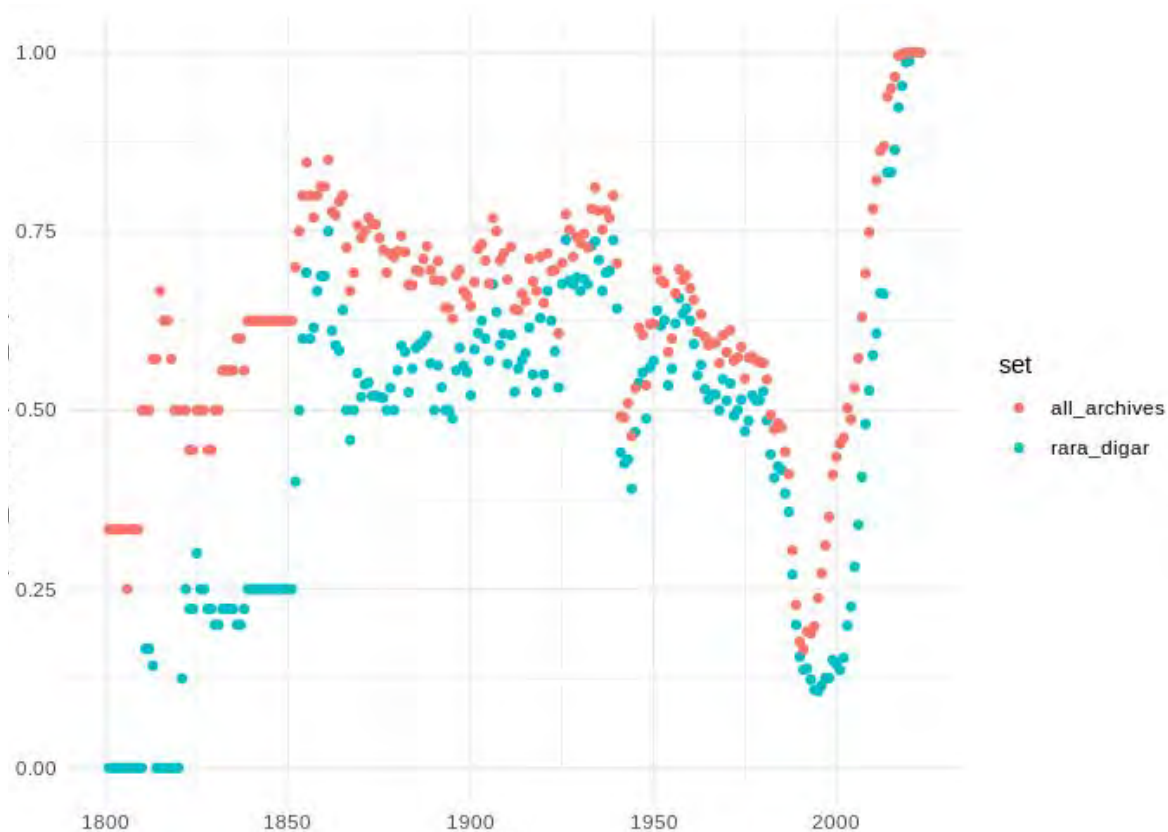
What is in the dataset?

Metadata explorer

Explore dataset with
your own parameters
in mind.

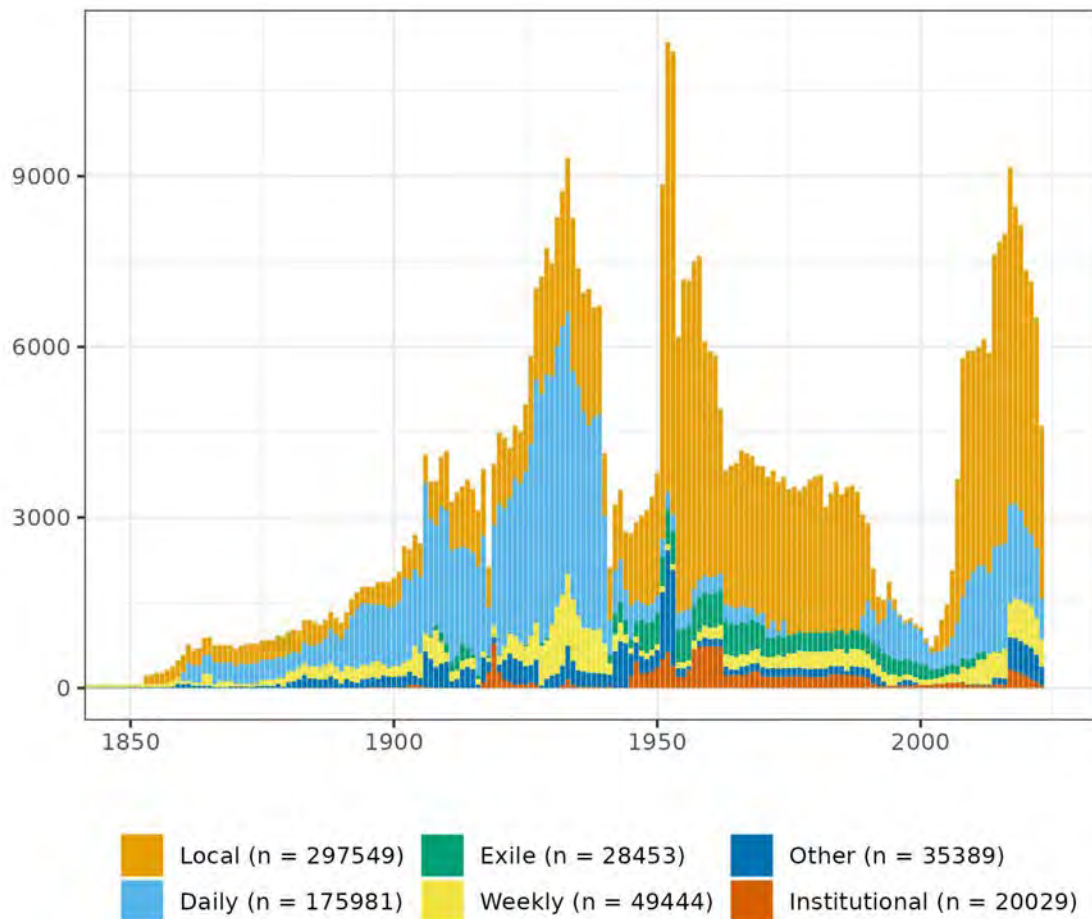Interactive tool
https://digilab.rara.ee/
en/tools/newspapers-
metadata-browser/

What's in the dataset?

Overview by publication type

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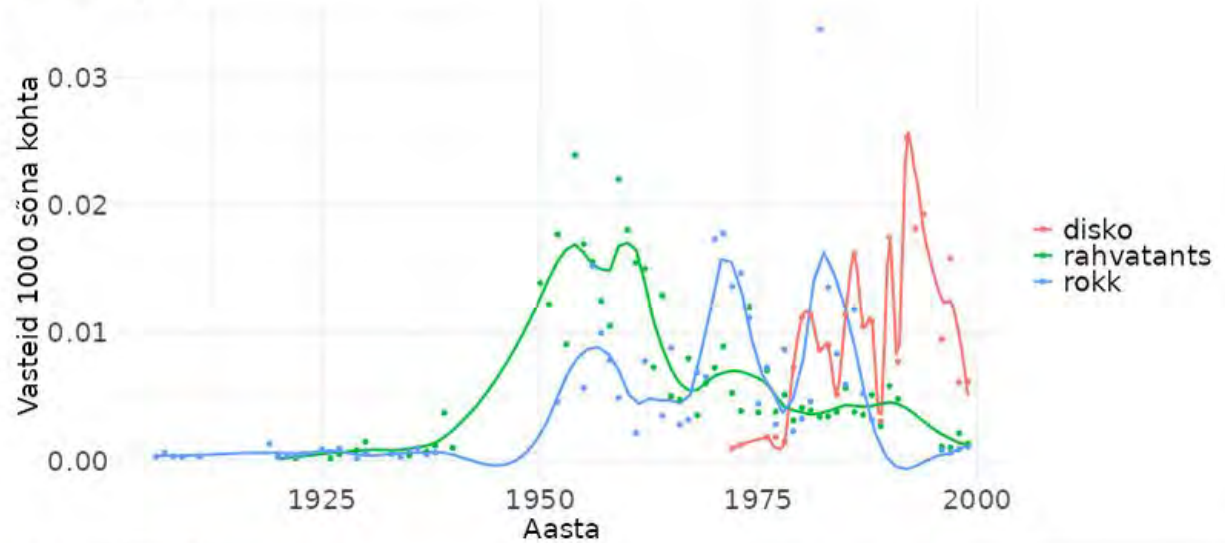- Exile newspapers
- Institutional newspapers

Local (n = 297549)   Exile (n = 28453)   Other (n = 35389)
Daily (n = 175981)   Weekly (n = 49444)   Institutional (n = 20029)

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

disko rokk rahvatants

Töötlus

- ◉ Sõnad (muutmata)
- ○ Lemmad (sõnade algvormid)

Aastad vaatluse all

| 1850 | **1900** | **2000** | 2023 |

1850 · 1868 · 1886 · 1904 · 1922 · 1940 · 1958 · 1976 · 1994 · 2012 2023

Joone kõverus

| 0 | **0.2** | 1 |

0 · 0.1 · 0.2 · 0.3 · 0.4 · 0.5 · 0.6 · 0.7 · 0.8 · 0.9 · 1

Lõika joonisel y-telg ära siin

| 0 | **100** |

0 · 10 · 20 · 30 · 40 · 50 · 60 · 70 · 80 · 90 · 100
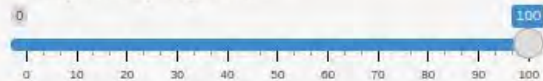
Märgi joonisel iga mitmes aasta

| 1 | **25** | 50 |
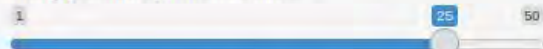
RaRa sõnamitmike loendaja koondab sõnade ja sõnamitmike sagedusi Rahvusraamatukogu ajalehekollektsioonides. Sisestades ülal otsingusse mõne sõna, näeb paremal joonisel selle sagedust kollektsioonides 1857-2023. Terminite eristamiseks kasuta koma, tab-i või enterit. Otsides sõnamitmikku eralda sõnad selle sees tühikuga. Kui teie otsitud sõna ei kuvata joonisel, ei ületanud ta vajalikku künnist, et
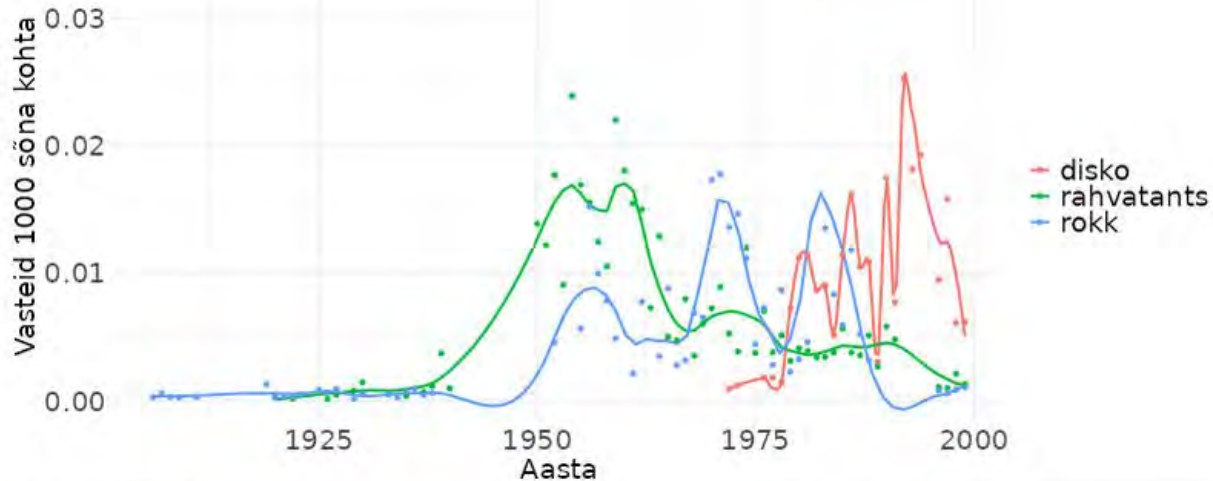
	Näita kirjeid	5	⌄	kaupa		Otsi (regex lubatud)	

	Sõna/Sõnamitmik	Tüüp	Mitu korda kokku	Mitmes dokumendis	Mitu sõna
	disko	All	All	All	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

Elektriuuringu 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 or

```
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"))
```

Andmestikus olevatest materjalidest saab ülevaate käsuga `get_digar_overview()` `digar.txts` paketist. 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")
```

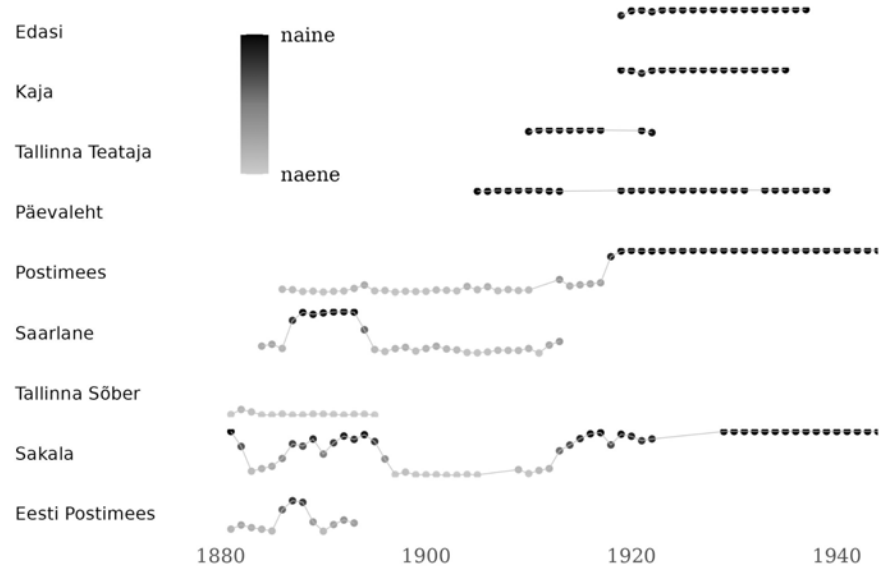Valimist ülevaate saamiseks vaatame kui palju on igal aastal valimis 1) artikleid ja 2) sõnu.

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.
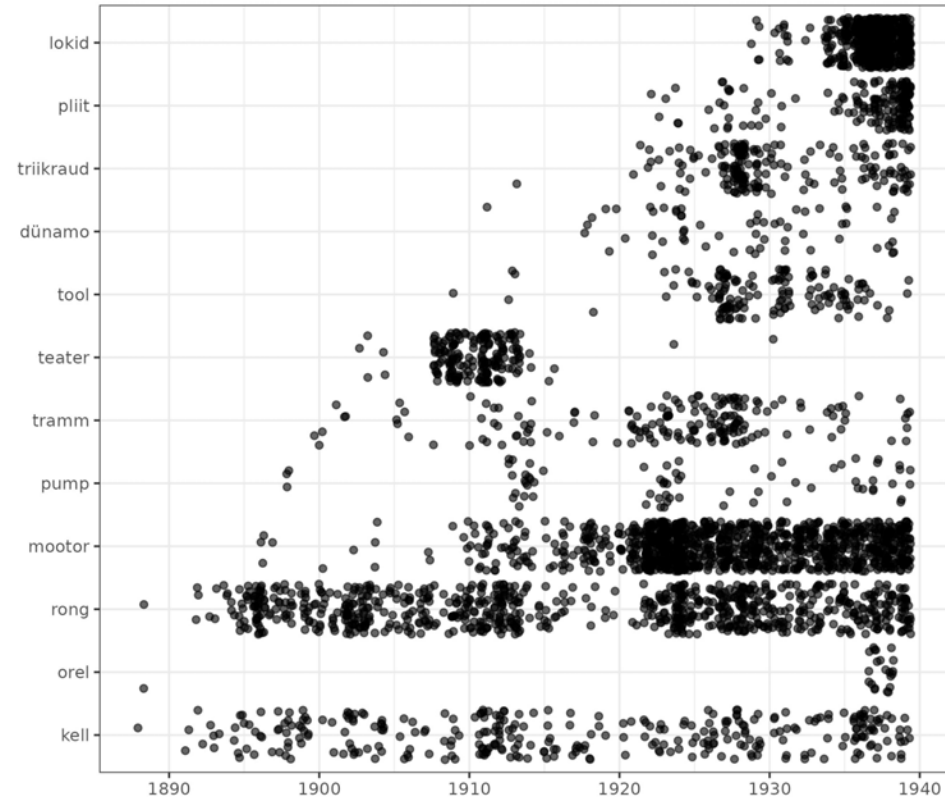
Full text access

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
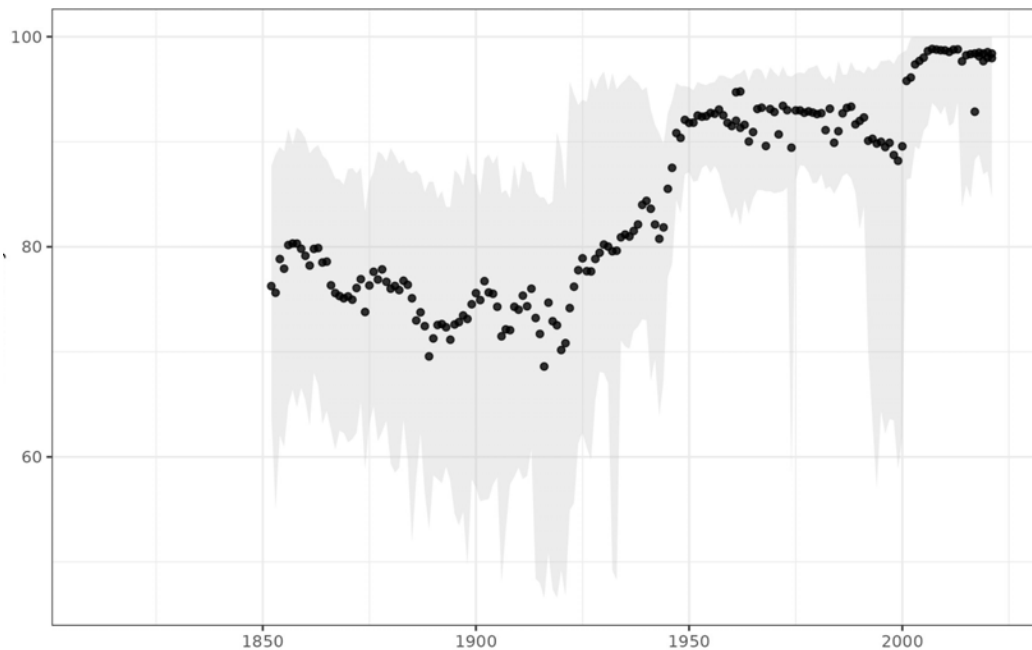
Digitized newspapers

Keeping the researcher in focus:

- Overviews
- Access
- Data explorers
- Data in context

Current topics:

- Data quality

2. Bibliographic data

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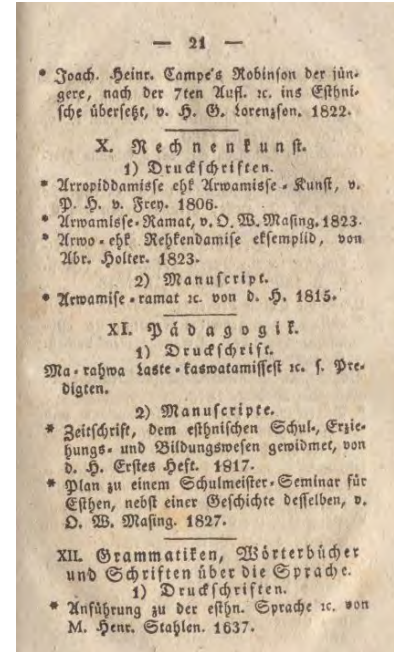
Excerpt from Rosenplänter's article
"Übersicht der estnischen Literatur" (1832)

Estonian National Bibliography

Consists of publications that are either:

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

Let's explore together!
digilab.rara.ee

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|adnnnaabn\\\\\\\\\\\\|/|a|\\\\\
=040 \\$aErRR$best$cErRR
=043 \\$cee
=100 1\$aAdojaan, Julius,$d1912-1994.
=400 1\$aАдоян, Юлиус,$d1912-1994.
=670 \\$aEesti elulood. Tallinn, 2000.
=680 \\$iEesti õigusteadlane, Aleksander Adojaani vend.

=LDR 00589nz a2200181n 4500
=001 a11124805
=003 ErRR
=008 981223|n|adnnnaabn\\\\\\\\\\\\|/|a|\\\\\
=040 \\$aErRR$best$cErRR
=043 \\$cee
=100 1\$aAdson, Artur,$d1889-1977.
=400 1\$aAdson, Karl Artur.
=400 1\$aAdson, Arthur.
=400 1\$aAdson, Karl Arthur.
=400 1\$aАдсон, Артур,$d1889-1977.
=500 1\$aBollmann, Peeter,$cpseud.,$d1889-1977.
=670 \\$aEesti kirjanike leksikon, 2000.
=680 \\$iEesti luuletaja, näitekirjanik ja kriitik. M. Undeni abika

The database format

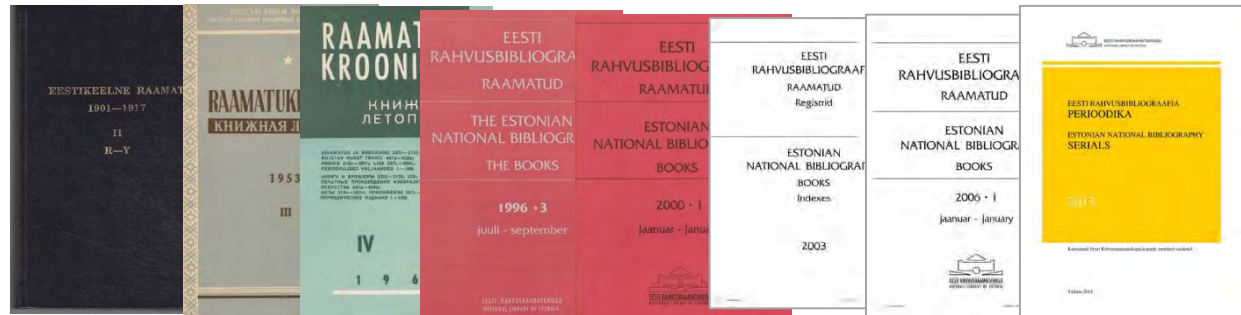- Has been maintained in Marc21, following library standards.
- Converted it into more compact tables for research purposes.
- Now as a .tsv in Zenodo
- We are working on a curated, simplified version of it.

```
=LDR 00376nz a2200133n 4500
=001 a11124799
=003 ErRR
=008 990402n|n|adnnnaabn|\\\\\\\\\\\|a|\\\\
=040 \\$aErRR$best$cErRR
=043 \\$cee
=100 1|$aAdojaan, Julius,$d1912-1994.
=400 1|$aАдоян, Юлиус.$d1912-1994.
=670 \\$aEesti elulood. Tallinn, 2000.
=680 \\$iEesti õigusteadlane, Aleksander Adojaani vend.

=LDR 00589nz a2200181n 4500
=001 a11124805
=003 ErRR
=008 981223n|n|adnnnaabn|\\\\\\\\\\\|a|\\\\
=040 \\$aErRR$best$cErRR
=043 \\$cee
=100 1|$aAdson, Artur,$d1889-1977.
=400 1|$aAdson, Karl Artur.
=400 1|$aAdson, Arthur.
=400 1|$aAdson, Karl Arthur.
=400 1|$aАдсон, Артур,$d1889-1977.
=500 1|$aBollmann, Peeter,$cpseud.$d1889-1977.
=670 \\$aEesti kirjanike leksikon, 2000.
=670 \\$iEesti luuletaja, näitekirjanik ja kriitik. M. Underi abikaasa. Kasutanud pseudonüümi Peeter Bollmann.
```

title_responsibility_statement	publication_place	publisher	publication_date
Anton Hansen-Tammsaare ; illustreerinud Herald Eelma	Tallinn	Eesti Raamat	1974
Mart Laar	Stockholm	Välis-Eesti & EMP	1997
Mihhail Makarov ; [Koes: E. Täll]	Tallinn	Eesti Raamat	1966
Pearl Buck ; inglise keelest tõlkinud A. Hanko ; [kujundanud...	Tallinn	Eesti Raamat	1994
Eesti Looduseuurijate Selts Eesti Teaduste Akadeemia j...	Tallinn	Valgus	1993
Knut Hamsun ; [tõlkinud Henrik Sepamaa ; kujundanud...	Tallinn	Eesti Raamat	1985
Alexandre Dumas (junior) ; [prantsuse keelest tõlkinud J...	Tallinn	Eesti Raamat	1992
Linda Petti	Tallinn	Valgus	1993
Heinrich Böll ; [saksa keelest tõlkinud M. Sirkel ; kujund...	Tallinn	Eesti Raamat	1993
Marina Tsvetajeva ; [koostanud, eestindanud ja järelsõ...	Tallinn	Eesti Raamat	1994
toimetajad: Raivo Kalmi, Sven Illiviäli, Ants Noormets,...	Melbourne	Eesti Meeslaulu Selts Melbourne'is	1994
U. Ekman	Uppsala	Elu Sõna	1993
Linda Petti	Tallinn	Valgus	1992
James Herriot ; [inglise keelest tõlkinud Jaagup Alands ; ka...	Tallinn	Valgus	1994
Juhan Jaik ; [illustreerinud H. Mugasto]	Tartu	Noor-Eesti	1934
A. H. Tammsaare	Tartu	Noor-Eesti	1917
Petev Bronk ; inglise keelest tõlkinud M. Patli ; Eesti Tea...	Tallinn	Eesti Raamat	1993
Mart Laar	Stockholm	Välis-Eesti & EMP	1994
laulanut ja Anter Valberna ; toimittanner by Paul Hagu...	Kuopio	Snellman-instituutti	1995
Harald Süülepv ; [kujundanud R. Pangsepp]	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">ﺳﺘﺎﻧﺪال ﺩﻩ ﻣﺎري ﻫﻨﺮي ﺑﻴﻞ</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">ﺳﺘﺎﻧﺪال ﺩﻩ ﻣﺎري ﻫﻨﺮي ﺑﻴﻞ</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|adnnnaabn|\\\\\\|\|\a|\\\\\
=040 \\$aErRR$best$cErRR
=043 \\$cee
=100 1\$aAdojaan, Julius.$d1912-1994.
=400 1\$aАдоян, Юлиус.$d1912-1994.
=670 \\$aEesti eluloog. Tallinn, 2000.
=680 \\$iEesti õigusteadlane, Aleksander Adojaani vend.

=LDR 00589nz a2200181n 4500
=001 a11124805
=003 ErRR
=008 981223|n|adnnnaabn|\\\\\\|\|\a|\\\\\
=040 \\$aErRR$best$cErRR
=043 \\$cee
=100 1\$aAdson, Artur.$d1889-1977.
=400 1\$aAdson, Karl Artur.
=400 1\$aAdson, Arthur.
=400 1\$aAdson, Karl Arthur.
=400 1\$aАдсон, Артур.$d1889-1977.
=500 1\$aBollmann, Peeter.$cpseud.$d1889-1977.
=670 \\$aEesti kirjanike leksikon, 2000.
=680 \\$iEesti luuletaja, näitekirjanik ja kriitk. M. Underi abikaasa. Kasutanud pseudonüümi Peeter Bollmann.
```

title_responsibility_statement	publication_place	publisher	publication_date
Anton Hansen Tammsaare ; illustreerinud Herald Eelma	Tallinn	Eesti Raamat	1974
Mart Laar	Stockholm	Välis-Eesti & EMP	1997
Mihhail Makarov ; [koos: E. Tali]	Tallinn	Eesti Raamat	1966
Pearl Buck ; inglise keelest tõlkinud A. Hanko ; [kujunda...	Tallinn	Eesti Raamat	1994
Eesti Loodusuurijate Selts Eesti Teaduste Akadeemia j...	Tallinn	Valgus	1993
Knut Hamsun ; [tõlkinud Henrik Sepamaa ; kujundanud...	Tallinn	Eesti Raamat	1985
Alexandre Dumas (junior) ; [prantsuse keelest tõlkinud j...	Tallinn	Eesti Raamat	1992
Linda Petti	Tallinn	Valgus	1993
Heinrich Böll ; [saksa keelest tõlkinud M. Sirkel ; kujund...	Tallinn	Eesti Raamat	1993
Marina Tsvetajeva ; [koostanud, eestindanud ja järelsõ...	Tallinn	Eesti Raamat	1993
toimetajad: Raivo Bähni, Sven Šviljuli, Ants Noormets...	Melbourne	Eesti Meeslaulu Selts Melbourne'is	1994
U. Ekman	Uppsala	Elu Sõna	1993
Linda Petti	Tallinn	Valgus	1992
Jamie Perriot ; [inglise keelest tõlkinud Jaagup Aluots ; ka...	Tallinn	Valgus	1994
Juhan Jaik ; [illustreerinud H. Mugasto]	Tartu	Noor-Eesti	1934
A. H. Tammsaare	Tartu	Noor-Eesti	1917
Peter Brook ; inglise keelest tõlkinud M. Pati ; Eesti Tea...	Tallinn	Eesti Raamat	1993
Mart Laar	Stockholm	Välis-Eesti & EMP	1994
laulusid by Anne Valberse ; toimittanene by Paul Hugu...	Kuopio	Snellman-instituutti	1995
Harald Süislepp ; [kujundanud R. Pangsepp]	Tallinn	Eesti Raamat	1984

```xml
<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">ﺳ, ﺮﺒﻳ‹«ﺓﺠ³ ˉ¾Œ</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">ﺳ, ﺮﺒﻳ‹«ﺓﺠ³ ˉ¾Œ</mx:subfield>
    <mx:subfield code="d">1783-1842</mx:subfield>
    <mx:subfield code="0">(NLR)RU NLR AUTH 775430</mx:subfield>
</mx:datafield>
```
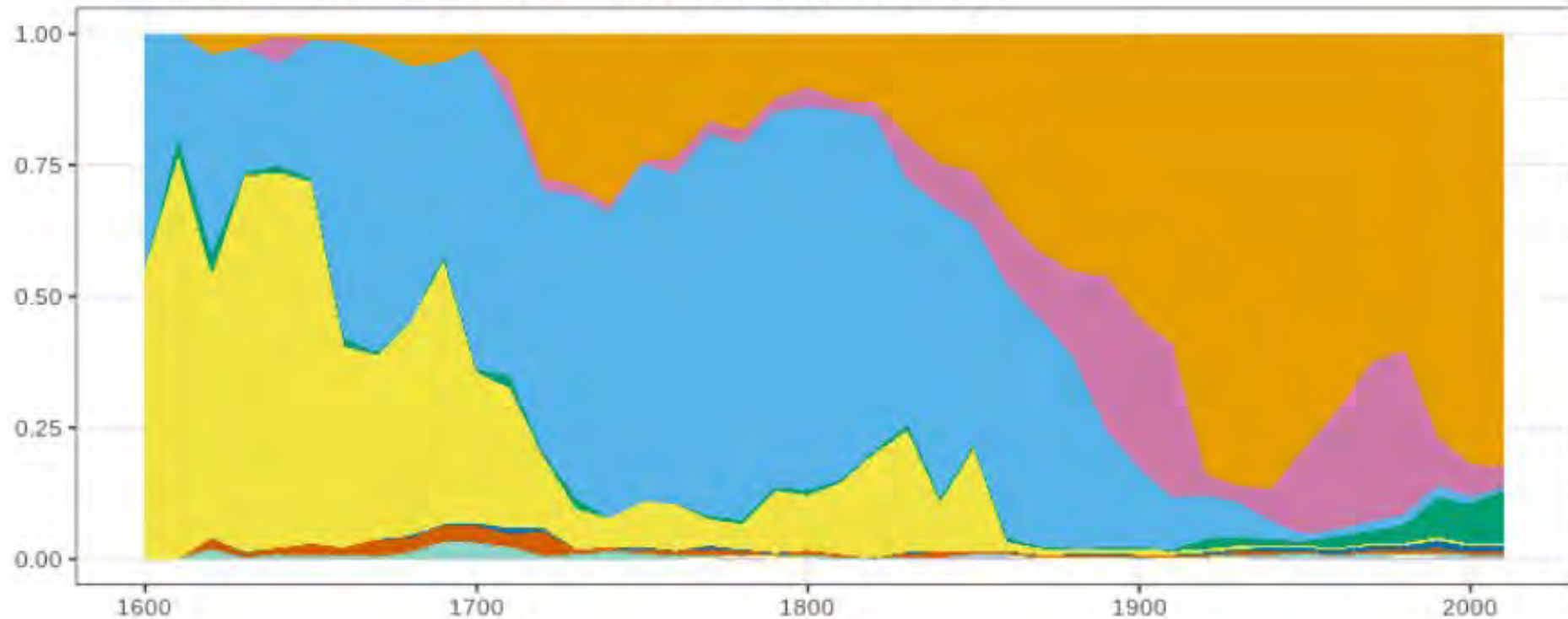
A look at the data

Books in ENB (1600-2020)
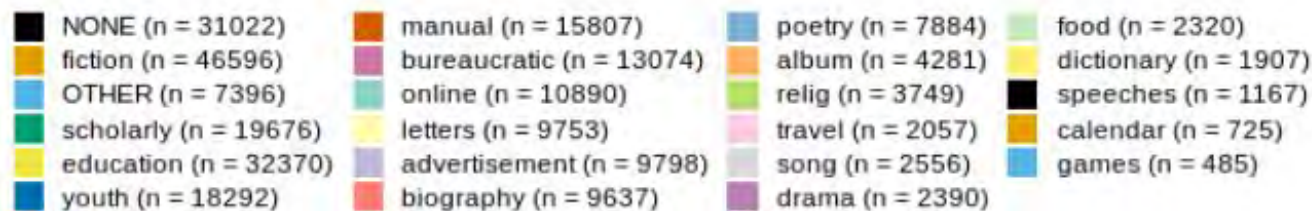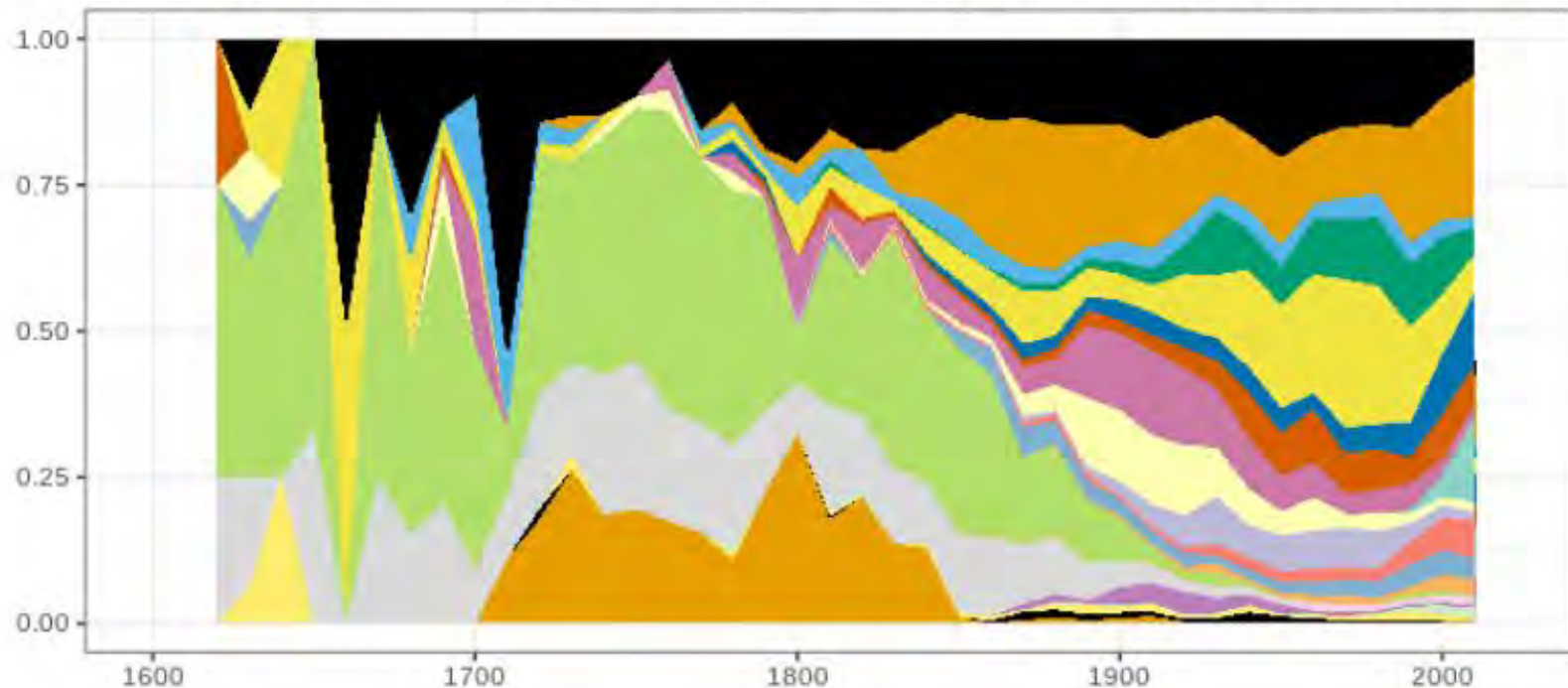
Some historical periods and events added for comparison

Legend: estonian (orange), other languages (grey)
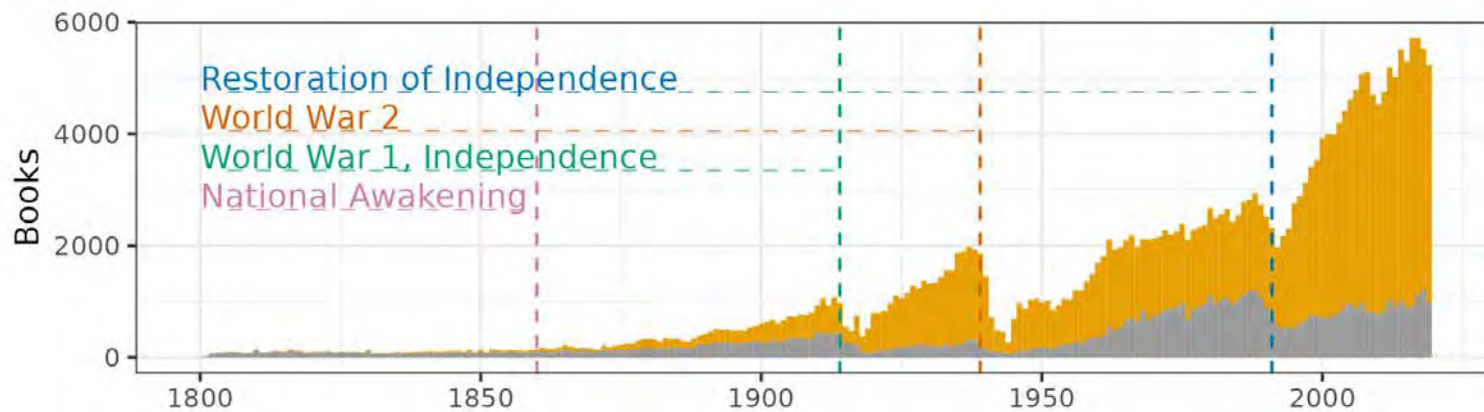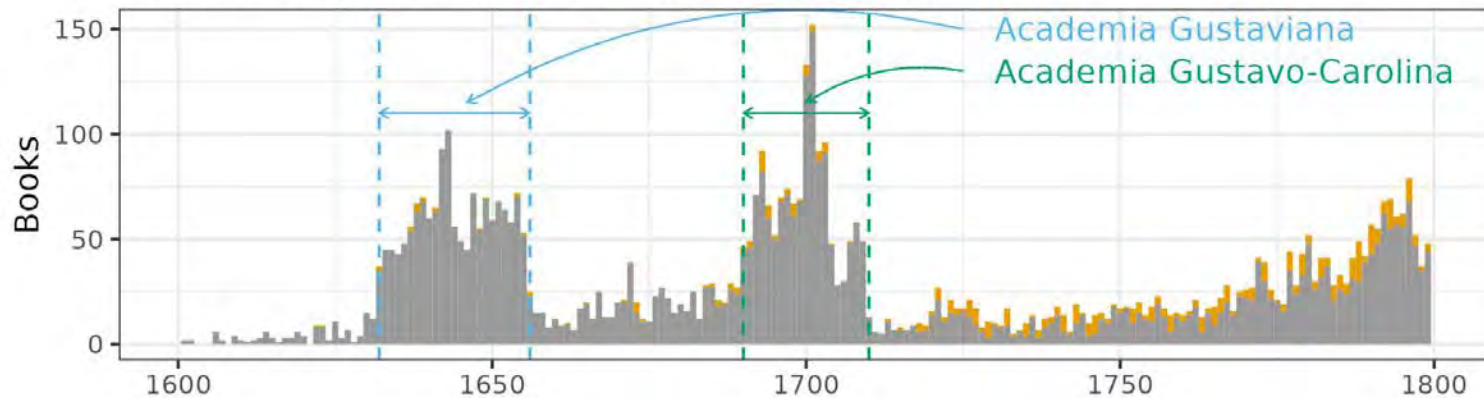
Languages in Estonian National Bibliography (1600-2020)

- 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)

■ NONE (n = 31022)	■ manual (n = 15807)	■ poetry (n = 7884)	■ food (n = 2320)
■ fiction (n = 46596)	■ bureaucratic (n = 13074)	■ album (n = 4281)	■ dictionary (n = 1907)
■ OTHER (n = 7396)	■ online (n = 10890)	■ relig (n = 3749)	■ speeches (n = 1167)
■ scholarly (n = 19676)	■ letters (n = 9753)	■ travel (n = 2057)	■ calendar (n = 725)
■ education (n = 32370)	■ advertisement (n = 9798)	■ song (n = 2556)	■ games (n = 485)
■ youth (n = 18292)	■ biography (n = 9637)	■ drama (n = 2390)	

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.

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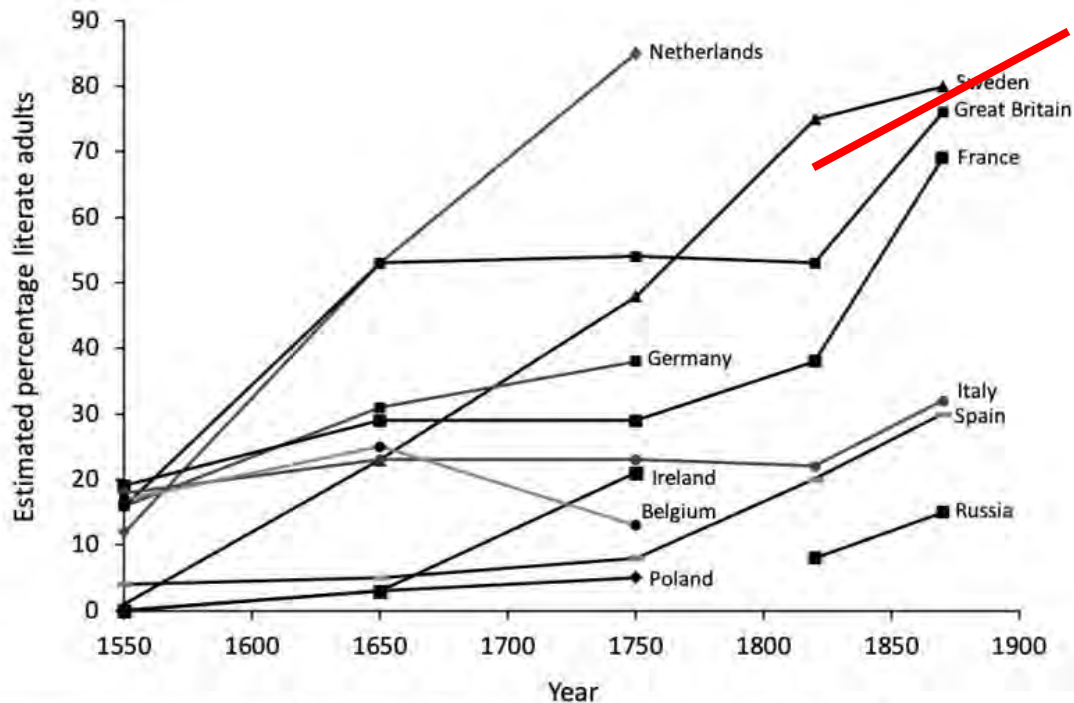
FIGURE P.1. Literacy rates for various European countries from 1550 to 1900. These estimates are based on book publishing data calibrated using more direct measures of literacy.[8]

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.

Genres of Estonian language books (1800-1940)

NONE (n = 6973)	advertisement (n = 2508)	drama (n = 1307)	letters (n = 384)
OTHER (n = 1592)	relig (n = 2598)	biography (n = 601)	calendar (n = 359)
fiction (n = 8211)	song (n = 1553)	speeches (n = 355)	food (n = 135)
scholarly (n = 2863)	manual (n = 1667)	album (n = 447)	
bureaucratic (n = 5540)	youth (n = 1644)	dictionary (n = 388)	
education (n = 4048)	poetry (n = 1089)	travel (n = 230)	
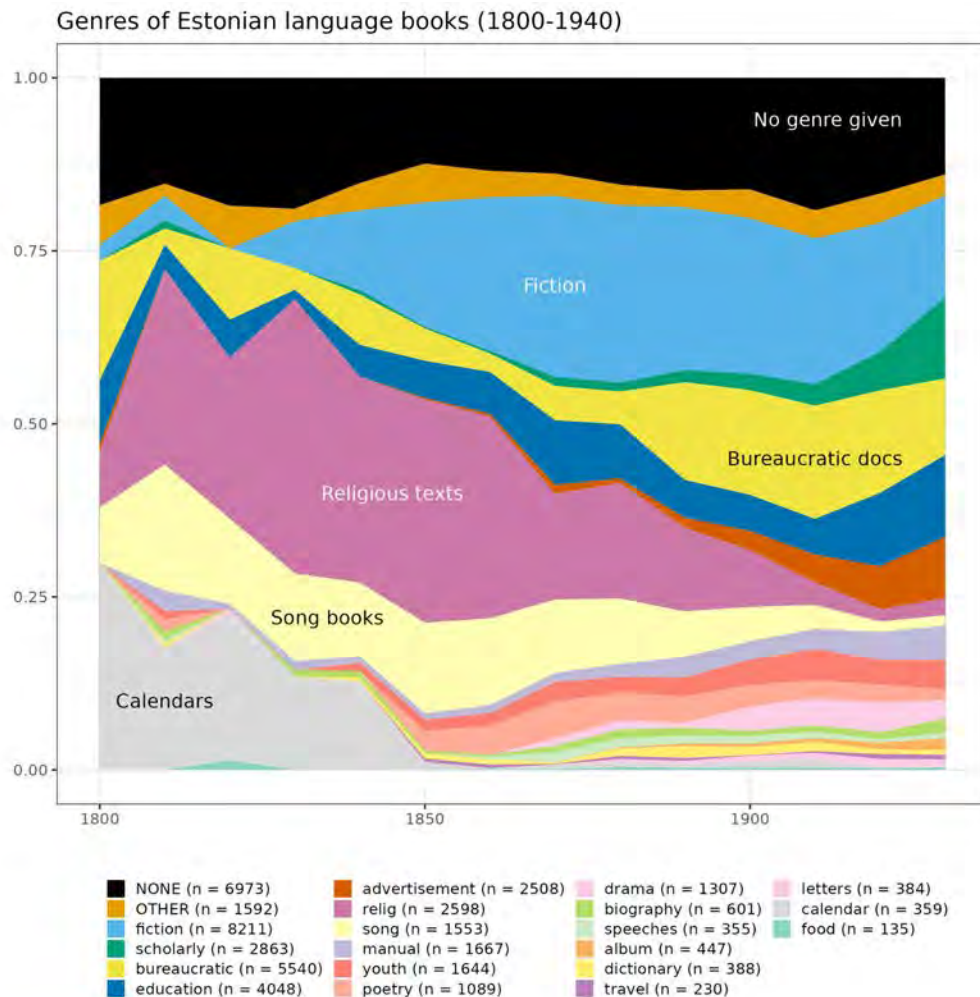
Emergence of the Estonian literary community

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)
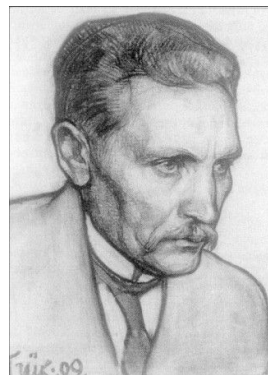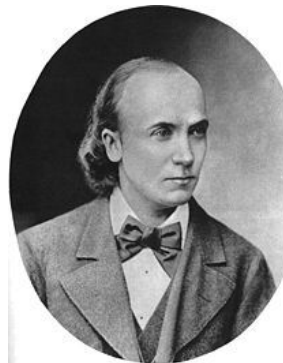
Population growth and publishing trends

Estonian native speakers (1800-1930)
Residing in modern Estonia
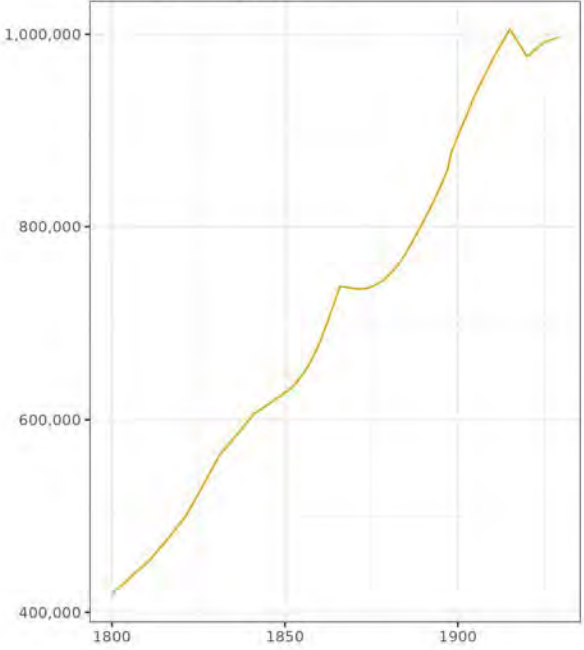
Estonian language books
Per million native speakers
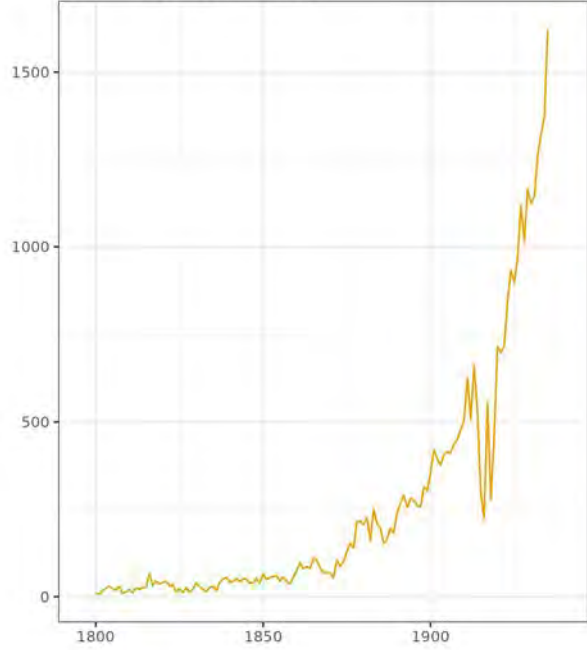
Estonian literary actors
Per million native speakers
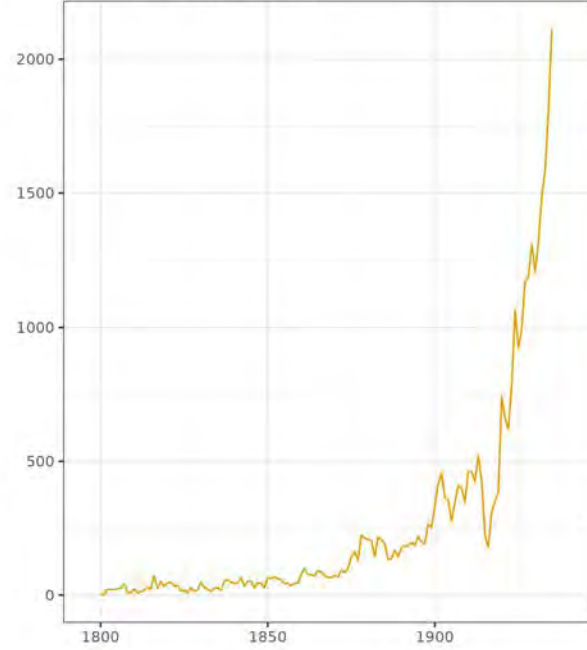
New people to written language community

Birthyears of actors active in 1800-1940
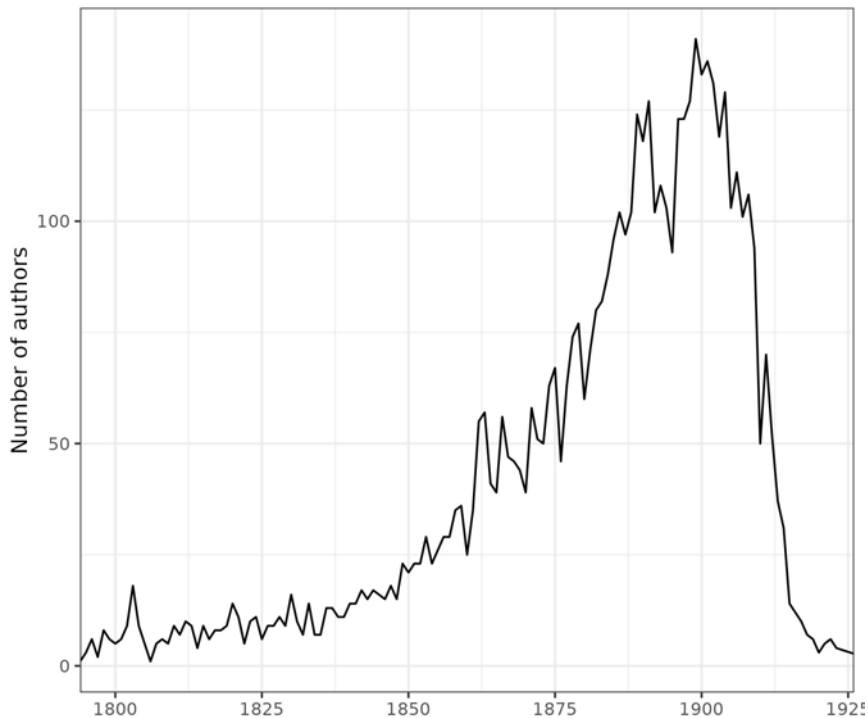
Age of average active living actor by language (1800-1940)
Across all books (——) and for first books published (- - -)
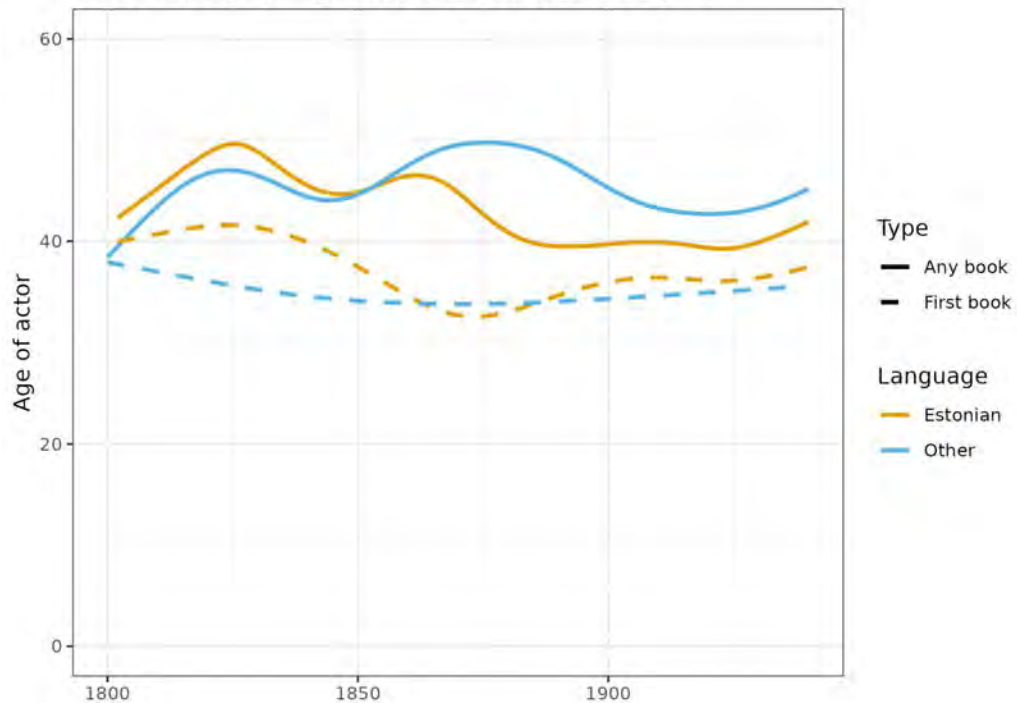
Language of books during actor's life
Actors with > 50 works, lifetime share on the right

Actor	Estonian %
Eisen, Matthias Johann,	100%
Tõnisson, Mats,	100%
Hermann, Karl August,	92%
Grenzstein, Ado,	95%
Jakobson, Eduard Magnus,	99%
Brandt, Friedrich,	100%
Gläser, Wilhelm,	6%
Schnakenburg, Heinrich,	15%
Gressel, Julius Hermann,	52%
Körber, Martin,	93%
Karow, Eduard Julius,	2%
Normann, Vilhelm,	100%
Schlater, Georg Friedrich,	2%
Kreutzwald, Friedrich Reinhold,	84%
Laakmann, Heinrich,	39%
Körber, Carl,	91%
Dullo, Carl,	3%
Henning, Carl Matthias,	84%
Gressel, Johann Hermann,	1%
Sonntag, Karl Gottlob,	1%
Kotzebue, August von,	0%
Grenzius, Johann Michael Gerhard,	0%

Estonian
- 100%
- 75%
- 50%
- 25%
- 0%

Language choice for multilingual actors (n = 711)

Lifetime share of language for actor vs median time of publication
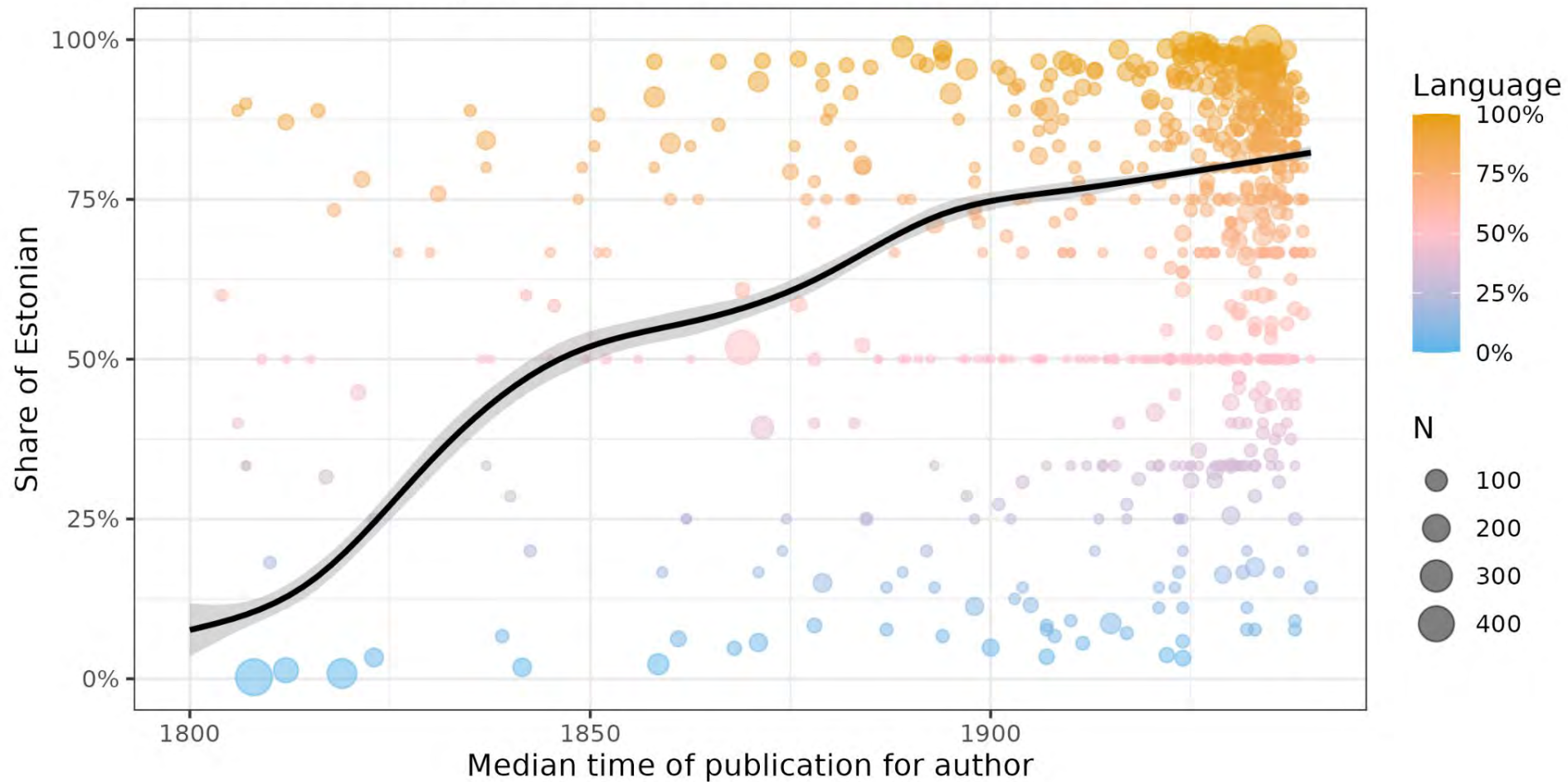
Networks

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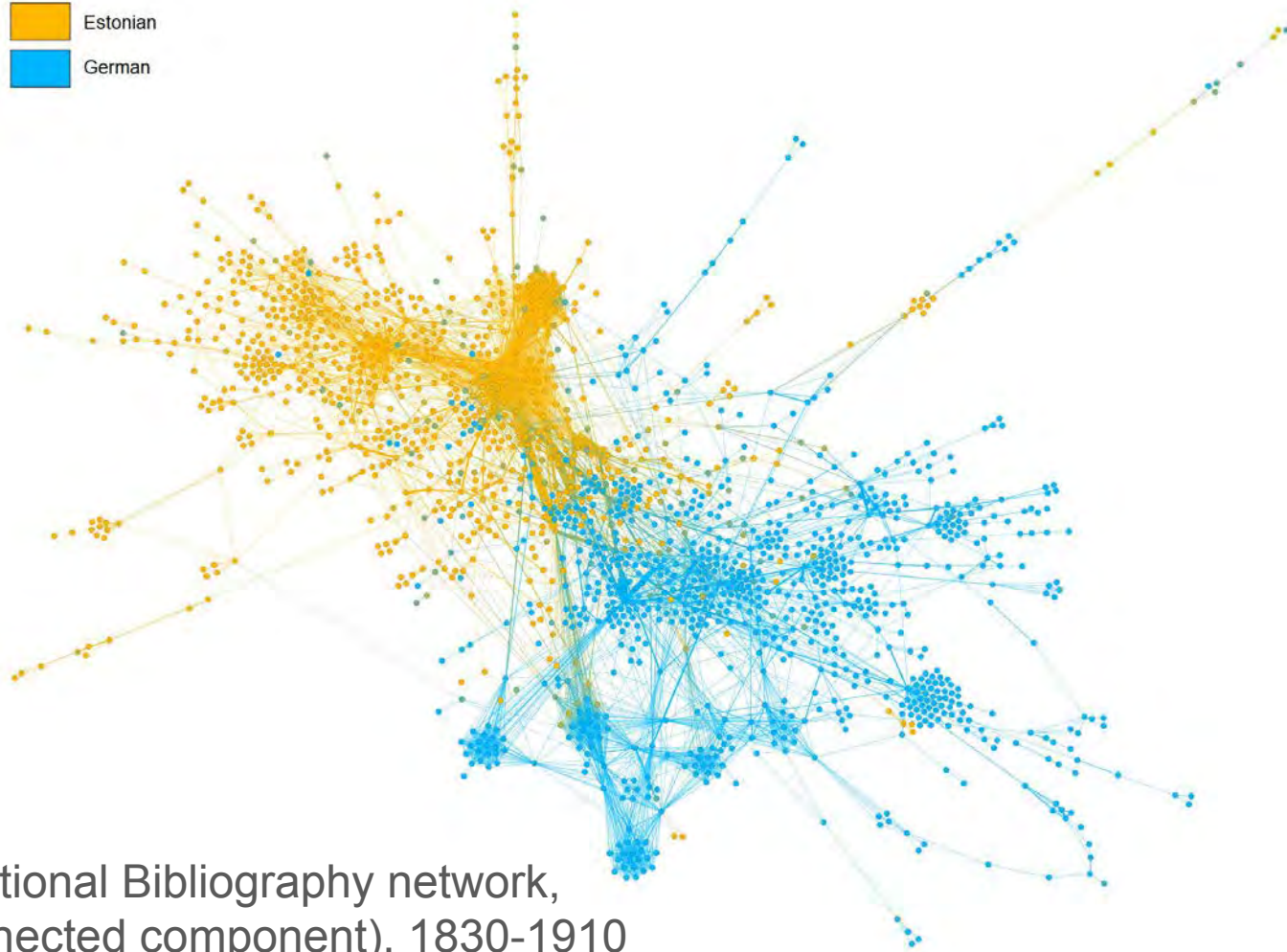
Estonian National Bibliography network,
(largest connected component), 1830-1910

Largest connected components
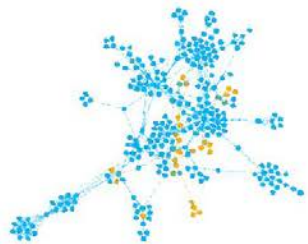
1830-1849   1850-1869   1870-1889   1890-1909
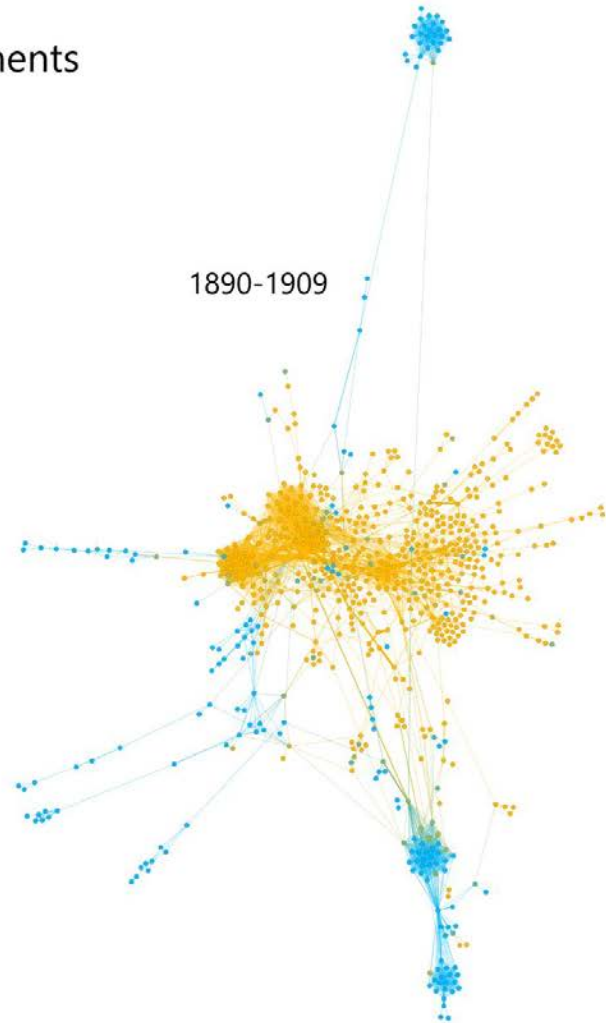
Estonian
German

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

1. Karow, Eduard Julius — P
2. Kluge, Franz — P
3. Schulz, Carl Anton — I
4. Gressel, Julius Hermann — P
5. Laakmann, Heinrich — P
6. Kreutzwald, Friedrich Reinhold — A

Ego networks (radius=1) of most central nodes by betweenness, 1850-1869
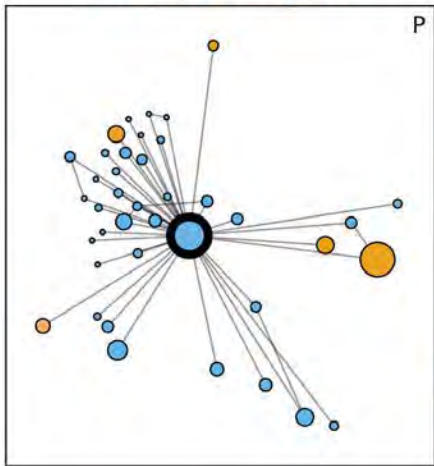
1. Karow, Eduard Julius
2. Kluge, Franz
3. Schulz, Carl Anton
4. Gressel, Julius Hermann
5. Laakmann, Heinrich
6. Kreutzwald, Friedrich Reinhold
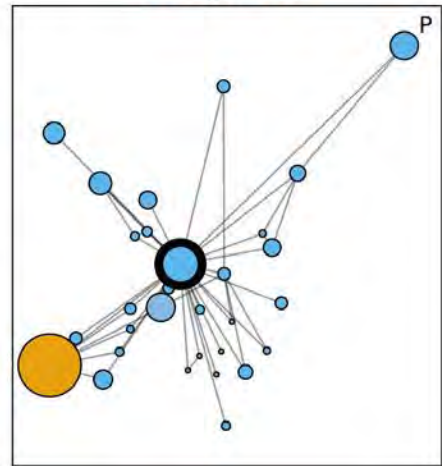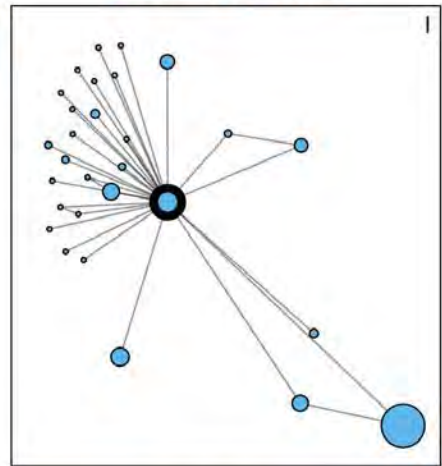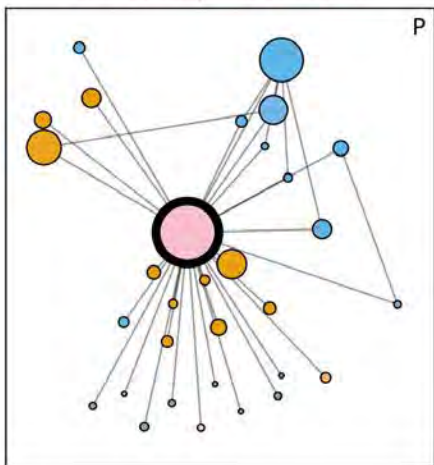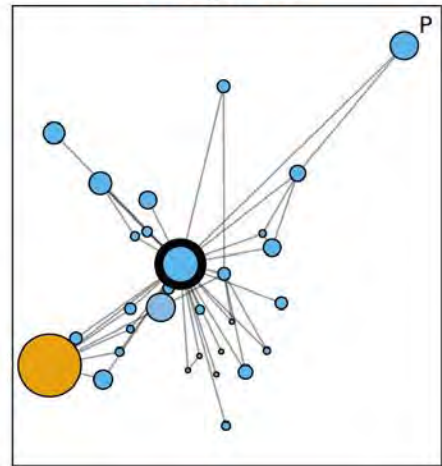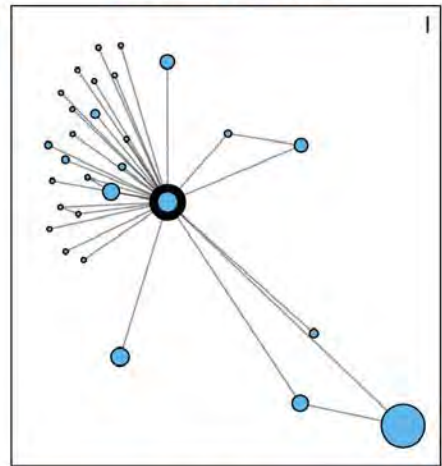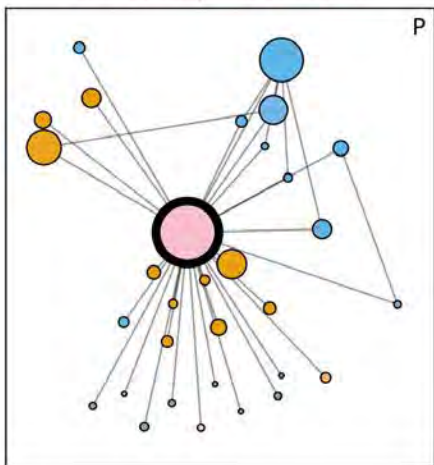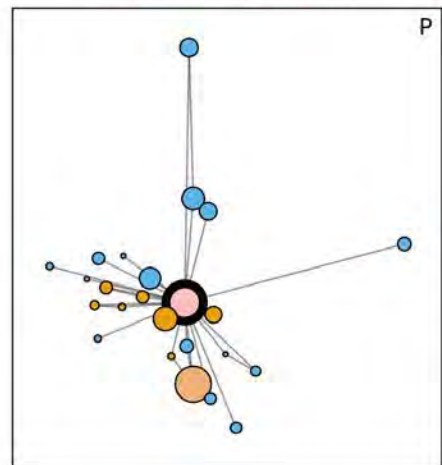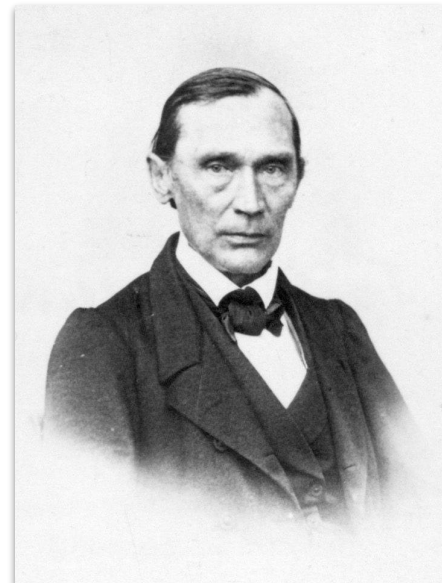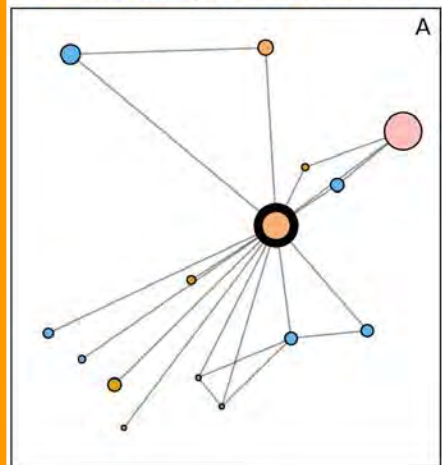
- Leading figure in the early National Awakening

- 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
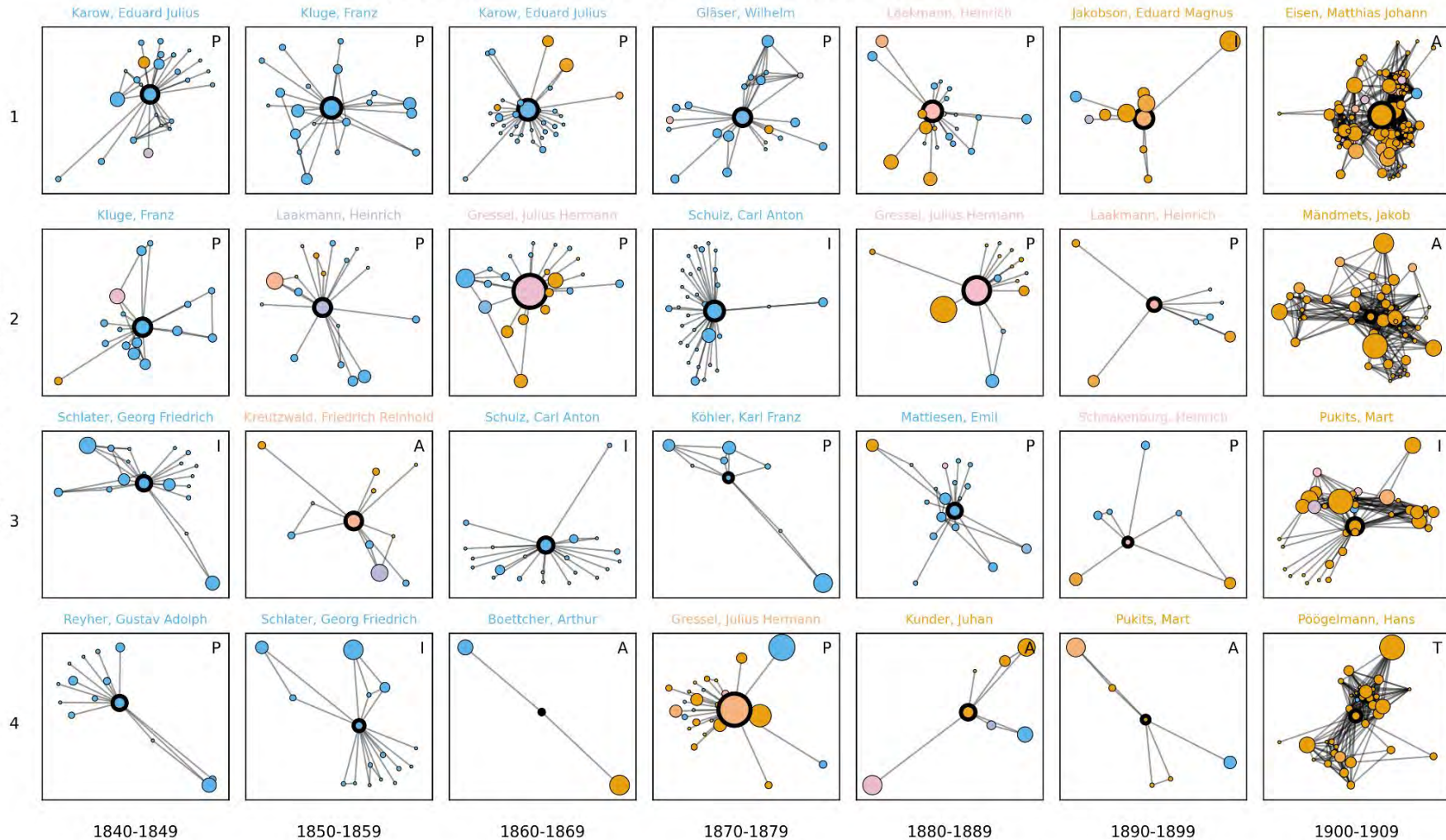
Top 50 most central nodes by decade, ranked by betweenness
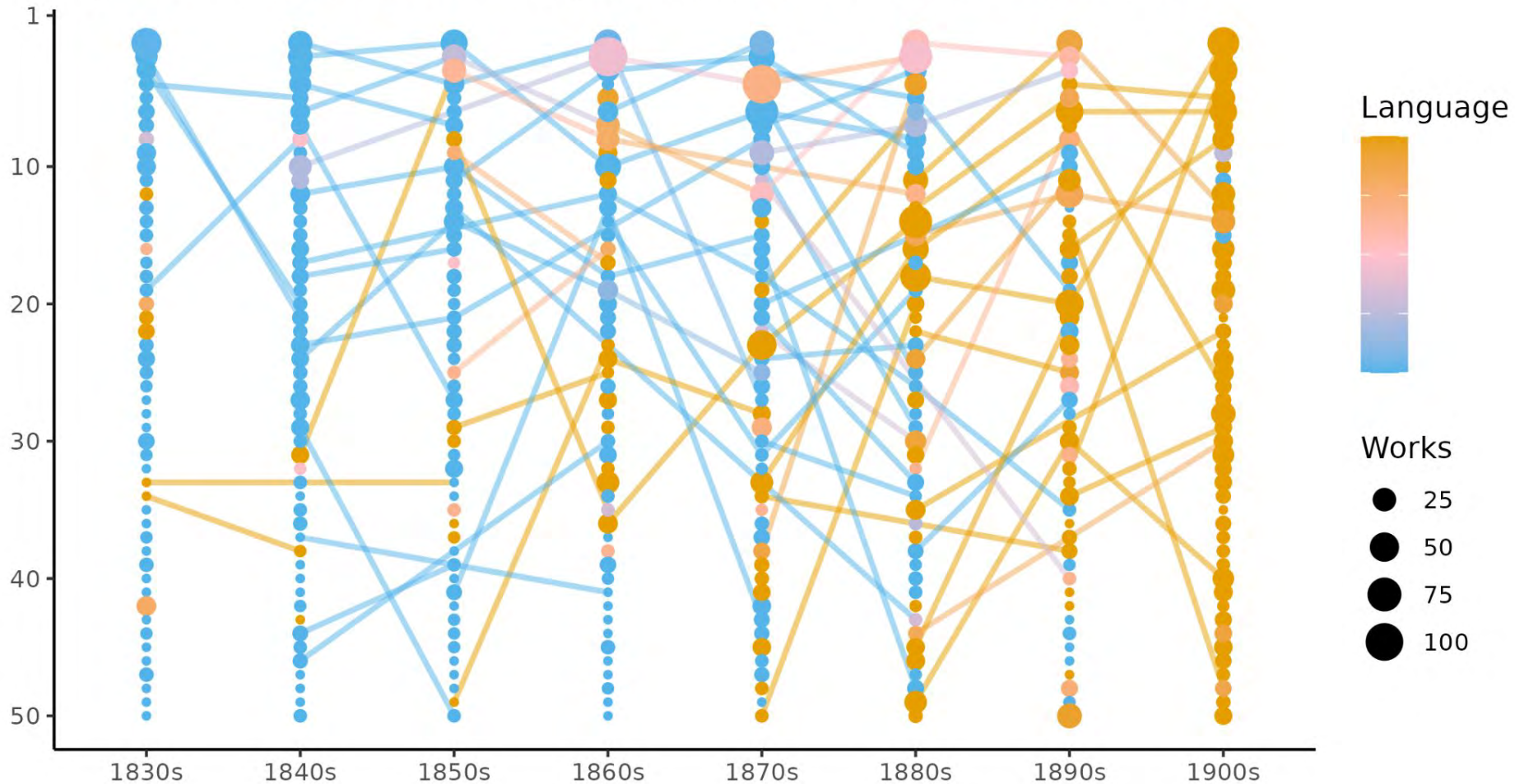
Top 50 most central nodes by decade, ranked by closeness
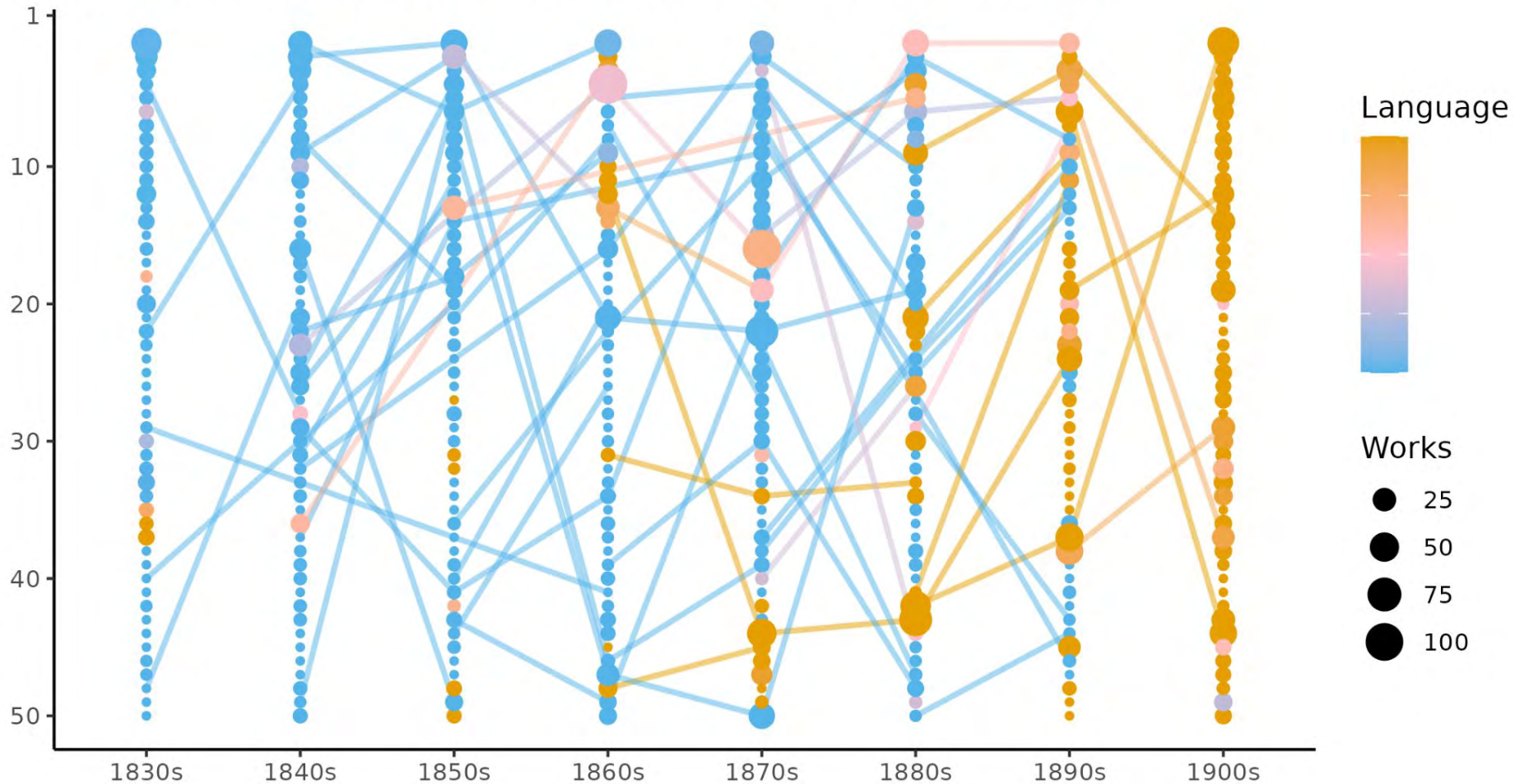
Linked biographies

Many contributors have a link to VIAF

Some more data in other datasets

Linked biographies

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

Mapped birthplaces of the contributors

Kartogramm 1. EBL-s käsitletud eesti rahvusest isikute sünnikohad linnade ja kihelkondade kaupa
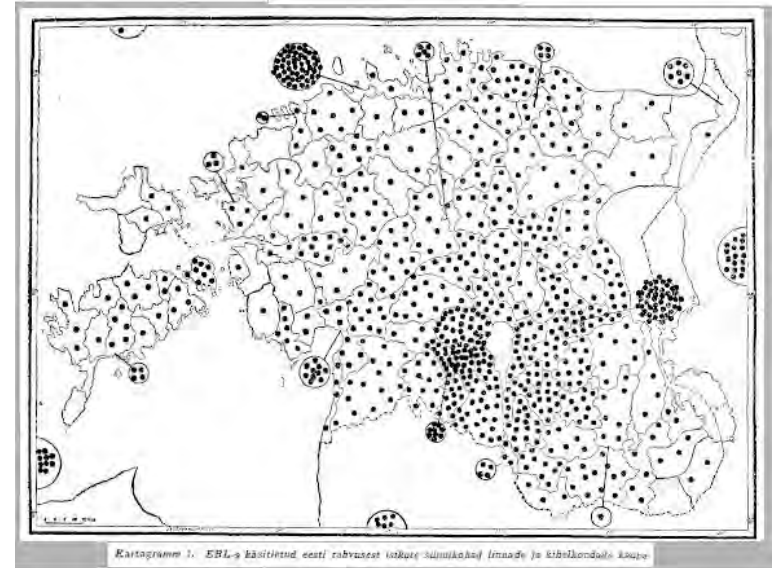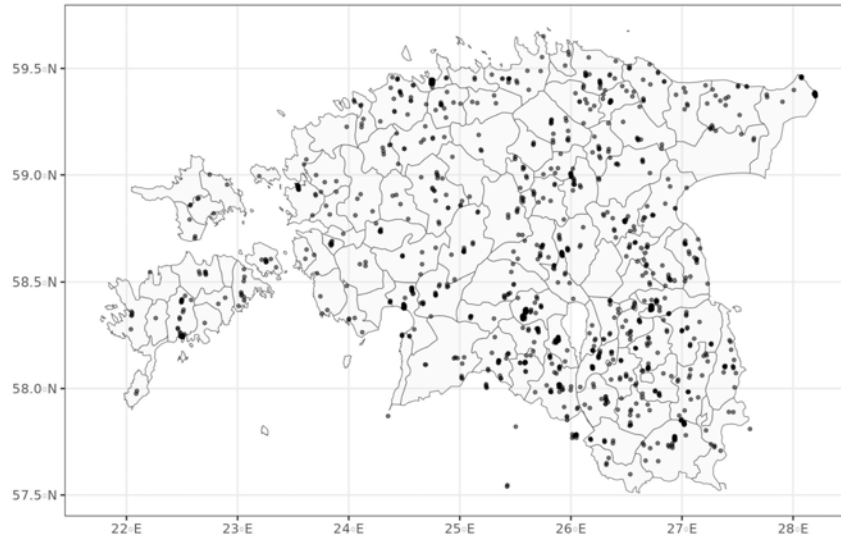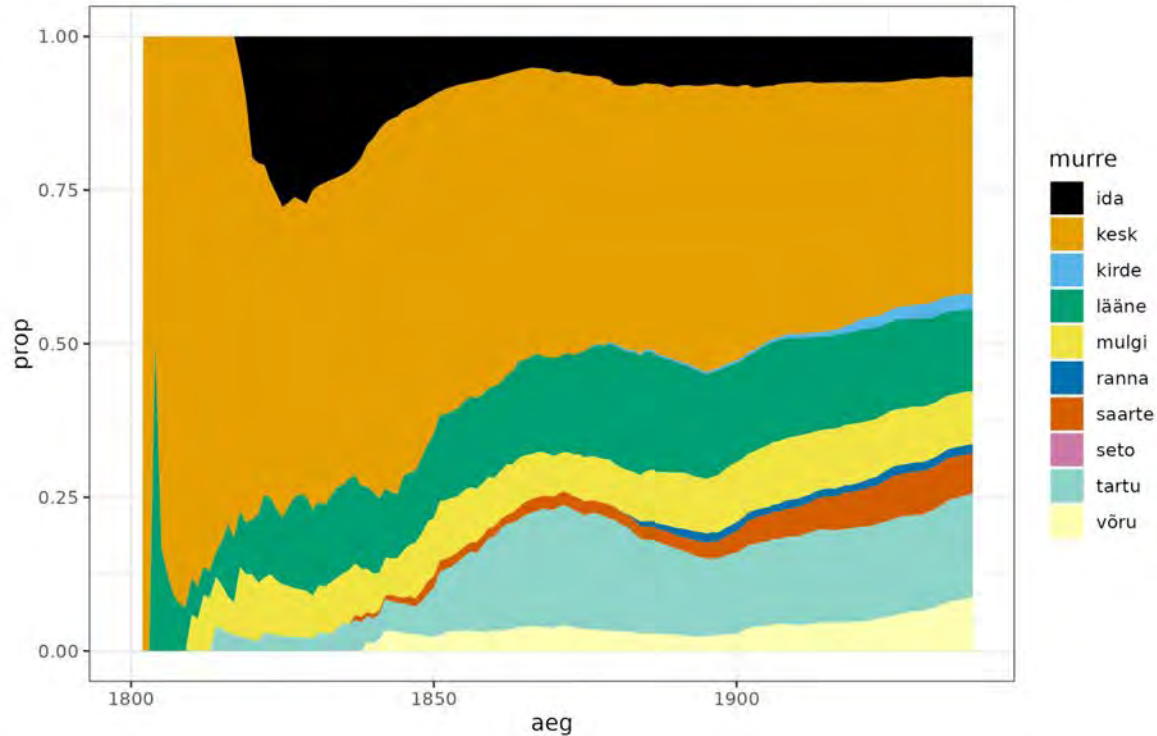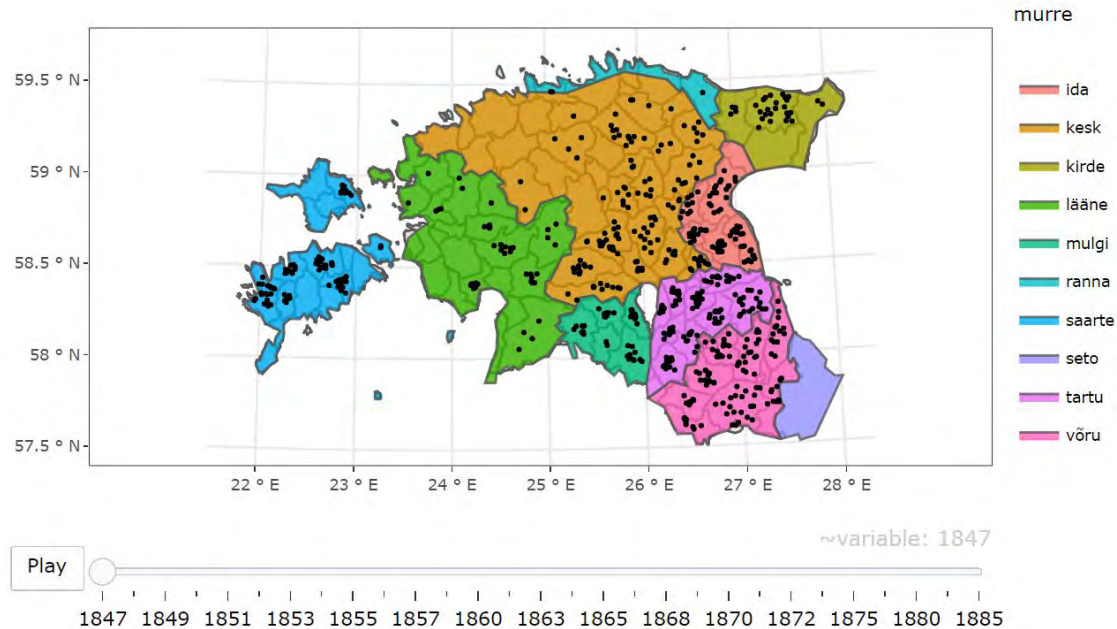
Mapped birthplaces and dialects
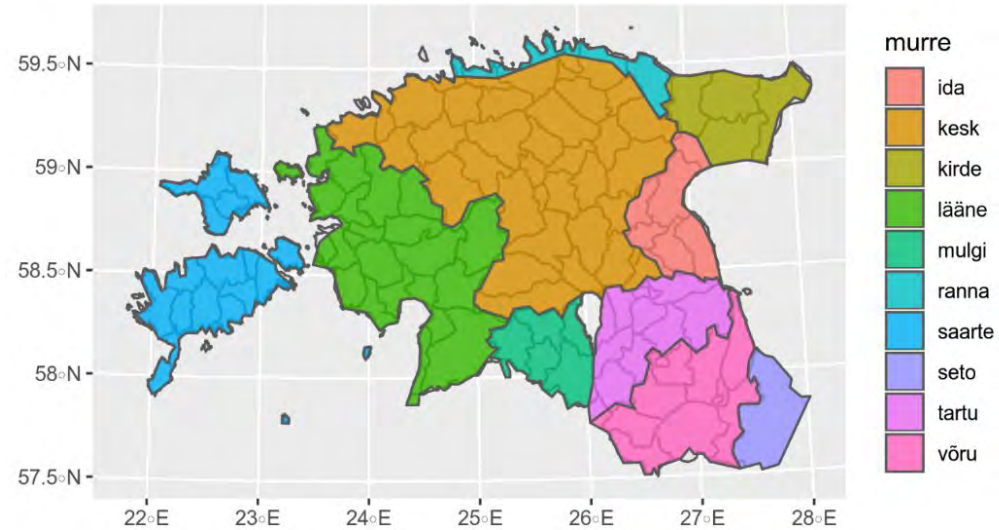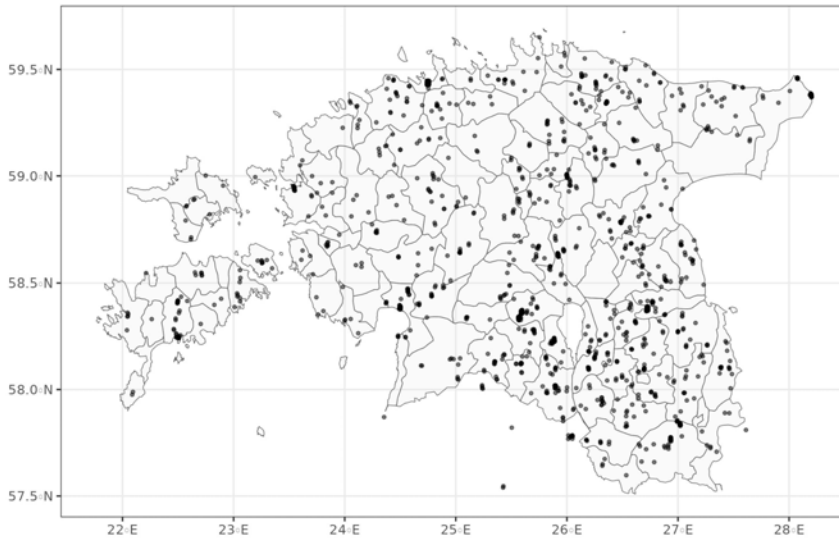
Share of dialects in the dataset

Cumulative share of different dialects of Estonian 1800-1940

Public schools in 19c

- Are founded from 1830s to 1880s

Schools visualization link

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

Literati per 1000 inhabitants

Schools per 1000 inhabitants
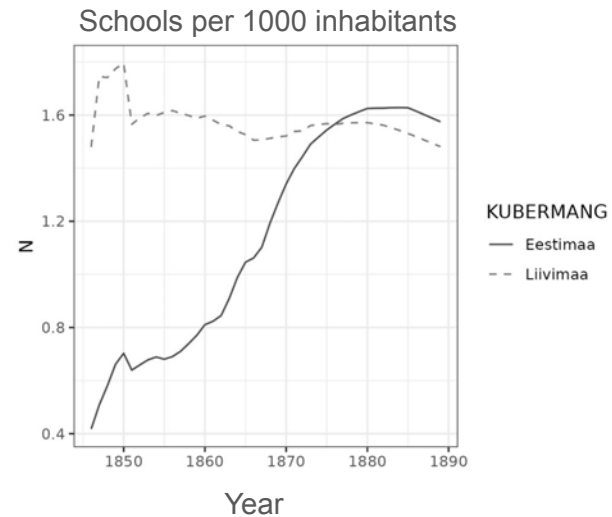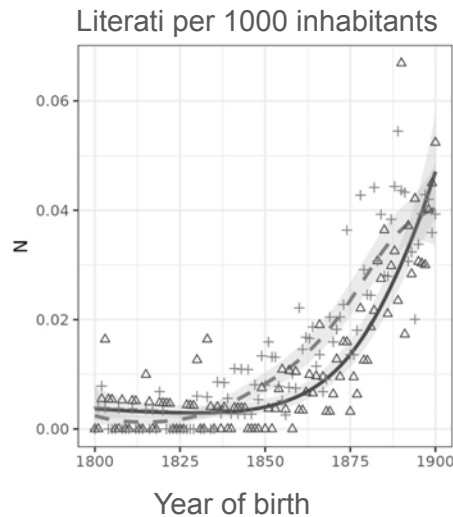
Results

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

Team in the National Library of Estonia

UNIVERSITY OF TARTU
Centre for Digital Humanities
and Information Society

Let's explore together!
digilab.rara.ee

RaRa

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

- 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.