Metadata from ERB

ERB metadata

Abstract

Estonian National Bibliography is a metadata set that aims to collect information on all publications written in any language in Estonia and all texts written in Estonian in whichever country. In this set, only the publications in Estonian language are used. The dataset has been compiled in digital format since 2002 and aggregates the work of multiple institutions and generations in collecting the publication information.

This dataset presents the Estonian National Bibliography dataset in wide instead of long format used in Marc21, with some of the variables that may be useful for text-mining studies. It includes the following information

- publication ID in National Library
- time of publication (aeg)
- place of publication (koht_raw)
- partly standardized place of publication (koht)
- publisher (kirjastus_raw)
- partly standardized publisher information (kirjastus)
- title (title, subtitle, comptitle)
- author (name, date of birth,id)
- other associated authors (translator, editor etc)
- number of copies printed
- font information
- keyword
- genre
- link to fulltext

Information on coding the variables can be found here: [meta file on github]

Helpful information on the metadata available can be found here: http://data.digar.ee/#page5

The rules followed in adding information on older books can be found here: http://www.elnet.ee/images/pdf/juhendid/vanaraamat_MARC21.pdf

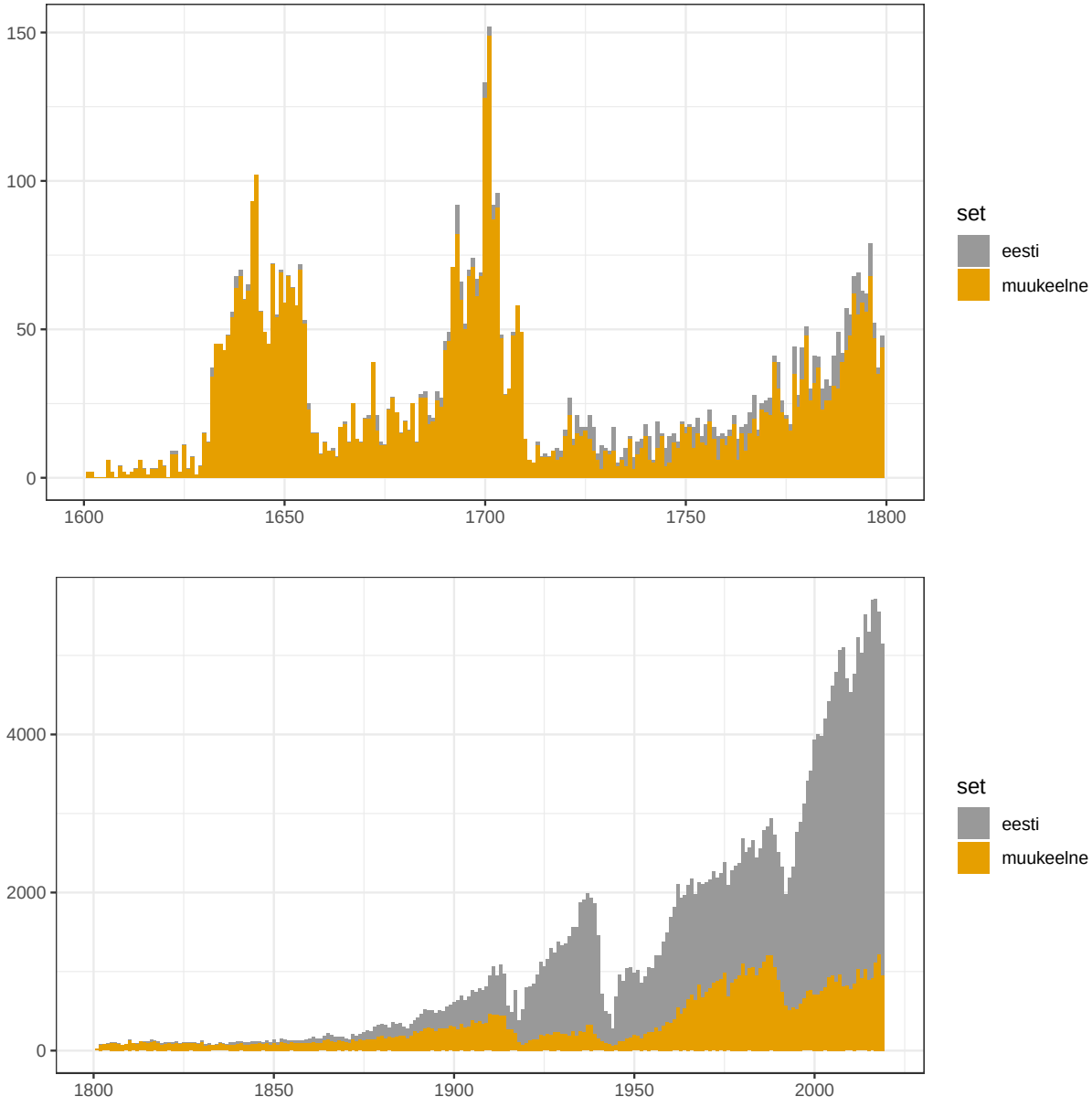
Intro

The current code organizes the ERB metadata into a tidy format: with one publication per line, and with the basic metadata information distributed into few relevant categories. This has been done to facilitate further data analytic steps, particularly when someone may be unfamiliar with Marc21 database structures.

Summary

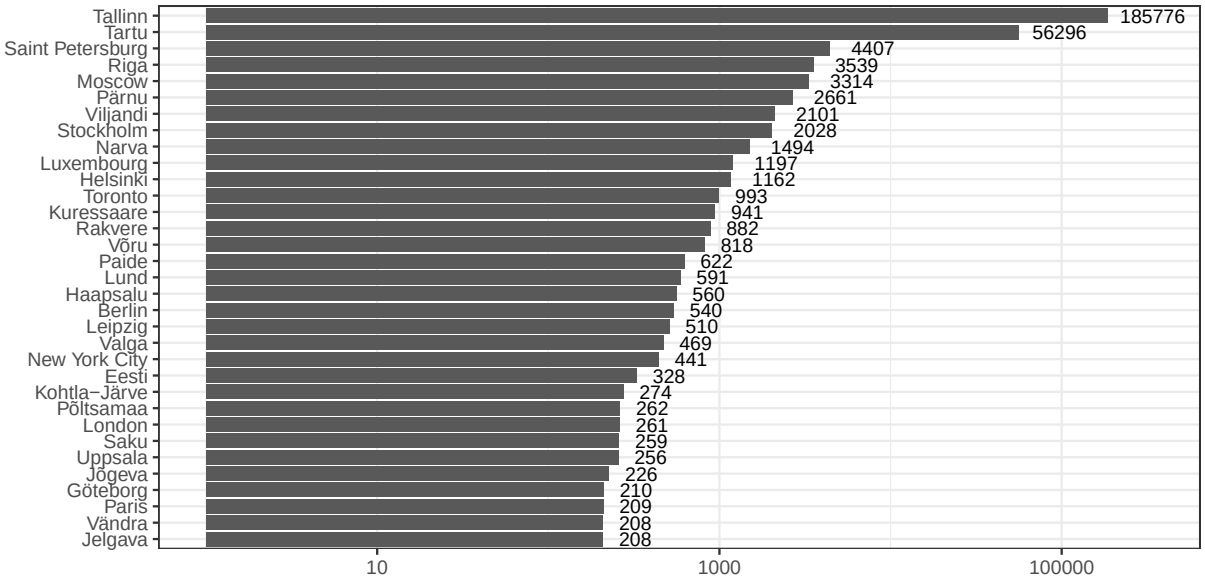
The dataset has altogether information on 291346 printed items. The coverage has been estimated to be better than 95% for all of the relevant published works.

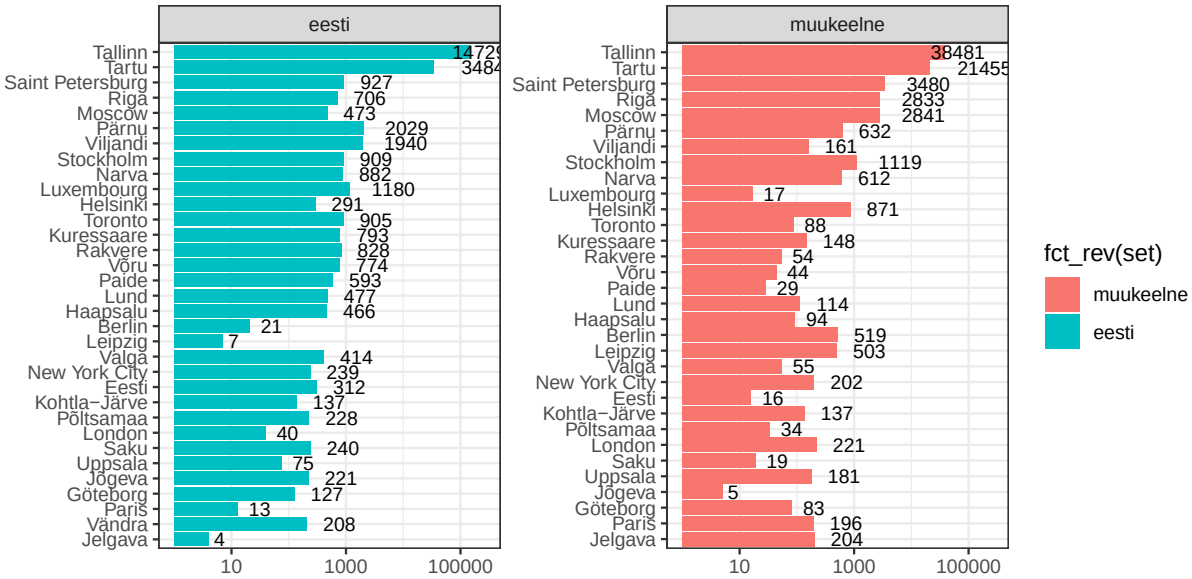
The dataset is divided into two: works in Estonian language, and works in other languages. The two sets are displayed over the year of publication here in different colors.



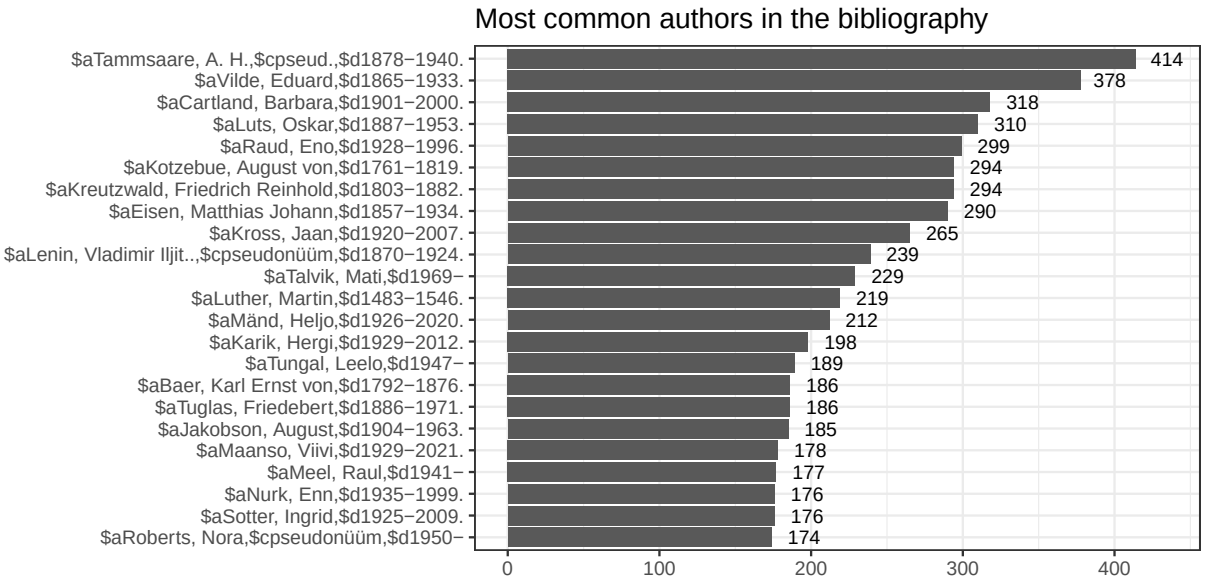
Cities

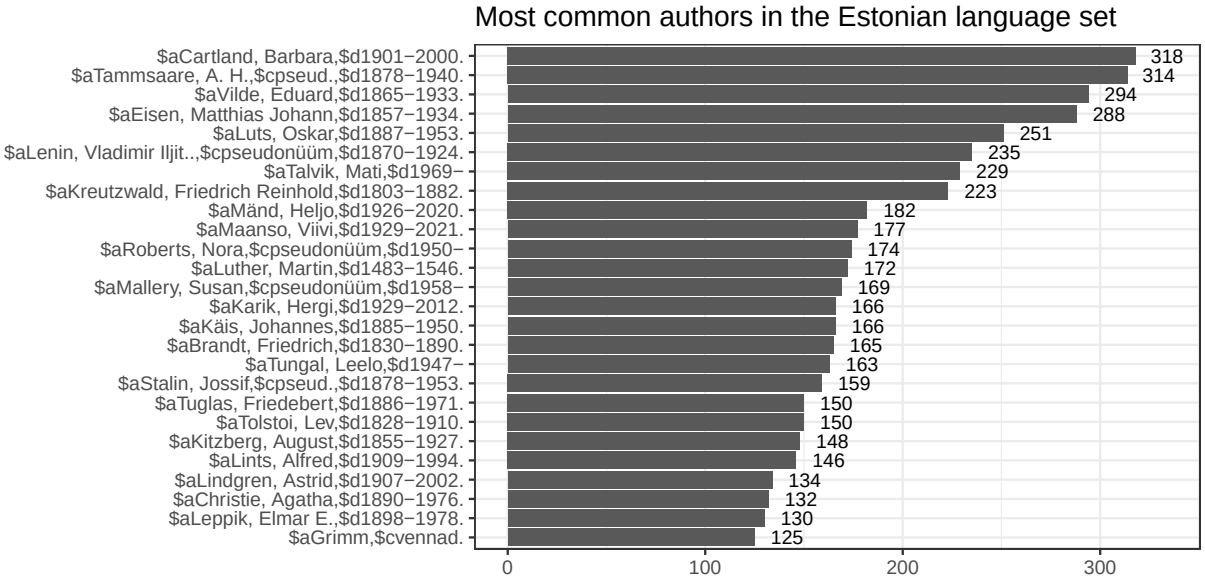
The city names where the works have been published have been harmonized manually and through a few heuristic algorithms. The tokens that appear more than 40 times should be mostly harmonized, while rarer tokens have been included only through algorithmic processing. Depicted below are the most common cities of publication and the number of publications in them separately for language sets.

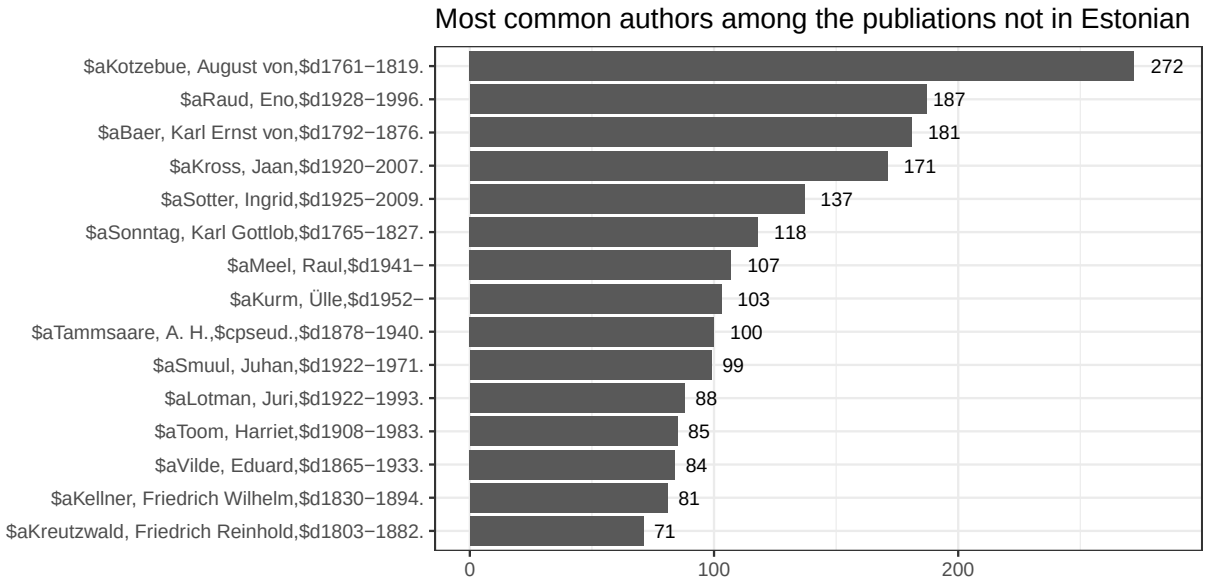

Tallinn 185776
Tartu 56296
Saint Petersburg 4407
Riga 3539
Moscow 3314
Pärnu 2661
Viljandi 2101
Stockholm 2028
Narva 1494
Luxembourg 1197
Helsinki 1162
Toronto 993
Kuressaare 941
Rakvere 882
Võru 818
Paide 622
Lund 591
Haapsalu 560
Berlin 540
Leipzig 510
Valga 469
New York City 441
Eesti 328
Kohtla−Järve 274
Põltsamaa 262
London 261
Saku 259
Uppsala 256
Jõgeva 226
Göteborg 210
Paris 209
Vändra 208
Jelgava 208
10
1000
100000


eesti
muukeelne
Tallinn 14729 38481
Tartu 3484 21455
Saint Petersburg 927 3480
Riga 706 2833
Moscow 473 2841
Pärnu 2029 632
Viljandi 1940 161
Stockholm 909 1119
Narva 882 612
Luxembourg 1180 17
Helsinki 291 871
Toronto 905 88
Kuressaare 793 148
Rakvere 828 54
Võru 774 44
Paide 593 29
Lund 477 114
Haapsalu 466 94
Berlin 21 519
Leipzig 7 503
Valga 414 55
New York City 239 202
Eesti 312 16
Kohtla−Järve 137 137
Põltsamaa 228 34
London 40 221
Saku 240 19
Uppsala 75 181
Jõgeva 221 5
Göteborg 127 83
Paris 13 196
Vändra 208 204
Jelgava 4
10
1000
100000
fct_rev(set)
muukeelne
eesti

Authors

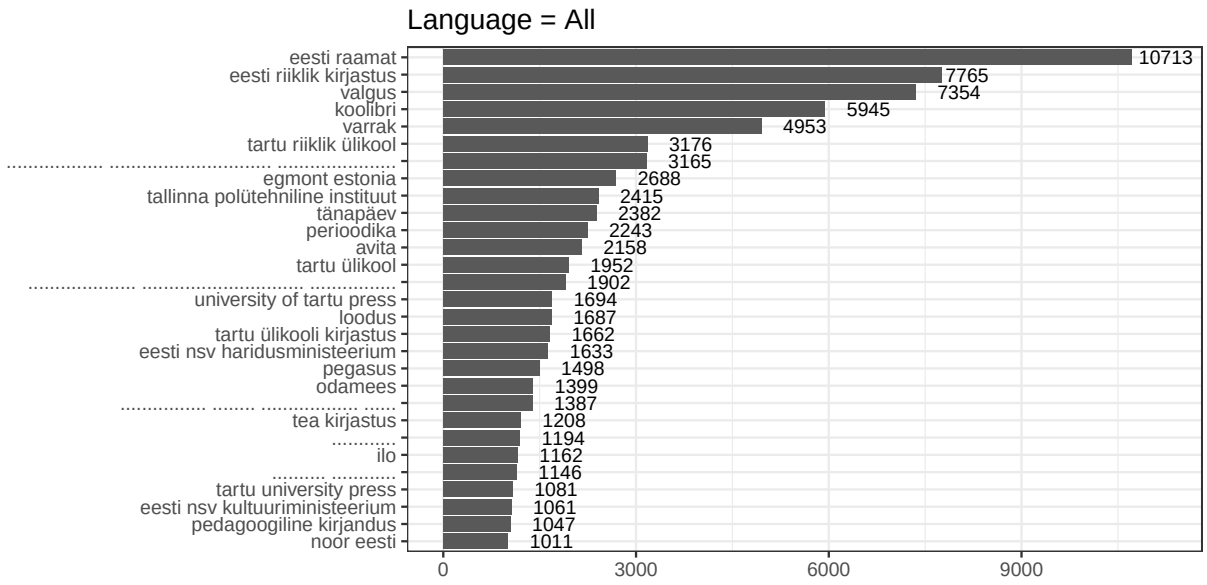

Most common authors in the bibliography
$aTammsaare, A. H.,$cpseud.,$d1878−1940. 414
$aVilde, Eduard,$d1865−1933. 378
$aCartland, Barbara,$d1901−2000. 318
$aLuts, Oskar,$d1887−1953. 310
$aRaud, Eno,$d1928−1996. 299
$aKotzebue, August von,$d1761−1819. 294
$aKreutzwald, Friedrich Reinhold,$d1803−1882. 294
$aEisen, Matthias Johann,$d1857−1934. 290
$aKross, Jaan,$d1920−2007. 265
$aLenin, Vladimir Iljit..,$cpseudonüüm,$d1870−1924. 239
$aTalvik, Mati,$d1969− 229
$aLuther, Martin,$d1483−1546. 219
$aMänd, Heljo,$d1926−2020. 212
$aKarik, Hergi,$d1929−2012. 198
$aTungal, Leelo,$d1947− 189
$aBaer, Karl Ernst von,$d1792−1876. 186
$aTuglas, Friedebert,$d1886−1971. 186
$aJakobson, August,$d1904−1963. 185
$aMaanso, Viivi,$d1929−2021. 178
$aMeel, Raul,$d1941− 177
$aNurk, Enn,$d1935−1999. 176
$aSotter, Ingrid,$d1925−2009. 176
$aRoberts, Nora,$cpseudonüüm,$d1950− 174
0
100
200
300
400

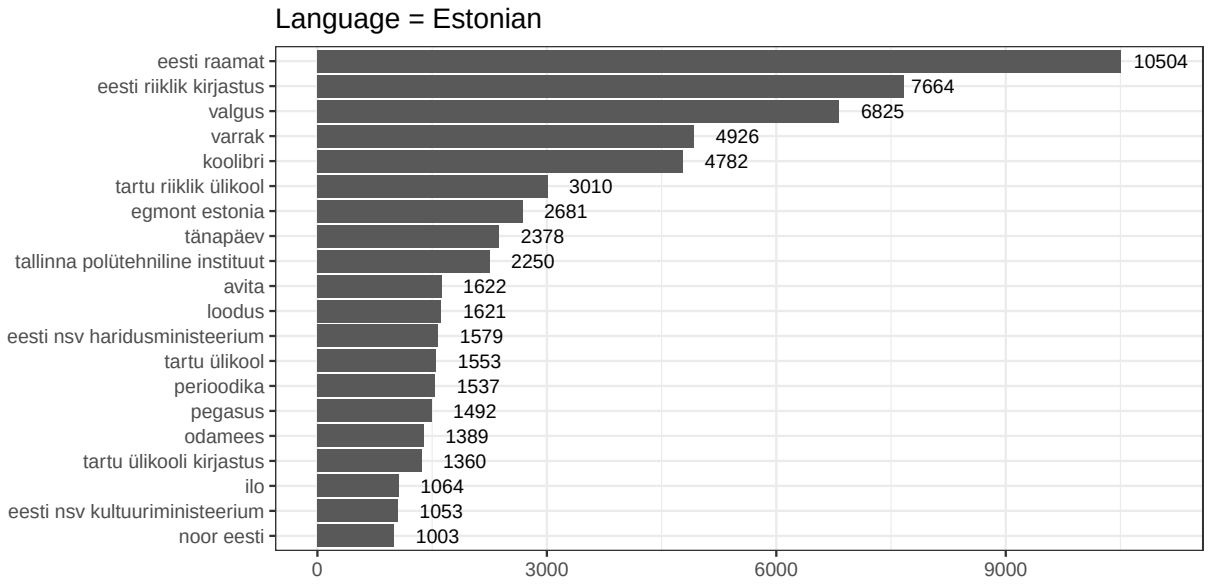

Most common authors in the Estonian language set
$aCartland, Barbara,$d1901−2000. 318
$aTammsaare, A. H.,$cpseud.,$d1878−1940. 314
$aVilde, Eduard,$d1865−1933. 294
$aEisen, Matthias Johann,$d1857−1934. 288
$aLuts, Oskar,$d1887−1953. 251
$aLenin, Vladimir Iljit..,$cpseudonüüm,$d1870−1924. 235
$aTalvik, Mati,$d1969− 229
$aKreutzwald, Friedrich Reinhold,$d1803−1882. 223
$aMänd, Heljo,$d1926−2020. 182
$aMaanso, Viivi,$d1929−2021. 177
$aRoberts, Nora,$cpseudonüüm,$d1950− 174
$aLuther, Martin,$d1483−1546. 172
$aMallery, Susan,$cpseudonüüm,$d1958− 169
$aKarik, Hergi,$d1929−2012. 166
$aKäis, Johannes,$d1885−1950. 166
$aBrandt, Friedrich,$d1830−1890. 165
$aTungal, Leelo,$d1947− 163
$aStalin, Jossif,$cpseud.,$d1878−1953. 159
$aTuglas, Friedebert,$d1886−1971. 150
$aTolstoi, Lev,$d1828−1910. 150
$aKitzberg, August,$d1855−1927. 148
$aLints, Alfred,$d1909−1994. 146
$aLindgren, Astrid,$d1907−2002. 134
$aChristie, Agatha,$d1890−1976. 132
$aLeppik, Elmar E.,$d1898−1978. 130
$aGrimm,$cvennad. 125
0
100
200
300



Publishers

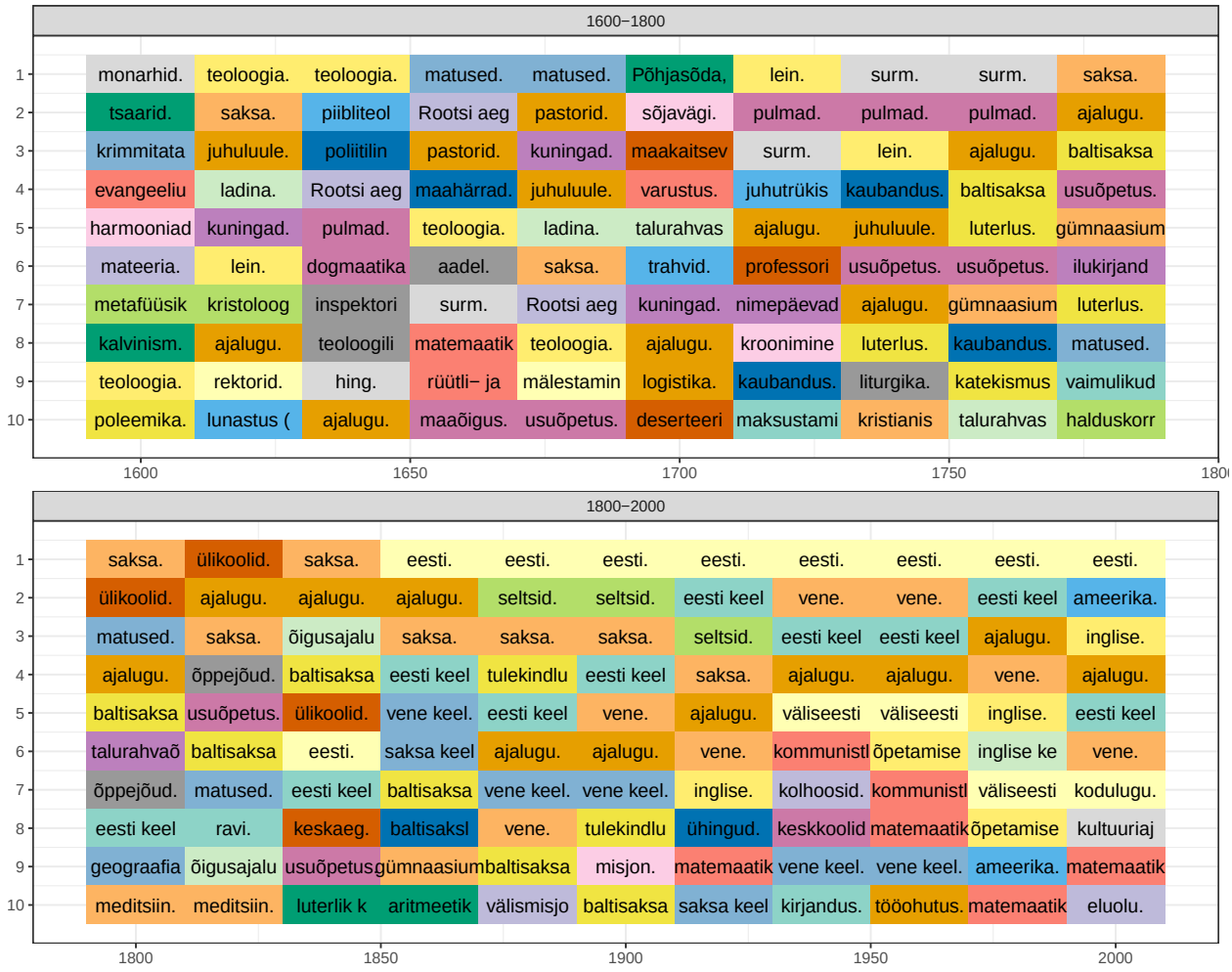
Publisher names have also been harmonized manually and algorithmically with the more frequent names in focus.





Topics

The bibliography has each publication marked by some general topics. Some works have no topic marking, many works have several topic markings. By 20-year intervals, here are the most common topics in the set. For visualization purposes, only the first 10 characters of the topic are shown.

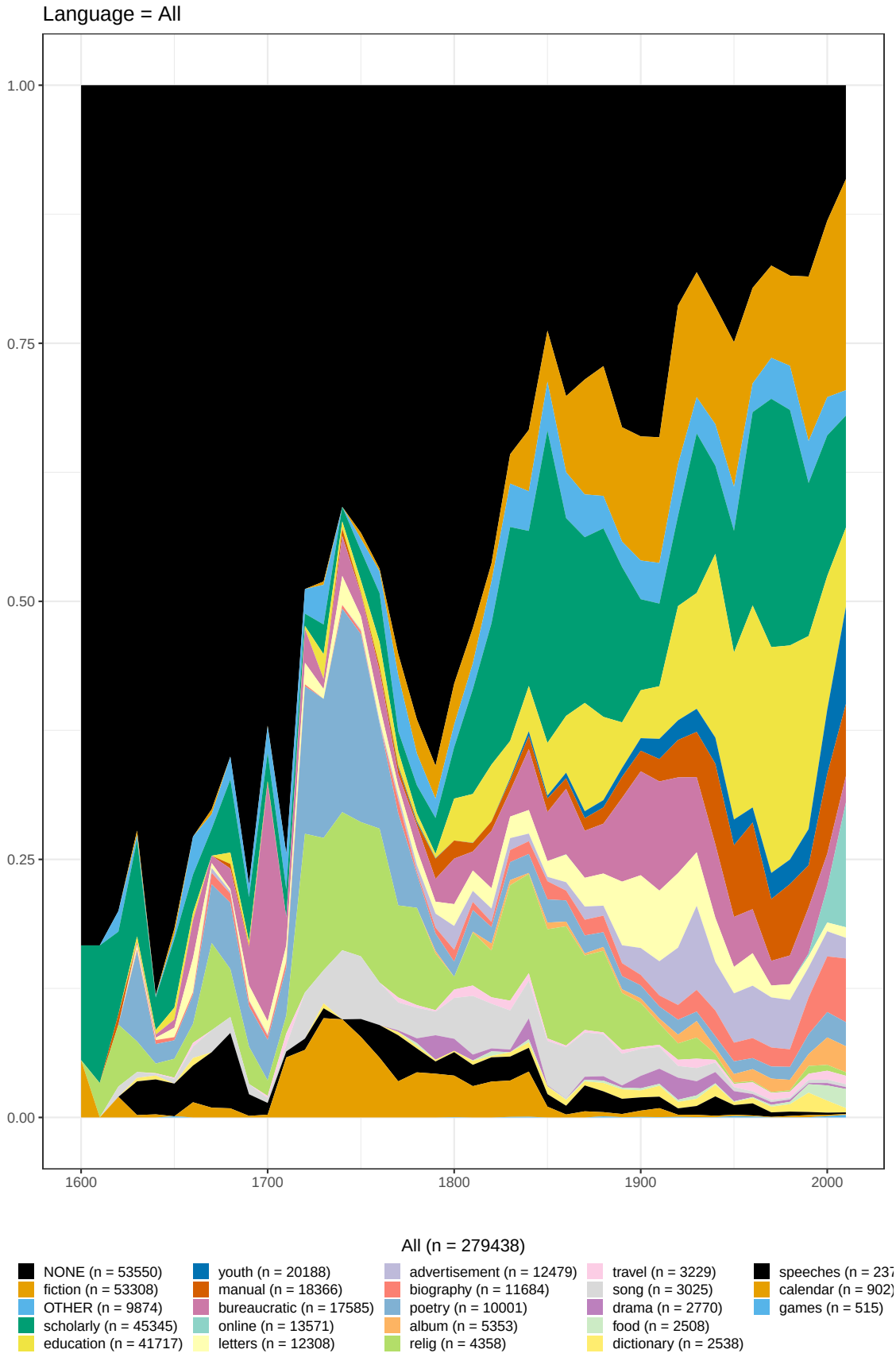


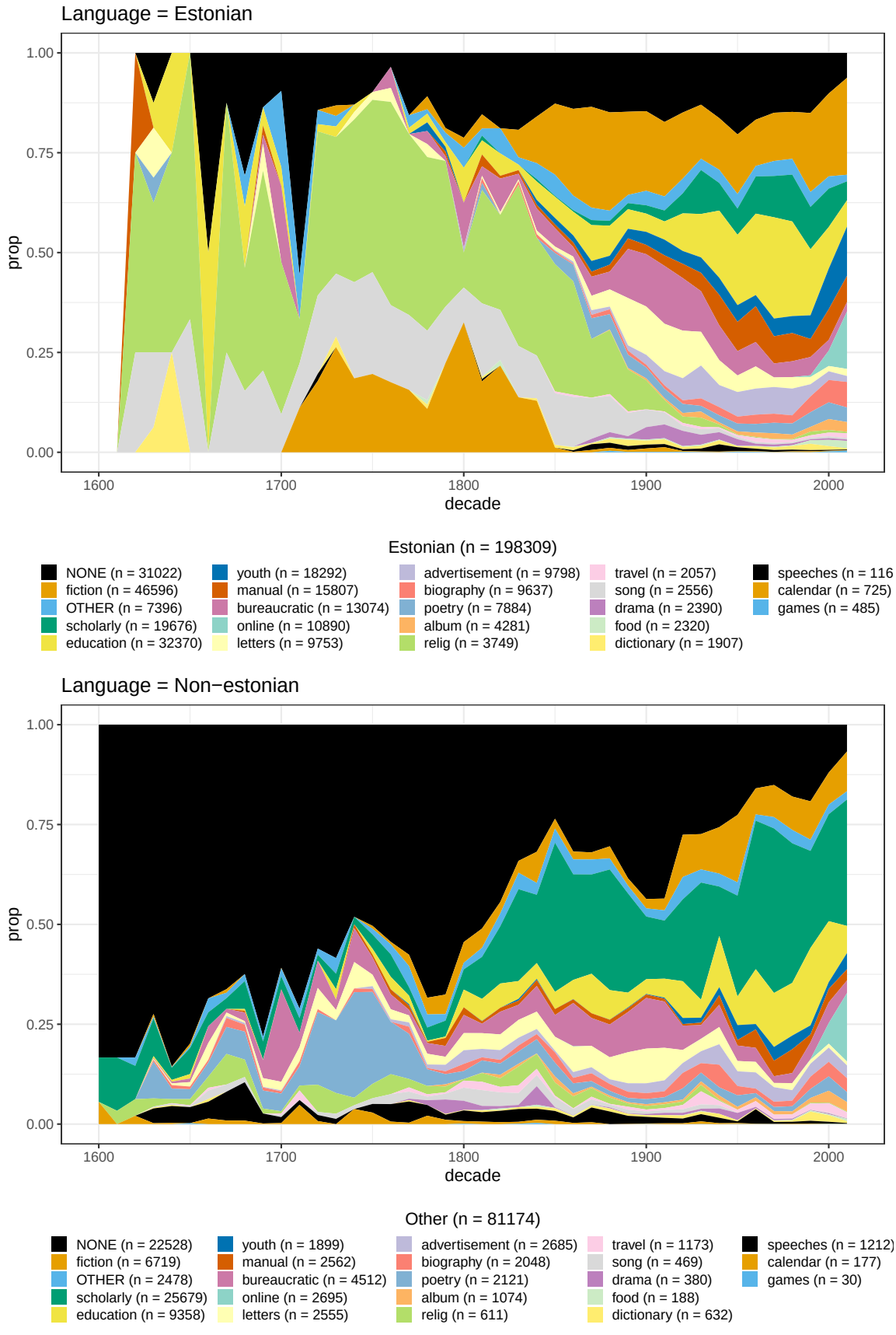
Genres

The bibliography also includes genre markers on many works. Some works have no markers, many have several. Here, we have built a few larger categories based on the dataset and visualized their frequencies over time. NONE means that the works did not have any genre markers.

Genres have been grouped into 21 major categories. Many remain uncategorized. A total of 474 unique keywords are currently classified as OTHER. Top genres there are given in the table below. 1 works in the dataset had no keyword marking.

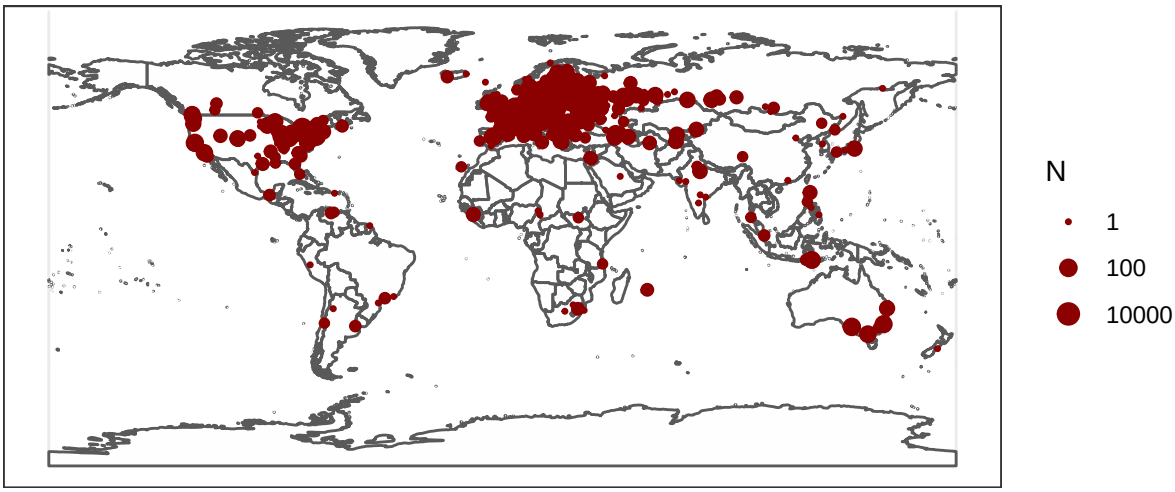
genres	N
annotatsioonid.	342
preprindid.	311
esseed.	306
harjutused.	282
õpetajaraamatud.	259
tabelid.	214
ehitusnormid.	197
anekdoodid.	188
kunstnikuraamatud.	185
aforismid.	184

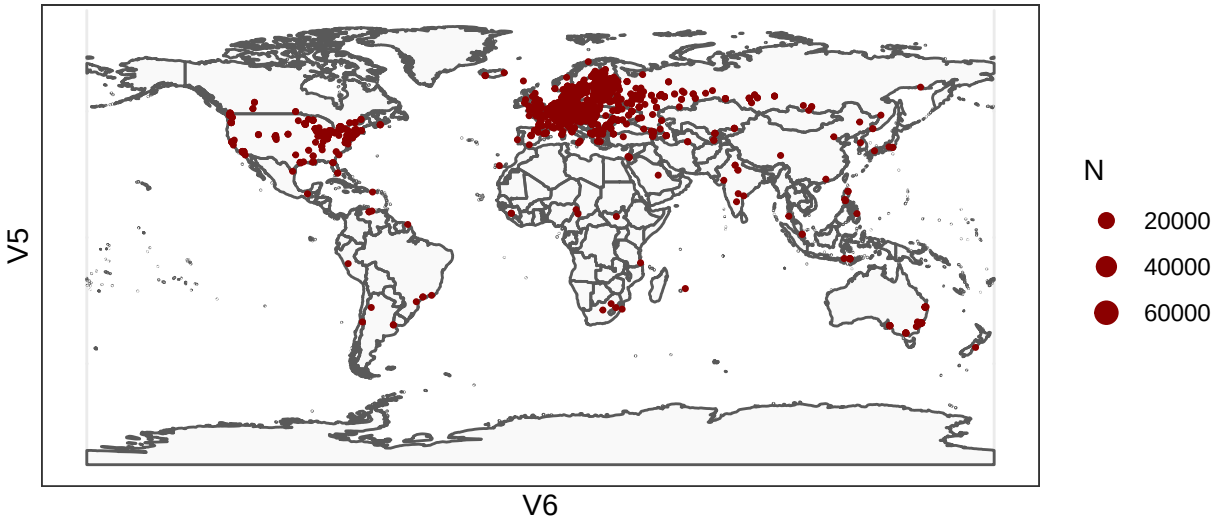

Language = All
1.00
0.75
0.50
0.25
0.00
1600
1700
1800
1900
2000
All (n = 279438)
NONE (n = 53550)
fiction (n = 53308)
OTHER (n = 9874)
scholarly (n = 45345)
education (n = 41717)
youth (n = 20188)
manual (n = 18366)
bureaucratic (n = 17585)
online (n = 13571)
letters (n = 12308)
advertisement (n = 12479)
biography (n = 11684)
poetry (n = 10001)
album (n = 5353)
relig (n = 4358)
travel (n = 3229)
song (n = 3025)
drama (n = 2770)
food (n = 2508)
dictionary (n = 2538)
speeches (n = 237
calendar (n = 902)
games (n = 515)

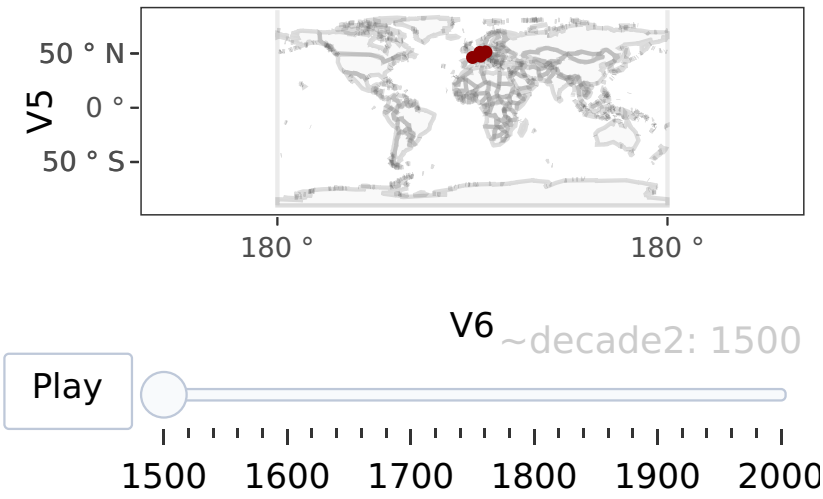

Language = Estonian
prop
1.00
0.75
0.50
0.25
0.00
1600
1700
1800
1900
2000
decade
Estonian (n = 198309)
NONE (n = 31022)
fiction (n = 46596)
OTHER (n = 7396)
scholarly (n = 19676)
education (n = 32370)
youth (n = 18292)
manual (n = 15807)
bureaucratic (n = 13074)
online (n = 10890)
letters (n = 9753)
advertisement (n = 9798)
biography (n = 9637)
poetry (n = 7884)
album (n = 4281)
relig (n = 3749)
travel (n = 2057)
song (n = 2556)
drama (n = 2390)
food (n = 2320)
dictionary (n = 1907)
speeches (n = 116
calendar (n = 725)
games (n = 485)
Language = Non−estonian
prop
1.00
0.75
0.50
0.25
0.00
1600
1700
1800
1900
2000
decade
Other (n = 81174)
NONE (n = 22528)
fiction (n = 6719)
OTHER (n = 2478)
scholarly (n = 25679)
education (n = 9358)
youth (n = 1899)
manual (n = 2562)
bureaucratic (n = 4512)
online (n = 2695)
letters (n = 2555)
advertisement (n = 2685)
biography (n = 2048)
poetry (n = 2121)
album (n = 1074)
relig (n = 611)
travel (n = 1173)
song (n = 469)
drama (n = 380)
food (n = 188)
dictionary (n = 632)
speeches (n = 1212)
calendar (n = 177)
games (n = 30)

Publication locations

Publication locations have also been joined with geographic information in the GeoNames database. This allows the publication locations also to be displayed on a map. To see the interactive version of the map, use the html overview.




V5
V6
N
20000
40000
60000


50 ° N
0 °
50 ° S
V5
180 °
180 °
V6
~decade2: 1500
Play
1500
1600
1700
1800
1900
2000