

Digiteeritud Eesti ajalehed uurimisallikana

Peeter Tinit (Rahvusraamatukogu, Tartu Ülikool)

EHAK 2024

Tänased küsimused

Digiteeritud Eesti ajalehed uurimisallikana

- Milleks ja kuidas kasutada digiteeritud ajalehti uurimistöös?
- Digiteeritud ajalehed ja nende representatiivsus
- Tööriistad ja vahendid
- Väljakutsed ja tulevik

1. Digiteeritud ajalehed uurimisallikana

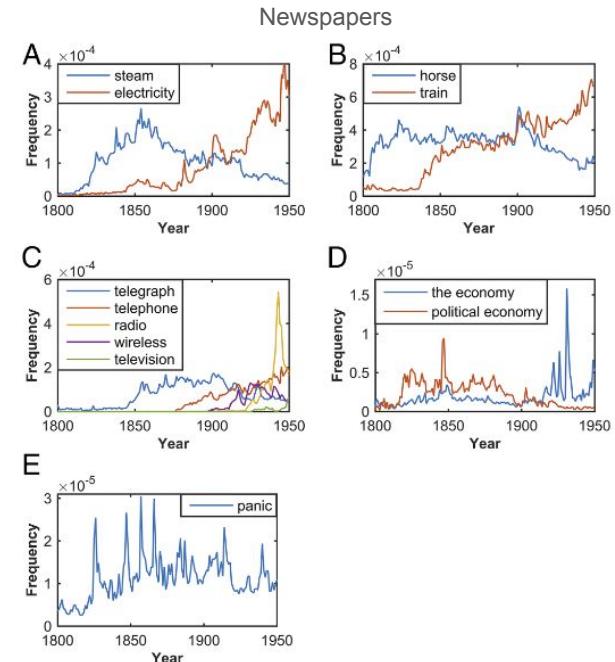
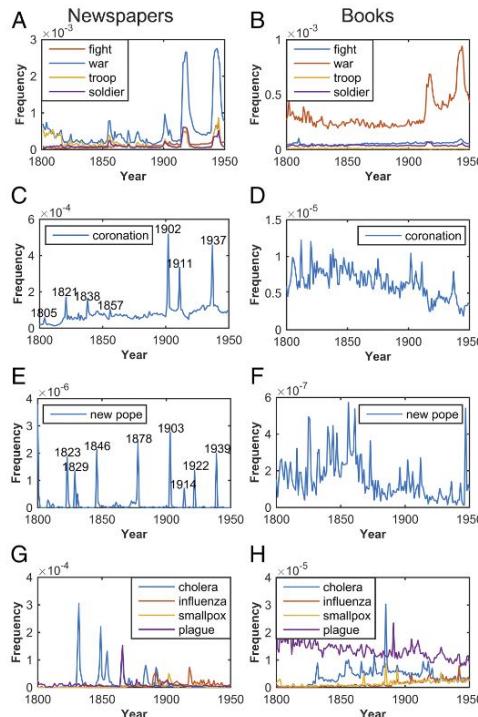
Digiteeritud lehed uurimisallikana

Ajalehed näitavad ajalisi trende

Andmed: 120 kohalikku lehte
Suurbritannias (14% ilmunust)

Vasakul: võrdlus ajalehtede ja
raamatute vahel

Paremal: mõned
sõnavõndlused (nt elektro vs
auruvedu, hobune vs rong)

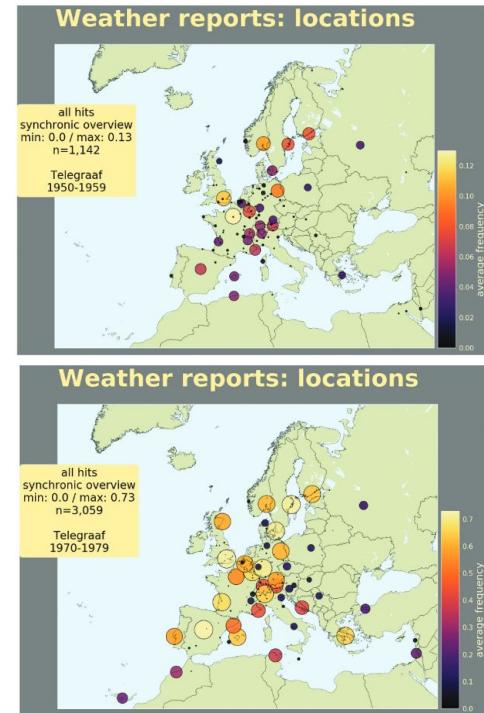
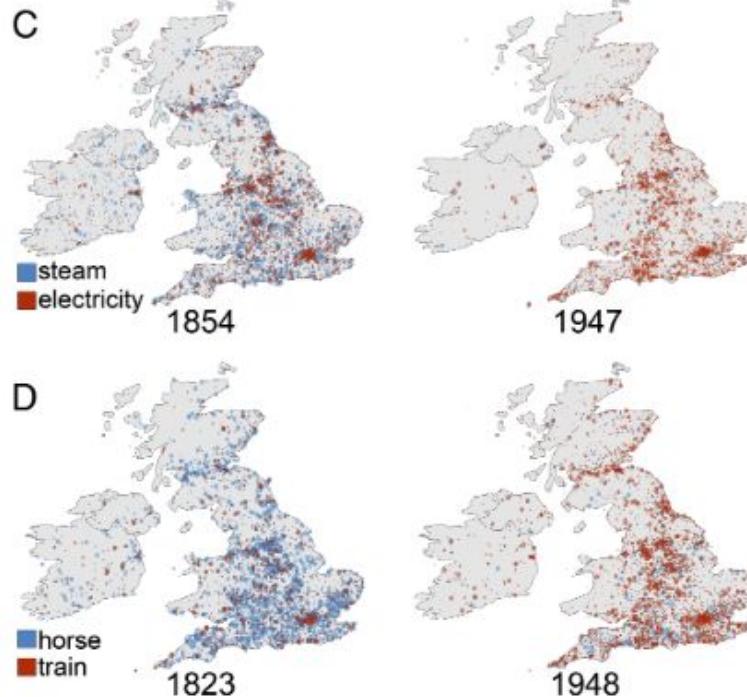


Digiteeritud lehed uurimisallikana

Ajalehed avavad sündmuste konteksti

Vasakul: kohanimed tehnoloogia ja transpordi kontekstis Suurbritannias

Paremal: ilmateates mainitud kohanimed Hollandi lehes Telegraaf 1950ndad ja 1970ndad



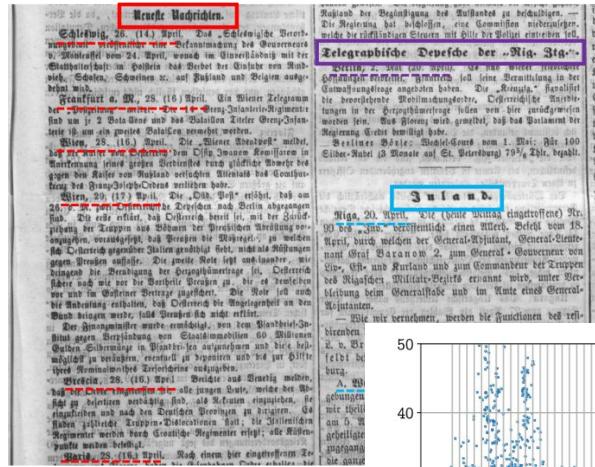
Digiteeritud lehed uurimisallikana

Ajalehed sisaldavad struktureeritud informatsiooni

Andmed: Riga sche Zeitung
1802-1888, 18,499 numbrit.

Uudised on kindla vormiga,
algul toimumiskohu ja kuupäev.

Saab vaadata mitu päeva läks
aega, kuni uudis jõudis Riiga.



Schleswig, 26. (14.) April.
Frankfurt a. M., 28. (16.) April.
Wien, 28. (16.) April.
Wien, 29. (17.) April.
Brescia, 28. (16.) April.
Paris, 28. (16.) April.
Berlin, 2. Mai (20. April).
Riga, 20. April.
A. Wenden, 17. April.

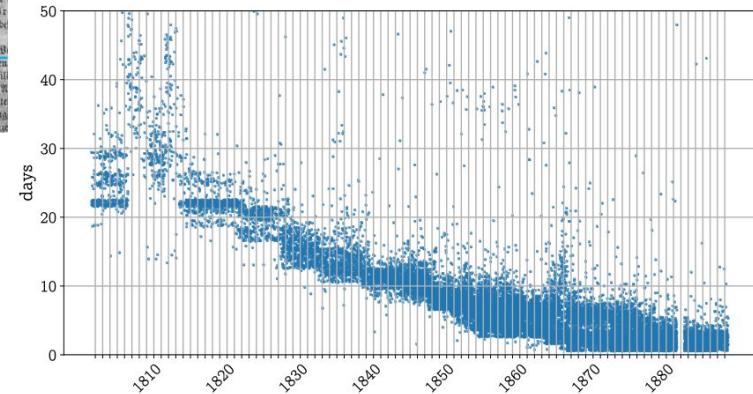


Figure 5. All news from London plotted by speed

2. Digiteeritud ajalehtede seis Eestis



JUVENILE READER. 37

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

23 After this long confinement, James could neither run nor play. With others, children, but he had to sit, all day long, wishing that he had been more attentive to his mamma's request.

24 It was a great treat to Mrs. Comedy to see her little son in this unhappy condition; and she often regretted that she had permitted him to grow up so obstinate and obstinate.

ABSURDITY OF PRIDE.

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

Varastati jalgratas
8. skp. oöl on Wöhmas warastatud Johan | Saewasterde kuulutust rauaaltööstusest, Edu | jalgratas. Waras on tulnud akna kaudu sisse ja | tahkunud uksest, seda seestpoolt awades.

Warastatud jalgratas oli toodud töökolla | paigutusseks ja kaubiks. Jõekalda, Imavere | ja Lihula linnadele.

Ilus piidupüür Karksi
Esimene luureküla luujenit muidulise ja telle | ümberlennuna taimole täielikult pildutavaooda, mit torraldatud Karksi liikloomina koolide poolt | teostatud läbirääkimistel lastest, mis olid | ehitatud ja kasutatud laste laste jumalateks | põimist. Nii püüdeti külalane ilmasit ali rohvalt | vabastada meeltutanud mägedest. Bildupäri | algas läbirääkimisega üldsegi lõpuks, ja par- | tiidil läks siis ümber lojundamise testidega. | Võistluseid, les pragu on full pensioni, hr. Roots | päästetehaseks läksid. Zelle järel tanti spilaste | poolt, kellel oli kõige suurimad ja võimelisemad | joonised, mis enam-mõnest ümberlennust. Edu | läitis läpsit halu muist.

Lutut olli alempe. Lutuleku läpsing uues ma- | jaanipäevale ei saanud läbirääkimisega poolt. Samuti | eette A. Saareni „Sakrament“.

Pauguts revolvr lätepael
Aleflikkonsülituse lätepaleku Karksi Lõõtmise

Digiteeritud ajalehtede seis Eestis



- 1993 mikrofilmidele, 2003 algab digiteerimine
 - Kõigepealt Perno Postimees (1857-1885), Postimees (1886-1920), Päewaleht (1905-1912)
- 2004 avatakse lugemisportaal DEA, lehed kuupäevade järgi
- 2014 täistekstiotsing DEAs (85 publications, 100 000 pages)

- Lehed DEAs, ETERAs, TÜ DSPACE-is, TTÜ rmtks, Kivikeses

- DEAs praegu 2787 publications, 5.8M pages, 15.6M articles

Digiteerituse seis

Kollektsiooni sisu: DEA + koondkorpus

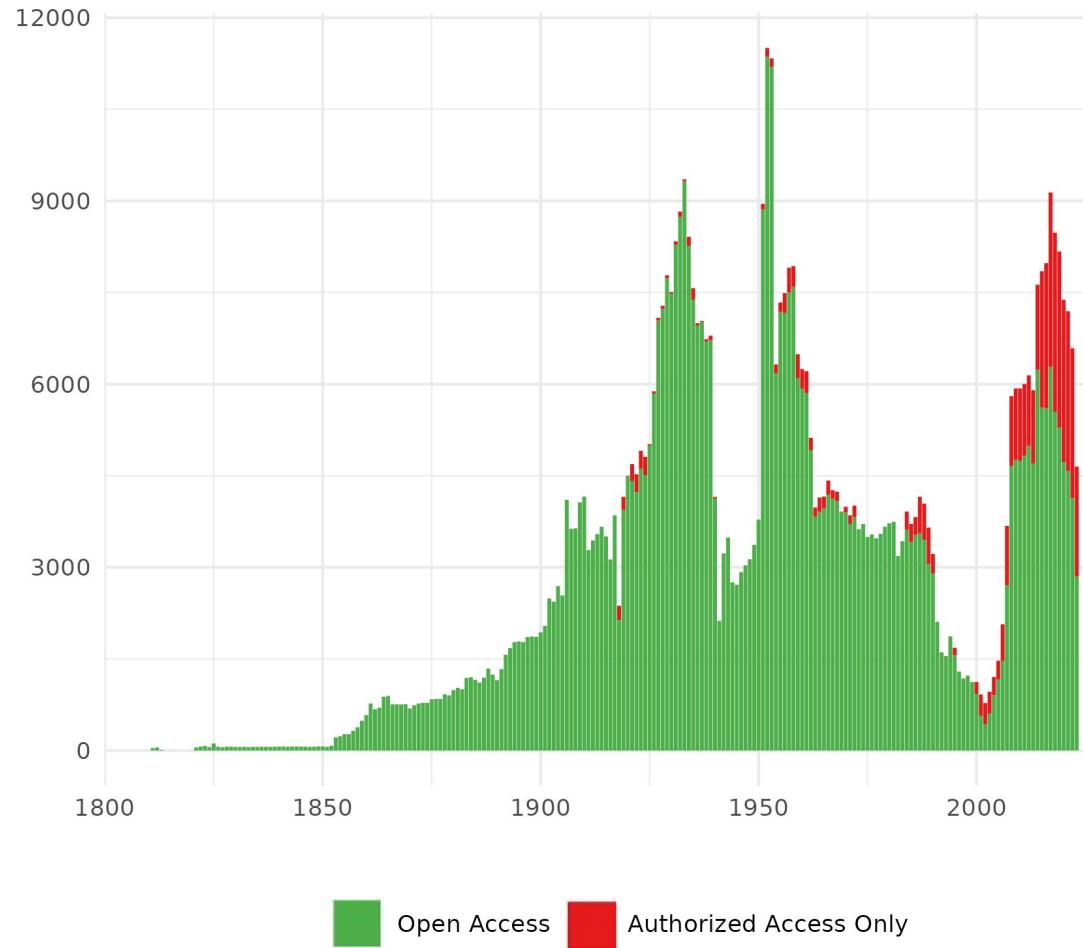
Väljaanded 1800-2023 (n = 622,061)

Roheline = avatud ligipääs (92.8%)

Punane = ainult autoriseeritud ligipääs
(7.2%)

Interaktiivne ülevaade:

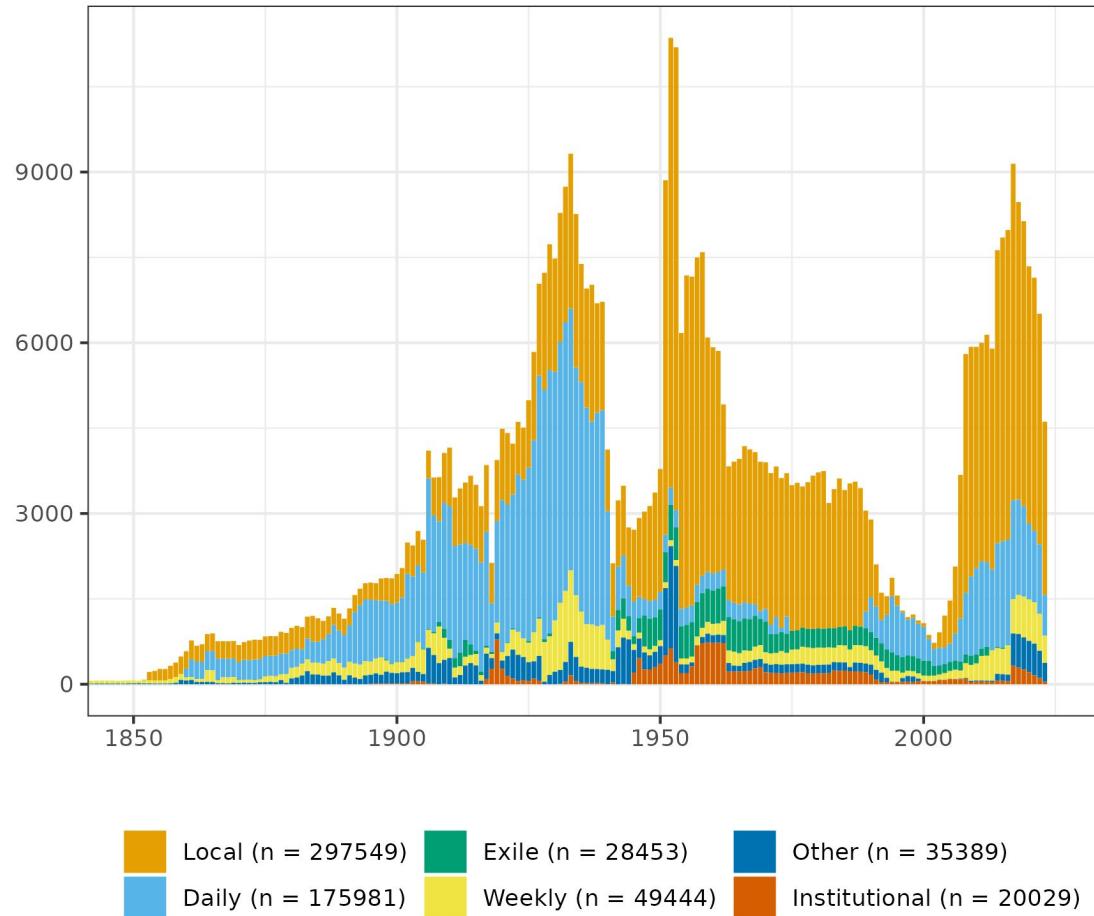
[https://peetertints.github.io/reports/nlib/
dea_info.html](https://peetertints.github.io/reports/nlib/dea_info.html)



Mis andmestikus on?

Ülevaade väljaande tüübi järgi

- Päevalehed
- Nädalalehed
- Kohalikud lehed
- Välis-Eesti lehed
- Asutuste lehed



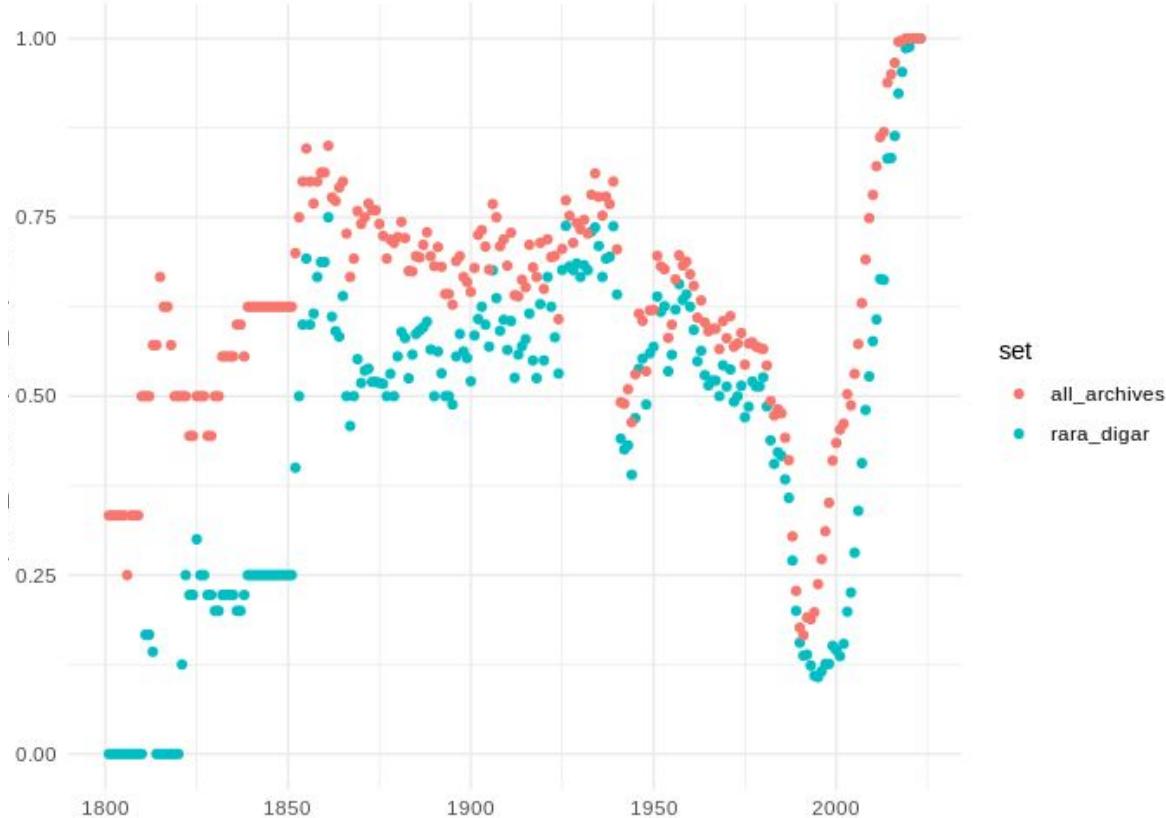
Digiteerituse seis

Digiteerituse kate

Ülevaade Eesti Rahvusbibliograafia põhjal: vaatame iga aasta kohta, kui suur osa lehti on digiteeritud.

Nt aastal 1900 on ERB-is 43 ajalehte,
27 neist on digiteeritud = 62.8%

1800–2020 on keskmiselt aastas
58.7% mingil kujul digitaalselt
ligipääsetav (vrd näites toodud 14%)



3. Kuidas digiteeritud ajalehti kasutada?

Näide

Näiteks tahame kasutada digiteeritud ajalehti,
et analüüsida suhtumist loodusesse 1920-tel.

Täistekstiligidipääs:

0 m., nahk- wöi ülem kalingorköites 1000 m. Eesti **loodus**. 8 iseseisvat artiklit Eesti ranniku, mere, geolo
i wärwikiillasemat, meeleteurikkamat ja sügawamat **loodus**ilu näinud. Ka kaasreisijate näod kõnelewad sedasa
See üsna magus lause on aga wabandataw: on nähtud **loodust** Antoni kaudu, ja tema psühholoogiale lause wastab
a oma office'ist wöi büroost wälja soidab wabasse **loodusesse** uut jõudu koguma. Kes kaugemale sõita ei jõua
ole suutnud köigutada isegi ilmasöda. Siin wabas **looduses** andub inglane täieliku innuga oma armsale lõbus
rk. Füüsiline maateaduse ülesannete kogu, I wihk «**Looduse maatlussd**"". Hind 85 markm Maateaduse ülesannete

Andmete tasakaal
(nt korpusse suurus)

Andmete kvaliteet
(nt OCR)

Esinduslikkus
(kas materjale on piisavalt?)

Uuringu
loomiseks on
vaja

Oleks hea teada juba planeerimise ajal

Uurijal oleks vaja teada

- Millised andmed on olemas (erinevates kogudes)
- Andmete kvaliteet (varieerub)
- Andmete esinduslikkus (digiteeritud vs avaldatud)

Oleks hea teada juba planeerimise ajal

Kogude arendamiseks

Püüda üles ehitada komplekti, mida oleks uurijale vaja

- Rahvusraamatukogu digilabor, avaandmed, juhendid ja näited



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

Tegemise käigus õppimine

- Juhtumiuuringud (nt 6 juhtumiuuringut 12 tudengilt TÜs, 5 juhtumiuuringut digilaboris)

Võimaluste läbi proovimise käigus selguvad vajadused

- Nt ligipääs täistekstidele, n-grami lugeja, ülevaated andmestikust

4. Vahendid ja tööriistad

Mida on ajalehtedes kirjutatud?

AVALEHT OTSING VÄLJAANDED ILMUMISAEG TEEMAD ABI Logisse ENG PYC

Otsi ainult kättesaadavaid Detailne otsing

OTS Täistekst Otsi 

Otsingute ajalugu

VALI KUUPÄEV

E	T	K	N	R	L	P
1	2	3	4	5	6	7
8 	9 	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

Valitud kuupäev
 Kättesaadav
 Juurdepääs piraatud

TERE TULEMAST!

"Kultuuripärandi digiteerimine 2018-2023" raames on digiteeritud 1,5 miljonit lehekülge tekste (sh 700 000 lk ajalehti). Suure osa projektist "Massdigiteerimine 2019-2022" moodustab nõukogudeaegne ajakirjandus, misläbi paraneb oluliselt 20. sajandi teise poole kate digitaalarhiivis.

Digiteerimist rahastas Euroopa Regionaalarengu Fond.

Ajalehed (1690)

Ajakirjad (600)

Jätkväljaanded (511)

NB! Ajalehed on portaalis alates 1811.a. Ajakirjad ja jätkväljaanded on portaalis aastast 2017 (varasemaid väljaandeid vaata digitaalarhiivis DIGAR).

Leitud vigadest saad teada anda, kui teed väljaande vaaturus avatud lehekülijel hiirega parem-klikk ja valid menüüst „Teata veast“. Teatele lisatakse automaatselt ka lehekülijel püsiliink.

Artiklite portaali täiendatakse iga päev ja see sisaldb hetkel: 639357 väljaannet, 5932121 lehekülge ja 15672415 artiklit.



HUVIPAKKUVAT

PARIMAD KORREKTORID

1. Marike V. 122,554
2. Herki H. 119,078
3. Meelik M. 72,401
4. Kalle G. 57,650
5. Maret M. 48,998

[Lisainfo...](#)

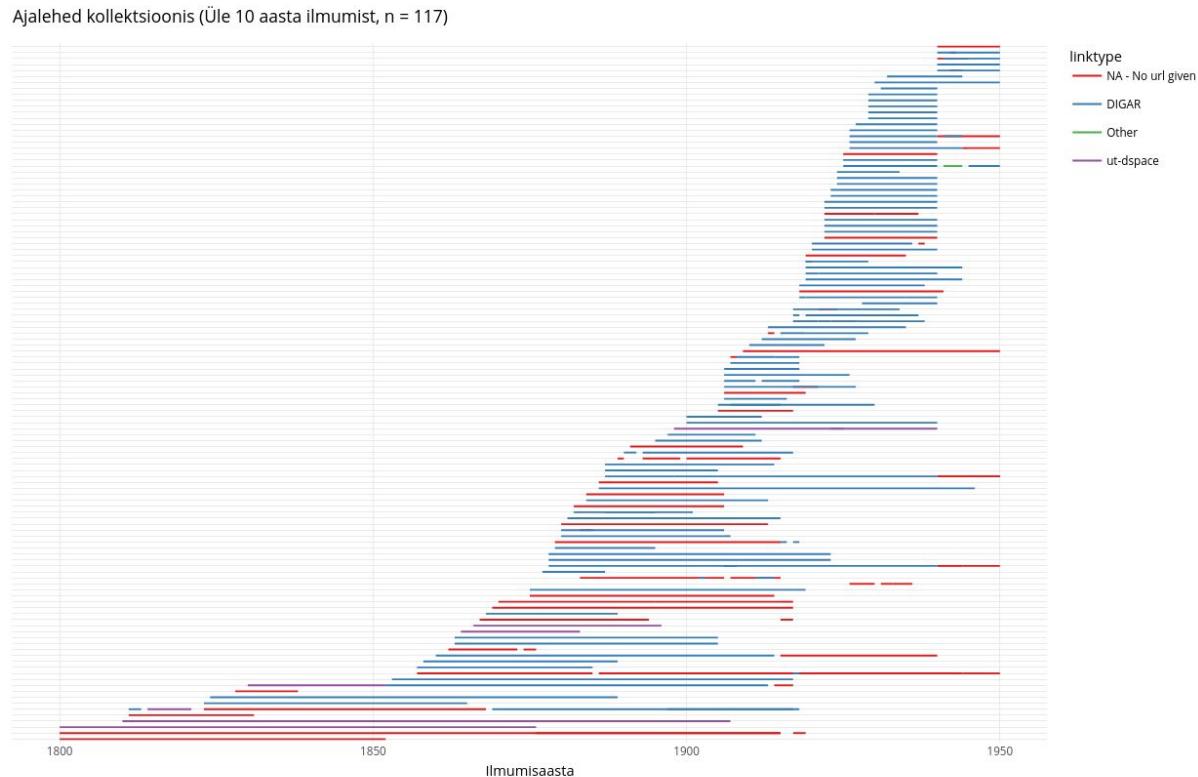
Millised lehed on digiteeritud?

Andmed kontekstis

Tööriist, millega saada ülevaadet lehtedest, mis on digiteeritud ja uuringutes kasutatavad.

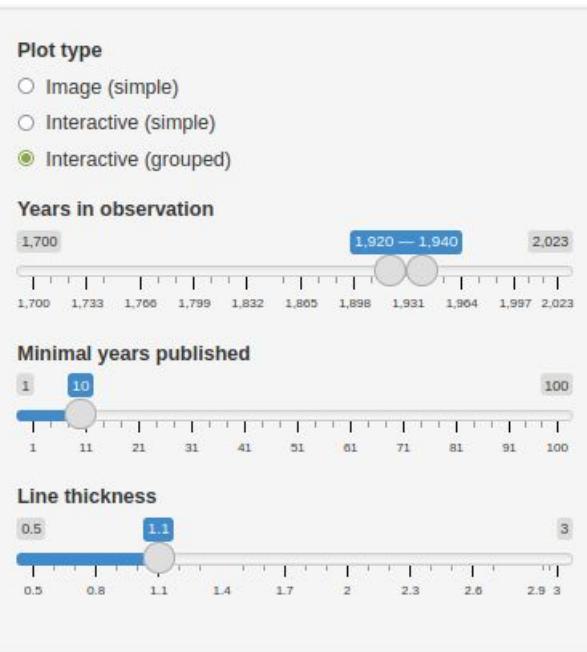
Põhineb Eesti Rahvusbibliograafia infol ja sellele lisatud dea.digar sisul.

<https://digilab.rara.ee/tooriistad/digiteeritud-ajalehed-eestis/>

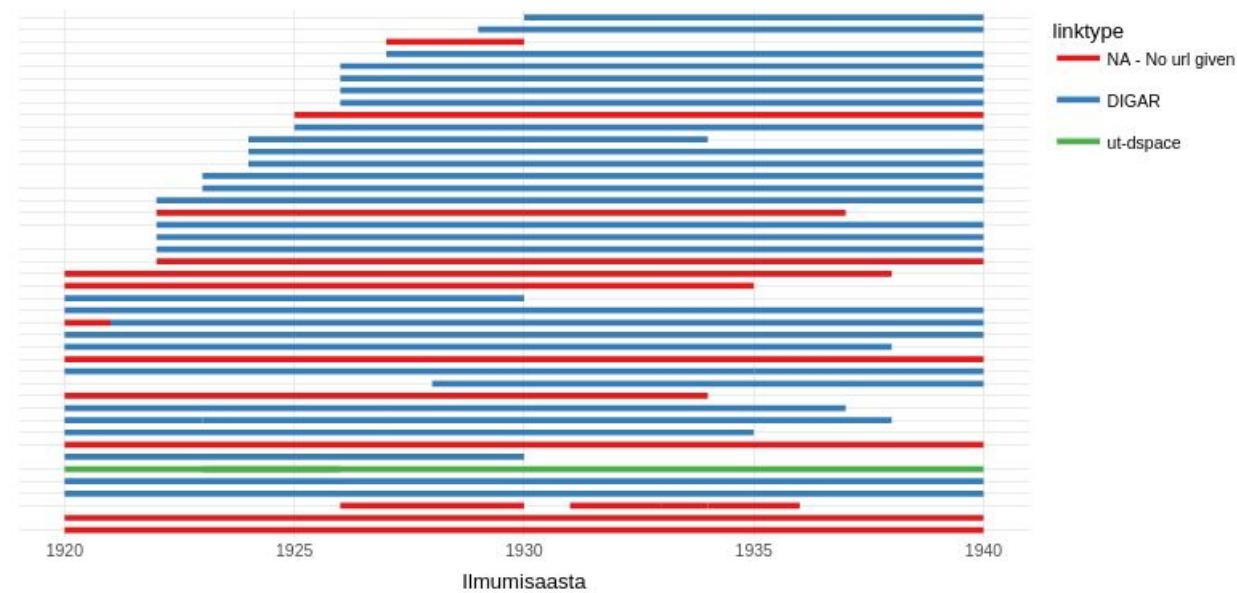


Millised lehed on digiteeritud?

Näidis - ajalehed 1920-1940, mis ilmusid kokku vähemalt 10 aastat



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



Milline on andmestiku sisu?

Metaandmete sirvija

Saab uurida andmestiku sisu teatud parameetrite järgi

Näiteks võrrelda kahe ajalehe mahtu või valida mõne riigi ajalehed

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



Explore DIGAR collection

HOME METADATA NEWSPAPER SEARCH TIMELINE LANGUAGES COUNTRIES TYPES

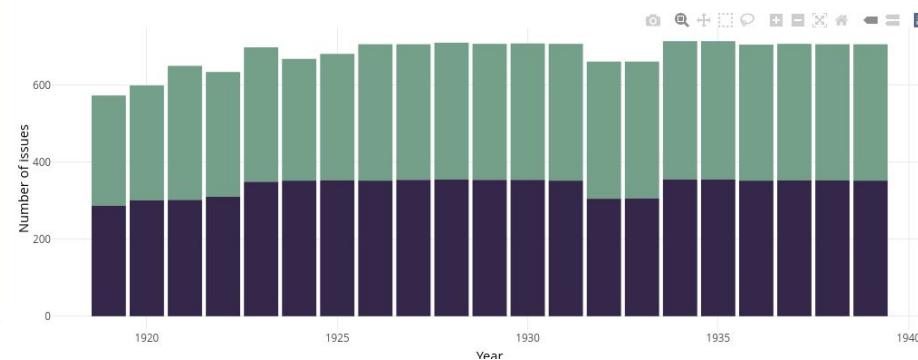
Search by title

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 shows summarised information for each newspaper.

Choose newspaper(s):

Päevalteht Postimees

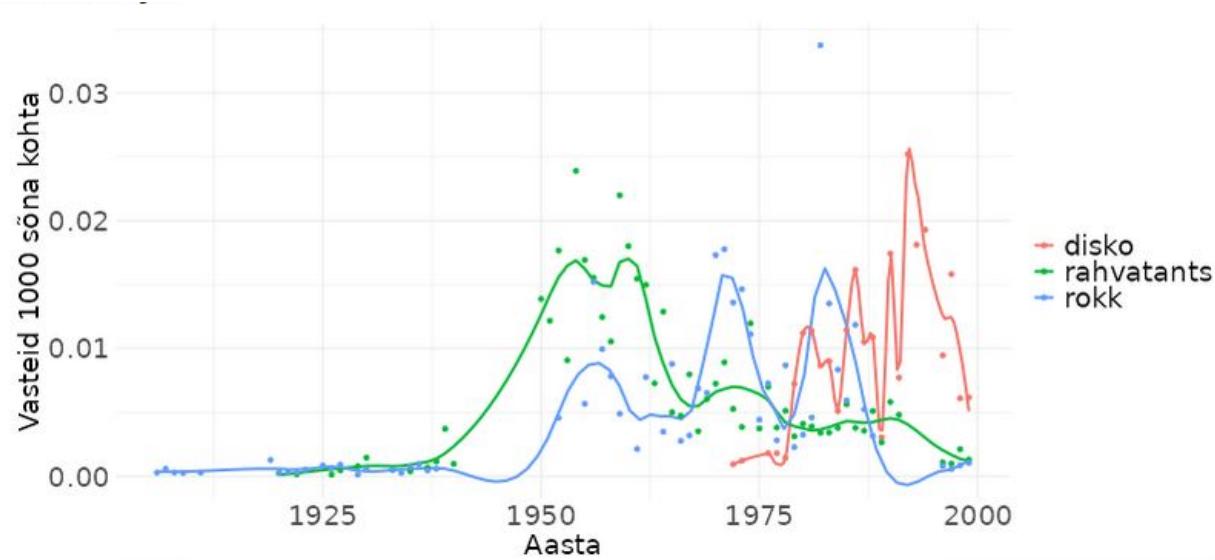


Mida ajalehtedes on kirjutatud?

Sõnamitmike loendaja

Sõnamitmike sagedused annavad kiire ülevaate tekstide sisust

Vaid valitud ajalehed, et valim oleks esinduslik



<https://digilab.rara.ee/tooriistad/sõnamitmikud-ajalehtedes/>

Rahvusraamatukogu sõnamitmike loendaja

Otsing

disko rokk rahvatants

Töötlus

Sõnad (muutmata)

Lemmad (sõnade algvormid)

Aastad vaatluse all

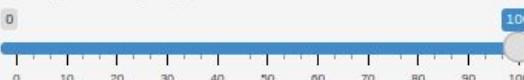
1850 1900 2000 2023



Joone kõverus



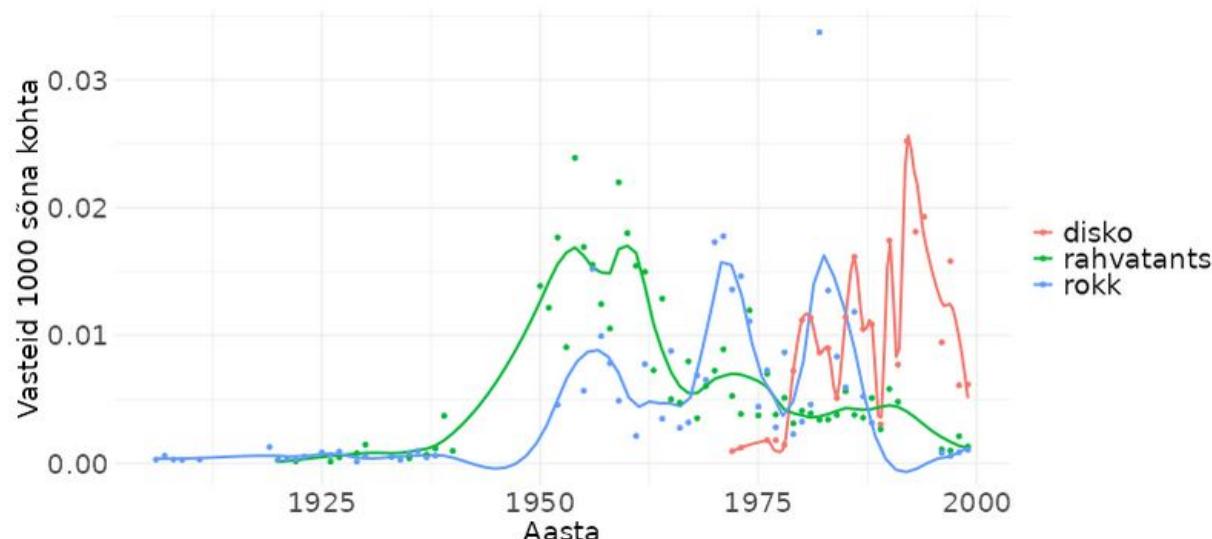
Lõika joonisel y-telg ära siin



Märgi joonisel iga mitmes aasta



RaRa sõnamitmike loendaja koondab sõnade ja sõnamitmike sagedusi Rahvusraamatukogu ajalehekollektioonides. Sisestades ülal otsingusse mõne sõna, näeb paremal joonisel selle sagedust kollektioonides 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 disk	lemma	3176	2519	1
85010 disk	word	1991	1676	1
99895 diskor	lemma	1652	1127	1
134944 retrodisko	lemma	1159	593	1
135147 retrodisko	word	1157	608	1

5. Töötamine täistekstidega

Otseligipääs täistekstidele

Eraldi keskkond vabakasutuses
olevate ajalehtede täistekstide
kasutamiseks

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. Lигипääsu vahendid on laiemalt lahti kirjeldatud [siin](#).

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

```
[1]: 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"))
```

Andmestikuks olevatest materjalidest saab ülevaate käsuga `get_digar_overview()` `digar.txts` paketist. Tulemuseks on tabel, kus on kõik

Andmed

```
[2]: all_issues <- get_digar_overview()
```

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

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

```
[3]: 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.

```
[4]: 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.

Otselipääs täistekstidele

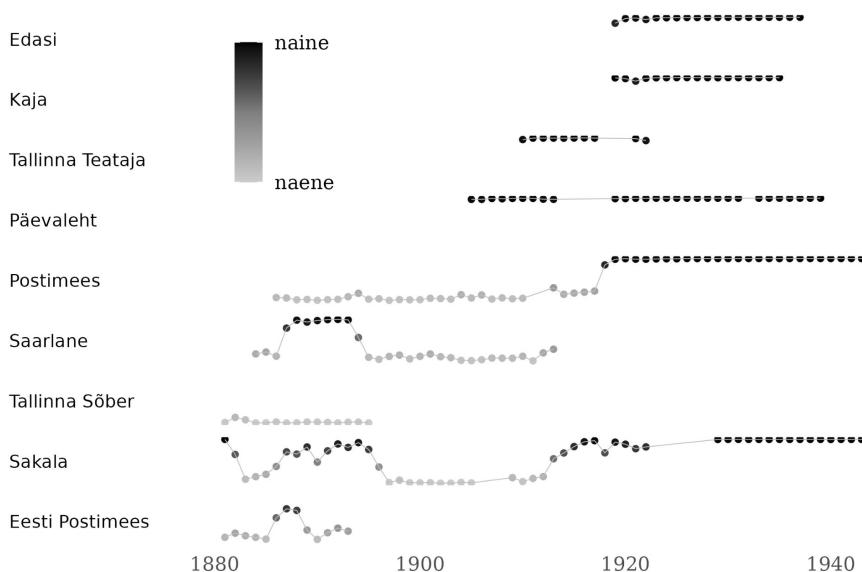
Täistekstide ligipääs võimaldab täpsemaid uuringuid.

Kui andmed on struktureeritud, võib vaadata näiteks komplekti sõnade sagedusi näiteks teatud lehtede sees.

Siin on näide keelekujude uuringust.

Näiteks saab lehtede kaupa vaadata toimetajate mõju keelele või keeledebattide jälgimist ajalehtedes

~ Keel ja ühiskond



Otseligipääs täistekstidele

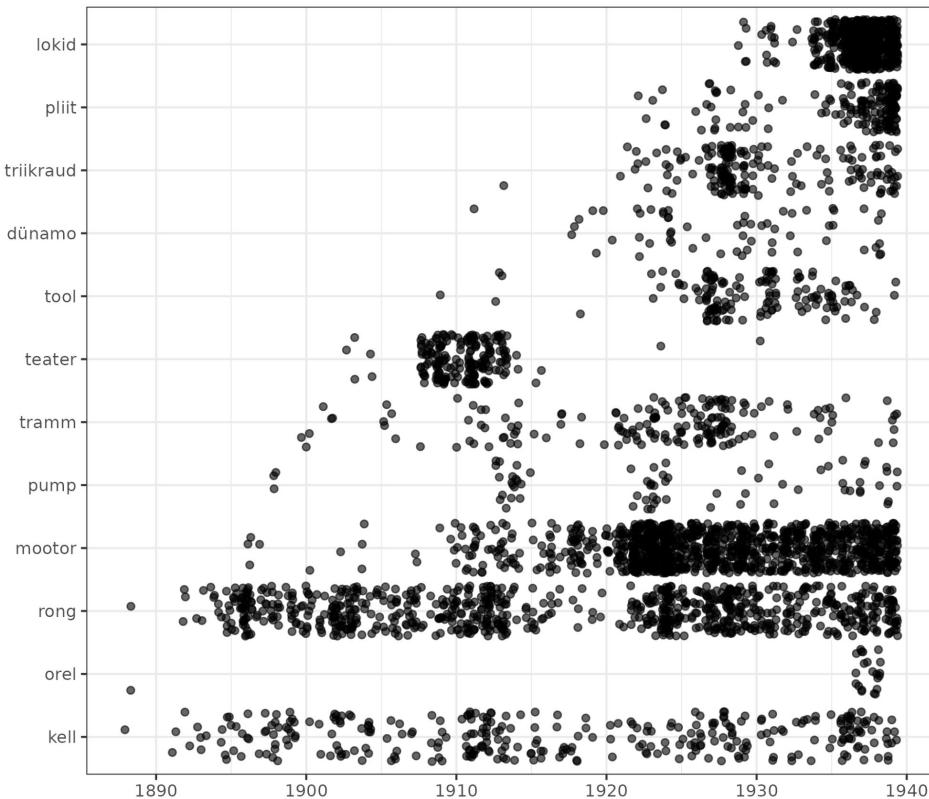
Täistekstide ligipääs võimaldab täpsemaid uuringuid.

Täistekste on võimalik paigutada uuringuteks spetsiaalselt loodud töövoogudesse.

Näiteks elektriga seotud tooded 20. sajandi algusest.

Siin me alguses ei tea täpseid märksõnu, aga korpuse abiga leiame üles tooted ning nendega seotud nimed.

~ Tehnoloogia ajalugu, reklaamid



6. Väljakutsed ja tulevik

Tärktuvastuse kvaliteet

Uue elektritrammi liini kalvatus.

| Nagst' juLa teada, wiiakse käeSolewal suweb Narva maantee tramm elektri peale" üle. Tööd on selleks juba käimas ja elektritrammi käimapaneku aeg tähendatud liinil oleneb ainult veel sellest, mil Jaani uulitsale mahapanekuks väljamaalt tellitud uued trammiroopad pärale jõuawad. Kõige. hiljem loodetakse siiski elektritramm Narva maanteel käima Panna juba septembri lõpuks. Selle järele tahetakse juba järgmisel suwel mõne teise trammiliini elektrofitseerimisele asuda.

KuuldaWasti kawatsetakse selleks esimeses järjekorras Pärnu maantee liim wöta, kuna see Narva maantee elektritrammi liini otsemaks jätkamiseks oleks. Nimetatud liini elektrofitseerimisega loodetakse ühtlasi ka hoogu anda linna kaSwamifele Pärnu maantee suunaS, kus suurem maa-alal ehituskruntideks planeeritud ja teatawad elawamab liikumisuulitsad ette nähtud. t

Uue elektritrammi liini kalvatus.

| Nagü juba teada, midaesse käeSolewal samel Narva maantee trammi elektri peale üle. Lõob on selleks juba läimas ja elektritrammi käimapaneku aeg tähendatud liinil oleneb ainult veel sellest, mil Jaani uulitsale mahapanekuks väljamaalt tellitud uued trammiroopad pärale jõuawad. Kõige. hiljem loodetakse siiski elektritramm Narva maanteel läima panna juba septembri lõpuks.

Selle järele tahetakse juba järgmisel suwel mõne teise trammiliini elektrofitseerimisele osuda. KuuldaWasti kawatsetakse selleks esimeses järjekorras Pärnu maantee liini wöta, kuna see Narva maantee elektritrammi liini otsemaks jätkamiseks oleks. Nimetatud liini elektrofitseerimisega loodetakse ühtlasi ka hoogu anda linna läbivamisele Pärnu maantee suunaS, kus suurem maa-alal ehituskruntideks planeeritud ja teatawad elawamab liikumisuulitsad ette nähtud. t

Läbi tööfides töökamatte tööfeltihidega on näitus-meessi juhatus määrenud väljas. poolt ilbauhindomist sohn rahalist autoju tööfeltihidele tööloomade väljapaneelu korraldamise eest. Väljapanef peals tööndamata ülemoottiflu pildi diagrammide, pilte ja seletuste hulgal vastava töör omadustest, seisuorras ja arenemise täigust Eesti ning riigiliblikkematest töör ehitajatest. Väljapanef tuuress onnachid seletusi voodajatele tööfeltihide ehitajeb. Autoelusõrahad on 25.000, 20.000 ja 15.000 marga suuruses.

Tööbörje tegewus juunis.

30. juunil tööta töölisti 833.

Registreeritud töötajaid 604 (272).

Ajutiselt töödelt tagasi 55 (49). Uusantud tööhad tööde 175 (279).

Lööndjate arv 56 (51). ? Löödele saadetud töö-

listi 289 (130). ? Löödele jääruud töölisti 250 (114). ? Lööta tööjõu hulgast välja-

astunud ja väljalangenuud 287 (344).

Tagasi töötub nõudmised 149 (143).

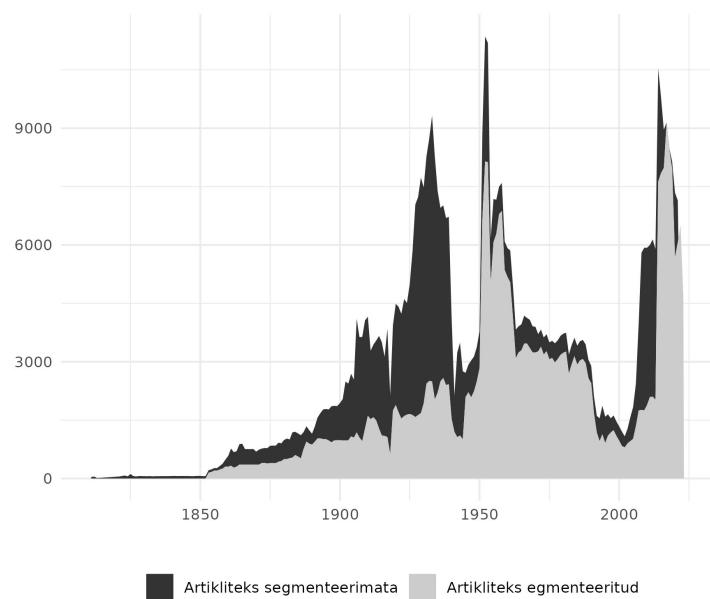
Tööta 30. juuni õhtul 833 (500). ? Wa-

bad tööhad 30. juuni õhtul 103 (34).

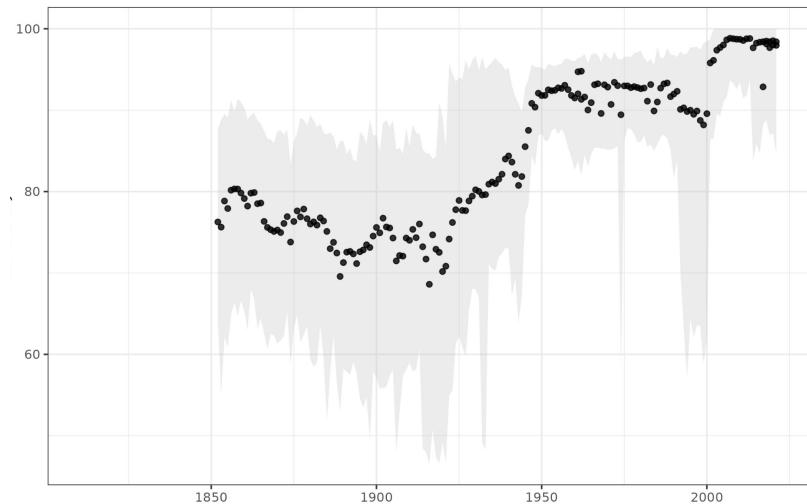
Siemenites töödub arvud läinud läbirub caasta juunilistu töhta.

Segmenteerimine artikliteks

Varem tehtud lehtede kvaliteet on kehvem



OCR kvaliteet, kus teada ja 95% tekstide ulatus



Digiteeritud lehed pakuvad uusi kasutusvõimalusi

- Lehti on digiteeritud väga palju
 - ~60% annab hea esinduslikkuse
- Suured andmehulgad avavad uusi võimalusi
 - Tööriistu saab teha erinevate oskustega huvilistele
- Saab ja tasub mõelda, mida võiks veel teha
 - Kasutamisest selguvad head ja vead ja ka kasutusvõimalused
- Kvaliteet hakkab mängima veel olulisemat rolli
 - Tasub mõelda digiteeritud materjalide parandamisele