

Digiteeritud Eesti ajalehed uurimisallikana

Peeter Tinitis 

Tartu Ülikool, eesti ja üldkeeleteaduse instituut, Jakobi 2, 50090
Tartu; Eesti Rahvusraamatukogu, Narva mnt 11, 15015 Tallinn;
peeter.tinitis@ut.ee

Kokkuvõte. Edusammud digihumanitaarias toetuvad tihti suurte andmes-
tike kasutuselevõtule uurimistöös. Võimaliku uurimisallikana analüüsitakse
artiklis neist üht – digiteeritud Eesti ajalehti. Eri digiarhiividest koondatud
andmete põhjal näidatakse, et suur osa Eesti trükipärandist on digiteeritud:
pea 60% viimase 200 aasta jooksul ilmunud lehtedest. Nende kasutamine
teadustöös eeldab uurija vajadusi arvestavaid erivahendeid. Artiklis tutvus-
tatakse kolleegide abiga loodud kolme tööriista: 1) digiteeritud ajalehtede
ülevaade, 2) sõnamitmike loendaja ja 3) ligipääsupunkt täistekstide kogumi-
seks. Arvestama peab, et digiarhiivide kasutusvõimalused sõltuvad materjali
kvaliteedist ja esitusviisist. Võimaluste täiendamiseks saab parandada varem
digiteeritud materjale, lisada infot kontekstist, aidata tekstiainest rikastada
ja luua ühenduspunkte digiarhiivide vahel. Olulist rolli saavad siin mängida
valdkondadevahelised juhtumiuuringud.

Märksõnad: ajalehed, tekstikorpus, metaandmed, digiteerimine, digi-
humanitaaria, digiajalugu, sõnamitmik

Paljude mälu- ja uurimisasutuste koostöös on suur osa Eesti kultuuri-
pärandist digiteeritud. Trükipärandi digiteerimine aitab säilitada
hapraid paberväljaandeid ning on parandanud ligipääsu trükistele ja
avanud uusi kasutusvõimalusi andmeteaduse ja andmepõhise humani-
taaria lähtekohast. Eelmise suure digiteerimiskava lõppanalüüsis
hinnati 2023. aasta lõpuks digitaalselt ligipääsetavaks 42% Eesti
mäluasutustes olevast kultuuripärandist, sh 4% dokumendipärandist,
42% esemepärandist, 56% filmipärandist, 78% fotopärandist ning
30% trükipärandist. Mäluasutuste juures asuva trükipärandi koguhulgaks
on märgitud 59,6 miljonit lehekülge, millest digitaalsena kättesaadav

on juba 18,1 miljonit lehekülge.¹ Järgmise digitegevuskava kohaselt on 2029. aasta lõpuks plaanis digitaalselt kättesaadavaks teha 52% Eesti kultuuripärandist. Tegevuskavas arvestatakse, et dokumentide, esemete, filmide, fotode ja trükiste digitaalkujutiste puhul on kriitiline mass juba digiteeritud ning uus rõhuasetus on seatud esemete, kinnismälestiste ja arheoloogilise materjali suures mahus 3D-digimisele, eesmärgiga hõlmata 21% prioriteetsetest kinnismälestistest.²

Kättesaadav on niisiis muljetavaldav osa Eesti trükipärandist ning see võimaldab nii paremat ligipääsu kui ka uute tehnoloogiate kasutuselevõttu. Olulise osa mäluasutuste digiteeritud trükipärandist moodustavad ajalehed ning siinses artiklis avangi nende numbrite taga peituvat ajalehtede kontekstis. Keskendun kolmele omavahel seotud küsimusele: 1. Milline on ajalehtede digiarhiivide sisu ja kui suur osa Eesti ajalehepärandist on digiteeritud? 2. Mida saab selle digimaterjaliga ette võtta ja kuidas sellele ligi pääseb? 3. Millised võiksid olla tuleviku väljakutsed kultuuripärandile digitaalse ligipääsu arendamisel? Küsimustele vastamiseks olen koondanud andmed ajalehti sisaldavatest digiarhiividest ning tutvustan ka ajalehtede digikollektsioonidele ehitatud tööriistu ja uurimisprojekte mujal maailmas. Digiarhiivide kasutamiseks olen Eesti Rahvusraamatukogu digilaboris kolleegide abiga loonud mõned rakendused, mis võiksid aidata digiteeritud ajalehti uurimistöös kasutada. Andmete põhjal tutvustan veidi ka tehnilisi piiranguid, millega materjali kasutades peaks arvestama. Toon esile enda hinnangul olulisemad aspektid, millele lähiaastatel digiarhiivide arendamise juures võiks keskenduda.

Digiteeritud ajalehtedest on räägitud kui võimalikust eldoraadost mineviku kogukondade uurimisel, mis võib avada nii ajaloolastele, sotsiaalteadlastele kui ka keeleteadlastele uurimistöös palju uusi võimalusi.³ Digiteeritud ajalehtede põhjal on näiteks välja pakutud võimalus mõisteajaloo andmepõhiseks uurimistööks. Nii analüüsi tekstitöötuse vahenditega rahvuse mõiste (*nation*) muutumist paralleelselt nelja piirkonna

1 „Kultuuripärandi digiteerimine 2018–2023“ tegevuskava lõpparuanne. Kultuuriministeerium, Tallinn, 2024. https://www.kul.ee/sites/default/files/documents/2024-06/LÕPPARUANNE_Kultuuripärandi_digiteerimine_2018-2023_tegevuskava_avalik.pdf (10.03.2025).

2 Kultuuripärandi digitegevuskava 2024–2029. Kultuuriministeerium, Tallinn, 2024. <https://www.kul.ee/sites/default/files/documents/2025-01/Kultuuripärandi%20digitegevuskava%202024-2029.pdf> (10.03.2025).

3 M. Ehrmann, E. Bunout, F. Clavert. Digitised Historical Newspapers: A Changing Research Landscape. – Digitised Newspapers – A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology. (Studies in Digital History and Hermeneutics, 3.) Ed. by E. Bunout, M. Ehrmann, F. Clavert. De Gruyter Oldenbourg, Berlin, Boston, 2023, 1–22.

(Soome, Rootsi, Hollandi ja Suurbritannia) ajalehtedes 18.–19. sajandil.⁴ Tekstide taaskasutus ajalehtedes on andnud võimaluse uurida info levikumustreid eri keelte ja kogukondade vahel – seda on kasutatud näiteks Rootsi ja Soome informuuri tiheda seotuse näitamiseks 19. sajandil.⁵ Ajalehed sisaldavad ka žanrispetsiifilist informatsiooni, mida teistes keelekorpustes ei ole, näiteks on täpselt ära märgitud sündmuse toimumisaeg ja infoallikas. 19. sajandi uudisartiklite struktuurile tuginedes on olnud võimalik näidata, kui kaua läks aega eri kohtade uudiste kajastamiseni Riias ja kuidas see muutus kommunikatsioonitehnoloogia arengu või rahvusvahelise poliitika mõjul.⁶ Ajalehtedes ilmunud info põhjal võib uurida mitmeid ühiskonna aspekte, näiteks on ajalehereklaamide põhjal uuritud sigaretituru ülesehitust Hollandis 20. sajandi teisel poolel⁷ ja noortele suunatud lihttöökohtade levikut Rootsis 20. sajandi vältel.⁸ Ajalehed võimaldavad uurida ka lugejate huvide peegeldusi. Näiteks ilmneb 20. sajandi teise poole Hollandi ajalehtede ilmatedetest, kuidas hollandlaste kujutus maailmast ümber kujunes: reisilendude kasv tõi ilmateatesse ka lõunamaised puhkusekohad, raudse eesriide varisemine Ida-Euroopa pealinnad.⁹

Ajalehtede digiteerimise ajalugu Eestis

Ajalehtede digitaalsel jäädvustamisel Eestis on praeguseks juba pikk ajalugu. Algasajaks võib pidada 1993. aastat, mil raamatukogud hakkasid hapral paberil, ent sagedasti loetavaid ajalehti salvestama mikrofilmidele. 2003. aastal viidi neilt mikrofilmidelt digitaalsete piltide kujule esimesed ajalehed – Perno Postimees (1857–1885), Postimees (1886–1920)

- 4 S. Hengchen, R. Ros, J. Marjanen, M. Tolonen. A Data-driven Approach to Studying Changing Vocabularies in Historical Newspaper Collections. – Digital Scholarship in the Humanities, 2021, 36, 2, iir09–iir26. <https://doi.org/10.1093/llc/fqab032>.
- 5 J. Marjanen, P. Paju, H. Rantala, H. Salmi. Crossing the border: A bird's-eye view on information flows between Sweden and Finland. – Information Flows across the Baltic Sea: Towards a Computational Approach to Media History. Ed. by P. Lundell, H. Salmi, E. Edoff, J. Marjanen, P. Paju, H. Rantala. Föreläsnings mediehistoriskt arkiv, Lund, 2023, 33–54. <https://doi.org/10.54292/s6au8axqht>.
- 6 K. Kruusmaa, K. Vanamölder. Positioning Riga in the 19th Century News Network: Tracing International News Flows in the Rigasche Zeitung Newspaper, 1802–1888. – Acta Historica Tallinnensia, 2023, 29, 1, 3–34. <https://doi.org/10.3176/hist.2023.1.01>.
- 7 M. Wevers. Mining Historical Advertisements in Digitised Newspapers. – Digitised Newspapers – A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology. (Studies in Digital History and Hermeneutics, 3.) Ed. by E. Bunout, M. Ehrmann, F. Clavert. De Gruyter Oldenbourg, Berlin, Boston, 2023, 227–252. <https://doi.org/10.1515/9783110729214-011>.
- 8 P. G. Håkansson, T. Karlsson, M. La Mela. Running Out of Time: Using Job Ads to Analyse the Demand for Messengers in the Twentieth Century. – Scandinavian Economic History Review, 2023, 71, 3, 299–318. <https://doi.org/10.1080/03585522.2022.2106300>.
- 9 J. van Eijnatten. Something about the Weather: Using Digital Methods to Mine Geographical Conceptions of Europe in Twentieth-Century Dutch Newspapers. – BMGN: Low Countries Historical Review, 2019, 134, 1, 28–61.

ja Päewaleht (1905–1912). 2004. aastal avati avalik veebikeskkond DEA – Digiteeritud Eesti Ajalehed, mis võimaldas ligipääsu ilmunisaja järgi indekseeritud digipiltidele. Järgneval kahel kümnendil on ajalehti digiteeritud suuremate ja väiksemate ettevõtmiste raames. 2004. aastal loodi kultuuriministeeriumi juurde ka kultuuripärandi digitaalse säilitamise nõukogu, nüüdseks digitaalse kultuuripärandi nõukogu, mis on koordineerinud tööloike suuremate Eesti mäluasutuste vahel. Ajalehtede digiteerimine on tihti toimunud mitme asutuse koostöös: üks asutus pildistab lehe üles, teine teeb tekstituvastuse ning kolmas võtab oma digiarhiivi hoiule.

Kuni 2014. aastani on kultuuripärandi digiteerimine toimunud peamiselt paljude väiksemate projektide koostöös.¹⁰ Alates 2015. aastast on olnud kandvaks kaks suuremat digiteerimisprogrammi: „Kultuuripärandi digimise rakenduskava 2015–2020“ ja „Kultuuripärandi digiteerimine 2018–2023“. Viimase raames on aastatel 2019–2023 toimunud massdigiteerimine suuremate mäluasutuste juhtimisel. Ajalehtede digiteerimisel on oluline olnud ka näiteks „Rahvuskaaslaste programm“, mille raames digiteeriti aastatel 2009–2013 ja 2014–2020 suur hulk Välis-Eesti ajalehti. Digiteerimisel on võetud põhimõtteks esimesena talletada kultuurilooliselt väärtuslikum aines. Valiku tegemisel on lähtutud seitsmest prioriteedist: „1) kasutatavuse potentsiaal, 2) riiklik, rahvuslik või kultuurilooline väärtus; 3) kollektsiooni terviklikkus, 4) metaandmete olemasolu, 5) autoriõiguste ja isikuandmete kaitse, muude juurdepääsupiirangute puudumine, 6) veebipõhise kasutuskeskonna olemasolu, 7) rareiteetsus“.¹¹

Digiteeritud ajalehed on jõudnud mitmesse portaali, suuremad neist on rahvusraamatukogu DEA, Tallinna Ülikooli akadeemilise raamatukogu Etera, Tartu Ülikooli raamatukogu DSpace (DSpace on avatud lähtekoodiga digiarhiiv, mis on kasutusel maailmas mujalgi, mistõttu viitan siin artiklis sellele kui TÜ DSpace) ja Eesti Kirjandusmuuseumi Kivike. Mitmest digiarhiivist on materjale koondatud ka mäluasutusteülesesse E-varamu portaali. Lisaks on ajalehetekste koondatud teaduskasutuseks mõeldud tekstikogudesse, näiteks on mitmed 1990–2010 ilmunud ajalehetekstid lisatud eesti keele koondkorpusesse. Praeguseks on kogunenud päris suur hulk ainet. 2025. aasta 13. jaanuari seisuga leiab E-varamust 1 300 276 ajalehele viitavat kirjet. DEA-s leidub perioodikat 662 429 numbrit, Eteras 57 294 numbrit, TÜ DSpace'is on

10 M. Laak, P. Viies. Digitaalkultuur Eesti kultuuriruumi osana 2004–2014: hetkeseis ja tulevikuprognoos. – Lõksudest välja: Eesti inimarengu aruanne 2014/2015. Koost. R. Vetik. Eesti Koostöö Koda, Tallinn, 2015, 226–236.

11 Samas, 231.

seda 2675 failis ning Kivikeses 195 kirjet. DEA keskkond ütleb lisaks, et kollektsoonis on kokku 6,4 miljonit lehekülge ja 17,3 miljonit artiklit – see hõlmab nii ajalehti kui ka ajakirju ja muid jätkväljaandeid.¹²

Ligipääs

Digiarhiivide puhul ei ole sugugi vähetähtis nende esitamiseks loodud virtuaalkeskkond – kasutajaliides, mille kaudu tekste loetakse. 2014. aastal loodud uus DEA keskkond laiendas oluliselt arhiivi kasutamismõimalusi. Kuupäevavaliku kõrval tekstiotsingus tekkis võimalus automaatselt tuvastatud teksti ka parandada, lisada tekstile märksõnu ja seeläbi tekstikollektsoone üles ehitada. Kasulikke võimalusi, mida kasutajaliidesesse sisse ehitada saaks, on hulgi. Maud Ehrmann ja tema kolleegid on välja toonud kuus peamist vajadust, mis kasutajal võiks olla: 1) allikakriitika – ülevaade sellest, kuidas materjalid on digiarhiivi kogutud ja millise osa kultuuripärandist need moodustavad, 2) sisuotsing (digiarhiivist õigete tekstijuppide leidmine), 3) materjalivalik (otsingutulemustele lisafiltrite paigaldamise võimalus), 4) uute leidude pakkumine (kas kasutajaliides oskab pakkuda tekste või otsinguid, mis on kasutajale olulised, aga mille peale ta ise pole tulnud), 5) võimalus luua isiklikke kollektsoone näiteks märksõnade kaudu digiarhiivi sees ja 6) ühenduspunktid teiste digiarhiivide sisuga (juhul kui vajalikud materjalid paiknevad mitmes virtuaalkeskkonnas).¹³ Iga Eesti digiarhiiv on nende seast valinud oma rõhuasetused.

Ehrmann ja tema kolleegid eristasid 24 digiarhiivi vaatluse põhjal lahenduste järgi nelja põlvkonda: 1. põlvkond keskendus tekstiotsingule, mis aitab lugeda ajalehe sisu, 2. põlvkond lisas interaktiivsust, võimaldades kasutajal teha isiklikke kollektsoone ja digiteeritud tekste parandada, 3. põlvkond lõi võimaluse tekste rikastada keeletehnoloogiliste vahenditega ja 4. põlvkond keskendub materjali läbipaistvusele, võimaldades tekstide avastamist nii, et nende päritolu, puudujäägid ja metainfo on samuti võimalikult kergesti ligipääsetavad.¹⁴ Samuti väärib märkimist, et kasutajaliideste loomisel on arenguhüpped tihti peale toimunud uurimisprojektide toel: juhtumiuuringud on viinud parema arusaamani digiarhiivide kasutusvõimalustest ja -vajadustest. Ka Eestis

12 DEA: <https://dea.digar.ee/>; Etera: <https://www.etera.ee/>; D-Space: <https://dspace.ut.ee/>; Kivike: <https://kivike.kirmus.ee/>; Koondkorpus: <https://www.cl.ut.ee/korpused/segakorpus/>; E-varamu: <http://e-varamu.ee/> (10.03.2025).

13 M. Ehrmann, E. Bunout, M. Düring. Historical Newspaper User Interfaces: A Review. – Proceedings of the 85th IFLA General Conference and Assembly. IFLA Library, Athens, 2019, 1–24. <https://doi.org/10.5281/zenodo.3404155>.

14 Samas, 15.

on ligipääsu arendamise juures olnud tihe seos uurimisprojektidega, sellest kirjutan põhjalikumalt artikli lõpuosas.

Lisaks kasutajaliidestele on viimasel aastakümnel muuseumides, raamatukogudes ja arhiivides arutelu alla tulnud ka avatud teaduse ja avatud andmete võimalused. Digiarhiivides peitub suur hulk informatsiooni, mida uurijad saaksid kasutada alusandmestikena. Üheks keskseks märksõnaks on siin saanud ingliskeelne akronüüm FAIR, mille põhimõtted soovitud andmestike haldajatel silmas pidada, et nende andmed oleksid leitavad (*findable*), ligipääsetavad (*accessible*), omavahel ühilduvad (*interoperable*) ja taaskasutatavad (*reusable*).¹⁵ FAIR-põhimõtted julgustavad andmete haldajaid otsima praktilisi lahendusi selles, kuidas oma andmestikud paremini nähtavaks teha ja milliseid ligipääsuvõimalusi neile saaks tekitada. Erinevad kasutajad vajavad erilaadi andmeid ja formaate. Näiteks võib andmete ühildumist parandada juba asutuse sees koondatud andmete esitamine lihtsas tabeliformaadis, mida saab Microsoft Excelis avada.¹⁶ Oluliseks on saanud ka andmetele selgete litsentside lisamine, mis selgitab, mida kasutaja võib andmetega peale hakata.

Eesti kontekstis on nii kasutajaliideseid kui ka avaandmeid käsitletud 2018.–2023. aasta kultuuripärandi digiteerimise tegevuskavas. Seal on lisaks andmete digiteerimisele esitatud ka nõue, et projekti tulemusena peaks olema võimalik olulise osaga Eesti kultuuripärandist virtuaal-keskkonnas tutvuda, toetades seeläbi kultuuriruumi loovat arengut.¹⁷ Eesmärkide seas on eraldi rõhutatud digihoidlate koondamist ühtsete mudelite alla ja pärandobjektide andmete avaldamist avaandmetena, kui võimalik. Olemasolevad digiarhiivide kasutajaliidesed on enamasti loodud juba varem ning seetõttu on need kasutamisevõimaluste poolest eespool toodud liigenduse järgi enamasti teise põlvkonna lahendused. FAIR-põhimõtete kohaselt on andmete ligipääsetavusega tegeletud pisuteliselt ja igas digiarhiivis omamoodi.

Eesti kontekstis tasub eraldi välja tuua, et kultuuripärandi, sealhulgas ajalehtede digiteerimisel on eestvedaja ja läbiviija olnud eelkõige avalik sektor. Eesmärgiks on enamasti olnud digiteerimise tulemused võimalikult laiale huviliste ringile avada, niivõrd kui autoriõigused seda lubavad. See ei ole kõikjal maailmas nõnda. Suurbritannias näiteks on ajalehti enamasti digiteeritud avaliku ja erasektori koostöös, mistõttu on

15 M. D. Wilkinson, M. Dumontier, U. J. Aalbersberg *et al.* The FAIR Guiding Principles for scientific data management and stewardship. – *Scientific Data*, 2016, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>.

16 G. Wilson, J. Bryan, K. Cranston, J. Kitzes, I. Nederbragt, T. K. Teal. Good Enough Practices in Scientific Computing. – *PLoS Computational Biology*, 2017, 13, 6, e1005510. <https://doi.org/10.1371/journal.pcbi.1005510>.

17 „Kultuuripärandi digiteerimine 2018–2023“ tegevuskava.

digiteerimise tulemused jäänud tihti erasektori hallata. Ülemaailmselt on digiteeritud ajalehekollektsioonid koondunud paari suurema erasektori teenusepakkuja kätte (nt Google, newspapers.com, Gale, Proquest), mõnel juhul seetõttu, et nood on digiteerimist korraldanud, mõnikord hilisemate kokkulepete alusel. Digiarhiivide omanikud ja vahendajad on materjalidest sageli teinud defitsiitse ressursi, küsides ligipääsu eest tasu, mida saavad lubada peamiselt hea rahastusega suuremad ülikoolid ja uurimisprojektid. Samuti dikteerivad erasektori digiarhiivide omanikud kasutusviise ja avatud teaduse põhimõtted seal eriti levinud ei ole. Mõned neist teenusepakkujaist on uue nõudluse tuules arendanud välja keskkondi, mis võimaldavad materjalidel andmekaevet, aga vajadus hoida andmeid suletud süsteemides piirab oluliselt andmeanalüüsiks kasutatavaid vahendeid. Selline olukord on mitmes uurijas tekitanud kõhkusi, kas massdigiteerimisega on ikka alati saavutatud soovitud tulemusi.¹⁸ Eesti digiarhiivide strateegiat – teha projektide tulemid nii vabalt kättesaadavaks kui võimalik – tasub väärtustada. Võib ka mõelda, et väiksel kultuuril ei pruugi vabale ligipääsule ollagi alternatiivi, kui eesmärgiks on kultuuriruumi loov areng.

KUI PALJU ON AJALEHTI DIGITEERITUD?

Pikaaegset tööd iseloomustavad teenitult uhked ja suured numbrid: digiteeritud on tuhanded väljaanded ja miljonid leheküljed. Sellest, mida need suured numbrid täpselt uurijale tähendavad, on veidi keerulisem aru saada. Artiklis püüan avada numbrite sisu ja loodud digikogude kasutusvõimalusi. Sellise ülevaate saamiseks koondasin andmeid digiarhiividest DEA, Etera, TÜ DSpace ja Kivike. Lisaks võtsin abiks Eesti rahvusbibliograafia (ERB) perioodikaandmed, kust leiab nii viiteid digiainese asukohtadele kui ka kasulikke metaandmeid digiteeritud andmehulga mõtestamiseks.¹⁹

Andmete kogumiseks tuli igale digiarhiivile kohandada eraldi lahendus. DEA kollektsioonis on andmed kättesaadavad avatud arhiivide initsiatiivi metaandmete kogumise protokollu kaudu (OAI-PMH) ning töövoog metaandmete koondamiseks on seal digiarhiivile ligipääsetavuse

18 L. Brake. Half Full and Half Empty. – *Journal of Victorian Culture*, 2012, 17, 222–229. <https://doi.org/10.1080/13555502.2012.683149>; P. Fyfe. Access, Computational Analysis, and Fair Use in the Digitized Nineteenth-century Press. – *Victorian Period Review*, 2018, 51, 716–737. <https://doi.org/10.1353/vpr.2018.0051>.

19 <https://digilab.rara.ee/andmestikud/perioodika/> (10.03.2025).

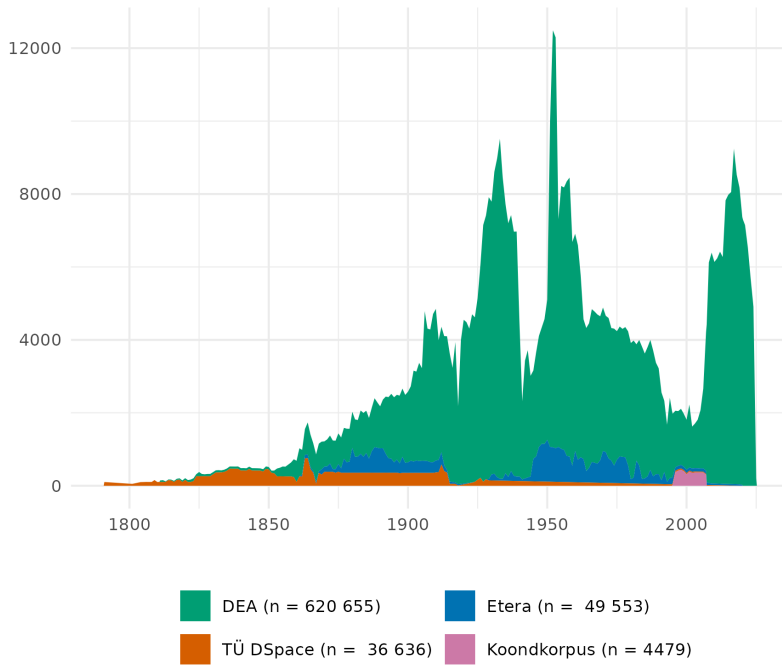
parandamise raames juba loodud.²⁰ Etera metaandmed olen kogunud veebilehe kaudu, kus on võimalik eri kategooriate ajalehti sirvida. TÕ DSpace'i perioodikast olen ajalehed eraldanud ERB registri abil ja metaandmestiku loonud digiteeritud ajalehtede esitusfailide põhjal. Need failid sisaldavad enamasti ühe väljaande kogu aastakäiku ja ilmunud numbrite arvu olen võtnud faili metaandmetest või sisust. Kivikese portaalist on võimalik teha väljavõtteid otsingus leitud säilikute metaandmete kohta ning ka need andmed koondasin ajalehtede, väikelehtede ja ERB-s toodud ajalehtede otsinguga. Kuna muudes kogudes on 1990.–2000. aastad vähem esindatud, olen lisanud ka eesti keele koondkorpuse ajalehenumbrid.

Milline on ajalehtede digiarhiivide sisu?

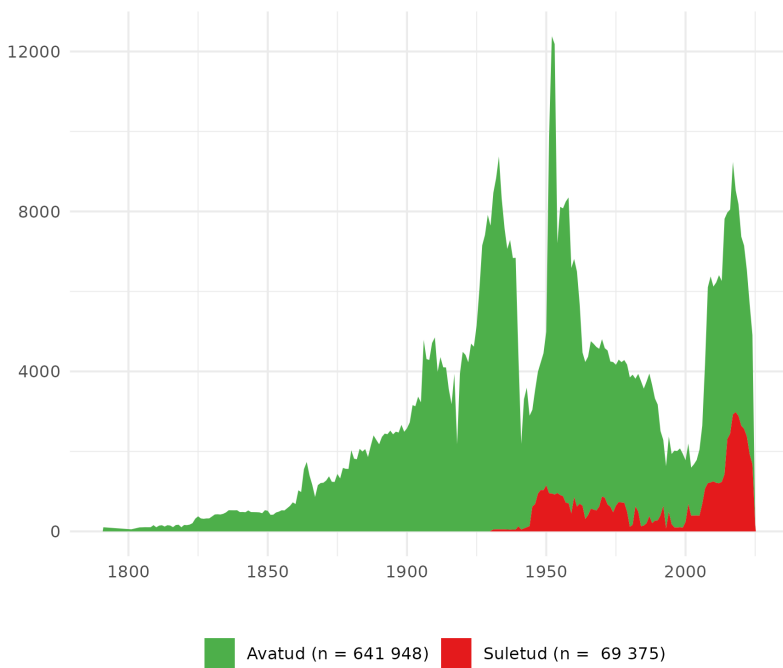
Kõigi digikogude peale sain nii kokku andmeid 711 531 digiteeritud ajalehenumbri kohta (DEA 620 655, Etera 49 553, TÕ DSpace 36 636, eesti keele koondkorpuse 4479, Kivike 208). Andmestiku sisu oleks siinkohal hea näidata eri viisidel visuaalselt. Joonisel 1 on näha digiteeritud ajalehenumbrite hulk ajas digiarhiivide kaupa. Joonisel on näha üldine digiainesega kaetus ja ka ajalehtede ilmumise hulga muutus. 19. sajandi esimesel poolel ei ilmunud veel palju Eestiga seotud ajalehti ning Vabadussõda ja II maailmasõda muutsid väljaannete koosseisu nii palju, et see ka joonisel välja paistab. Üldiselt võib graafiku põhjal öelda, et enamasti on viimase kahesaja aasta kohta võimalik digiteeritud uurimismaterjali leida. Veidi üllatava erandina paistavad siin ehk silma 1990-ndad ja varased 2000-ndad, mis pole veel digiteerimise fookusesse jõudnud. Joonisel peegeldub ka digiarhiivide erinev kujunemine: TÕ DSpace on keskendunud enne 1920. aastat ilmunud allikatele ja enne 1850. aastat ilmunud trükiseid leiab peaaegu et ainult sealt. Etera kollektsioonis on samuti enne 1920. aastat ilmunud ajalehti ning lisaks hulk nõukogudeaegset ainet. DEA-s on ainet kõige rohkem: sõdadevahelise aja ja aastate 2005–2025 kohta on see peaaegu ainus allikas. Eesti keele koondkorpuse ajaleheartiklid aga täidavad olulist lünka 1990.–2000. aastate veel digiteerimata ajalehtedest.

Teine oluline teema on ligipääs, nii lubade kui ka kasutajaliidesega seotud piiranguid. 9,8% materjalist on piiratud ligipääsuga ja 90,2% avatud ligipääsuga. Piiranguga materjalid paiknevad DEA-s ja Eteras – nimelt on osa materjalist õiguslike piirangute tõttu kättesaadavad küll

20 P. Tinitis. Eesti Rahvusraamatukogu digilabori tööriistad tekstimaterjali ligipääsuks ja töötlemiseks. – Zenodo, 2020, 21.07. <http://doi.org/10.5281/zenodo.3953795>; <https://digilab.rara.ee/tooriistad/ligipaas-dea-tekstidele/> (10.03.2025).



Joonis 1. Digiteeritud ajalehenumbrite arv 1780–2025 digiarhiivide kaupa



Joonis 2. Digikollektsioonide ajalehenumbrite ligipääsetavus aastate lõikes

raamatukogus kohapeal või mõnel puhul ka välisveebis, kuid mitte edasiseks töötluseks. Joonisel 2 on kujutatud ka neid; piiranguga materjal pärineb peamiselt viimasest kahekümnest aastast ja mingil määral ka Nõukogude ajast. Kasutajaliidesed pakuvad digiteeritud ajalehtede teksti erinevates formaatides. Etera ja TÜ DSpace pakuvad peamiselt PDF-faile, milles on tihti ka masintöödeldav tekst. DEA-s võimaldab OAI-PMH samuti otseligiipääsu masintöödeldavale tekstile ja sellega seotud metainfole, mille põhjal on loodud uurijale eraldi ligipääsupunktid.²¹

Visuaalne kujutis annab andmestiku sisust hea sissevaate. Samal põhimõttel on olemas ka detailsem ülevaade DEA sisust ja on loodud ka tööriist andmete uurimiseks.²²

Kui esinduslikud on digikogud?

Trükipärandi digiteerimine on toimunud alati piiratud ressursside tingimustes. See on lähtunud protsessi juhtide prioriteetidest ja sageli on digiteeritud ka n-ö hoogtöö korras: mõne projekti raames on keskendutud just üht tüüpi tekstidele. Selle tulemusel on paratamatu, et digikogu ei esinda võrdselt kogu trükipärandit, vaid mingi materjal on paremini esindatud. Näiteks on võidud trükipärandi digiteerimisel eelistada tekste, mis tänapäeval on veel tuntud, ning kõrvale jäetud tekstid, mis küll omal ajal olid populaarsed, kuid on vahepeal kultuurimälust kadunud.²³ Tekstikorpuste esinduslikkuse uurimisest on saanud omaette uurimisteema, näiteks on varauusaja ingliskeelse tekstikogu ECCO juures hinnatud põhjalikult eri žanride ja tekstitüüpide esinduslikkust.²⁴ Eestis olen ise esinduslikkuse hindamiseks kasutanud ERB ülevaadet aastatel 1800–1940 ilmunud raamatute kohta loodud tekstikogu.²⁵

Digiteeritud ajalehtede esinduslikkuse uurimiseks on metaandmeid kasutatud Suurbritannias. Selleks digiteeriti kaasaegseid ajalehti dokumenteerinud kataloog „Mitchell’s Press Directory“ aastatest 1846–1920.²⁶ Selles kataloogis on esitatud ülevaade ajalehe poliitilistest

²¹ Samas.

²² P. Tinitis. Rahvusraamatukogu Eesti artiklite andmebaasi (dea.digar.ee) ülevaade. – OSF, 23.01.2025. <https://doi.org/10.17605/OSF.IO/3GZXE>; https://peetertinitis.github.io/reports/nlib/dea_info.html (10.03.2025); A. Martynenko. DEA Metadata Explorer. – OSF, 14.11.2023. <https://doi.org/10.17605/OSF.IO/MDRX7>; <https://digilab.rara.ee/tooriistad/ajalehtede-metaandmete-sirvija/> (10.03.2025).

²³ M. Algee-Hewitt, S. Allison, M. Gemma, R. Heuser, F. Moretti, H. Walser. Canon/Archive: Large-scale Dynamics in the Literary Field. – Pamphlets of the Stanford Literary Lab 11. Stanford Literary Lab, Stanford, 2016, 1–13.

²⁴ M. Tolonen, E. Mäkelä, L. Lahti. The Anatomy of Eighteenth Century Collections Online (ECCO). – *Eighteenth-century Studies*, 2022, 56, 1, 95–123.

²⁵ P. Tinitis. Stratified historical corpus of Estonian 1800–1940. – Eesti Rakenduslingvistika Ühingu aastaraamat, 2023, 19, 175–194. <https://doi.org/10.5128/ERY19.11>.

²⁶ K. Beelen, J. Lawrence, D. C. S. Wilson, D. Beavan. Bias and Representativeness in

elistustest, ilmumis- ja levitamiskohad, ilmumissagedus ja hinnad. Kataloogist selgub, et näiteks 1850. aasta paiku oli umbes 400 ajalehte, 1900. aastal aga juba ligi 2000 ajalehte. Kataloogi koostajad püüdsid kataloogi põhjal hinnata Briti Raamatukogu JISC-kollektsiooni esinduslikkust, keskendudes maapiirkondade ajalehtedele. Kollektsioonis on maapiirkondade 19. sajandi ajalehtedest esindatud umbes 30–40 väljaannet aasta kohta. See tähendab, et 1850. aasta ajalehtedest sai kollektsiooni kaudu ligi umbes 8–10%-le trükipärandist, 1900. aasta ajalehtedest aga juba vähem kui 2%-le.

Eestis on meil võimalik võrrelda oma digiteeritud ainekogu Eesti rahvusbibliograafiaga (ERB). ERB koondab informatsiooni kõikide Eesti alal, eesti keeles või Eestit käsitlevate trükiste kohta. ERB-s on perioodika all kirjeldatud 18 316 väljaannet, sh 5084 väljaannet, mida on iseloomustatud kui ajalehte või väikelehte. Link perioodika digiteeritud versioonile on lisatud 38%-le väljaannetest ($n = 7011$), ajalehtede ja väikelehtede puhul 33%-le ($n = 1687$). Viimastest 78% ($n = 1311$) on koondatud DEA digiarhiivi, teised on muudes digikollektsioonides. Kuna digiteerimine toimub etapiti ja mitmes asutuses, ei ole sellist infot lisatud päris kõikidele digitaalselt kättesaadavatele lehtedele. Siinses artiklis lisan ERB-s toodud linkidele ka DEA-sse hiljuti lisatud materjali, tuginedes Esteri kirjetele ja ERB enda linkidele. Ka teistes mäluasutustes võib veel olla ERB-sse lisamata digiteeritud materjali.

Rahvusbibliograafias esindatud perioodilised väljaanded erinevad suuresti nii ilmumisaasta, mahu kui ehk ka olulisuse poolest kultuuri-pärandi seas. Samuti sisaldab rahvusbibliograafia suurt hulka eriala- ja nišiajakirju, mis ilmusid ainult paar korda, nt okultistlik ajakiri Fakiir (1924), või koolide ja asutuste aastaraamatuid, nt Hansapanga toimetised (1991–1993). Ka ajalehtede seas on palju väljaandeid, mis on seotud mõne konkreetse sündmusega, nt Eesti Põllumeeste Seltsi uue platsi avamisnäituse erileht (1930), või ilmusid muul põhjusel lühikest aega, nt tarbijate kuukiri Boonus (1995). Kui kasutada digiteeritud pärandit eesti kultuuriloo mõtestamisel, on osa allikaid ilmselt tähtsamad kui teised, kuigi olenevalt uurimisküsimusest võivad kedagi huvitada ka ehk lühiajalised nišiväljaanded.

Püüan siinkohal rahvusbibliograafia abil hinnata digikogude ulatust. Bibliograafias on väga paljudel väljaannetel (aga mitte kõigil) kirjas ilmumisaasta kestus. Näiteks ilmus Postimees 58 aastat (1886–1944), samas kui noorteleht Noor-Eesti Sõnumid (1995) ilmus ainult ühel aastal.

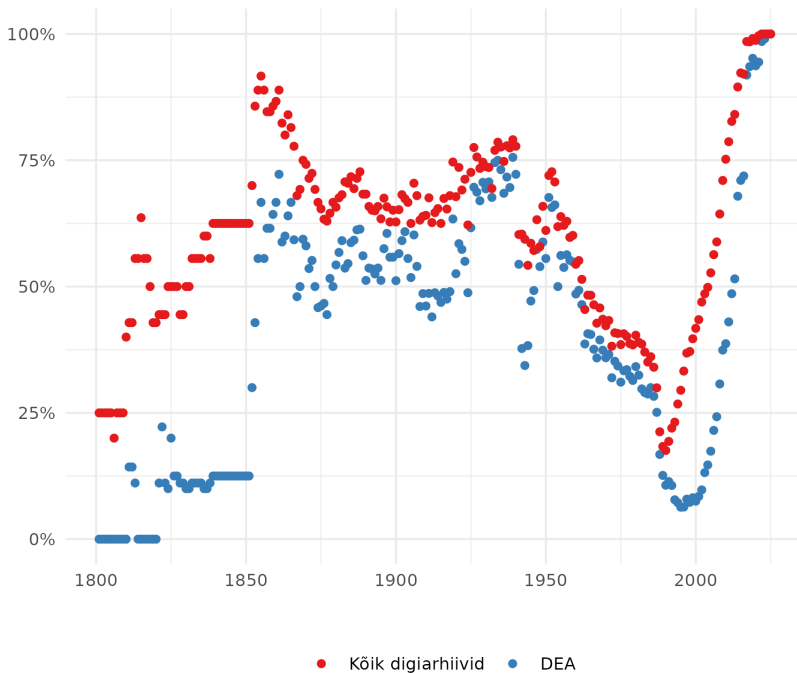
Seda informatsiooni saab esinduslikkuse hinnangus arvestada ja iga aasta kohta arvutada, kui suur osa väljaannete koguhulgast on digiteeritud. Täpsemaks hinnanguks peaks arvestama ka võimalusega, et ajaleht on digiteeritud osaliselt – mõni aastakäik on digiteerimata või digiteeritud osaliselt. Ajalehenumbrite ülevaade tuleb koondada igast digiarhiivist eraldi. Koondinfo alusel eristan info iga lehe kohta, vaadeldes, millised aastakäigud on kollektsioonis esindatud ning mitu numbrit aastakäigus on. Info numbrite kohta saab kergesti kätte Digarist, mis on märksõnade ja Esteri kirje alusel hõlpsasti bibliograafiaga seostatav. Teiste arhiivide puhul eeldan praegu lihtsustatult, et kõik väljaanded on digiteeritud täisulatuses. Neis arhiivides esindatud digiteeritud perioodika ilmus keskmiselt ainult viis aastat ja pikemat aega ilmunud väljaannetel, nt Rigaische Stadtblätteril (1807–1910) ja Dörptsche Zeitungil (1789–1878), on vähemalt enamik aastakäike digiteeritud. Selline lihtsustus muudab digiteerituskatte veidi optimistlikumaks, aga mitte palju. Analüüsi saaks täpsustada, lisades Esteri kirjed ka teiste kollektsioonide ülevaadetele.

Hindamisel teen veel mõned lihtsustused. ERB-sse on mõned väljaanded sisse kantud mitu korda, üks kord paber- ja teist korda võrguväljaandena. Digiteeritust hinnates jätan analüüsist võrguväljaanded välja. Mõnel väljaandel ei ole ERB-s antud ilmumise algus- ega lõppaastat (333 ajalehel 4574-st). Kui leht on digiarhiivis esindatud, siis määrان sel juhul ilmumisaaja alguse ja lõpu digiarhiivi varaseima ja hiliseima ilmumisaasta järgi. Kui vahepealsed aastad on kõik digiteeritud, siis loen ajalehe täismahus digiteerituks, kuigi on võimalik, et ajaleht ilmus ka enne ja pärast esimest digiarhiivi eksemplari. Alles jäi 100 nimetust, mille puhul ei ole andmestikus määratledud ilmumisaega ning mida pole õnnestunud seostada ka ühegi digiteeritud nimetusega. Need 100 ajalehte jäävad analüüsist välja. Teiselt poolt on DEA digiarhiivis veel 528 väljaannet, mis ei vasta ERB kirjetele. Enamik neist on suuremate ajalehtede lisalehed, mis on DEA-s esitatud eraldi, ERB-s aga hõlmatud lehe kirjesse. On siiski võimalik, et mõni neist on ajaleht, mida ei õnnestunud seostada. Lisaks on ERB-s ka mõned väljaanded, mis on ilmunud väljaspool Eestit ja digiteeritud mujal. Näiteks on Rigasche Zeitung täismahus ligipääsetav Läti Rahvusraamatukogu veebilehel.²⁷ Need seosed on praegu ERB-ga loomata.

Arvestades lihtsustustega, mis mõnel pool teevad hinnangu optimistlikumaks (puuduva metainfoga ajalehed on kõrvale jäetud), teisel pessimistlikumaks (puudu on näiteks väljaspool Eestit digiteeritud

27 <http://www.periodika.lv> (10.03.2025).

ajalehed), võib välja arvutada, kui palju ajalehti on praegu nende andmebaaside järgi digiteeritud. Joonis 3 võtab selle võrdluse kokku. Toon kõigi kollektsioonide kõrval eraldi välja DEA kollektsiooni, kus siinses artiklis tagapool mainitud tööriistade abil on võimalik ligi pääseda ka ajalehtede täistekstidele. Nende andmete põhjal võib digiteeritud materjali hulka hinnata veelgi suuremaks kui väljaandepõhiselt – pärast 1800. aastat ilmunud lehtedest on digitaalselt ligipääsetavad 60,4%, sh 40,3% DEA-s täistekstidena. Alates 1850. aastast on hinnang veelgi kõrgem: ligipääsetavad on 63,7% digiteeritud lehtedest ja vähemalt 49,5% DEA-s täistekstidena. Näiteks on ERB-s registreeritud 1850. aastast 8 ajalehte, millest 5 ehk 62,5% on andmestikuga seotud (2 DEA-s, 3 TÜ DSpace'is). 1900. aastast on ERB-s registreeritud 43 ajalehte, millest 27 ehk 62,8% on andmestikuga seotud (25 DEA-s ja 2 TÜ DSpace'is). Seejuurest on 1850. aasta kolmest puuduvast ajalehest kaks (Rigasche Zeitung ja Rigische Anzeigen) digiteeritud Läti Rahvusraamatukogus. Välisarhiividega sidumine võib siinset hinnangut veelgi optimistlikumaks muuta, eriti seoses baltisaksa trükipärandiga.



Joonis 3. Digiteeritud ajalehtede osakaal ERB-s registreeritud ajalehtedest aastate lõikes

Trükipärandit digiteeritakse hoogsalt ka teistes riikides, kuid head ülevaadet digiteeritud ainese hulgast on raske leida.²⁸ Uuringuid ja ülevaateid on tehtud perioodiliselt, enamasti uute digiteerimisprojektide alguses. Seega saab võrrelda pisut varasemaid andmeid. 2018. aastaks oli Eestis digiteeritud 9% raamatutest ja 26% perioodikast (hinnang on antud tõenäoliselt ERB ühenduste põhjal, mis on nüüd u 20% raamatutel ja 38% perioodikal). Euroopa kontekstis on võimalik neid numbreid võrrelda Europeana raportiga 2014. aastast, mil selles kajastatud Euroopa raamatukogudest ainult veerand oli digiteerinud rohkem kui 10% oma ajalehekogudest. Suuremate riikide puhul, kus ajalehti on ka ilmunud määratult rohkem, on hea digiteerituskatte saavutamine palju keerulisem. Pärast mitmeid mahukaid digiteerimisprojekte jõudis Briti Raamatukogu 2019. aastaks ajalehtede digiteerimisel 5%-ni ja 2015. aastal oli USA ajalehtedest digiteeritud 4%. Praeguseks on kindlasti kogude esinduslikkus paranenud, aga arvestades sealse ilmunud trükipärandi suurt koguhulka, siis tõenäoliselt mitte hüppeliselt.²⁹

Eesti ajalehtedest on nüüdseks digiteeritud oluliselt rohkem kui kümme aastat tagasi väga heaks tulemuseks loetud 10%. Praegune 60%-le liginev kate on teiste Euroopa riikidega võrreldes kui mitte parem, siis vähemalt väga heal tasemel. See on hea uudis uurijatele, kes meediatekstide põhjal püüavad kätte saada teatud esinduslikku hulka domineerivast meediadiskursusest, näiteks laia tellijate ringiga päevalehtedest või kohalikest ajalehtedest. Võimalik, et kaugel ei ole aeg, mil saab öelda, et kõik olulised materjalid laiapõhjalisteks meediauuringuteks on digiteeritud ja analüüsiks kättesaadavad.

Millised ajalehed on digiteeritud?

Mahule lisaks võib ühe uuringu tegemisel oluliseks osutuda ka teadmine, millised ajalehed on digiteeritud ja millised mitte, kas digiteeritud on kõik ajastu olulised lehed ja kas andmekogu hõlmab piisavalt erinevaid poliitilisi eelistusi ja geograafilisi piirkondi. Uurimist plaanides võib kasulik olla ka põgus ülevaade sellest, kui sageli leht ilmus ja kui suur oli näiteks selle maht lehekülgedes või artiklites.

28 Hea ülevaade on nt Tšehhis: <https://www.registrdigitalizace.cz/> (10.03.2025).

29 „Kultuuripärandi digiteerimine 2018–2023“ tegevuskava; http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/ENP-Deliverable_4.1_final.pdf; <https://blogs.bl.uk/thenewsroom/2019/01/heritage-made-digital-the-newspapers.html>; <https://web.archive.org/web/20230926014457/>; <https://web.archive.org/web/20230926014457/>; https://www.crl.edu/sites/default/files/d6/attachments/events/ICON_Report-State_of_Digitization_final.pdf (kõik 10.03.2025).

Niisuguse ülevaate saamiseks olen kolleegide abiga loonud rakenduse „Digiteeritud ajalehed Eestis“,³⁰ mis annab visuaalse ülevaate ajalehtede digiteerimise seisust Eestis ja aitab leida ajalehti, mis vastavad uurija vajadustele. See rakendus põhineb samuti ERB andmetel, kust on saadud ilmumisaajad ja ajalehtede omavahelised seosed, ning DEA metaandmetel. Rakenduses on välja toodud ajalehed, aga jäetud kõrvale väikelehed, et hoida allikate hulk hoomatavana.

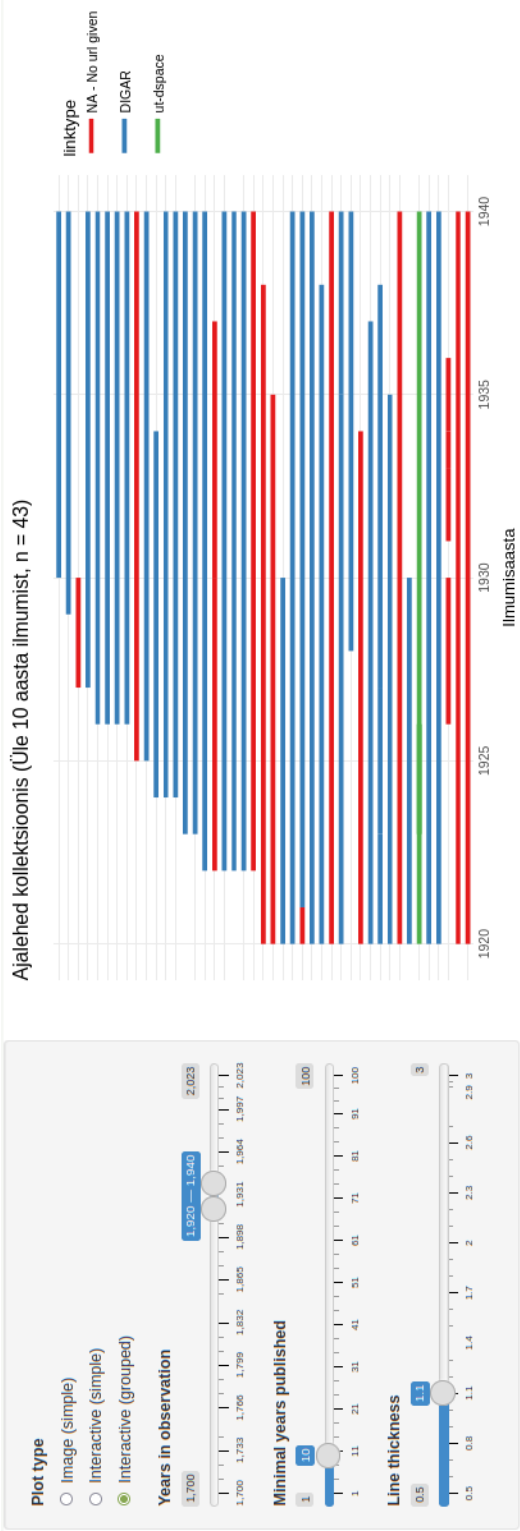
Rakendus võimaldab valida ajaperioodi, mil soovitud ajalehed ilmusid (nt 1920–1940), ning valida nende seast näiteks lehed, mille ilmumisaeg ületab kümme aastat. Rakendus kuvab seejärel kõik kriteeriumidele vastavad ajalehed ajajoonel ja digiarhiivi, millest neid võiks leida. Vajutades lehe ajajoonele, viib rakendus sellega seotud digiarhiivi veebilehele. Joonisel 4 on kujutatud üht selle rakenduse näidisvaadet. On näha, et perioodil 1920–1940 on üle kümne aasta ilmunud 43 lehte ja enamik neist on digiteeritud: üks neist TÜ DSpace'i kollektsioonis, ülejäänud DEA kollektsioonis. ERB sisaldab informatsiooni ka ajalehtede kohta, mis on nime vahetanud või on mõnel muul viisil eelmise väljaande jätkaja (nt Perno Postimehest sai 1886. aastal Postimees). Rakenduses on võimalik vaadata digiarhiivides eristatud ajalehti ja nende jätkväljaandeid nii koos kui eraldi.

Visuaalsele ülevaatele on võimalik lisada kiht DEA sisust, kus on näha juba täpsemalt, millised aastakäigud ja mitu numbrit igast aastast digikogus olemas on. Seda arvu saab võrrelda eeldatava ilmunud numbrite hulgaga, mis on loodud ERB-s esitatud ilmunussageduste põhjal. DEA digikogus olevate ajalehtede puhul näeb ka iga aasta lehenumbrite ja -külgede arvu ning seda, kas väljaanne on artikliteks segmenteeritud või mitte. Juhul kui see on teada, näeb ka digiteerimisprogrammi antud hinnangut OCR-i kindlusele.

Uurimistööd plaanides on rakenduse abiga võimalik vaadata täpsemalt, milliseid ajalehti on mõistlik ja vajalik uuringusse kaasata või mis ulatuses on teemale olulised ajalehed juba digiteeritud. Samuti saab mõne täpsema alamhulga raames hinnata, kui suure osa avalikust diskursusest digiteeritud materjalid katavad ja kui palju jääb materjalipuuduse tõttu teadmata.

Sarnaseid katsetusi uurimistöö plaanimise hõlbustamiseks on tehtud mõnel pool veel. Näiteks lõi Briti Rahvusraamatukogu rakenduse „Press Picker“, mis aitab sättida digiteerimise prioriteete, näidates ülevaadtlikult, kui palju mingi ajalehe aastakäike on raamatukogus olemas

30 P. Tinitis. Digiteeritud ajalehed Eestis. – OSF, 11.14.2023. <https://doi.org/10.17605/OSF.IO/B2HPX>; <https://digilab.rara.ee/tooriistad/digiteeritud-ajalehed-estis/> (10.03.2025).



Joonis 4. Kuvatõmmis rakendusest „Digiteeritud ajalehed Eestis”

paberkujul, kui palju mikrofilmina. Rakendus aitab leida lehti, millega saaks väheste kuludega kõige paremini täiendada uurimiseks kasutatavat materjali, eelistades digiteerimisel mikrofilme.³¹ Kogude hindamise meetodeid on erinevaid: digikogu esinduslikkust saab näiteks hinnata ka žanrilise mitmekesisuse kaudu – kas see vastab meie ootustele teadaoleva trükitoodangu kohta?³² Samuti on ERB ja loodud rakenduse kaudu võimalik mõne konkreetse uurimuse puhul välja selgitada, milliseid materjale oleks selle jaoks vaja digiteerida või mille kvaliteeti parandada.

KUIDAS SAAB AJALEHTEDE DIGIKOLLEKTSIOONE KASUTADA?

Ajalehtede digiarhiivid pakuvad uurijale üsna mitmekesiseid võimalusi avaliku arutelu uurimiseks. Üksjagu huvitavat selgub juba uurides, kui tihti sõnad teatud tekstikogudes esinevad või kuidas see on ajas muutunud. Veel rohkem võimalusi avavad sõnadevahelised seosed, mille kaudu on võimalik analüüsida ka sõnade tähendusi või avalike arutelude konteksti. Üha olulisemaks on saanud suured keelemudelid, mis koondavad sõnadevahelisi seoseid hiiglasuurtes tekstikogudes. Järgnevalt esitlen kaht abivahendit, mis on loodud tekstipõhiste uuringute toetamiseks Eesti ajalehtede digikogudes. Sõnamitmike loendaja aitab teha kiireid ülevaateid sõnasageduste muutusest ilmunud ajalehtedes. Ligipääs täistekstidele võimaldab arendada välja konkreetse uuringu kontekstis vajalikke töövoogusid. Toon paar näidet kummagi kasutusalaadest.

Sõnamitmike otsimine

DEA veebilehel (dea.digar.ee) on võimalik tekstiotsingute kaudu leida kollektsioonist lugemiseks artikleid. See otsing töötab nagu tavaline otsimootor, mis püüab leida kõige asjakohasemaid vasteid ning toob sobivamad vasted eespool esile. Kui meid huvitab digiteeritud ajalehetekstide

³¹ G. Tolfó, O. Vane, K. Beelen, K. Hosseini, J. Lawrence, D. Beavan, K. McDonough. Hunting for Treasure: Living with Machines and the British Library Newspaper Collection. – Digitised Newspapers – A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology. (Studies in Digital History and Hermeneutics, 3.) Ed. by E. Bunout, M. Ehrmann, F. Clavert. De Gruyter Oldenbourg, Berlin, Boston, 2023, 25–46. <https://doi.org/10.1515/9783110729214-002>.

³² L. Lahti, E. Mäkelä, M. Tolonen. Quantifying Bias and Uncertainty in Historical Data Collections with Probabilistic Programming. – CHR2020 Computational Humanities Research 2020: Proceedings of the Workshop on Computational Humanities Research (CEUR workshop proceedings, 2723). Ed. by F. Karsdorp, B. McGillivray, A. Nerghe, M. Wevers. CEUR-WS.org, Aachen, 2020, 280–289.

põhjal Eesti ühiskonna uurimine, siis on lisaks artiklite lugemisele võimalik püüda teha üldistusi ka artiklis esinevate sõnade kohta. Kõige lihtsamalt on võimalik vaadata mingil perioodil sõnade esinemissagedust, mis näitab, millal meid huvitavad teemad ajakirjandusse ilmusid või millal said neist juba ühiskonnas olulised teemad, mis leidsid palju kajastust. Millal näiteks hakati ajalehtedes rääkima raudteest või hambapastast? Millal pääses mõni poliitik või laulja avalikkuse tähelepanu alla ning kuidas ajalehtede tähelepanu on nihkunud?

Varasemad seesugused uuringud on näiteks raamatukorpuste põhjal leidnud, kuidas üksnes sõnasageduse järgi ajas on võimalik jälgida kuulsate isikute või ka mingite toitute populaarsust.³³ Ajalehetekstidest on aga võimalik juba täpsemalt teada saada näiteks epideemiade või poliitiliste sündmuste kajastust, samuti tehnoloogiate levikut.³⁴ Teooria-põhiselt otsides on sõnade esinemissageduse põhjal välja toodud ühiskonnas räägitud lugude olulisus majandusajaloos: sõnasageduste kasv ja kahanemine viitavad järgnenud majandusbuumidele või -kriisidele.³⁵ Täiendades sõnamitmikl põhinevat märksõnaotsingut täpsustavate filtritega, oleme kolleegidega saanud näidata keskkonnateemade ilmumist ajalehetekstidesse üsna erineva ajalooga riikides suhteliselt samal ajal 1960-ndatel.³⁶

Suurte organiseeritud tekstikorpuste juures on sellisteks analüüsideks välja mõeldud lisalahendus – *n-gram*’ide ehk sõnamitmike lugeja. Selleks, et andmestikud oleksid tasakaalus ja oleks võimalik arvestada ka aja jooksul toimunud mahumuutusega, on enamasti tarvis andmeid eeltöödelda. Tavalistes otsinguportaalides on sellega raskem arvestada. Näiteks näeb DEA otsinguportaalil küll leidude jaotust ajateljel, aga kuna kollektsiooni maht ja koostis ajas oluliselt muutub, võib leidude arv peegeldada hoopis kasutatud korpuse muutumist, mitte teema olulisust. Sageduspõhiste otsingute jaoks on loodud mitmeid lahendusi, näiteks Google’i Google Books Ngram Viewer³⁷ või prantsuskeelsete ajalehtede põhjal GallicaGram.³⁸

33 J.-B. Michel Y. K. Shen, A. P. Aiden *et al.* Quantitative Analysis of Culture Using Millions of Digitized Books. – *Science*, 2011, 331, 6014, 176–182. <https://doi.org/10.1126/science.1199644>.

34 T. Lansdall-Welfare, S. Sudhahar, J. Thompson *et al.* Content Analysis of 150 Years of British Periodicals. – *Proceedings of the National Academy of Sciences*, 2017, 114, 4, E457–E465. <https://doi.org/10.1073/pnas.1606380114>.

35 R. J. Shiller. *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*. Princeton University Press, Princeton, 2019.

36 L. Kanger, P. Tinitis, A.-K. Pakker *et al.* Deep Transitions: Towards a Comprehensive Framework for Mapping Major Continuities and Ruptures in Industrial Modernity. – *Global Environmental Change*, 2022, 72, 102447. <https://doi.org/10.1016/j.gloenvcha.2021.102447>.

37 <https://books.google.com/ngrams/> (10.03.2025).

38 B. Azoulay, B. de Courson. Gallicagram: un outil de lexicométrie pour la recherche. – *OSF*, 08.12.2021. <https://doi.org/10.31235/osf.io/84bf3>.

Kollegide abiga olen DEA tekstide põhjal loonud ka Eesti ajalehtede jaoks sarnase rakenduse.³⁹ Sõnamitmike loendaja võimaldab otsida korpusest levinud sõnajärgendeid ehk sõnamitmikke ja vaadelda nende sagedust aja jooksul. Andmestikku on kaasatud ka sõnade lemmatiseeritud kujud, kus erinevad käände- ja pöördvormid on koondatud ühe algvormi alla. Otsida saab kuni kolme sõna pikkuseid ühendeid ja valida saab perioodi, millele keskenduda. Tööriist loendab esinemiskordi 1000 sõna kohta igal aastal. Rakendusest võib eelkõige kasu olla uuringu esimestes faasides, kui on oluline saada ülevaade sellest, millistest teemadest Eesti ajalehetekstides teatud perioodil räägiti. Joonisel 5 on näiteks kujutatud kolme spordiala – korvpalli, jalgpalli ja võrkpalli – muutuvat populaarsust ajalehtede kajastuses veidi rohkem kui saja aasta jooksul. Nõukogude ajal kajastati neist enim võrkpalli, iseseisvusperioodidel jalgpalli. Otsingute lihtsustamiseks on lisatud ka ülevaade sõnamitmike üldsagedustest, mis võimaldab otsustada, milliseid otsinguid tasub korpusest teha. Ülevaate juures on võimalik kasutada ka regulaaravaldisi.

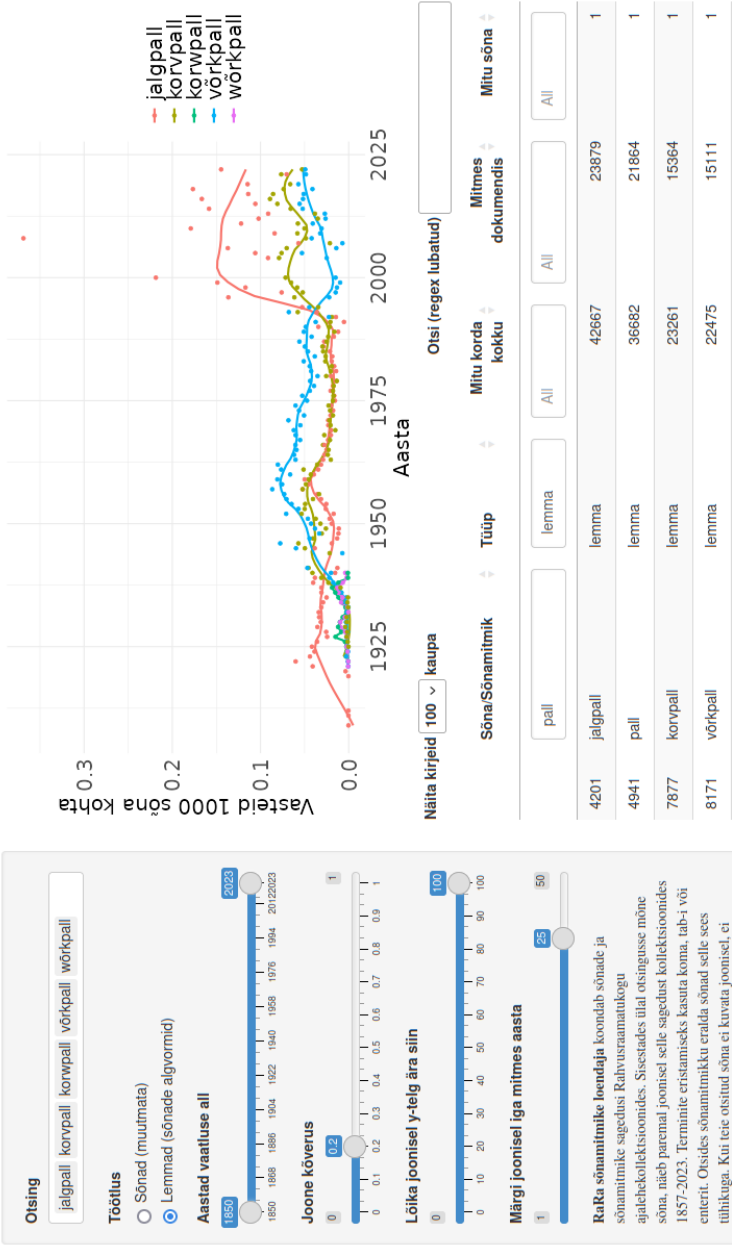
Selliste andmestike tähenduslikkuse jaoks on oluline ka andmete tasakaalustatus. Näiteks on jõuliselt kritiseeritud Google Booksi korpuse sisu, mille žanriline koostis erineb periooditi põhjalikult, sest on püütud hõlmata kõiki teoseid. Kui aga sagedused peegeldavad muutusi pigem andmestiku sisus kui aegade vältel kirjutatud tekstides, võib sageduste jälgimine osutuda tähendusetuks.⁴⁰ Seejuures on Google Booksist ka suhteliselt raske saada ülevaadet, kuidas korpus aja jooksul on muutunud. Sellest tulenevalt pole DEA sõnamitmike puhul andmestike aluseks võetud kõiki ajalehekogu andmeid, vaid välja on valitud pikema aja-perioodi jooksul ilmunud päevalehed ja kohalikud lehed, mis ilmusid kõik tänapäevase Eesti alal. Selliselt saab esinduslikud tulemused just Eesti alal kajastatud teemade kohta. Välis-Eesti ajalehtede kaasamine võiks Nõukogude aja tulemusi tugevalt kallutada, lühemat aega ilmunud lehed aga võivad tuua ootamatuid kõikumisi sõnasagedustes. Väljavalitud ajalehtede seas on ajavahemikust 1857–2023 kokku 39 Eestis ilmunud päevalehte ja kohalikku lehte, millest leidub vähemalt viis digiteeritud aastakäiku.

Sõnamitmike koondandmestikel on ka teisi rakendusi: see, mitu korda on mainitud hambapastat, raudteed või metsaloomi 1000 sõna

39 P. Tinitis. DEA Ngram Viewer. – OSF, 14.11.2023. <https://doi.org/10.17605/OSF.IO/XHU2A>, <https://digilab.rara.ee/tooriistad/sonamitmikud-ajalehtedes/> (10.03.2025).

40 E. A. Pechenick, C. M. Danforth, P. S. Dodds. Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-cultural and Linguistic Evolution. – PLoS One, 2015, 10, 10, e0137041. <https://doi.org/10.1371/journal.pone.0137041>.

Rahvusraamatukogu sõnamitmike loendaja



kohta, võib olla tähenduslik mõne tekstiliigi või ajastu iseloomustamisel. Eesti keeles on praegu hõlpsasti kasutatavad ilukirjanduse põhjal tehtud sõnamitmike kogud.⁴¹ Kokku koguti 92 ilukirjandusteost ning neis loendati ühe, kahe ja kolme sõna pikkuseid järjendeid ehk sõnamitmikke tavalisel ja lemmatiseeritud kujul. Sõnamitmike sageduse alusel saab näiteks iseloomustada, kuidas erineb teatud tekst tüüpilisest ilukirjandustekstist. DEA sõnamitmike puhul saab vaadelda ka seda, kuidas erinevad valitud artiklid mõne perioodi ajaleheartiklite keskmisest. Selle tarvis on võimalik sõnamitmike loendeid ka eraldi alla laadida.⁴²

Andmekaeveligipääs terviktekstidele

Veel suurema uurimispotentsiaaliga digiarhiivide tekstiaineses on võimalus moodustada temaatilisi tekstikogusid. Seni on raamatukogud avanud ajalehetekste peamiselt lugemiseks ja vahest ka tekstiotsinguks (nt Eestis DEA ja Etera otsinguportaal⁴³). Nende kaudu saab tekstidele ligi kas ühe artikli kaupa või laadides PDF-formaadis alla terve ajalehe numbri, kust on võimalik tekst välja võtta. Püüdes teha suuremaid üldistusi, tuleb kokku koguda hulk artikleid, mille jaoks on tarvis juba eraldi kasutajaliidest.

Selle probleemi lahendamiseks olen kolleegide abiga arendanud välja lihtsa vahendi, mille abil pääseb DEA kollektsooni tekstidele ligi otse ja suuremas mahus.⁴⁴ Tekstidele ligipääsuks tuleb sisse logida JupyterHubi serverikeskkonda, mis võimaldab andmetöötlust ETAIS-i arvutuskeskusesse salvestatud tekstimaterjali peal. JupyterHubi keskkonnas saab töötada Pythoni, R-i või Linuxi käsurea vahenditega ning kollektsooni tekste töödelda ja alla laadida.

Lahenduses on R-i-põhised näidiskäsud tekstide metainfost ülevaate saamiseks, metaandmete põhjal kollektsoonist alamhulga valimiseks, tekstiotsingu tegemiseks ja leitud tekstide väljavõtmiseks. Lisaks on seal võimalik töötada ka lemmatiseeritud tekstidega ning on tehtud paar vahekäsku, et toortekste saaks ühendada lemmatiseeritud tekstidega ja tõsta tekstidest välja sõnade esinemiskontekste. Digiarhiivi materjalide kasutusvõimalused laienevad tohutult, kui lisaks otsinguaknale pääseb ligi ka artiklite täistekstidele.

41 U. Raudvere, K. Uihoaed. Uuema eesti ilukirjanduse mitmikute loendid. – DataDOI, 2018, 01.04. <https://doi.org/10.15155/re-8>.

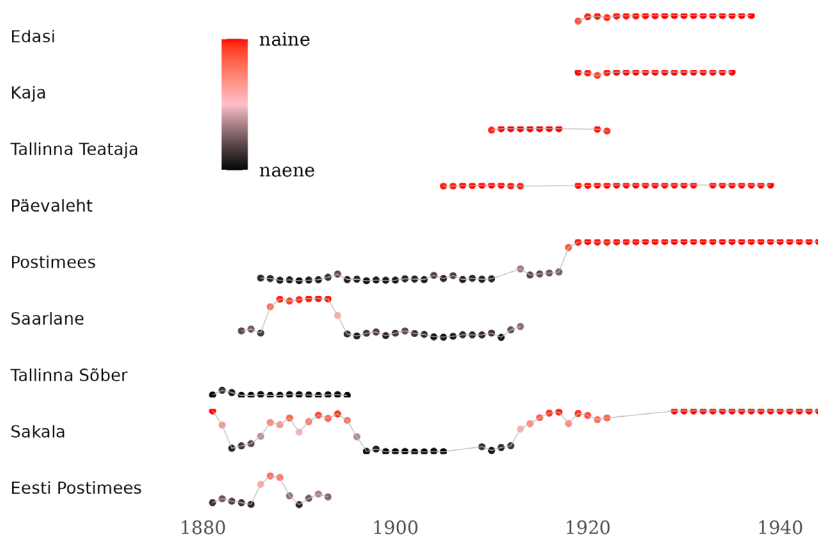
42 P. Tinitis. DEA Ngram Viewer.

43 <https://dea.digar.ee/>; <https://www.etera.ee/> (10.03.2025).

44 P. Tinitis. Eesti Rahvusraamatukogu digilabori tööriistad tekstimaterjali ligipääsuks ja töötlemiseks. <https://digilab.rara.ee/tooriistad/ligipaas-dea-tekstidele/> (10.03.2025).

Näiteks on võimalik digiteeritud ajalehetekste kasutada keeleajaloo uurimiseks. Keeleteadlastel on üldiselt teada, et kirjakeelt ajalehtedes on juba varakult mõjutanud toimetajate otsused ja eelistused. DEA kollektsiooni põhjal on võimalik täpsemalt uurida, millal ja kuidas näiteks toimetajavahetus ajalehe keelekasutust mõjutas. Joonis 6 näitab kahe konkureeriva sõnakuju kasutust DEA-s kõige paremini esindatud ajalehtedes aastatel 1880–1940. Iga aasta juures on märgitud, kumb sõna oli ajalehetekstis levinum, kas tänapäevane *naine* või varem domineerinud *naene*. Näeme, et Saarlane, Sakala ja Eesti Postimees kasutasid põgusalt sõna *naine* ka 1890-ndatel, s.o ajal, mil Postimees ja Tallinna Sõber kasutasid ainult sõna *naene*. Alates 1910. aastast asutatud suuremad ajalehed asusid kohe kasutama sõna *naine*, samas kui Sakala ja Saarlane tulid sellega aeglasemalt kaasa. Kõige konservatiivsemalt kasutas sõna *naene* Postimees – 1918. aastani.

Huvitav võib olla ka artiklite sisu. Näitena toon siin uuringu elektritoodetest 20. sajandi algul. Võrreldavuse tagamiseks piirdun analüüsis ühe digikogu väljaandega: Postimees 1880–1940. Otsisin tähekombinatsiooni *elekt* sisaldavaid tekste. Tulemuseks sain suure hulga mitmesuguseid elektriga seotud sõnu nende esinemiskontekstidega. Valisin nendest välja paar kujukamat näidet: „elektrilokid“, „elektripliit“, „elektritramm“, „elektripump“, „elektrirong“, „elektrikell“. Joonis 7

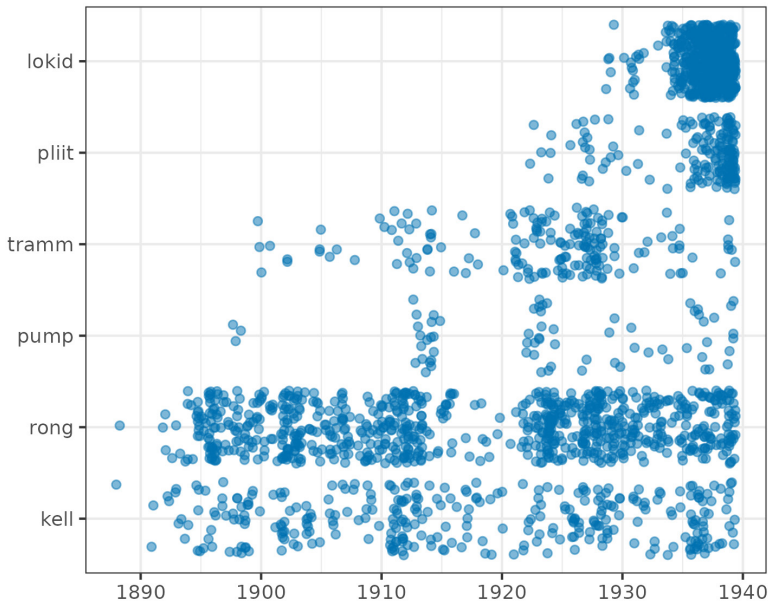


Joonis 6. *Naene* ja *naine* esinemise osakaal suuremates ajalehtedes ajavahemikus 1880–1940. Kõrgemal asuv punkt viitab 100% sõna *naine* kasutamisele; allpool asuv punkt viitab 100% sõna *naene* kasutamisele; vahepealne asukoht näitab varieeruvust

esitab aasta täpsusega kõikide leitud artiklite ilmumisaegu. Näidet ja kasutatud analüüsivahendeid on pikemalt kirjeldatud peatselt ilmuvas korpuslingvistika õpikus.⁴⁵

Ilmumisaja põhjal saab kohe mõningase ülevaate teemast. On näha, et sõnad „elektrirong“ ja „elektrikell“ olid pidevalt tarvitusel 1890. aastatest saadik, elektripumbast räägiti mõnel üksikul aastal, aga sel juhul enamasti korduvalt, näiteks kui oli tarvis üleujutuste kahjusid kõrvaldada. Elektritrammidest räägitakse vähem kui elektrirongidest, alates 1929. aastast iseäranis vähe – võib arvata, et siis hakkas juurduma sõna „tramm“ ilma viiteta elektrile. 1930-ndatest on märgata elektripliidi ja elektrilokkide märgatavat lisandumist Eesti inimese ellu: mõlema puhul on paar varajast mainimist ja seejärel juba ulatuslikke reklaamikampaaniaid.

Kui on ligipääs täistekstidele, saab uurida ka otsingusõnade esinemiskontekste. Võtsin korpusest välja 30 tähemärgi pikkuse konteksti otsisõna eest ja tagant ning vaatasin, millised sõnad nende kõrval paiknevad. Sisulistele küsimustele keskendumiseks jätsin mõned sõnad kõrvale. Kõigepealt eemaldasin hulga tavalisi sõnu, mis eesti keeles on sagedased, aga sisulist tähtsust analüüsis tihtipeale ei oma, kasutades selle otstarbeks



Joonis 7. Elektriga seotud tooteid või nähtusi sisaldavate artiklite esinemine aastate lõikes (Postimees 1880–1940)

45 P. Tinitis. Korpuslingvistika rakendamisest digihumanitaarias: Elektrii saabumine Eesti aladele 20. sajandi algul. – Korpused lingvistilises uurimistöös: Praktiline käsiraamat. Koost. L. Lindström, A. Veismann, M.-L. Pilvik, K. Muischnek, J. Klavan. Ilmumas.

loodud keeleressurssi – stoppsõnade nimekirja.⁴⁶ Seejärel jätsin kõrvale ka loodud tekstikogu tüüpilised sõnad ja tõin esile iga teemaga seotud eristavad märksõnad. Selleks kasutasin TF-IDF-algoritmi R-i paketi *Tidytex*.⁴⁷ Joonisel 8 on iga nähtuse juures näha 15 sõna, mis seda teistest teemadest eristavad.

Nähtustega kaasneva sõnu vaadates on võimalik kergesti aimu saada ka nende ühiskondlikust rollist. Näiteks olid elektrilokid suunatud eelkõige daamidele, aga vahest ka härradele, elektripliidi kasutamise kohta tehti kursuseid ja elektripumpa kasutati liigvee kontekstis, näiteks põrmivabriku päästmisel. Elektrirongidest kirjutati konkreetsete jaamade kontekstis ja esines ka õnnetusi, elektritrammidega seoses räägiti enamasti trammiteede ehitamisest. Selgus ka sõna „elektrikell“ varasem tähendus – nimelt kasutati seda uksekella tähenduses ja sellest räägiti palju korterite müümise ja üürimise kontekstis.

Otsisõnu ümbritsev konteksti on uurimistöös mitmeti kasutatav. Näiteks põhineb kontekstil eespool mainitud uuring, mis käsitles sõna „rahvus“ tähenduse muutumist 19. sajandi ajalehtedes.⁴⁸ Kontekste saab kasutada ka otsingutulemuste täpsustamiseks, kui märksõnad on mitmetähenduslikud (nt sõna „uurimine“ võib ilmned nii teadus- kui politseiuurijatega seoses, aga huvi pakub ainult üks kahest).⁴⁹ Samuti

	kell	lokid	pliit	pump	rong	tramm
1	uul	aurulokke	reisikohver	tänavale	jäi	tramvai
	kellad	aparaadil	kursused	rbl	nõmme	eile
	helisema	rullid	kursuste	pumbata	pääsküla	tramvaid
	ülesseadmine	tehakse	toitude	kiviõli	raudtee	tramm
5	tarbeasju	daamide	keetmine	pumbamaja	jaamas	ehitamine
	korterisse	daamid	köök	põrmivabriku	rongi	ehitamise
	majale	vello	toite	pump	lähedal	liini
	päevapildi	teen	kasutamine	vett	rong	trammi
	telefo	aurulokid	kompanii	eeloleval	elanik	roobasteta
10	telefonide	välismaa	muuseumi	liigvee	raudteed	trolley
	marconi	juuksetööstus	näitusel	pange	juhtus	asjus
	mehaanika	lp	õmblusmasin	jõkke	õnnetus	liikumise
	mässitud	saate	keedukohaga	kaevu	raudteel	maanteel
	nikeldamine	härade	keedukursused	kalda	liikumine	milj
15	plattesid	kestvaid	kööki	uulitsa	ehitada	ehitada

Joonis 8. Elektriga seotud toodete või nähtuste lähikontekstis esinevad sõnad, mis eristavad neid teistest leitud nähtustest (Postimees 1880–1940)

46 K. Uiboed. Eesti keele stoppsõnad. – DataDOI, 2018, 19.04. <https://doi.org/10.1515/re-48>.

47 J. Silge, D. Robinson. tidytext: Text Mining and Analysis Using Tidy Data Principles in R. – Journal of Open Source Software, 2016, 1, 3, 37. <https://doi.org/10.21105/joss.00037>.

48 S. Hengchen, R. Ros, J. Marjanen, M. Tolonen. A Data-driven Approach to Studying Changing Vocabularies in Historical Newspaper Collections.

49 S. Oberbichler, E. Pfanzelter. Topic-specific Corpus Building: A Step Towards a

on konteksti kasutatud valitseva meelsuse või hoiakute tuvastamiseks.⁵⁰ Täistekstide kasutusvõimalused töötavad lähitulevikus suuresti laieneda, pakkudes nii uusi meetodeid tekstidest info leidmiseks kui ka efektiivsemaid lahendusi juba olemasolevatele rakendustele.⁵¹

DEA täistekstide materjali on mõnes teadusartiklis juba kasutatud. Keeleteaduses on DEA korpus avanud pikaajalisi trende sõnakasutuses. Näiteks on sealse keelekasutuse põhjal näidatud sõnatähenduse allumatust keelele seatud normingutele sõnade „vabandama“⁵² ja „kostuma“⁵³ näitel. Keele ja ühiskonna seoseid uurides on ajalehekorpusete toel kirjeldatud trende sõnade „esinaine“ ja „esimees“ kasutuses: „esinaist“ kasutati algselt ainult naisorganisatsiooni juhi tiitlina, aga tänapäeva tekstides on see aina enam kasutusel ka segasoolise organisatsiooni naisjuhi nimetusena.⁵⁴ Ligi-pääs viimase aja artiklitele uurimisallikana on avanud võimaluse uurida keskkonnakonfliktide kajastamist ja meedia toimemehhanisme.⁵⁵ Veel mõned juhtumiuuringud on loodud tudengiprojektidena, mõned neist EKKD72 projekti „Tekstiainese kasutusvõimalused digihumanitaaria juhtumiuuringutes Eesti ajalehekolleksioonide (1850–2020) näitel“ raames.⁵⁶ Nendega saab tutvuda Eesti Rahvusraamatukogu digilabori blogikeskkonnas, mis on loodud Eesti mäluasutuste andmetega tehtud katsetuste ja juhtumiuuringute esiletoomiseks ja jäädvustamiseks.⁵⁷

Representative Newspaper Corpus on the Topic of Return Migration Using Text Mining Methods. – Journal of Digital History, 2021, 1, 1. <https://doi.org/10.1515/JDH-2021-1003>;
P. Tinitis. Finding Environmental Discourse in Historical Newspapers: A Topic Model Workflow for Query Disambiguation. – Digital Humanities in the Nordic and Baltic Countries Publications, 2023, 5, 1, 344–354. <https://doi.org/10.5617/dhnpub.10663>.

- 50 L. Kanger, P. Tinitis, A.-K. Pahker *et al.* Deep Transitions: Towards a Comprehensive Framework for Mapping Major Continuities and Ruptures in Industrial Modernity; M. Mets, A. Karjus, I. Ibrus, M. Schich. Automated Stance Detection in Complex Topics and Small Languages: The Challenging Case of Immigration in Polarizing News Media. – PLoS One, 2024, 19, 4, e0302380. <https://doi.org/10.1371/journal.pone.0302380>.
- 51 A. Karjus. Machine-assisted Mixed Methods: Augmenting Humanities and Social Sciences with Artificial Intelligence. – Humanities and Social Sciences Communications, 2025, 12, 277. <https://doi.org/10.1057/s41599-025-04503-w>.
- 52 L. Raadik. Sõnatähenduste normimisest eesti keelekorralduses verbi *vabandama* näitel. – Keel ja Kirjandus, 2020, 63, 10, 853–874. <https://doi.org/10.54013/kk755a3>.
- 53 L. Risberg. Jonnakas *kostuma*. – Keel ja Kirjandus, 2021, 12, 1078–1085. <https://doi.org/10.54013/kk768a3>.
- 54 E. Kaukonen. Kes on esinaine? – Keel ja Kirjandus, 2023, 66, 3, 328–336. <https://doi.org/10.54013/kk783a4>.
- 55 T. Jonuks, L. Päll, A. Remmel, U. Kadakas. Sõjakas kaitse – konfliktid loodus- ja kultuuripärandi hoiu kujundamisel. – Methis: Studia humaniora Estonica, 2022, 24, 30, 156–183. <https://doi.org/10.7592/methis.v24i30.22112>.
- 56 P. Tinitis, K. Vaik. Digihumanitaaria juhtumiuuringute projekti failid (EKKD72). – OSF, 2023, 14.11. <https://doi.org/10.17605/OSF.IO/HBFBMY>.
- 57 <https://digilab.rara.ee/blogi/> (10.03.2025).

DIGIKOLLEKTSIOONIDE TULEVIK

Arvestatav osa Eesti ajalehepärandist on nüüdseks digiteeritud. See on digiarhiivide kaudu ligipääsetav ning teatud ulatuses ka edasisteks analüüsideks kokkukogutav ja allalaetav. Juba on kehtima hakanud uus tegevuskava, mille raames on plaanis jätkata ulatuslikku digiteerimist, parandada digiarhiivide kasutusmugavust ja laiendada materjali kasutajate ringi. Digiteeritud ajalehtede kohta on tegevuskavas kirjas mitu olulist arendust. Rahvusraamatukogu digiarhiiv Digar saab endale uue kasutajaliidese, kus on erilist rõhku pandud andme- ja kodanikuteaduse suuremale kaasamisele. Digarile lisatakse ka teadlase vaade, mis võimaldab arhiivi sisu kergemat visualiseerimist, väiksemate andmehulkade väljavõtmist, ainese märksõnastamist ja väiksemaid ühisloomeprojekte. Lisaks tehakse andmeteadlastele ja digihumanitaaridele eraldi koodipõhine ligipääs, mille abil saab ehitada suuremaid tekstikorpusi ning siluda, väärindada, analüüsida ja taaskasutada digiarhiivi andmeid. Piltide esitlemisel võetakse aluseks IIIF-raamistik, mis on mõeldud materjalide asutuseüleseks ristikasutamiseks. Eraldi suurema ettevõtmisena luuakse lahendus digiarhiivide automaatseks märksõnastamiseks, rikastades seeläbi andmekogu veelgi.⁵⁸

Plaanitud arenduste valguses tasub veel korra peatuda olemasoleva sisu ja selle kasutusvõimaluste hindamisel. Digiainese kasutusvõimalus sõltub oluliselt digiteerimise kvaliteedist ja esitusvormist. Kasutajaliideste arendamisel tasub eeskujuna võtta muu maailma kogemustest, et liikuda n-ö uue põlvkonna abivahendite poole. Mitmed sellised sammud on juba töös. Viimaks toon esile valdkondadevaheliste uurimisprojektide erilise rolli arenduste juures. Et digiainese kasutusvõimaluste leidmiseks on tarvis nii uurijate sisulist huvi ja kompetentsi kui ka tehnilisi oskusi ja tausta, on paljud arengud maailmas toimunud suurte koostööprojektide raames, kus digiainese kasutusvõimalused proovile pannakse. Ka Eestis on mõningane kogemus siin olemas.

Digiainese kvaliteet ja vorm

Digiteerimise tehnoloogiat on järk-järgult parandatud ja täiendatud ning kahekümne aasta jooksul on selle võimekus hüppeliselt kasvanud. Digiarhiividesse kogunenud materjal on seetõttu üpris erineva kvaliteedi ja ülesehitusega. Kehvemas seisus on enamasti varem digiteeritud materjal,

⁵⁸ Kultuuripärandi digitegevuskava 2024–2029.

mida aga digiteerimise eesmärkide kontekstis on peetud iseäranis tähtsaks. Digiteeritud ajalehtede puhul on üks kesksemaid tehnilisi probleeme tärgtuvastus (s.o optiline tähemärkide tuvastus – *optical character recognition*, OCR). Pildilt teksti äratundmine on olnud üpris keeruline tehnoloogiline ülesanne ning varasemad lahendused on palju või vahest isegi enamuse tähti tuvastatud valesti. Seda võis omakorda võimendada ka ebapiisav pildikvaliteet. Uuema tehnoloogiaga on kaod pigem väiksed, segi võivad minna mõned üksikud tähed igas lauses ja sedagi tüüpkohtades. Hea pildikvaliteedi või digitaalselt sündinud materjali puhul võib OCR pakkuda tänapäeval peaaegu veatut tekstituvastust. Tärgtuvastuse kvaliteet mõjutab aga olulisel määral seda, mida andmetega edaspidi saab või ei saa teha.⁵⁹ Seejuures võib ebaühtlane kvaliteet ka tulemusi oluliselt kallutada, näiteks võib mingi perioodi või ajalehe analüüs tulemusi mitte anda puuduliku tekstikvaliteedi tõttu. Digiarhiivides on ka nii madala pildikvaliteediga ajalehti, mida on isegi uurijal raske oma silmaga lugeda. Niisiis on uurijal oluline saada ülevaade digiteerimise kvaliteedist.

Digiarhiivide kasutajaliidestest ei ole üldiselt eriti palju informatsiooni konkreetse ainese digiteerimismeetodi kohta. Sellise konteksti loomine võiks kuuluda nn neljanda põlvkonna kasutajaliidestest soovinimekirja.⁶⁰ Tehnoloogiliselt ei oleks tegelikult väga keeruline jäädvustada, millal (kas või aasta täpsusega) ja mis tarkvara ning seadetega aines on digiteeritud.⁶¹ Nende andmete põhjal saaks juba üpris hästi aimata lõpptulemuse kvaliteeti ja näiteks välja valida ainult viimasel viiel aastal lisatud materjali, kus tärgtuvastus on juba üpris täpne. Lisavõimalus, mis sellest hargneda võib, on juba digiteeritud ainese uus töötlus. Uuemate tärgtuvastustehnoloogiatega saab ka kakskümmend aastat tagasi üles pildistatud lehtedest parema tekstikvaliteedi kätte.

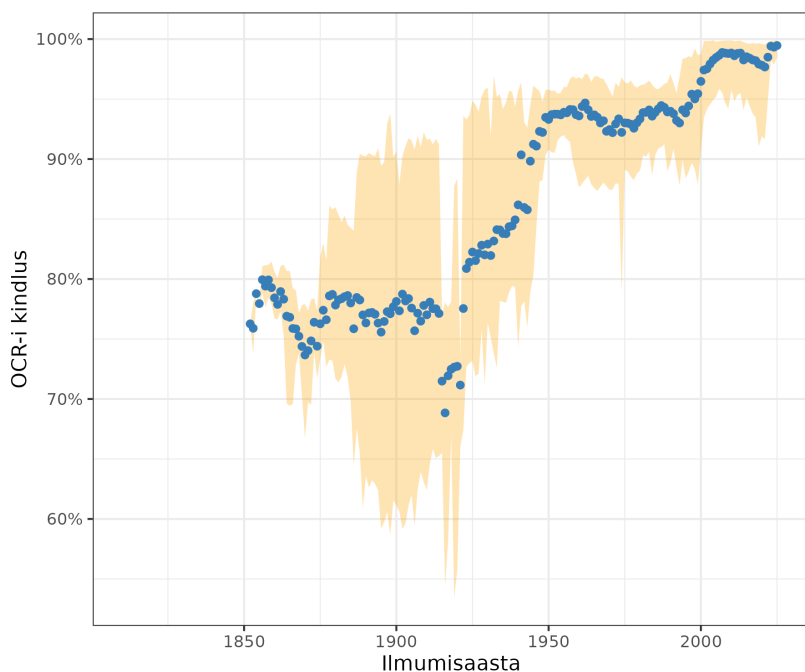
59 Nt M. J. Hill, S. Hengchen. Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study. – Digital Scholarship in the Humanities, 2019, 34, 4, 825–843. <https://doi.org/10.1093/llc/fqz024>; D. van Strien, K. Beelen, M. C. Ardanuy, K. Hosseini, B. McGillivray, G. Colavizza. Assessing the Impact of OCR Quality on Downstream NLP Tasks. – Proceedings of the 12th International Conference on Agents and Artificial Intelligence, Vol. 1: ARTIDIGH. SciTePress: Science and Technology Publications, 2020 484–496. <https://doi.org/10.5220/0009169004840496>; C. Resch. Volltextoptimierung für die historische Wiener Zeitung. – Digitised Newspapers – A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology. (Studies in Digital History and Hermeneutics, 3.) Ed. by E. Bunout, M. Ehrmann, F. Clavert. De Gruyter Oldenbourg, Berlin, Boston, 2023, 89–111. <https://doi.org/10.1515/9783110729214-005>.

60 M. Ehrmann, E. Bunout, M. Düring. Historical Newspaper User Interfaces: A Review.

61 A. J. Torget. Mapping Texts: Examining the Effects of OCR Noise on Historical Newspaper Collections. – Digitised Newspapers – A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology. (Studies in Digital History and Hermeneutics, 3.) Ed. by E. Bunout, M. Ehrmann, F. Clavert. De Gruyter Oldenbourg, Berlin, Boston, 2023, 47–66. <https://doi.org/10.1515/9783110729214-003>.

DEA-s on OAI-PMH kaudu tehtud ligipääsetavaks info tähtsuse tarkvara usaldusväärsuse kohta. Täpsemat infot selle kohta, millal ja millise tarkvaraga mõni ajaleht on digiteeritud, ei ole, aga võrdlusaluse annab seegi. Usaldusväärsuse info põhineb tarkvara enda hinnangul tuvastatud tähe märkide kindlusest ning oleneb tähtsuse tehnoloogia seisust digiteerimise ajal ja digipildi kvaliteedist. Hinnanguid saaks siin pakkuda erinevaid, näiteks saaks täpsust hinnata, kui võrrelda tähtsust käsitsi sisse trükitud tekstidega või vaadelda eri tähtsustatarkvarade tulemusi sama pildi juures. Samuti võiks analüüsida tulemusi mõne keeletehnoloogilise vahendi abil, näiteks vaadata, kui suur osa sõnadest on äratuntavad eestikeelsed sõnad.⁶²

DEA-s on OCR-i kindluse hinnang eelkõige uuematel tekstidel, kokku 381 885 (62%) ajalehenumbri, mistõttu on alust arvata, et selles kogumis on ilma hinnanguta numbrid enamasti kehvema kvaliteediga. Ka hinnanguga tähtsuse kvaliteet jaotub ebaühtlaselt. Joonisel 9 on kujutatud tähtsuse usaldusväärsuse keskmine hinnang, lisaks



Joonis 9. Keskmine OCR-i kindluse hinnang (punktid) ja vahemik millesse langeb 95% artiklitest (toonitud ala)

62 M. L. Pilvik, K. Muischnek, G. Jaanimäe, L. Lindström, K. Lust, S. Orasmaa, T. Tärna. *Mõistus sai kuulotedu*: 19. sajandi vallakohtuprotokollide tekstidest digitaalse ressursi loomine. – Eesti Rakenduslingvistika Ühingu aastaraamat, 2019, 15, 139–158. <http://doi.org/10.5128/ERYa15.08>; P. Tinitis. Stratified Historical Corpus of Estonian 1800–1940.

vahemik, kus paikneb 95% ajalehtedest. Joonise aluseks on võetud ajalehenumbrite keskmine, mis on arvutatud lehekülgede hinnangute pealt. Ajavahemikus 1850–1950 kõigub näidik suures ulatuses 70–80% ümber, Nõukogude ajast on keskmine 90–95% ja viimasel kahekümnel aastal on hinnang juba 98–99% kandis. See näitab nii digiteerimise prioriteete kui ka aluseks olnud ainest. Esimene digiteerimisperiood oli suhteliselt prioriteetne ja võeti ette ajal, kui tehnoloogia ei olnud veel nii heal tasemel kui praegu. OCR-hinnanguga nõukogudeaegsed ajalehed on digiteeritud suuresti viimasel perioodil (2019–2023). 2000-ndatest alates on tuvas-tamiseks sageli kasutatud juba ajalehtede trükifaile, kus pildikvaliteet on kõige parem. Kõige varem digiteeritud ajalehtedel puudub OCR-i kindluse hinnang.

Arvestades, et hinnang erineb märgatavalt varem ja hiljem digiteeritud materjali puhul ning puuduva hinnanguga materjal on ilmselt veel kehvema kvaliteediga, võiks otsida võimalusi varem digiteeritud materjali kvaliteedi parandamiseks. See on oluline asjaolu tõttu, et just varem digiteeritud materjali on peetud kultuurilooliselt prioriteetseks. Ebaühtlane arhiiv võib aga tähendada, et halvema kvaliteediga materjal ei tule otsingutes ja analüüsides välja.⁶³ Ühe lahendusena võiks proovida moodsa tehnoloogiaga uuesti läbi käia varem üles pildistatud ajalehed.⁶⁴ Maailmas on katsetatud ka võimalust parandada tärgtuvastust puhtalt tekstimudelite põhjal.⁶⁵ Viimastel aastatel üha rohkem kasutatavad suured keelemudelid võivad siinkohal pakkuda üht praktilist alust teksti-analüüsi parandada.⁶⁶ Sel moel ei pruugi tulemuseks olla täpselt õige, vaid suhteliselt samaväärne parandatud sõnastusega tekst, mida saaks samuti analüüsideks kasutada. DEA korpusel on võimalus toetuda veel ühele võimalikule abiallikale: alates 2014. aastast on lugejad käitsi parandanud enam kui 800 000 rida teksti rohkem kui 26 000 artiklis. Võimalik, et

63 Nt J. Burchardt. Are Searches in OCR-generated Archives Trustworthy? An Analysis of Digital Newspaper Archives. – *Jahrbuch für Wirtschaftsgeschichte/ Economic History Yearbook*, 2023, 64, 1, 31–54. <https://doi.org/10.1515/jbwg-2023-0003>.

64 K. Kettunen, M. Koistinen. Open Source Tesseract in Re-OCR of Finnish Fraktur from 19th and Early 20th Century Newspapers and Journals- Collected Notes on Quality Improvement. – *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, 2019, 2, 1, 270–282. <https://doi.org/10.5617/dhnpub.1102>.

65 H. Hammarström, S. M. Virk, M. Forsberg. Poor Man's OCR Post-Correction: Unsupervised Recognition of Variant Spelling Applied to a Multilingual Document Collection. – *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage (DATeCH)*, 2017, 71–75; M. Härmäläinen, S. Hengchen. From the Past to the Future: A Fully Automatic NMT and Word Embeddings Method for OCR Post-Correction. – *Proceedings of Recent Advances in Natural Language Processing*, Ed. by G. Angelova, R. Mitkov, I. Nikolova, I. Temnikova. INCOMA, Shoumen, 2019, 431–436.

66 A. Karjus. Machine-assisted Mixed Methods.

sellist andmekogu, n-ö paranduste kuldstandardit, saab aluseks võtta ka automaatsete tekstiparandamise lahenduste väljatöötamisel.⁶⁷

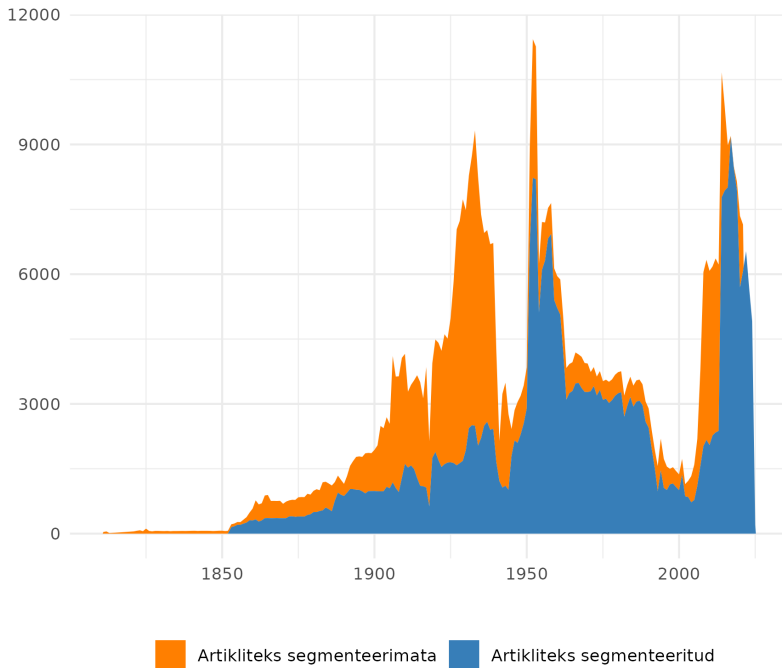
Teine koht, kus tehnoloogia uusi võimalusi pakub, on teksti struktuuri esitamine. Nimelt on ajalehtedes enamasti hulk suuremaid ja väiksemaid temaatilisi artikleid, mis võtavad enda alla mingi osa leheküljest, aga võivad ulatuda ka üle mitme lehekülge. Tekstiosade piiride visuaalne ülesmärkimine võimaldab digiteerimise käigus ajaleheteksti artikliteks jaotada. Digiteerimises nimetatakse seda teksti segmenteerimiseks, mida saab teha käsitsi, uuema tehnoloogiaga mingil määral ka automaatselt.⁶⁸ Eestis on artiklite piirid enamasti digiteerijate tõmmatud (tehnoloogia mõningase abiga). Digiteerimise tavaliseks osaks on see saanud umbes viimase kümne aasta jooksul, varasem materjal on salvestatud lehekülgede kaupa. Hea ligipääsu artiklitele on tekitanud DEA, kus ajalehti ja ajakirju on võimalik sirvida artiklite kaupa ja OAI-PMH kaudu analüüsiks välja võtta. Nii on artiklid lisatud ka Jupyteri täistekstide liidesesse.

Kui ajalehed on artikliteks segmenteeritud, avaneb mitmeid uusi tekstianalüüsi võimalusi.⁶⁹ Kuna ühel leheküljel on ajalehtedes tüüpiliselt mitu artiklit, on selle peale raske ehitada teemamudeleid või sõnavektoreid, mis toetuvad teksti õigele järgnevusele ja sisu koostõlale. Artiklite kaupa on tekstist võimalik kätte saada hulk temaatiliselt terviklikke üksusi. Lisaks võivad artiklid olla eri liiki, näiteks võimaldab reklaamide eraldi märkimine, nii nagu seda ka DEA-s on tehtud, korraldada ainult reklaamižanri puudutavaid uuringuid. Pealkirjadega võivad eristuda veel näiteks spordiküljed, ilmateade või naljad. DEA-s on segmenteeritud ajalehed, mis sisaldavad ka OCR-i kindluse hinnangut, s.o praegu 62% ajalehenumbritest. Joonis 10 näitab DEA kollektsiooni artikliteks segmenteeritud ja segmenteerimata ajalehenumbrite osakaalu. Siin tuleb taas välja, et paljudel varem digiteeritud ajalehtedel (aastatest 1850–1950) puudub oluline infokiht. Lisaks on tihti segmenteerimata ka

67 K. Kruusmaa. Estonian Historical Newspaper Crowdsourced OCR Corrections [Data set]. National Library of Estonia. – Zenodo, 2024, 15.08. <https://doi.org/10.5281/zenodo.13325713>.

68 N. Girdhar, M. Coustaty, A. Doucet. STRAS: A Semantic Textual-Cues Leveraged Rule-Based Approach for Article Separation in Historical Newspapers. – International Conference on Asian Digital Libraries. Ed. by D. H. Goh, S. J. Chen, S. Tuarob. Springer Nature, Singapore, 2023, 89–105. https://doi.org/10.1007/978-981-99-8085-7_8; W. Sun, H. T. H. Tran, C. E. González-Gallardo, M. Coustaty, A. Doucet. LIAS: Layout Information-Based Article Separation in Historical Newspapers. – International Conference on Theory and Practice of Digital Libraries: Linking Theory and Practice of Digital Libraries. Ed. by A. Antonacopoulos, A. Hinze, B. Piwowarski, M. Coustaty, G. M. Di Nunzio, F. Gelati, N. Vanderschantz. Springer Nature, Cham, 2024, 256–272. https://doi.org/10.1007/978-3-031-72437-4_15.

69 L. Nemvalts. Kus on kodukoht? Analysing the Meaning of Home and Improving OCR Quality in Estonian Exile Newspapers Published in Sweden. Linnaeus University, Växjö, 2024. <https://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-132420>.



Joonis 10. DEA digiarhiivi ajalehenumbrite artikliteks segmenteeritus

Välis-Eesti lehed. Kui DEA-s saab teksti segmentidele ligi päringusüsteemi ja kasutajaliidese kaudu, siis mõnel pool on materjal segmentideks jaotatud ka digiteeritud failide sees. Nendest on pealkirjastatud artiklite väljavõtmine küll võimalik, aga võib nõuda juba erilisi tehnilisi oskusi.

Mida ette võtta järgmiseks?

Mida võiks ette võtta, et liikuda nn neljanda põlvkonna lahenduste poole ja ehk sealt juba ka edasi, nii et ajalehekogudest saaks tõesti hästi kasutatav uurimisaines? Siinkohal on abi samuti Ehrmanni ja tema kolleegide lähtekohtadest.⁷⁰

Neljanda põlvkonna lahenduste defineerivaks tunnuseks on sisu läbipaistvus. See tähendab, et huviline peaks iga hetk aru saama, kust ta viibib ja mis laadi ainesega parasjagu tegeleb. Väga kergesti peaks kättesaadav olema kontekstivaate, s.o kuhu leitud ajaleht või lehekülg paigutub. Samuti võiks kergesti teada saada, millised olid sama ajastu sarnased lehed ja kas ehk mõni oluline väljaanne neist võib olla

70 M. Ehrmann, E. Bunout, M. Düring. Historical Newspaper User Interfaces: A Review.

digiteerimata. Selleks on tähtis, et digiarhiivide metaandmed oleksid võimalikult hõlpsasti kättesaadavad, et uurija saaks neist ise vajaliku info leida. Kasutajaliides võiks aga pakkuda võimalikult mitmekesiseid visualisatsioone. Analüüside jaoks võib päris oluline olla teadmine, millal ja mis tehnoloogiaga on materjal digiteeritud: 2008. ja 2023. aastal digiteeritud tekstid ei pruugi olla kvaliteedilt võrreldavad ega pruugi seega toimida sama andmeanalüüsi sisendina. Tekstivaate kõrval võiks näidata ka OCR-i kvaliteedihinnanguid, nii tarkvara enda pakutud numbritega kui ka järelanalüüsi tulemi oma. Väike osa sellest andmekogumisest on siinse artikli tarvis ära tehtud: väljaandepõhiselt on koondatud info DEA, Etera, TÜ DSpace'i ja Kivikese kollekttsioonist ning ühendatud see ERB perioodikaosas pakutud kontekstiinfoga. DEA OAI-PMH-liidese kaudu oli võimalik osale tekstile pakkuda ka OCR-kvaliteedi esialgset hinnangut. Siinse artikli elektroonilises lisas esitatud andmeid saab kasutada ülevaate saamiseks mõne väiksema juhtumiuuringu puhul.

Oluline on parandada digiainese kvaliteeti ja ligipääsuvõimalusi. Varem ja hiljem digiteeritud ajalehtede OCR-kvaliteet kõigub oluliselt, tekst on artikliteks jaotatud vaid mõnes digiarhiivis. Tehnoloogia areng pakub siin head ja tõenäoliselt suhteliselt taskukohast lahendust. Varem üles pildistatud tekste on võimalik uuesti tärgtuvastada ja võimalik, et ka loomuliku keele töötuse ja tehisintellekti vahenditega parandada. DEA OAI-PMH-liides, mis võimaldab tekste koguda ühe kaupa, ja selle baasil tehtud Jupyteri lahendus suurema hulga teksti korraga väljavõtmiseks on kasutamiseks siiski tehniliselt suhteliselt keerulised ning nõuavad mõningast programmeerimise ja andmetöötuse oskust. Eesti Rahvusraamatukogu Digari uues digiainese analüüsi kasutajaliideses on plaanis välja arendada ligipääsupunktid nii tavakasutajatele kui ka andmeteadlastele.⁷¹ Mitmes kasutajaliideses on praegu ligipääs failide ja lehekülgede kaupa, mis tähendab, et suurema hulga ajalehtede analüüsimiseks tuleb kokku koguda päris mahukas andmestik. Mõnel pool, nt Etera kollekttsioonis, on ajalehed tihtipeale küll artikliteks segmenteeritud, aga kasutajaliides ei paku võimalust neid artiklite kaupa lugeda. Artiklikihti on siiski võimalik kasutada, aga selle toorfailidest väljavõtmine nõuab tehnilisi erioskusi. Mitmel pool näib olevat võimalusi väikeste muudatuste abiga tekstiainese kvaliteeti ja ligipääsetavust oluliselt parandada.

Digiarhiivide arendamisel on üpris palju teha ka nn kolmanda põlvkonna⁷² tekstide rikastamise suunal. Loomuliku keele töötlusvahendid

⁷¹ Kultuuripärandi digitegevuskava 2024–2029.

⁷² M. Ehrmann, E. Bunout, M. Düring. Historical Newspaper User Interfaces: A Review.

on eesti keele analüüsis juba mõnda aega suhteliselt võimekad⁷³ ja saavad tehisintellekti jätkuva arenguga täita üha keerulisemaid rolle. Praegu on nende vahendite põhjal Jupyteri tekstikogusse lisatud lemmatiseeritud tekste. Suurem arendus toimub automaatse märksõnastamise projekti käigus, kus tekstikihti töödeldakse ja klassifitseeritakse mitmel viisil.⁷⁴ Selle ja järgmiste arendustega on võimalik lisada digiarhiividele tuvastatud olemite kiht, kust võiksid näiteks tekstides käsitletud isikud, kohad ja teemad olla eraldi kättesaadavad.⁷⁵ Tekstide metaandmetest võiks tekkida omaette allikas kultuuripärandi sisuliseks analüüsiks.⁷⁶ Lisaks salvestatud andmekihile võib ette kujutada, et tulevikus oleks tekstikogu koostades üpris kerge tehisintellekti abiga lisada tekstide sisu põhjal märksõnad.⁷⁷

Oluline proovikivi Eesti kontekstis on ka ühenduse loomine eri asutustes eri aegadel ja erinevate eelistustega loodud digiarhiivide vahel. Kui huvipakkuv aines paikneb mitmes arhiivis, on suhteliselt keeruline seda ühendada, isegi kui igas arhiivis eraldi võib olla häid analüüsivahendeid. Ka siin on Digari uue kasutajaliidese arendus võtnud enda kanda keerulise ülesande kaasata sisu ka mitmest väiksemast arhiivist.⁷⁸ Kuna asutuste vaatepunktid oma arhiivide arendamisel on väärtuslikud ja ühise suure platvormi tegemisel on omad riskid (haldamine võib osutuda keeruliseks ja uute mõtete suhtes paindumatuks), võiks kaaluda välja arendada teatud vahejuppe, mis haakuks ühtviisi hästi nii Digari, Kivikesse, Etera, TÜ DSpace'i kui ka väiksemate kogudega. Samasuguseid plaane on proovitud ellu viia ka E-varamu juures, püüdes ühendada informatsiooni eri arhiividest. Praegust lahendust täiendades oleks tarvis, et otsingust saaks lisaks asukohtadele arhiivis kätte ka digimaterjali sisu ja väljavõtted. Võimalik, et selle jaoks peaks digiarhiivide ligipääsuvõimalusi veidi parandama, aga töö mahult oleks see palju kergem kui igale neist mitmekülgsed visualiseerimise ja andmetest koondülevaadete tegemise võimalused sisse ehitada. Kui seda teha väikeste ja võimalikult avatud

73 S. Orasmaa, T. Petmanson, A. Tkachenko, S. Laur, H.-J. Kaalep. EstNLTK-NLP Toolkit for Estonian. – Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Ed. by N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. P. European Language Resources Association (ELRA), Portorož, 2016, 2460–2466.

74 Kultuuripärandi digitegevuskava 2024–2029.

75 A. Kõnno, K. Allkivi-Metsoja, J. Kippar, T. Kamarik, H. Petrov, I. Ibrus. Projekti EKKD119 andmestik: ERR-i raadioarhiivi kultuurisaadete süntaktiline märgendus, tuvastatud nimeolemid, märksõnad, teemad ja nendevahelised seosed, 2003–2021. – DataDOI, 2024, 17.06. <https://doi.org/10.23673/re-469>.

76 I. Ibrus, M. Ojamaa. Newsreels Versus Newspapers Versus Metadata: A Comparative Study of Metadata Modelling the 1930s in Estonia. – VIEW Journal of European Television History and Culture, 2018, 7, 14, 123–137. <https://doi.org/10.18146/2213-0969.2018.jethc157>.

77 A. Karjus. Machine-assisted Mixed Methods: Augmenting Humanities and Social Sciences with Artificial Intelligence.

78 Kultuuripärandi digitegevuskava 2024–2029.

ühendustlülide kaudu, oleks neid kerge täiendada ja kombineerida ning loota saaks ka uurijate panusele.

Valdkondadeülene koostöö ajalehtede uurimisel

Huvitavate rakenduste leidmine digiteeritud ajalehtedele nõuab nii digikompetentsi kui ka sisulisi teadmisi. Seetõttu on valdkonna suured arenguhüpped toimunud just suurte koostööprojektide raames. Näiteks koondas projekt „NewsEye“ (2018–2022), mille eesmärk oli koostöös raamatukogudega arendada ajalehekogude kasutamismõimalusi uurimistööks, 35 uurijat (ajaloolased, ühiskonnateadlased, digihumanitaarid, arvutiteadlased). Projektis arendati töövoogusid materjalide kvaliteedi parandamiseks, pakuti lahendusi kasutajaliidestele ja täiendati meetodeid ajalehtede sisuliseks analüüsiks.⁷⁹ Projekt „Impresso“ (2017–2020, jätkuprojekt 2023–2027) koondas 13 uurijat (ajaloolased, disainerid, arvutuslingvistid, arvutiteadlased), et välja töötada keeletehnoloogilistele lahendustele toetuv digiteeritud ajalehtede kasutajaliides.⁸⁰ „Living with Machines“ (2018–2023) koondas 34 spetsialisti (raamatukogunduse, ajaloo, keeletehnoloogia ja andmeteaduse taustaga) ning püüdis ajalehtede digiarhiivide kaudu mõtestada masinate rolli 19. sajandi tööstusühiskonna kujunemisel.⁸¹ „Oceanic Exchanges“ (2017–2019) koondas 41 uurijat, et uurida riikidevahelist infolevikut aastatel 1840–1914 ilmunud ajalehtede põhjal.⁸² Püüdlustes kasutada olemasolevaid ressursse sisuliste uurimistööde tarvis selguvad uurijate tegelikud vajadused ja täienevad asjakohaste meetodite komplektid.

Ka siinses artiklis toodud lahendused on välja kasvanud seesugusest koostööprojektidest. Ligipääsupunkt täistekstidele kasvas näiteks välja soovist kasutada pikaajaliste ühiskondlike muutuste uurimiseks ajalehetekste Tartu Ülikooli projektis PRG346.⁸³ Selles projektis töötas

79 S. Oberbichler, E. Boros, A. Doucet, J. Marjanen, E. Pfanzelter, J. Rautiaine, H. Toivonen, M. Tolonen. Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists, and Librarians. – Journal of the Association for Information Science and Technology, 2021, 73, 2, 225–239. <https://doi.org/10.1002/asi.24565>.

80 M. Ehrmann, M. Romanello, S. Clematide, P. Ströbel, R. Barman. Language Resources for Historical Newspapers: the Impresso Collection. – Proceedings of the 12th Language Resources and Evaluation Conference. Ed. by N. Calzolari, F. Béchet, P. Blache *et al.* European Language Resources Association (ELRA), Marseille, 2020, 958–968.

81 R. Ahnert, E. Griffin, G. Tolfo, M. Ridge. Collaborative Historical Research in the Age of Big Data: Lessons From an Interdisciplinary Project. Cambridge University Press, Cambridge, 2023. <https://doi.org/10.1017/9781009175548>.

82 Oceanic Exchanges Project Team. Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840–1914. – OSF, 11.10.2019. <https://doi.org/10.17605/OSF.IO/WA94S>.

83 L. Kanger, P. Tinitis, A.-K. Pahker *et al.* Long-Term Country-Level Evidence of Major but Uneven Ruptures in the Landscape of Industrial Modernity. – Environmental Innovation

koos 14 uurijat (ühiskonnateadlased, ajaloolased, andmeanalüütikud), teiste ülesannete seas püüdsime digiteeritud ajalehtede abiga mõõta ka ühiskondlike hoiakute muutumist. Digiteeritud tekstide ülevaade, sõnamitmike loendaja ja ka metaandmete sirvija on ajendatud projektist EKKD72, kus 12 tudengi (ajaloolased, keeleteadlased, kultuuriteadlased, arvutilingvistid, andmeteadlased) ja kahe juhendaja koostöös viisime läbi juhtumiuuringuid, mis panid proovile DEA kasutusvõimalused ja töid esile arendusvajadused.⁸⁴ Rahvusraamatukogu digilaboris oleme teinud ka mitu juhtumiuuringut, millesse on kaasatud ajaloo, keeleteaduse, digihumanitaaria ja andmeteaduse eksperte ning mille tulemusena oleme põhjalikult arendanud uuringutes kasutatud andmestikke ja tööriistu, mida on eespool juba kirjeldatud.⁸⁵

Digihumanitaarias on tüüpiline, et kogude arendamine toimub praktilise kasutamise käigus järk-järgult tehtud arendustöö kaudu.⁸⁶ Rahvusvaheline kogemus näitab, et olulisi edusamme aitab siin teha just tihe koostöö humanitaarteadlaste, arvutiteadlaste ja mäluasutuste vahel.⁸⁷ Sama on märgatud mäluasutustes, eesotsas GLAM Labsi liikumise kaudu: digikogude arendamine töötab hästi just juhtumiuuringute kaudu, kui sisuliste küsimustega uurijad otsivad koos mäluasutuste töötajatega võimalusi digiarhiivist olulist infot kätte saada.⁸⁸ Valdkondadevaheliste koostööprojektide kaudu on võimalik teha sihikindlat tööd, mille tulemusena laienevad digiarhiivide kasutusvõimalused, et peegeldada uurijate mitmekesiseid huvisid.

KOKKUVÕTE

Artiklis tutvustasin digiteeritud Eesti ajalehtede kasutamisevõimalusi uurimisallikana. Allikate digiteerimiskaetust hinnates võib öelda, et digitehnoloogia abiga on võimalik uurida suurt hulka ajalehepärandist. Ligi kaks kolmandikku või rohkemgi ainesest on siinsel hinnangul

and Societal Transitions, 2023, 48, 100765. <https://doi.org/10.1016/j.eist.2023.100765>.

84 P. Tinitis, K. Vaik. Digihumanitaaria juhtumiuuringute projekti failid (EKKD72).

85 <https://digilab.rara.ee/blogi/> (10.03.2025).

86 E. Mäkelä, K. Lagus, L. Lahti *et al.* Wrangling with Non-Standard Data. – Proceedings of the Digital Humanities in the Nordic Countries 5th Conference. (CEUR workshop proceedings, No. 2612). Ed. by S. Reinsone, I. Skadiņa, A. Baklāne, J. Daugavietis. CEUR-WS.org, Aachen, 2020, 280–289.

87 M. Tolonen, E. Mäkelä, J. Marjanen, T. Tahko. Arvutiteaduse kaasamine humanitaarharidusse. – Methis: Studia humaniora Estonica, 2020, 26, 35–50. <https://doi.org/10.7592/methis.v2i12.16909>; S. Oberbichler, E. Boroş, A. Doucet *et al.* Integrated Interdisciplinary Workflows for Research on Historical Newspapers.

88 M. Mahey, A. Al-Abdulla, S. Ames *et al.* Open a GLAM Lab. Digital Cultural Heritage Innovation Labs, Doha, 2019.

digiteeritud ja suur hulk sellest juba kergesti ligipääsetav ka täistekstidena. Seda võib pidada väga heaks uudiseks kohalikule uurijale, aga see pakub konkurentsieelist ka maailmas, kus kogukast trükipärandist on väga keeruleine enamikku kogukonnas toimunud avalikust arutelust kätte saada.

Lähemas tulevikus jäävad uurijate valikuid kindlasti mõjutama konkreetsete digikogude ligipääsuvõimalused ja kvaliteet. Uuringuid tuleb planeerida selle ümber, mis on kättesaadav, tänapäeval tuleb seda teha juba uurimisprojekti koostamise etapis. Uurimistöö teemavalikute tarvis on koostatud digiteeritud ajalehtedest ülevaade nii üldnumbriti kui ka interaktiivse rakendusena, mis aitab ehk kaasa digiteeritud kultuuripärandi ulatuslikumale kasutuselevõtule teadusprojektides.

Niisuguste vahendite kasutuselevõtuks on oluline keskenduda uurijale suunatud rakenduste loomisele. Uuringute esimesteks sammudeks on saadaval sõnamitmike sageduse kuvaja, mis annab kiire ülevaate digikogude sisust, ning täistekstide otsingusüsteem, kust võib kätte saada täpselt selle osa tekstikogust, mida tarvis. Selleks saab digiarhiivist välja valida sobiliku ainese, otsida sealt teatud tähekombinatsioone sisaldavaid tekste ning rakendada ka muid koodipõhiseid analüüse.

Järgmises etapis tasub tulevase uuringu tarvis kriitiliselt üle vaadata olemasolev digiteeritud aines ja välja selgitada, kas oleks ehk võimalik juba praegu kättesaadavate lihtsate vabavaraliste lahendustega parandada varem digiteeritud ainese kvaliteeti. Kasutajaliidesed võiks tööle panna andmetele ligipääsu võimaldamise nimel ning püüda teha võimalikult läbipaistvaks ja selgeks, kuidas leitud ajalehenumbreid seostuvad nii teiste numbritega digikogus kui ka muu seotud meediaruumiga, mis ei pruugi asuda samas digikogus. Otsida võiks ka võimalusi olemasolevate, ent teistel platvormidel paiknevate digikogude ühendamiseks, et kogu ressurss oleks ühtsel viisil ligipääsetav.

Väheoluliseks ei tohiks digiteeritud kultuuripärandist väärtuse loomisel pidada ka valdkondadevahelist koostööd ja juhtumiuuringupõhist katsetamist. Rahvusvaheline kogemus näitab, et just koos konkreetseid uurimisküsimusi lahendada püüdes tehakse märkimisväärsed edusamme nii andmete täpsustamise, infrastruktuuri edendamise kui ka analüüsimeetodite loomisel. Digiteeritud Eesti ajalehed on praegu nii esinduslikkusest kui ka ligipääsetavusest üpris eeskujulikus seisus ning ootavad järgmisi uuringuid ja katsetusi.

Andmed ja kood

Analüüsi käigus kogutud andmed ja kood nende esitlemiseks asuvad andmehoidlas <https://doi.org/10.17605/OSF.IO/ZUT9F>.

Tänuavaldus

Uurimistööd on finantseerinud Haridus- ja Teadusministeeriumi programm „Eesti keel ja kultuur digiajastul“ (EKKD72) ja Eesti Teadusagentuur (PRG346, PRG2160, PUTJD1251). Tööriistade valmimisel on olulist rolli mänginud Eesti Rahvusraamatukogu digilabor, eriti Jane Makke, Urmas Sinisalu, Laura Nemvalts ja Krister Kruusmaa. Tänan artiklit retsenseerinud ja kommenteerinud kolleege, eriti Krister Kruusmaad, Laura Nemvaltsi, Kaisa Langerit, Liina Lindströmi, Tambet Muidet, Erki Tammiksaart, Aro Velvetit, Olaf Mertelsmanni, Lydia Risbergi ja Elisabeth Kaukoneni.

DIGITISED ESTONIAN NEWSPAPERS AS A RESEARCH SOURCE

Peeter Tinit

Large quantities of printed heritage have been digitized in Estonia allowing computer assisted research in digital history. By official estimates, at least 42% of Estonian cultural heritage has been digitized, including 30% of printed heritage. In this article, I approach three interrelated questions on the state of the archives: 1) What do the digital newspaper archives contain and how representative are they of the historical media publishing? 2) How could we use this material in research and what opportunities does it present? 3) What could we pay attention to in developing the state of the art further? I address these questions by collecting data from several archives, building tools to assist research, and situating the current state in the context of general developments in digital archives.

For a general overview, I collected data from four different digital archives that contain newspapers (Digar, Etera, UT DSpace, Kivike). Altogether more than 700 000 newspaper volumes have been digitized and made available. I created visual overviews to see their distribution in time, along with their access rights (90% are openly accessible). Estimating year to year representativeness based on bibliographic data, I find that as much as 59% of the relevant newspapers' issues printed could be already digitized.

I introduce three tools that I have developed with the help of colleagues to address the researcher's needs. 1) A visual overview of digitized newspapers in the context of the newspapers registered in the Estonian National Bibliography, 2) A visual presentation of n-gram frequencies in selected digitized historical newspapers in Estonia, 3) An access point for aggregated full text data from the newspaper archive of the National Library of Estonia. The tools have been built to be small and flexible to allow researchers to gain initial access, with larger platforms being built in the same line in the National Library of Estonia.

Finally, I discuss future directions in developing the digitized newspapers here as a research resource. The choices available for researchers are limited by the quality of the digitized materials and the way that they are presented. Often the works that were considered important were digitized early and therefore have a poorer quality. I present some summary data from Digar on this. There may be a need to improve the form and quality of the materials already digitized.

In the near future strive towards 'fourth generation' digital archives and beyond, I see the goals for research transparency, quality improvement, data enrichment and linking of archives as the most relevant. For transparency, an interface should try to provide an understanding of the digitization and historical context of the visible materials, including metadata on when and how they were digitized and how this relates to the contemporary media landscape. For quality improvement, semi-automatic workflows can be relied upon for re-OCR-ing old images, correcting OCR mistakes in text or segmenting pages into articles. In data enrichment, natural language tools and artificial intelligence can be sought to integrate with the text access workflows to simplify text classification or entity tagging. Finally, intermediary links between archives could be built to easily combine data from different sources.

Special emphasis in developing the resource here could be placed on large interdisciplinary collaborative groups that gather both content and technical specialists. Here, specific case studies can bring out both the value from the archives as well as the requirements for the studies to be made. International experience here shows that advances for the archives often come out of these collaborations. The progress made in this article also comes from similar sources.

In sum, the digital newspaper collections in Estonia have good historical coverage, are available for access and can be used for data-intensive research. There are steps ahead in the near future for quality improvement, user experience and an increase in transparency, with some

developments already in motion. This article aims to contribute to it with the collected data on the contents of the archives and the simple tools made to visualize and provide access to the available materials. Digitized newspaper data in Estonia is ready to be adopted for research use and offers opportunities for various case studies to be made.