

Eesti mäluasutuste andmebaasid: metaandmete kasutusvõimalused ajalooallikate analüüsis

Liisi Taimre

Tartu Ülikool, ajaloo ja arheoloogia instituut, Ülikooli 18,
50090 Tartu; liisi.taimre@ut.ee

Kokkuvõte. Artikkel uurib, mil määral saab Eesti mäluasutuste andmebaaside metaandmeid uurimisprojektides taaskasutada. Kasutades Asutava Kogu ja Tartu linnavolikogu protokolle juhtumiuuringuna, analüüsitakse arhiivi infosüsteemi AIS ja muuseumide infosüsteemi MuIS metaandmete võimekust toetada ajalooliste allikate andmeanalüüsi. Kuigi andmebaasid pakuvad väärtuslikke võimalusi, piirab ühtlustamata isiku- ja organisatsioonikirjete struktuur nende efektiivset kasutust. Lahendusena pakutakse välja luua eraldiseisev linkandmete portaal, mis võimaldaks paremat andmete ühendamist ja taaskasutust.

Märksõnad: metaandmed, linkandmed, andmete taaskasutus, kultuuri-pärandi andmebaasid

Eesti mäluasutuste – muuseumide, arhiivide ja raamatukogude – andmebaasid sisaldavad miljoneid kirjeid, mis on väärtuslikud nii kogude haldamiseks kui ka nende avamiseks laiemale kasutajaskonnale. Andmebaasid pakuvad huvilistele teavet üksikute museaalide, arhivaalide ja teavikute kohta ning võimaldavad kogusid sirvida. Kasutajad saavad otsida ja filtreerida andmeid erinevate parameetrite alusel. Lisaks võimaldavad mitmed andmebaasid interneti kaudu ligipääsu museaalide, säilikute ja teavikute digitaalsetele kujutistele, vähendades vajadust füüsiliselt muuseumi, arhiivi või raamatukokku kohale minna.

Astutud on ka samme andmebaaside koosvõime tagamiseks. E-varamu portaal tutvustab end lausega „Eesti kultuuripärand ühes kohas, ühe otsinguga“.¹ Sinna on koondatud 12 asutuse 19 andmebaasis leiduv info.² Samuti on tänu ühtlustatud metaandmetele eesti kultuuripärand kättesaadav sellistes üle-euroopalistes veebikeskkondades nagu Europeana³ ja European Archives Portal.⁴

Kõik need sammud on aga eelkõige astutud silmas pidades kasutajat, kelle jaoks on oluline info võimalikult kiiresti ja mugavalt kätte saada. Kui soovida kasutada andmebaasides leiduvat infot andmeanalüüsiks, siis on esimeseks kriteeriumiks see, et võimalikult suur osa infot oleks võimalikult hästi struktureeritud ja metaandmetega varustatud. Kohati need eesmärgid muidugi kattuvad, sest mida rohkem on andmed struktureeritud, seda paremad on otsivõimalused, seda nii avaliku veebiliidese kasutaja kui ka andmestikku töötleva teadlase jaoks.

Selliseid ühtlustatud andmetel põhinevaid otsingu- ja sirvimissüsteeme on Soome arvutiteadlane Eero Hyvönen nimetanud esimese põlvkonna süsteemideks. Teise põlvkonna lahendused keskenduvad digihumanitaaria uurimisprobleemide lahendamisele, integreerides andmeanalüüsi- ja visualiseerimise tööriistu otse veebirakendustesse. Kolmanda põlvkonna süsteemide fookus on juba tehisaru võimalustel.⁵ Oluline on Hyvöneni jaotuse puhul tähele panna, et esimese põlvkonna süsteemide puhul räägib ta ühtlustatud andmetest, teise ja kolmanda puhul tööriistadest, mida on võimalik andmeanalüüsiks kasutada – on ju ka tehisaru oma olemuselt tööriist. Pole aga mõtet kasutada andmeanalüüsi tööriistu ilma andmeteta, seega on teise või kolmanda põlvkonna süsteemide eelduseks piisava hulga hästi struktureeritud andmete olemasolu, mida analüüsiks rakendada.

Siinjuures on oluline eristada uurimisandmeid ehk (digiteeritud) originaalallikaid ja uurimismaterjali, mis hõlmab nii allikaid endid kui ka nendega seotud metaandmeid. Ilma korralike metaandmeteta on digitaalsed andmekogud vaid osaliselt kättesaadavad ja analüüsitavad.⁶ Metaandmete kvaliteet ja nende kasutusvõimalused on teema, mille

1 E-varamu. <https://www.e-varamu.ee/et/> (03.01.2025).

2 E-varamu partnerid. <https://www.e-varamu.ee/et/partnerid/> (12.12.2024).

3 Europeana. <https://www.europeana.eu/et> (01.12.2024).

4 Archives Portal Europe. <https://www.archivesportaleurope.net/> (01.12.2024).

5 E. Hyvönen. Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-Analysis and Serendipitous Knowledge Discovery. – Semantic Web – Interoperability, Usability, Applicability, 11, 1, 2020, 2–4. <https://doi.org/10.3233/SW-190386>.

6 K. Elo. Big Data, Bad Metadata: A Methodological Note on the Importance of Good Metadata in the Age of Digital History. – Digital Histories: Emergent Approaches within the New Digital History. Ed. by M. Fridlund, M. Oiva, P. Paju. Helsinki University Press, Helsinki, 2020, 104, 103–111. <https://doi.org/10.33134/HUP-5-6>.

olulisusele on tähelepanu juhitud nii riiklikes strateegiates⁷ kui ka digitaalsete kultuuripärandi kasutusvõimalustele pühendatud uuringutes.⁸

Teadusandmete haldamisel on hea tava lähtuda FAIR-põhimõtetest. FAIR on akronüüm, mis tähistab nelja põhimõtet: leitavus (*findability*), ligipääsetavus (*accessibility*), koostalitlusvõime (*interoperability*) ja taaskasutatavus (*reusability*).⁹ Kuigi mäluasutuste andmebaasid pole küll klassikalises mõttes teadusandmebaasid, tasub enne nende kasutamist uurimistöoriistana analüüsida nende vastavust FAIR-põhimõtetele. Eriti huvipakkuv on seejuures metaandmete taaskasutuse võimalus, mis aitab uurimisprojektide puhul märkimisväärselt aega kokku hoida.

Käesoleva artikli peamiseks uurimisküsimuseks on see, kas ja mil määral saab Eesti mäluasutuste andmebaasides olevaid metaandmeid uurimisprojektides taaskasutada. Analüüsin küsimust kahe Rahvusarhiivis hoiul oleva allika, Asutava Kogu protokollide ja Tartu linnavalikogu protokollide näitel. Tartu linnavalikogu protokollid aastatest 1918–1940 on hoiul Tartu linnavalitsuse fondi ERA.2966 linnavalikogu koosolekute protokollide sarjas.¹⁰ Tegu on umbes 1,7 miljonist sõnast koosneva tekstiga, mille nimeüksused on pilootkatses osaliselt märgendatud. Asutava Kogu protokollid¹¹ asuvad fondis ERA.15 ja sisaldavad umbes kaks miljonit sõna. Allikate iseloomust lähtuvalt sisaldavad mõlemad isiku- ja organisatsiooninimesid. Isikud ja organisatsioonid ning nendega seonduvad metaandmed on ka Rahvusarhiivi arhiivi infosüsteemi ja muuseumide infosüsteemi andmemudelite olulised osad ja see lubab oletada, et neid metaandmeid on võimalik taaskasutada.

Kultuuripärandi andmete taaskasutuse mõõtmine on keeruline, kuna standardiseeritud meetodid on alles kujunemas.¹² Varasemates Eesti mäluasutuste andmebaase käsitlevates analüüsides on juhitud tähelepanu andmebaasides esinevale ebaühtlasele kirjelduskvaliteedile

7 Eesti avaliku teabe masinloetava avalikustamise roheline raamat, 2014, 12. <https://www.aki.ee/sites/default/files/documents/2024-03/avaliku-teabe-masinloetava-avalikustamise-roheline-raamat-20141125.odt> (03.01.2025).

8 M. Lorenzini, M. Rospocher, S. Tonelli. On Assessing Metadata Completeness in Digital Cultural Heritage Repositories. – Digital Scholarship in the Humanities, 36 (Supplement 2), 2021, iii82–iii88. <https://doi.org/10.1093/lc/fqab036>; B. Whearty. Interoperable Metadata and Failing Toward the Future. – Digital Codicology: Medieval Books and Modern Labor. Redwood City, CA, 2022, 168–212. <https://doi.org/10.1126/stanford/9781503632752.003.0005>.

9 M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg *et al.* The FAIR Guiding Principles for Scientific Data Management and Stewardship. – Scientific Data, 3, 2016, 160018. <https://doi.org/10.1038/sdata.2016.1>.

10 Linnavalikogu koosolekute protokollid. 1918–1940. (Sari.) Rahvusarhiiv (RA). ERA.2966.

11 Asutava Kogu protokollid. 1919–1922. (Sari.) RA, ERA.15.

12 N. McNeilly, S. Vasileva, M. Brambilla, P.-Y. Lin, E. Hamhoughia. Measuring the Impact of Reuse of Digital Heritage. – Europeana Pro, 19.09.2024. <https://pro.europeana.eu/post/measuring-the-impact-of-reuse-of-digital-heritage> (02.12.2024).

ja tehnilistele raskustele masspäringute tegemisel.¹³ Siinse artikli fookuses on andmebaaside teoreetiline võimekus pakkuda võimalust metaandmete taaskasutamiseks ning järelduste tegemisel lähtutakse andmemudelitest ja andmebaasi tootejuhtide kommentaaridest.

Artikli esimeses osas keskendun sellele, kuidas linkandmete ja ontoloogiatega abil muuta metaandmed masinale mõistetavaks ning seejärel toetada ajalooliste tekstide andmepõhist analüüsi. Seejärel käsitlen konkreetsete allikate, Asutava Kogu ja Tartu linnavolikogu protokollide põhjal loodavat võimalikku andmemudelit ning arutlen, millisel määral võimaldavad AIS-i ja MuIS-i metaandmed neid allikaid uurimisprojektides rikastada ja taaskasutada.

LINKANDMETE ROLL MASINALE MÕISTETAVA INFO LOOMISEL

Selleks, et arvuti abiga andmeid taaskasutada, on ülimalt oluline, et informatsioon oleks lisaks inimesele tähenduslik ka arvutile. Siinjuures on oluline eristada mõisteid „masinloetav“ ja „masinale mõistetav“.¹⁴ Esimene neist viitab andmete vormingule, mida masin suudab tehniliselt lugeda ja töödelda.¹⁵ Määratlus „masinale mõistetav“ rõhutab seevastu andmete semantilist struktuuri, mis võimaldab arvutil andmeid lisaks töötlemisele ka omavahel tähenduslikult seostada. Kui eelteadmisi omava inimese jaoks pole oluline, kas kasutatakse erakonna täisnimetust Eesti Sotsiaaldemokraatlik Töölise Partei või lühendit ESDTP, siis andmete automaatse analüüsi puhul jäävad ilma lisainfota need kaks nimetust omavahel seostamata. Samuti on andmete masstöötluse puhul keeruline tuvastada, et Tartu ja Dorpat viitavad samale linnale. Ei saa eeldada isegi seda, et arvuti käsitleks neid kohanimenä, kui kontekst pole masinale mõistetaval viisil määratletud.

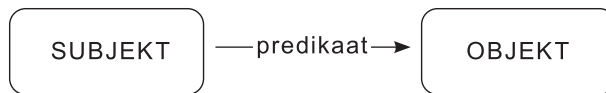
Üks võimalik tehnoloogia andmetele konteksti loomiseks on linkandmete kasutamine. Linkandmed on struktureeritud andmed, mis on seotud unikaalsete identifikaatoritega ning võimaldavad erinevate

13 J. Ariva, V. Puik, T. Sepp. Kratt Folli: digitaalsete objektide kirjeldamine tehisintellekti abil: Eelanalüüs. STACC OÜ, Tartu, 2022 (käskikiri); L. Taimre. Museaalide päritolukaart: muuseumite infosüsteemi avaandmeid kasutav kaardirakendus: Bakalaureusetöö. Tallinna Tehnikaülikool, Infotehnoloogia teaduskond, 2020 (käskikiri).

14 Eesti keeles on laialdasel kasutusel termin „masinloetav“ (*machine readable*). Määratlusele „masinale mõistetav“ (*machine understandable*) veel üldtunnustatud vastet pole.

15 Üks levinumaid masinloetavaid andmevorminguid on CSV (*Comma-Separated-Values*) – failivorming, mis on olemuselt lihtne tabel, kus andmed on eraldatud komadega.

andmekogude tähenduslikku seostamist ja kontekstualiseerimist.¹⁶ Struktureeritud andmete loomine eeldab standardiseeritud mudeleid, mis võimaldavad informatsiooni üheselt mõistetaval ja jagataval viisil esitada. Üks levinumaid tehnoloogiaid linkandmete realiseerimiseks on RDF (*Resource Description Framework*). RDF-i kõige olulisemaks omaduseks on võimalus defineerida seos kahe URI (*Uniform Resource Identifier*, protokoll infoallika üheseks määratlemiseks veebis) vahel. Kaks ressursi ja nendevaheline seos moodustavad RDF-kolmiku – subjekt, predikaat ja objekt. Joonisel 1 on kujutatud RDF-kolmikut. Oma olemuselt on see struktuur suunatud siltidega graaf, mille servad esitavad seost kahe ressursi vahel.

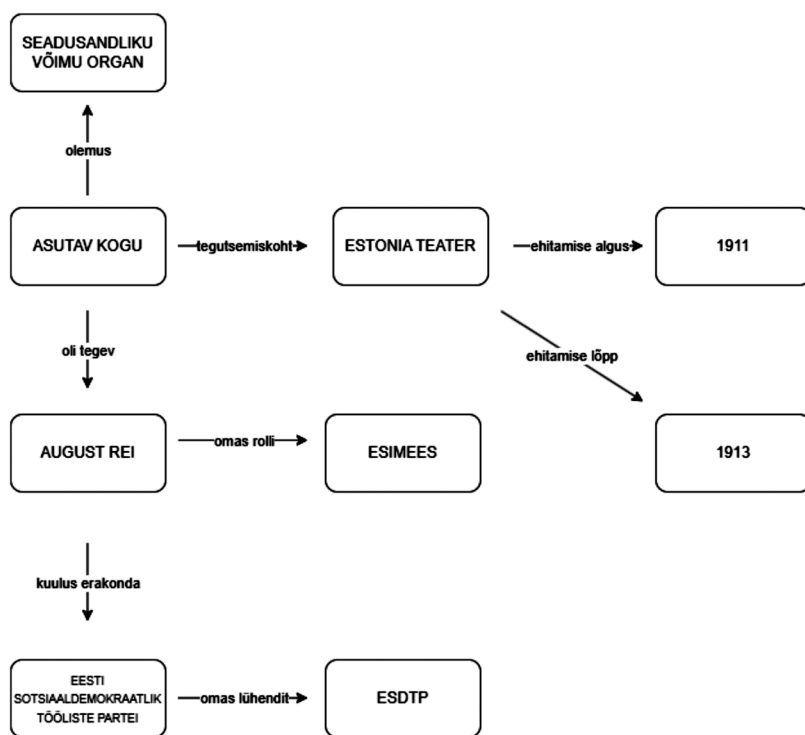


Joonis 1. RDF-kolmik

Joonisel 2 on lihtsustatud mudel, kus on esitatud hulk Asutavat Kogu puudutavaid seoseid. Nagu jooniselt näha, siis seoseid on võimalik luua lõpmatult ning sel viisil seostada inimesi, organisatsioone ja daatumeid. Tegu on näidismudeliga, mille aluseks on võetud Asutava Kogu kohta käivad üldtuntud faktid ja neid on kujutatud RDF-kolmikutena. Asutav Kogu on sel joonisel kolme RDF-kolmiku subjektiks ning seotud kolme omaduse (predikaadi) abil kolme objektiga. Esimene predikaat defineerib Asutava Kogu olemuse (seadusandliku võimu organ), teine defineerib tegutsemiskoha ja kolmas („oli tegev“) loob seose konkreetsete isikutega. Nagu näha, saab ühe kolmiku objektist järgmise kolmiku subjekt. Tegutsemiskohale ehk Estonia teatrile on omakorda võimalik lisada lugematu arv omadusi.

Kui joonisel esitatud kolmikud on lihtsuse eesmärgil esitatud vabatekstina, siis andmebaasides kasutatakse andmete üheselt mõistetavuse ja kasutatavuse tagamiseks standardiseeritud sõnastikke ja ontoloogiaid. Ontoloogiad on formaalsed andmemudelid, mille abil on võimalik määratleda mõistete omavahelisi seoseid, mis omakorda parandab andmete otsivõimalusi. Ontoloogiad on oma olemuselt hierarhilised ning seetõttu saab näiteks määrata, et Estonia teater kuulub laiemasse kategooriasse „ooperiteater“, mis omakorda on kultuuriasutuste alaliik. Nii saab suuri andmekogusid ühendades teha näiteks päringu, millised seadusandliku võimu organid on tegutsenud kultuuriasutustes, mis on ehitatud enne I maailmasõda.

¹⁶ Linked Data. <https://www.w3.org/DesignIssues/LinkedData> (02.12.2024).



Joonis 2. Näiteid Asutava Koguga seotud RDF-kolmikutest

Üks levinumaid ontoloogiaid kultuuripärandi ja ajalooliste andmete modelleerimisel on CIDOC¹⁷ CRM (*Conceptual Reference Model*). Erinevalt teistest ontoloogiatega keskendub see mäluasutuste eripäradele, arvestades nii materiaalse kui ka vaimse kultuuripärandiga ning nendega seotud sündmuste, kohtade ja isikutega.¹⁸

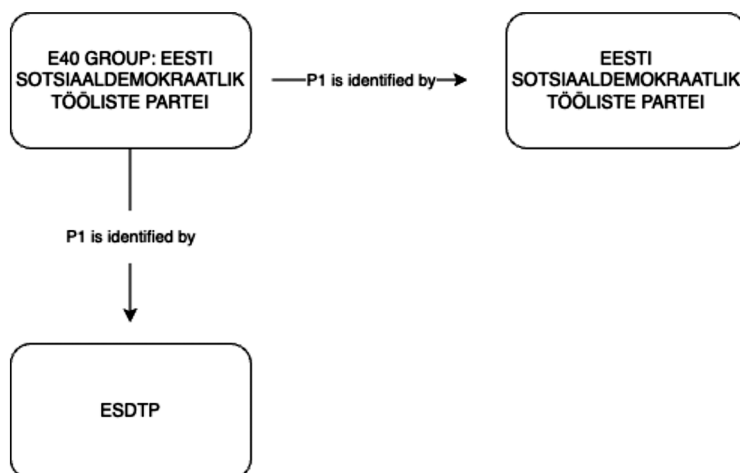
Joonistel 3 ja 4 on kirjeldatud CIDOC-i nimeruumis Eesti Sotsiaaldemokraatliku Tööliste Partei ja Tartu linna alternatiivseid nimekujusid. Eesti Sotsiaaldemokraatlik Tööliste Partei on defineeritud CIDOC-i klassina „E40 Group“, mis tähistab inimeste või inimrühmade kogunemisi või organisatsioone, kes tegutsevad ühiselt või sarnasel viisil, ja mis omaduse „P1 is identified by“ kaudu kirjeldab, kuidas see grupp (või mõni muu entiteet) on identifitseeritav oma nime, lühendi, koodi või muu tunnuse kaudu.¹⁹ Kui siinsel juhul on lisatud vaid kaks nimekuju, siis mudel seoste arvule piiranguid ei sea.

17 CIDOC (Rahvusvaheline Dokumentatsiooni Nõukogu) on osa Rahvusvahelisest Muuseuminõukogust (ICOM).

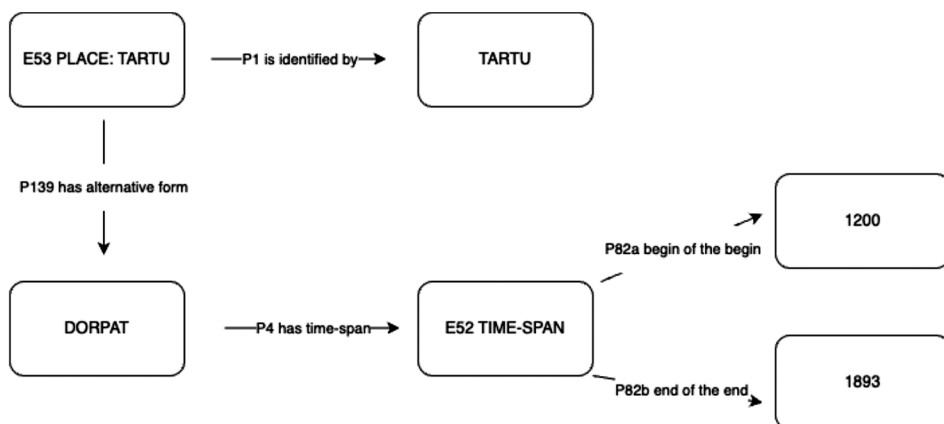
18 CIDOC CRM. <https://cidoc-crm.org/> (05.12.2024).

19 C. Bekiari *et al.* Definition of the CIDOC Conceptual Reference Model. Versioon 7.1.3. 2024.

Tartu alternatiivsete nimekujude kirjeldamise detailsuses on astunud veel samm edasi ja omadusele „P139 has alternative form“ (alternatiivne nimekuju) on lisatud ka ajavahemik, mil alternatiivne nimekuju kasutuses oli.



Joonis 3. Eesti Sotsiaaldemokraatliku Tööliste Partei nimekujud kirjeldatuna CIDOC-i numeruumis



Joonis 4. Tartu nimekujud kirjeldatuna CIDOC-i numeruumis

ANDMEMUDEL TARTU LINNAVOLIKOGU JA ASUTAVA KOGU PROTOKOLLIDE ANALÜÜSIKS

Ajalooallikate analüüs digitaalsete vahenditega tähendab pikka tegevuste jada: materjali digiteerimine, tekstialade tuvastamine, teksti tärgtuvastus, vajadusel üksuste märgendamine ja teksti tükeldamine osadeks ning lõpuks teksti sisuanalüüs. Kõikide nende tegevuste käigus tekib rohkelt metaandmeid, mis on konkreetse allikaga otseselt seotud, alates digiteeritud failide vormingust ja tuvastatud tähemärkide arvust kuni tekstis esinevate nimeüksuste ja sisuanalüüsi tulemusena loodud märksõnadeni. See, millised konkreetsete tegevused teostatakse ja millised metaandmed nende raames tekivad, sõltub uurimisprojekti spetsiifikast.

Käesoleva artikli fookuses on metaandmed, mis ei pärine otseselt allikast endast ega ole tekkinud selle sisuanalüüsi raames, vaid on seotud allikas märgendatud nimeüksustega, näiteks isikute, organisatsioonide, seaduste ja rahavääringutega seotud väärtuslik lisainfo. Tegu on andmetega, mida ei ole mõistlik iga kord uuesti sisestada, vaid neid on mõistlik taaskasutada.

Seonduvad metaandmed muutuvad oluliseks uurimistöö faasis, kui esmane infoeraldus allikatest on juba tehtud ning allikate teksti on võimalik arvuti abil töödelda, märgendada ja vastavalt vajadusele osadeks jagada. Asutava Kogu ja Tartu linnavolikogu protokollidel on mitu eripära, mis lihtsustavad nende analüüsiks vajalikke eeltöid.

1. Allikad on trükimasinal trükitud. Seetõttu on olnud võimalik neile rakendada tärgtuvastust, mis on võimaldanud kätte saada kogu tekstikorpuse suhteliselt lihtsa vaevaga.²⁰

2. Allikad on selgelt struktureeritud. Protokollide formaat võimaldab masintöötuse tööriistadega liigendada teksti sõnavõttude kaupa.

3. Tekstis esinevate unikaalsete isikunimede hulk on pigem väike.²¹ Samuti on Asutava Kogu ja Tartu linnavolikogu liikmete hulgas mitmeid kultuurilooliselt tuntud inimesi, kelle kohta leidub informatsiooni ka muudes allikates ja andmekogudes. Seega pole isikusamasuse tuvastamine probleemiks. Kuna liikmete ja erakondade nimed on leitavad ka

20 Käsikirjaliste arhiiviallikate puhul sõltub infoeralduse edukus praegu tihti veel vabatahtlike panusest. Üheks näiteks sellest on vallakohtute projekt (www.ra.ee/vallakohtud), kus 2019. aastast alates on vabatahtlikud käitsi ümber kirjutanud protokollide täistekste ja seejärel need märgendanud. Kuigi ka Eestis on käsikirjaliste tekstide tärgtuvastuse algoritmeid juba mitmes projektis edukalt kasutatud (vt nt <https://rahvusarhiiv.transkribus.eu/>), siis nõuab nende kasutamine oluliselt rohkem tööd kui standardseid trükitud tähti sisaldavate tekstide tärgtuvastus.

21 Asutava kogu liikmed oli 121, Tartu linnavolikogu liikmete arv jääb alla 300.

muudest allikatest, siis on võimalik juba enne märgendamisprotsessi lõppu hakata koondama täiendavaid isikute ja organisatsioonidega seotud metaandmeid.

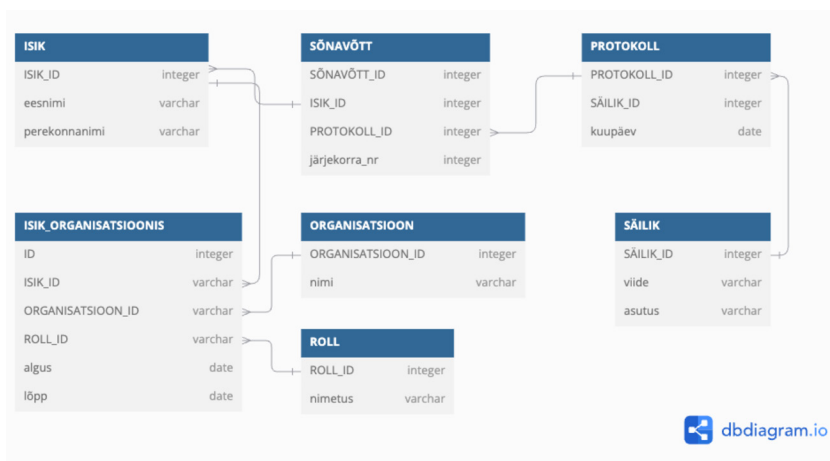
2025. aasta jaanuari seisuga käib Tartu linnavolikogu protokollide tekstitöötamise kontroll ja Asutava Kogu protokollide märgendamisprotsess. Tüüpiline nimeüksuste liigitus sisaldab klasse „Isik“, „Koht“ ja „Organisatsioon“. See jaotus on välja töötatud eelkõige tänapäeva meediatekstide märgendamiseks ja nendest info saamiseks.²² Sellele jaotusele on lisatud veel mitmeid allikaspetsiifilisi märgendusi: isiku amet või roll, poliitiline üksus, geopoliitiline üksus, nimelised sündmused ja tähtpäevad, seaduste täisnimetused, rahasummad.

Kuigi märgendamisprotsess annab nii uurijale kui ka arvutile infot, milline tähekombinatsioon tähistab isikut, ametit või kohanime, siis laiemat konteksti see siiski ei taga. Selleks, et teha päringuid, mis nõuavad mitme parameetri arvesse võtmist – näiteks kuvada kõik sõnavõtted, mille autor on noorem kui 30 või võrrelda Põhja- ja Lõuna-Eestist pärit isikute sõnakasutust –, on vajalik, et iga märgendatud isikunimi oleks seostatud ka konkreetse isikukirjega, mis sisaldab infot sünniaja ja -koha kohta.

Mida rohkem andmeid on võimalik omavahel automaatselt ühendada, seda rohkem uurimisvõimalusi on. Kui eesmärgiks on pakkuda kasutajatele, kel puuduvad sügavad teadmised andmeteadusest, võimalust andmeid tõhusalt kasutada, on vajalik, et eri andmekogudes asuv info oleks automaatselt omavahel kombineeritav. Sisese juhatuses mainitud Hyvöneni jaotuse järgi tähendaks see teise põlvkonna süsteemi loomist, mis võimaldab kasutajatel otse keskkonnas rakendada andmeanalüüsi ja visualiseerimise vahendeid. See omakorda tähendab väga erilaadsete andmehulkade ühendamist.

Joonisel 5 on esitatud relatsioonilise andmebaasi mudel, mis toetab ajalooallikate analüüsi, mille keskmes on organisatsioonide või poliitiliste üksuste tegevus ning kus olulised sündmused või otsused on jäädvustatud protokollide vormis. See toetab võimalust sortida ja töödelda sõnavõtte isikute kaupa, mis omakorda võimaldab põhjalikumalt uurida osalejate rolli ja panust poliitilise üksuse tegevusse. Andmemudel on kokku pandud lähtudes eeldusest, et andmete automaatne taaskasutamine pole praegu võimalik ja kogu info tuleb lisada manuaalselt. Kuna säilikud, protokollid ja sõnavõtted on omavahel hierarhilises seoses, siis on valitud relatsiooniline andmebaas, mis toetab fikseeritud seoseid.

22 S. Orasmaa, K. Muischnek, K. Poska, A. Edela. Named Entity Recognition in Estonian 19th Century Parish Court Records. – Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 2022, 5304, 5304–5313. European Language Resources Association.



Joonis 5. Andmemudel Asutava Kogu ja Tartu linnavolikogu protokollidega seonduva andmestiku kirjeldamiseks

Iga asutuse kogusse kuuluv säilik sisaldab ühte või mitut protokoll. Iga protokoll koosneb ühest või mitmest sõnavõtust. Iga sõnavõtuga on seotud üks isik, sõnavõtul on ka järjekorranumber. Isik on seotud ühe või mitme organisatsiooniga. Ühe organisatsiooni raames võib tal olla erinevaid rolle.

Sõnavõttude tabel koosneb peamiselt viidetest teistele tabelitele: protokollide tabelile ja isikute tabelile (kes on sõnavõtu autor). Ainsaks erandiks on sõnavõtu järjekorranumber, mille abil on võimalik järjestada sõnavõttud ühe protokoll raames. Sõnavõttude tekste on mõistlik vähemalt uurimistöö faasis eraldi andmebaasis hoida. Kui soovida rakendada teksti sisuanalüüsi, siis on analüüsitavaid faile mugavam hallata giti-põhistes keskkondades.²³ Välja SÕNAVÕTT_ID abil on võimalik vajadusel sõnavõtu tekst metaandmetega seostada. Kui analüüsi käigus peaks tekkima uusi metaandmeid, on need võimalik sõnavõttude tabeliga seostada.

Tabelid „Isik“, „Isik organisatsioonis“ ja „Roll“ võimaldavad saada infot selle kohta, millist organisatsiooni kõneleja esindas ning millises rollis ta oli. Iga kirje tabelis on varustatud võimalusel teabega, mis kuupäevast mis kuupäevani isik selles rollis oli. Nii on võimalik kerge vaevaga teha päring, kus küsitakse välja kõik sõnavõttud, mida isik ühe rolli esindajana tegi, sest ka sõnavõttud on kuupäevaga varustatud. Või teha teistpidi päring – millistes rollides isik mingi sõnavõtu tegemise ajal oli.

²³ Git on versioonihalduse süsteem, mis võimaldab salvestada failidest eri versioone ja jälgida neis tehtud muudatusi.

Relatsiooniline andmebaas tähendab aga seda, et tabelid ja nende vahelised seosed on andmemudelil selgelt fikseeritud ning uute seoste lisamine võib olla keeruline. Üheks lahenduseks on relatsioonilise andmebaasi ja linkandmete omavaheline kombineerimine,²⁴ kus näiteks isiku või organisatsiooni tabeli kirjet saab käsitleda kui RDF-kolmiku subjekti. Sel viisil on võimalik säilitada ühelt poolt dokumendi enese kirjeldamiseks vajalikud fikseeritud seosed ja teisalt rikastada neid suuremat paindlikkust ja lihtsaid kombineerimisvõimalusi pakkuvate linkandmetega.

AIS JA MUIS KUI POTENTSIAALSED TAASKASUTATAVAD ANDMESTIKUD

Üheks allikaks, kust leida seonduvaid metaandmeid, mida uurimisprojektides taaskasutada ja olemasolevate andmetega kombineerida, on Eesti mäluasutuste andmebaasid. Järgnevalt analüüsin neid isiku- ja organisatsioonikirjete näitel. Potentsiaalsete taaskasutatavate andmestikena vaatlen arhiivi infosüsteemi AIS²⁵ ja muuseumide infosüsteemi MuIS²⁶ metaandmeid. Tegu on kahe valdkonna – muuseumide ja arhiivinduse – kesksete andmebaasidega, mis tähendab, et need sisaldavad kõige eripalgelisemat materjali nii sisu kui ka infokandjate mõttes. Mõlema kirjete arv ulatub miljonitesse. AIS-ist võib lisaks dokumentidele leida ka kaarte, pitsereid, fotosid ja pärgamente.²⁷ MuIS-is on näiteks Eesti Ajaloomuuseumist kirjeldatud 29 esemetüübi järgi eristatud alamkogu.

Teiseks oluliseks sarnasuseks on asjaolu, et mõlemad koondavad andmeid objektide kohta, mille nimetus või pealkiri võib esmapilgul vähe konteksti avada – näiteks aruanne, protokoll või kruus –, kuid mille tähenduslikkuse avavad metaandmed. Museaalide kultuuriväärtuslikkus peegeldub tavaliselt nende valmistamise, kogusse jõudmise konteksti või isikute ja organisatsioonidega seotud andmetes. Kirjalike dokumentide puhul on sageli vaja lisada metaandmeid, mis kirjeldavad säiliku sisu täpsemalt, kui seda võimaldavad pealkiri ja daatumid. Heaks näiteks on nimekirjad, mis võivad sisaldada tuhandeid isikukirjeid. Konteksti avamiseks vajalik metaandmete rohkus eristab neid ka raamatukogude keskandmebaasist Ester, mis toimib tõhusalt märksa väiksema hulga

²⁴ N. Konstantinou, D.-E. Spanos. *Materializing the Web of Linked Data*. Springer, Cham, 2015, 73–102. https://doi.org/10.1007/978-3-319-16074-0_4.

²⁵ Rahvusarhiiv. AIS on Eesti arhiivinduse infosüsteem. <https://ais.ra.ee/>.

²⁶ Muuseumide veebivärav. <https://www.MuIS.ee/>.

²⁷ Neid on suurema detailisusega kirjeldatud vastavates andmebaasides. Täieliku nimekirja Rahvusarhiivis olevatest andmebaasidest leiab aadressilt www.ra.ee/vau.

metaandmetega. Olukord erineb veelgi Eesti Rahvusraamatukogu artiklite andmebaasist Digar, mis võimaldab ajalehtede ja ajakirjade täistekstiotsingut, pakkudes sellega täiesti teistsuguseid uurimisvõimalusi.

Oma universaalse iseloomu tõttu on tulnud siiski ka AIS-il ja MuIS-il metaandmete detailsuses järeleandmisi teha. Näiteks on Rahvusarhiivi Meediateek²⁸ või kaartide andmebaas²⁹ küll oluliselt rikkalikuma metaandmestikuga, aga need on vaid mõne meediumi põhised ja seetõttu oluliselt väiksema kasutusareaaliga.

Arhiivi infosüsteem AIS

Mõlema andmebaasi puhul on oluline tunda nende saamislugu. Rahvusarhiivi arendatav arhiivi infosüsteem AIS koondab infot arhiivisäilike kirjelduste kohta. AIS-i arendustega alustati 1998. aastal. Avalik veebiliides valmis 2004. aasta lõpuks, selleks ajaks olid kirjeldatud kolme miljoni säiliku andmed.³⁰ Suurem osa dokumentidest pärineb Rahvusarhiivist, samuti hoiavad oma kogude infot AIS-is Tallinna Linnaarhiiv ja Eesti Ajaloomuuseum.³¹ Kokku on 2025. aasta alguse seisuga AIS-is 11 miljonit kirjet.³² AIS-i andmed on kättesaadavad ka avaandmetena vormingus apeEAD XML.³³

AIS-i ülesehitus lähtub eelkõige säilike kirjeldusest ning metaandmed puudutavad eelkõige dokumendi paiknemist struktuuris ja selle piirdateid. Siinse artikli kontekstis on kõige huvipakkuvam väli „Arhiivimoodustaja“,³⁴ sest see on praegu ainus väli, mis pakub struktureeritud infot isikute ja organisatsioonide kohta. Üks lihtne päring, milleks seda kasutada saaks, on näiteks uurida, millistel Asutava Kogu liikmetel (või ükskõik millise inimgrupi liikmetel) on Rahvusarhiivis isikufond. Kui tavakasutaja jaoks tähendaks see kõikide nimede ükshaaval otsilahtrisse sisestamist, siis arvuti jaoks on tegu üherealise päringuga – seda tänu asjaolule, et tal on olemas kontekst, milline väli mida tähendab. Mitmed isikufondid sisaldavad ka fondimoodustajate elulugu ja fondi kirjeldust, aga kuna peale info, kas tegu on asutuse või isikuga ning isiku

28 Rahvusarhiiv. Meediateek. <https://www.meediateek.ee/>. Meediateek sisaldab Rahvusarhiivis säilitatavaid Eesti-ainelisi foto-, filmi- ja helikogusid.

29 Rahvusarhiiv. Kaardid. <https://www.ra.ee/kaardid>. Kaartide andmebaas sisaldab Rahvusarhiivis ja Tallinna Linnaarhiivis hoitavate kaartide kirjeldusi ja digikujutisi.

30 Rahvusarhiiv. AISi ajalugu. <https://ais.ra.ee/et/about/history> (03.01.2025).

31 Täieliku nimekirja leiab AIS-ist. <https://ais.ra.ee/et/about/overview>.

32 Rahvusarhiiv. AISi ajalugu. <https://ais.ra.ee/et/about/history>.

33 apeEAD XML (*Encoded Archival Description in the Archives Portal Europe*) on standardiseeritud XML-formaat, mida kasutatakse arhiivikirjelduste struktureerimiseks esitamiseks ja jagamiseks. Formaat põhineb EAD-standardil, mis on kohandatud Euroopa Arhiiviportaali vajadustele.

34 Asutus või isik (sh nimevormid), kelle tegevuse käigus dokumendid loodi.

sünni- ja surma-aastate, on need sisestatud vabatekstina, siis ilma põhjalikuma töötluseta neid andmeid kasutada ei saa.

Rahvusarhiivis on astunud esimesed sammud, et tulevikus tekiks võimalus luua arhiiviallikas esinevale isikunimele unikaalne identifikaator ehk ID, mis võimaldaks isikusamasuse tuvastamisel seda teiste kirjete ja arhiiviallikatega ühendada. Ühisloomeprojektide raames on vabatahtlike abiga arhiiviallikates märgendatud tuhandeid nimeüksusi, eelkõige isikunimesid, aga ka kohanimesid ja andmeid kinnistute³⁵ kohta. Nimeüksuste märgendamine on aga alles esimene etapp, sest isikusamasuse tuvastamine on väga keeruline ja vahel ka võimatu.

Eriti keeruline on see näiteks 20. sajandist varasemate arhiiviallikate puhul, kus isikutel perekonnanimed puudusid või neid tuvastati talunimede järgi. Vallakohtute protokollide puhul on uurijatele pakkunud keerukusi olukorrad, kus mehel on väga levinud eesnimi (nt Juhan, Hans, Jüri), või juhtumid, kus inimest on nimetatud ainult talunime järgi (nt Kirikumäe tüdruk).³⁶ Mõnel juhul on isikute tuvastamisel abiks rahvastikuloo allikad, nagu näiteks vallaelanike nimekirjad või koguduste personaalraamatud, aga esineb ka juhtumeid, kus isikusamasuse tuvastamine ei osutugi võimalikuks.³⁷

Üks isikusamasuse tuvastamise mõttes kõige suurema potentsiaaliga projekte on Rahvusarhiivi ühisloomealgatus „Eestlased Esimeses maailmasõjas“. Alates 2014. aastast on kasutajad sisestanud 198 022 isikukirjet 901 säilikust. Teadmata on aga see, mitu kirjet käib ühe ja sama isiku kohta, sest üks ja sama inimene võib figureerida mitmes dokumendis.³⁸ Rahvusarhiivi Tartu kasutusosakonna juhataja Tõnis Tärna sõnul on siiski lootust tulevikus need andmed vähemalt osaliselt tehisaruga ühendada, kuna lisaks isikunimedele on allikatest välja toodud ka sünniajad ja sissekirjutused ning kõigi kattuvate tunnustega isikud on väga tõenäoliselt identsed.³⁹

Muuseumide infosüsteem MuIS

Muuseumide infosüsteem MuIS on Muinsuskaitseameti arendatav infosüsteem, mis on mõeldud muuseumikogude haldamiseks ja

35 Rahvusarhiiv. Tartu 1867. aasta kaart. <https://www.ra.ee/tartu1867>.

36 K. Lust, S. Orasmaa, M.-L. Pilvik. Kes kellega kohut käis? Vallakohtuprotokollide analüüs. – *Acta Historica Tallinnensia*, 29, 1, 2023, 35–64, siin 42.

37 Samas, 42–43, 35–64.

38 Arhiiviallikate valik on lai, nimesid on sisestatud väkke võetute nimekirjadest, teenistuslehtedelt ja meditsiinidokumentidest. <https://www.ra.ee/ilmasoda/index.php/site/documents>.

39 Tõnis Tärna, telefonivestlus, 08.01.2025.

tutvustamiseks. Praegu on Eesti muuseumides ligikaudu 7 miljonit museaali,⁴⁰ millest MuIS-i on sisestatud rohkem kui 4 miljonit.⁴¹ MuIS-i on koondunud rohkem kui 60 muuseumi. MuIS-i eellane, kultuuri-väärtuste infosüsteem KVIS valmis 1997. aastal. See võimaldas jälgida museaalide liikumist ja andis ülevaate muuseumide kogudest, kuid ei pakunud avalikkusele veebipõhist ligipääsu. 2003. aastaks kasutas KVIS-i 42 muuseumi.⁴² MuIS-i hakati arendama 2005. aastal ja töökeskkond valmis 2008. aasta suveks.

MuIS-i andmete struktuur on suhteliselt hea detailsusega lahti kirjutatud ja see on leitav MuIS-i avaandmete lehelt.⁴³ MuIS avaldab avaandmeid RDF+XML-formaadis. Kasutusel olevad nimeruumid on CIDOC CRM, Dublin Core, MuIS, OWL ja FOAF. Kuigi arenduses on MuIS 2.0, mis kasutaja jaoks tähendab muu hulgas mugavamad veebikeskkonda ja lisafunktsionaalsusi,⁴⁴ siis lisandused andmete struktuuri oluliselt ei muuda.

Üks põnevamaid MuIS-i andmemudeli osi siinse artikli vaatenurgast on sündmused ja neist lähtuvad seosed. Museaaliga seotud sündmus võib olla nii administratiivne (arvelevõtmine, näitusel osalemine, kogumistegevus) kui ka ajalooline (valmistamine, fotol kujutatu).⁴⁵ Iga museaali on võimalik sündmuses osalenud predikaadi (CIDOC. P12_occured_in_the_presence_of) abil siduda sündmusega (CIDOC. E5_Event). Sündmusega omakorda on võimalik liita osaleja (CIDOC. P11_had_participant), mis võib olla nii isik kui ka organisatsioon.

Museaaliga seotud sündmustes osalenud isikuid ja organisatsioone saab MuIS-i avalikus liideses leida, kasutades otsinguvormi lahtrit „Isik“. See on hea näide sellest, et kasutajale kuvatavad otsilahtrid ei pruugi peegeldada andmete tegelikku struktuuri. Andmemudeli tundmine võimaldab aga teha oluliselt paremini sihitud päringuid. Sisestades isikulahtrisse organisatsiooni Tartu linnavolikogu, jõuame Saaremaa Muuseumis arvel oleva 1934. aastal välja antud trükiseni „Eesti Sotsialistlik Tööliste Partei: töökava ja nõuded Tartu linnavolikogus“.⁴⁶ Seda tänu asjaolule, et koguhoidja on märkinud seotud organisatsiooniks

40 Muinsuskaitseamet. Muuseumide infosüsteem MuIS. <https://www.muinsuskaitseamet.ee/kultuuriparand-eesis/muuseumid/muuseumide-infosusteem-MuIS> (10.12.2024).

41 Samas.

42 Eesti muuseumide infosüsteem: KVIS-i tutvustus. https://vana.muuseum.ee/erialane_areng/muuseumitootajale/muuseumide_infosyste/kvisi_tutvustus/index.html (10.12.2024).

43 MuIS püsiviited ja avaandmete liides. <https://opendata.MuIS.ee/> (14.12.2024).

44 Muinsuskaitseamet. MuIS 2.0. <https://www.muinsuskaitseamet.ee/amet-koostoo-uudised-ja-kontakt/arendus-ja-koostooprojektid/MuIS-20#2024> (15.12.2024).

45 MuIS püsiviited ja avaandmete liides: Sündmus. <https://opendata.MuIS.ee/#opendata-event> (14.12.2024).

46 Museaali detailvaade. <https://www.MuIS.ee/museaalview/2101366> (08.01.2025).

Tartu linnavolikogu. Üldotsingust see trükkis välja ei tule, sest „Tartu linnavolikogu“ on pealkirjas käändes. Nii vajadus kasutada seotud organisatsioonide leidmiseks otsinguvormi lahtrit „Isik“ kui ka otsimootori suutmatus leida käändes märksõnu viitavad kasutajaliidese probleemidele. Andmete taaskasutamist teistes keskkondades need ei takista.

Kui esmapilgul tunduvad käesoleva artikli kontekstis olevat huvipakkuvad ainult MuIS-is kirjeldatud ajaloolised sündmused, siis tegelikult võivad infot ajalooliste isikute kohta pakkuda ka administratiivsed sündmused. Sisestades isikulahtrisse Asutava Kogu liikme Villem Ernitsa nime, kuvatakse 119 museaali kirjet, millest rohkem kui pooled on Villem Ernitsa Ingerimaalt kogutud ja Eesti Rahva Muuseumisse üle antud rahvarõivaste kohta.⁴⁷

Sarnaselt AIS-iga ei ole ka MuIS-i isiku- ja organisatsioonikirjed ühtlustatud ning ühe ja sama isiku või organisatsiooni kohta on kindlasti rohkem kui üks kirje. Praeguseks on MuIS-is kirjeldatud 270 678 isikut ja 48 176 organisatsiooni.⁴⁸ Muinsuskaitseameti pärandihalduse osakonnajuhataja Kadri Nigulase sõnul on probleemi teadvustatud ja kui seni oli isiku- või organisatsioonisamasuse tuvastamise ainus võimalus ühildada kirjeid otse andmebaasitabelites, siis arendusplaanides on võimalus seda teha töötajaliideses käsitsi. Kuna töötaja näeb korraga rohkem konteksti – näiteks detailsemaid metaandmeid, seotud kirjeid ja muutmisloogisid –, siis muudab see tegevuse oluliselt vähem veaohlikuks kui kirjete ühendamine poolautomaatselt. Kõikide isiku- ja organisatsioonikirjete ühendamine ei pruugi siiski võimalik olla, sest samasust tuvastada võimaldavaid toetavaid metaandmeid on liialt vähe.⁴⁹

KOKKUVÕTTEKS

Kui tulla tagasi sissejuhatuses püstitatud küsimuse juurde, mil määral saab MuIS-is ja AIS-is olevaid metaandmeid uurimisprojektides taaskasutada, siis vastus on, et mõlemal andmebaasil on tugev potentsiaal pakkuda tuge andmeanalüüsiks. Kuna mõlemad kasutavad avatud standardeid, peaks piisava tehnilise toe korral olema andmevahetus kolmandate osapooltega sujuv.

47 Tulemusi on võimalik küll alla laadida rdf/xml vormingus, aga otsingu püsilinki MuIS ei toeta.

48 K. Nigulas, Skype'i-vestlus, 13.01.2025.

49 K. Nigulas, Skype'i-vestlus, 09.01.2025.

Mõlema andmebaasi kitsaskohad tulenevad sellest, et nende juured ulatuvad 1990. aastatesse ja neid on arendatud eelkõige kogude haldusega seonduvaid vajadusi silmas pidades. MuIS-i andmemudel on väga paindlik ja erinevate metaandmete struktuuridega arvestav. Kuna aga näiteks isiku- ja organisatsioonikirjeid on algusest peale sisestatud isikusamasust kontrollimata, siis teeb praeguseks kogunenud kirjete suur hulk nende ühtlustamise keeruliseks. AIS-is on struktureeritud metaandmete hulk, mis ei puuduta otseselt säilikute paiknemist hierarhias, pigem väike. Rahvusarhiivi ühisloomeprojektide raames on ajalooallikates märgendatud rohkelt isiku- ja kohanimesid. Erinevate andmehulkade ühtlustamistöö seisab alles ees.

Vaadeldes AIS-i ja MuIS-i kujunemislugu ja struktuuri, on näha, et mõistetavalt on neid kujundanud valdkondade eripärad. Kui aga vaadelda andmeid, mida oleks vaja Asutava Kogu ja Tartu linnavolikogu protokollide uurimiseks arvutite toel, siis esimese sammuna oleks vaja isikute- ja organisatsioonide sõnastikku, mis sisaldaks piisavalt lisaandmeid, et samasuse tuvastamine oleks võimalik. Ideaalis oleksid sellega lingitud andmed Eesti mäluasutuste andmebaasidest, sest nagu käesolevast artiklist selgub, tekib kultuurilooliste objektide kirjelduste andmebaasidesse sisestamise käigus lisaandmeid ka isikute ja organisatsioonide kohta.

Üheks lahenduseks oleks soomlastest eeskuju võttes⁵⁰ luua eraldi linkandmete portaal, mis koondab kokku kultuuripärandi andmed, mida saaksid kõik teised andmebaasid kasutada ja ka oma andmetega täiendada. Linkandmete oluliseks eeliseks võrreldes traditsioonilise relatsioonilise andmebaasiga on võimalus lisada piiramatult huvipakkuvaid seoseid ning kombineerida omavahel erinevaid andmekogumeid. Valdkonnaülesed andmestikud aitaksid andmebaase ühtlustada ja pakuksid tuge erinevatele uurimisprojektidele. Mugav võimalus andmeid taaskasutada hoiaks kokku olulise hulga ressursse. Linkandmete portaali loomine ei pakuks lahendust üksnes andmete ühtlustamisele, vaid looks ka uued võimalused AIS-i ja MuIS-i andmete kasutamises. Tuleb aga rõhutada, et linkandmete portaali loomine on mahukas ja keerukas ettevõtmine, mis nõuab märkimisväärt tehnilist ja organisatoorset ressursi. Seetõttu on otstarbekas alustada katsetustega väiksemahulistest pilootprojektidest, mille käigus on võimalik testida andmemudeleid, töövooge ja koostööpraktikaid.

50 Datasets. Linked Data Finland. <https://www.ldf.fi/datasets.html> (15.01.2025).

Tänuavaldus

Tänan heade nõuannete eest Kadri Nigulast, Tõnis Türrnat, anonüümseid retsensente ja teemanumbri toimetajaid. Artikli valmimist on toetanud Haridus- ja Teadusministeeriumi projekt EKKD-TA10.

DATABASES OF ESTONIAN MEMORY
INSTITUTIONS: THE POTENTIAL USES
OF METADATA IN THE ANALYSIS
OF HISTORICAL SOURCES

Liisi Taimre

Estonian memory institutions – museums, archives, and libraries – maintain extensive databases containing millions of records that serve both collection management and user accessibility. These databases allow users to browse and search collections based on various parameters and, in many cases, provide access to digital representations of objects, reducing the need for physical visits. Efforts have also been made to enhance database interoperability, exemplified by the E-varamu portal, which aggregates data from 19 different databases across 12 institutions, making Estonian cultural heritage accessible through a unified search. Furthermore, standardised metadata have facilitated the integration of Estonian cultural data into broader European platforms such as Europeana and the European Archives Portal.

While these advancements primarily focus on improving user access, the reuse of metadata for research purposes presents distinct challenges. To effectively support data analysis, metadata must be highly structured and comprehensive. The Finnish computer scientist Eero Hyvönen categorises search and browsing systems into three generations: first-generation systems rely on harmonised data, second-generation systems integrate data analysis and visualisation tools, and third-generation systems incorporate artificial intelligence (AI) functionalities. The reuse of metadata for research requires robust structured data to facilitate meaningful analysis, particularly in the transition from first-generation to more advanced research-driven systems.

This study examines the potential for reusing metadata in Estonian memory institutions' databases within research projects, focusing on two historical sources housed in the Estonian National Archives: the records

of the Constituent Assembly and the Tartu City Council protocols. These sources, containing millions of words and various named entities, are well-structured and partially annotated, making them suitable for data analysis. The study explores the extent to which metadata associated with these sources, particularly information about individuals and organisations, can be leveraged for research purposes.

A key consideration is distinguishing between research data – digitised original sources – and research materials, which encompass both primary sources and their associated metadata. Without high-quality metadata, digital collections remain only partially accessible and analysable. FAIR principles (Findability, Accessibility, Interoperability, and Reusability) serve as a useful framework for evaluating metadata quality and reusability in research contexts. The study investigates how well Estonian memory institutions' databases align with these principles, with a particular focus on metadata reuse.

Two major databases – AIS (the archival information system) and MuIS (the museum information system) – are analysed in detail. AIS primarily manages archival records and provides structured metadata, although its structured metadata fields are limited mainly to document hierarchy and location. Meanwhile, MuIS offers a more flexible data model and detailed metadata structures, but the lack of standardised identifiers for individuals and organisations poses challenges for data integration. Both databases rely on legacy systems developed in the late 1990s and early 2000s, which were primarily designed for collection management rather than research-oriented data analysis.

The study highlights the importance of linked data in creating machine-readable and machine-understandable metadata. While current databases facilitate machine-readability, machine-understandability requires semantic structures that enable meaningful data relationships. Technologies such as Resource Description Framework (RDF) and ontologies like CIDOC CRM can enhance metadata interoperability and reusability. The research emphasises the need for a centralised linked data portal that integrates metadata from different cultural heritage databases, improving data uniformity and supporting diverse research projects.

In conclusion, while Estonian memory institutions' databases hold valuable metadata, their reuse for research remains hindered by inconsistencies in metadata quality, lack of standardisation, and technical constraints. Addressing these challenges through enhanced metadata structures, linked data technologies, and interdisciplinary collaboration

can significantly improve the potential for metadata-driven research. Establishing a centralized linked data repository would not only facilitate metadata harmonisation but also open new avenues for digital humanities research in Estonia.