JYX JYVÄSKYLÄN YLIOPISTO UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Kyläkoski, Kaisa

Title: Historiantutkija verkon ääressä : hakee, hakee ja soisi löytävänsä

Year: 2022

Version: Published version

Copyright: © 2022 Kaisa Kyläkoski

Rights: CC BY 4.0

Rights url: https://creativecommons.org/licenses/by/4.0/

Please cite the original version:

Kyläkoski, K. (2022). Historiantutkija verkon ääressä : hakee, hakee ja soisi löytävänsä. Ennen ja nyt, 22(4), 33-47. https://doi.org/10.37449/ennenjanyt.122103

CC BY

KATSAUS

Kaisa Kyläkoski

Historiantutkija verkon ääressä: hakee, hakee ja soisi löytävänsä

ABSTRAKTI / ABSTRACT

Jo vuosikymmenen ajan on Suomessa keskusteltu "digitaalisesta historiantutkimuksesta" ja sen rajauksesta, joka on useimmiten tiivistetty laskennallisuuteen. Huomattavasti vähemmän on mietitty digitaalisuuden merkitystä muuhun historiantutkimukseen. Tässä katsauksessa keskitytään perusasioihin historiantutkijoiden tietokantahauissa. Esimerkein osoitetaan, että on helppoa muodostaa virheellinen kuva tietokannan sisällöstä. Käy myös ilmi, ettei totuuden selvittäminen ole aina edes mahdollista. Lopuksi annetaan pari käytännön ohjetta käyttäjille ja toiveita hakuliittymien kehittäjille.

tietokanta, data, metodit

Kaisa Kyläkoski, Jyväskylän yliopisto, Kaisa.Kylakoski@iki.fi

Kun kolmekymmentä vuotta sitten mietittiin mittavan bibliografian julkaisemista, todettiin, että tietokanta on hankalampi jakaa kuin kirjasarja. Lisäksi kirjasarjan käyttö olisi helppoa eikä tietokannan tapaan vaatisi paksun käyttöoppaan lukemista tai erityistä kurssitusta.[1] Tilanne muuttui tämän jälkeen hyvin nopeasti, kun internet tuli laajemman yleisön käyttöön ja näin tietokantojen jakamisen suurimmat käytännön esteet poistuivat. Suhteellisen pian niin yleisten kuin yliopistojen kirjastojen luettelot olivat Suomessa käytettävissä kotikoneelta ja syksyllä 1997 Arkistolaitos (nyk. Kansallisarkisto) avasi verkkoon ensimmäisen version arkistotietokannastaan. Useimmat tai ainakin useat tiedonetsijät siirtyivät käyttämään näitä hakuihinsa edes vilkaisematta ohjesivua.[2]

Tämä käänne ja myöhemmät internetin mahdollistamat hakumahdollisuudet ovat muuttaneet historiantutkijoiden työtä merkittävästi,[3] sillä historiantutkimuksen perustoimi ennen analyysiin ja johtopäätöksiin pääsyä on haku. Haetaan kirjallisuutta, haetaan kontekstuaalista tietoa, haetaan aineistoa, haetaan tutkittavia tapauksia. Analyysin ja johtopäätösten teon lomassa yhden asian selvittyä syntyy uusia kysymyksiä ja taas haetaan.

Tiedonhaku ei kuitenkaan ole historiantutkimuksen metodi vaan "käytäntöä". Tällainen arkinen tekeminen jätettiin esimerkiksi tuoreessa suomalaisessa historiantutkimuksen metodioppaassa lyhyen maininnan varaan ja aloittelijaa kehotettiin haastattelemaan "kokeneempia tutkijoita".[4] Sillä näin on aina toimittu?

> *"Se menee sellasella oppipoika-kisälli-systeemillä, että joku sulle neuvoo, että näissä kohden ne on muuten sit näitä anomusdiaareja. Sit sen niin kun tajuaa ja kun sen on muutaman kerran kattonu, niin alkaahan sen sit näkeen, et mitä ne suurin piirtein on."*
> Haastateltu historiantutkija[5]

Kahdenkeskiset ja hierarkkiset keskustelut eivät kuitenkaan ole omiaan tehokkaaseen eivätkä välttämättä rehelliseenkään tiedonvälitykseen. Historiantutkija Jorma Kalela totesi kollegoistaan, että "Mitä alan edustaja *käytännössä tekee* ei ole sitä, mitä hän *sanoo tekevänsä* eikä kumpikaan näistä ole sitä, mitä *metodiesitykset edellyttävät* hänen tekevän."[6] Suomalaisen historiantutkijan tiedonhaulle perusteellisin, mutta niukka, opastus sisältyy historiantutkija Jussi T. Lappalaisen kirjaan *Haluatko historiankirjoittajaksi?* (2002).[7] Siitä, onko neuvoja seurattu tai parempia kahdenkymmenen vuoden aikana keksitty, ei ole dokumentaatiota. Historiantutkimuksen omissa esityksissä työn kuvaus kuuluu asiaan, mutta kyseessä ei ole tehdyn työn raportointi.[8] Kirjallisuuden etsinnän kuvausta myöten selostettu tekeminen olisikin työekonomisesti järjetön hanke, kun edes tarkan työpäiväkirjan pito ei ole historiantutkijoiden yleisenä tapana.[9] Kun hakutoimet eivät näy lopputuloksessa eikä niistä tohdi verbaaliakrobatiatta puhua, on mahdollista ja jopa todennäköistä, että historiantutkijan reflektointi jää tekemättä myös yksityisesti.[10] Työprosessiin ei silloin kohdistu kriittistä ja kehittävää katsetta sisältä eikä ulkoa päin. Myöskään yksilöille kertynyt tieto toimivista työtavoista ei siirry eteenpäin. Puhumattakaan tehdyistä erehdyksistä ja virheistä.

Ulkoapäin suomalaisten historiantutkijoiden hakutapoja ovat tarkastelleet ensisijaisesti hakukäyttöliittymiä ja niiden kautta löytyvää digitaalista aineistoa tarjoavat tahot, mutta myös ryhmä informaatiotutkijoita. Suuria hakuliittymiä hallinnoivat tahot, kuten Kansallisarkisto ja Kansalliskirjasto, tekevät käyttäjätyytyväisyyskyselyitä, joiden tuloksia raportoidaan myös julkisesti.[11] Kansalliskirjasto on kerännyt käyttäjäkokemuksia ja toiveita lisäksi esimerkiksi tutkimushankkeessa *Digitaalinen avoin muisti* (2019–2021).[12] Käyttöliittymiä kehittäessä käyttäjiä, mukaan lukien historiantutkijoita, on pyydetty prototyyppien testaukseen. Tällaisen havainnoinnin tarkoitus on operationaalinen, mutta joistakin prosesseista on julkaistu opinnäytetöitä, joista on löydettävissä historiantutkijoiden tarpeita, tapoja ja mieltymyksiä.[13]

Yleisemmät tutkimukset ovat olleet harvinaisempia. Varhaisimpia selvityksiä oli Outi Hupaniitun vuonna 2011 toteuttama kyselytutkimus, jossa tutkijoilta kertyi myös näkemyksiä hakujen tekemisestä.[14] Vastaava myöhempi tutkimus rajattiin Kansalliskirjastoon ja Kansallisarkistoon.[15] Arkistoaineiston hakijoina historiantutkijoita on haastateltu kahdessa pro gradu -työssä. Näihin sisältyvät kansainvälisen tutkimuksen kartoitukset paljastavat, että tutkimusta on useimmiten tehty kyselyillä tai haastatteluilla, mikä ei ole rikkain ja tarkin mahdollinen tapa kerätä tietoa.[16] Viime vuosina Tampereen yliopiston viestintätieteissä on tehty tutkimusta informaatiovuorovaikutuksesta historiantutkimuksen kontekstissa myös varjostamalla eli tarkkailemalla historiantutkijan työtä.[17]

Ulkopuolelta tehtyjä havaintoja käytetään tässä katsauksessa tukemaan ja yleistämään henkilökohtaisia kokemuksiani verkkohauista historiantutkijana.[18] Päähuomio on "googlauksen" sijaan hauissa organisaatioiden verkon kautta tarjoamiin kokonaisuuksiin, tietokantoihin ja tekstimassoihin.[19] Sekä Suomessa että ulkomailla on n. s. digitaalisen historiantutkimuksen piirissä keskusteltu verkkolähteiden vaatimasta "uudesta" lähdekritiikistä, jota on usein määritelty pitkin listoin kysymyksiä.[20] Näistä käsittelen tässä artikkelissa kahta pääteemaa, jotka ovat tuottaneet itselleni toistuvasti vaikeuksia: tietokannan rajauksen ja sisällön hahmotus. Molempien taustalla on tietokannan muodostava prosessi, joten lähden liikkeelle siitä, miten historiantutkijoiden käyttämistä aineistoista syntyy dataa. Lopuksi esitän ajatuksiani kehitystarpeista.

Historiantutkijan data

Perinteiseksi mielletty historiantutkimuksen aineiston keruu on "arkistossa istumista", jossa usein käydään läpi joku yhtenäinen tai melko yhtenäinen aineisto hakien tutkittavan asian esiintymiä: esimerkiksi mainintoja toimijasta tai toimintamuodosta. Tällaista tutkimusta voisi tehdä pitäen tukkimiehenkirjanpitoa ja arkistopäivän lopuksi todeta vaikkapa löytäneensä lasten leikkimisestä 45 mainintaa, joista 12 käsitteli hippaa ja muut kuuluivat luokkaan "muu". Käytännössä tämä olisi järjetöntä, sillä tällaisten muistiinpanojen varassa ei voisi vastata myöhemmin herääviin lisäkysymyksiin suoraan eikä niistä myöskään saisi tukea kyseisten leikkien pariin paluuseen. Työläämpi, mutta järkevämpi ratkaisu on tehdä jokaisesta esiintymästä vakiomuotoinen kirjaus listaksi tai taulukoksi.[21] Ennen tietokoneita analyysimahdollisuudet olivat suurimmillaan, kun kullakin kirjauksella oli oma korttinsa, joita voitiin myöhemmin järjestellä ja ryhmitellä. Lisäksi kortit mahdollistivat haut joko järjestyksensä perusteella tai neulakorttien tapauksessa muillakin kriteereillä.[22] Tietokoneajan koitettua kirjauksia alettiin tehdä tekstinkäsittely- tai taulukkolaskentaohjelmaan. Tämä ei tunnu kaipaavan ohjeistusta, vaan jokainen keksii omaan tutkimukseensa sopivan muodon.[23] Enemmän tukea on tarjolla kirjallisuusviitteiden ja näistä tehtyjen muistiinpanojen hallintaan, mutta näidenkin osalta moni historiantutkija päätyy omiin ratkaisuihinsa.

> *"Nyt mä meinaan jatkaa tätä, tämmösen tietokannan keräämistä, mihin mä laittelen, mä vasta hiljattain oon alottanut tän, niin sit mä laitan tänne niinkun niit mun alkuperäisaineistooni. Ja sit mä oon luonut tähän semmosen niinku pohjan, että sit mä pystyn helpommin myöhemmin löytään niit tuloksia ja käsitteleen ja vertaileen niitä."*
> Varjostettu historiantutkija[24]

Suuri osa datan käyttökelpoisuutta syntyy tietomallista eli siitä rakenteesta, johon poimitaan alkuperäisaineiston tiedosta tulkinta. Yhdysvaltalainen historian ja digitaalisen humanismin tutkija Sharon Leon arvelee, että tietorakenteen suunnittelun taitoja ei historiantutkijoille opeteta eikä minulla ole syytä

uskoa, että tilanne olisi kaikissa suomalaisissakaan yliopistoissa kunnossa. Jos tietorakenne syntyy sattumalta työn ohessa ja lähteen ohjaamana, se ei välttämättä ole paras mahdollinen. Jopa alkuesimerkin leikkitutkijan tukkimiehen kirjanpito on tietomalli. Kun tutkijalle sitten pälkähtää päähän vasta useamman arkistopäivän jälkeen, että leikin paikka ja aika sekä leikkijöiden määrä ja sukupuoli olisivat hyödyllisiä tietoja, hänen pitää istua arkistossa päiviä, joiden käytön olisi voinut välttää paremmalla suunnittelulla. Mutta ei ole olemassa mitään oikeaa tai luonnollista rakennetta ja tietomalli – vaikka se olisi asiakirjan puhtaaksikirjoitus sanasta toiseen – on uusi esitys menneisyydestä.[25] Aineistonhallinnan koulutusten opit datan dokumentoinnista on otettava tosissaan. Jotta historiantutkijan luomaa dataa voidaan käyttää uudelleen ja itsenäisesti, tarvitaan tieto tutkijan tekemistä valinnoista ja päätöksistä.[26]

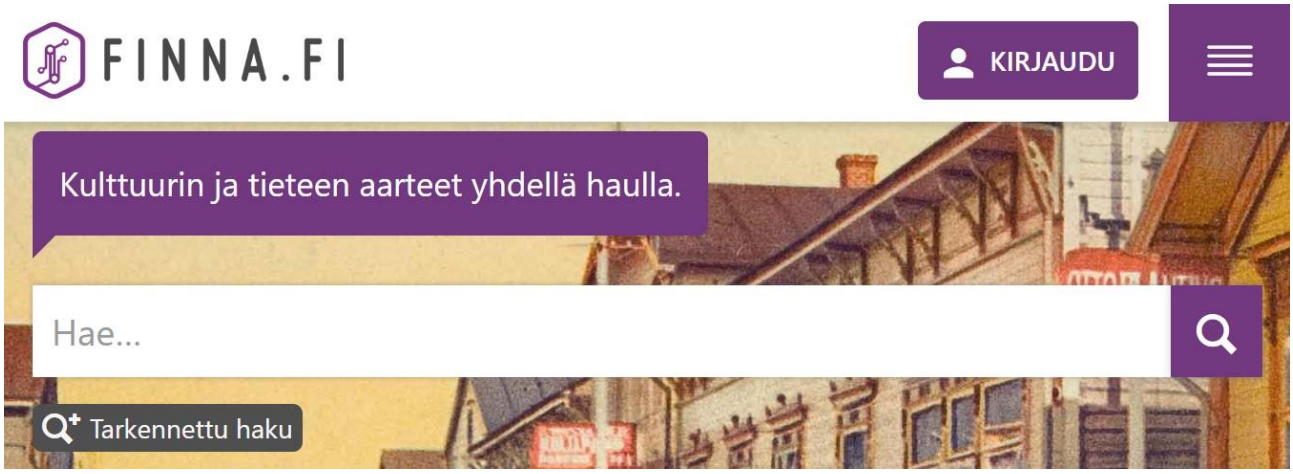
Leon epäilee, että kaikki historiantutkijat eivät hahmota muistiinpanojaan tai niistä luokitteluilla ja tiedon yhdistelyllä luotuja tutkimuksen välituloksia dataksi, joka olisi jaettavissa muille.[27] Suomalaisessa yliopistomaailmassa tietoisuus mahdollisesta jakamisesta on syntynyt viime vuosina aineistonhallinnan kontekstissa. Jos on osunut sopivaan tiedotustilaisuuteen tai koulutukseen, on saanut ohjeita suunnitelmallisuudesta ja dokumentoinnista sekä yliopiston tarjoamista säilytyspalveluista. Esimerkit jaetuista aineistoista ovat kuitenkin vähäisiä ja vastaavasti todennäköisesti vain murto-osa menneisyyden listoista ja kortistoista on päätynyt arkistoihin jatkokäyttöön.

Käynnissä olevan väitöskirjatutkimukseni puitteissa olen omakohtaisesti tunnistanut kaksi syytä jättää taulukkoni kovalevylleni. Ensinnäkin minulta puuttuu akateemista altruismia jakaa satojen tuntien työn tulos tuottamaan muille hyötyä minuuteissa. Toiseksi tuntuu siltä, että taulukoiden pitäisi olla täysin koneluettavia eli niistä pitäisi siivota pois kaikki epäjohdonmukaisuudet ja ylimääräiset merkinnät, jotta taulukkoni näyttäisivät ”oikeammalta” datalta.

Tästä päästään olennaiseen lähtökohtaan: kaikki data on ihmistyön tulosta. Jos data kuvaa reaalimaailmaa, se on syntynyt joko “käsin” tai ihmisen suunnittelemin instrumentein ja algoritmein. Data ei laskeudu virheettömänä ja kaikkiin käyttötarpeisiin sopivana mihinkään tietokantaan. Data on luotu tietylle yleisölle ja tiettyyn tarkoitukseen. Sitä pitäisi tulkita samalla lähdekriittisellä otteella kuin humanistisessa tutkimuksessa tekstejä eli huomioiden tekijöiden tarkoitusperät. Kyseessä ei ole neutraali tai objektiivinen tietojoukko.[28] Kuitenkin dataa helposti lähestytään kuten 1800-luvun historiantutkijat virallisia asiakirjoja: totuuden kertovana lähteenä.[29]

Historiantutkijoiden data on tulkintaa alkuperäislähteistä, joihin kohdistettu lähdekritiikki on historiantutkimuksen keskiötä. Aivan liian helposti taulukoitu tieto alkaa kuitenkin muistuttaa “oikeampaa” tai “luonnollista” dataa. Datan lähtökohta hämärtyy erityisesti, kun sitä käsitellään ja esitetään eli normalisoidaan, luokitellaan, lasketaan tai visualisoidaan. Puutteet ja tulkinnanvaraisuudet jäävät näkymättömiin ja syntyy riski, että tehdään johtopäätöksiä, joita alkuperäiset lähteet eivät tue. Tämä riski kasvaa huomattavasti, kun datan käyttäjä ei tunne alkuperäisaineistoa, historiallista kontekstia tai historiantutkimuksen perusteita.[30] Kansainvälisiä esimerkkejä paremmin on jäänyt itselleni mieleen Suomen sukututkimusseuran *HisKi-tietokantaa* louhinut, jolle tavallista suurempi kuolleisuus 1860-luvun lopulla oli merkittävä havainto.

Niin kauan kuin historiantutkija käsittelee omaa dataansa, hänellä on ymmärrys siitä, että se edustaa vain tiettyä aineistoa, jonka sisällöstä on poimittu tulkiten ja tietomallin vaatimin yksinkertaistuksin osa. Vaikka joku valinta ja sen dokumentaatio voi työn touhussa unohtua, edellä käytetyn esimerkin leikkien tutkija tietää, että hippa-tietueidensa joukossa on sekä tapauksia, joissa sanottiin leikityn hippaa, että tapauksia, joissa kuvaus tutkijan mielestä riitti leikin tunnistamiseen. Nämä datan muodostuksen realiteetit kuitenkin aivan liian helposti unohtuvat, kun vieras data on siistissä käyttöliittymässä, joka lupaa “Kulttuurin ja tieteen aarteet yhdellä haulla”.[31]



Kuva 1. Kuvakaappaus Finna.fi-hakuportaalin etusivulta. Kuvalähde: https://finna.fi/ 30.9.2022.

Tietokannan rajauksen hahmottaminen

> *"Kaikista tehokkain totta kai on kääntyä arkiston henkilökunnan puoleen. Sen on oppinut tässä vaiheessa, että pelkkiin luetteloihin ei kannata luottaa, sieltä ei välttämättä löydy ihan kaikkea. Ainakin ne arkistot, joita olen itse käyttänyt, esimerkiksi Työväenarkisto, niin siellä on henkilökunta, joka tietää ihan tasan tarkkaan, mitä siellä arkistossa on. Siellä on aika paljon sellaistakin, mitä ei ole vielä luetteloitu."* Haastateltu historiantutkija[32]

Oma toistuva ajatusvirheeni verkon tietokantojen äärellä on olettaa, että tietokannassa on "kaikki". Ei se "kaikki", jonka tekijät ovat määritelleet tietokannan sisällöksi, vaan se "kaikki", joka minusta pitäisi olla mukana. Ainakin Suomen sukututkimusseuran *HisKi-tietokannan* osalta tiedän muidenkin sortuneen vastaavaan, joten se sopii aloitukseksi.

Tutkijan oman aineistoläpikäynnin korvaavista tietokannoista Suomen vanhin ja yksi käytetyimmistä on Suomen sukututkimusseuran *HisKi*.[33] Sen taustalla on Suomen sukututkimusseuran 1900-luvun alkupuolella järjestämä seurakuntien historiakirjojen kopiointi, jossa jo alun perin suhteellisen määrämuotoisesta kirjauksesta sukututkijoille keskeisimmät tietoelementit jäljennettiin taulukkomuotoon. Hankkeen ensisijaisena tavoitteena ei ollut tiedon löydettävyyden parannus vaan tiedon saavutettavuus, sillä kirkonkirjat olivat tuolloin edelleen omissa seurakunnissaan.

Niin kauan kuin historiakirjojen kopiot olivat sarakkeellisia taulukoita paperilla tai näiden mikrofilmijäljenteitä, käyttäjillä ei ollut mitään vaikeuksia ymmärtää sitä, mitä vuosia ja seurakuntia oli mukana. Tilanne muuttui merkittävästi, kun 1990-luvulla kirjauksia alettiin viedä digitaaliseen muotoon ja niille rakennettiin käyttöliittymä. Kun siinä tekee nopeasti ja helposti haun "kaikkiin" seurakuntiin, jää ymmärtämättä, että kaikista seurakunnista tietoja ei ole ollenkaan ja jokaisen seurakunnan tallennustilanne eri historiakirjojen ja näiden rajavuosien suhteen on erilainen. Sadat sukututkijanoviisit ovat tuloksettoman haun jälkeen kuvitelleet, että tarvitsemaansa tietoa ei ole "missään". Kokeneemmat käyttäjät ja toivottavasti myös koulutetut historiantutkijat puolestaan tietävät, että tietokannassa voi tarkistaa kunkin seurakunnan tallennustilanteen. Tallennettuihin tietoihin voi myös tehdä eksploratiivisesti eli sisältöä kartoittavasti pelkän aikahaun, jolloin tuloksena on alkuperäistä jäljennettä vastaava lista, josta voi arvioida pap-

pien kirjausten laatua ja kattavuutta. Lisäksi seurakuntakohtaisten lisätietojen vanhat luettelot mikrofilmeistä ja kerätyt tiedot kirkonkirjojen tuhoutumisesta mahdollistavat nopean tarkastuksen siitä, onko tietoa "jossain".

Puutteita ei ole kuitenkaan yksinkertaista tuoda näkyviin yleishakuun. Tämä on ongelma esimerkiksi Kungliga Biblioteketin *Svenska Dagstidningar* -kokoelmassa.[34] Sivusto ei lupaa mitään erityistä kattavuutta, mutta ei tee myöskään helpoksi ymmärtää, mitä lehtiä on jo digitoitu. Tosiasiallisesti prioriteettina on Kungliga Biblioteketin omien kokoelmien digitointi, mikä tarkoittaa, että varsinkin varhaisista nimikkeistä voi puuttua kokonaisia vuosikertoja tai numeroita sieltä täältä, vaikka näitä olisi olemassa toisaalla. Tällaiset puutteet hahmottaa vain käyden kyseistä lehteä järjestelmällisesti läpi. Digitointi on edelleen käynnissä, joten täydellisyyteen pyrkivän pitäisi tehdä tarkistus sisällöstä ainakin nimikkeiden ja vuosikertojen tasolla jokaisella käyttökerralla. Tämä on tietenkin epärealistista.

Vastaava oman kokoelman digitointi on lähtökohtana myös Kansalliskirjaston lehtidigitoinneissa. Kansalliskirjasto luo mielikuvan täysin kattavasta tarjonnasta sanoillaan "on digitoinut kaikki Suomessa vuosina 1771–1929 ilmestyneet sanomalehdet".[35] Todellisuudessa Kansalliskirjasto on digitoinut sanomalehdiksi luokittelemansa periodiset julkaisut siinä laajuudessa kuin ne ovat kokoelmissaan. Kattavuus on lähes täydellinen, mutta digitoidusta materiaalista puuttuu numeroita, sivuja ja liitteitä, minkä käyttäjä huomaa vasta etsiessään juuri niitä. Tällaisten puutteiden merkitys on useimmiten merkityksetön, mutta erityisesti nollatuloksia tulkitessa se voi olla syytä muistaa. Lisäksi liitteiden puuttuminen tarkoittaa, että tutkija ei voi tavoittaa menneisyyden lukukokemuksen raamejakaan.

Oman kokoelman perusteella tehtiin Kansalliskirjastossa myös Turun akatemian väitöskirjojen digitointi. Vuonna 2010 tiedotettiin, että "Kansalliskirjasto digitoi Turun akatemian väitöskirjat" ja vain tiedotteen tekstiin tutustumalla ymmärsi, että digitointi kattoi ainoastaan sen osan opinnäytteistä, jotka kuuluvat alkuperäisinä Kansalliskirjaston kokoelmiin.[36] Kun vuoden 2019 lopulla olisin halunnut ymmärtää kattavuuden paremmin eli saada listan digitoimattomista töistä, kysyin neuvoa kirjastolta. Tiedusteluuni vastattiin, että voisin saada "vihjeen tilanteesta" tekemällä Kansalliskirjaston Finnassa haun 'turun akatemia väitöskirja puuttuu kansalliskokoelmasta'. Hakutulos kuitenkin vastaajan mukaan tarvitsi tulkintaa, sillä ainakin yksi työ oli digitoitu Kansalliskirjaston ulkopuolelta ja väitöskirjoista on variantteja.[37] Totesin, että tutkimuskysymykseni puitteissa työtä ei kannattanut tehdä.

Kansalliskirjasto on tarkempi sanoissaan esitellessään verkkosivuillaan (nykyään Finnan kautta tavoitettavaa) Suomen kansallisbibliografia *Fennicaa*. Kiistatta *Fennica* "sisältää tietoja kirjoista".[38] Mutta jos lukija muodostaa käsityksen, että *Fennicassa* on haettavissa ja näkyvissä kaikki Suomessa julkaistu kirjallisuus, hän erehtyy. Alaotsikon Taustatieto alta on tarkasti lukien hahmotettavissa se, minkä neuvonta kertoi minulle suorasanaisemmin: "1800–1900-luvun *Fennican* luettelointitiedot on konvertoitu painetuista kansallisbibliografioista: Suomalainen kirjallisuus/Suomen kirjallisuus sekä Katalog över den svenska litteraturen i Finland. Mitä niissä ei ole, puuttuu periaatteessa tietokannasta (tosin 2000-luvulla on tehty täydennysluettelointia). Eikä tämä koske pelkästään ruotsinkielistä 1800-luvun aineistoa, olkoonkin, että ruotsinkielisen kirjallisuuden Katalogin alkuvuosi on 1886, joten 1810–1885 ruotsinkielisen aineiston vajaus on huomattava."[39] Puutteita *Fennicaan* syntyy jatkuvasti, sillä sisältö perustuu julkaisijoiden ilmoituksiin, joita jää tekemättä.

Suomen kielen mahdollistamaa sijamuotohäivytystä käytetään myös Kansallisarkiston *Diplomatarium Fennicum* -tietokannan etusivun esittelyssä: "verkkopalvelu keskiaikaisista Suomea koskevista asiakirjoista." Jos jaksaa jatkaa lukemista seuraavaan virkkeeseen saa ilmaisusta "kuuluu tällä hetkellä" vihjeen siitä, että tietokannassa ei ole kaikkia mahdollisia asiakirjoja, vaikka toisinaan näin kuulee sanottavan.[40] Itse, keskiajan tutkimusta osaamattomana, ymmärsin lisäyspotentiaalin määrän kuitenkin vasta Eljas Orrmanin katsauksesta.[41]

On ymmärrettävää, että tietokantojen tai digitointiprojektien julkaisujen tiedotuksessa ja etusivuilla käytetään ilmaisuja, jotka häivyttävät ikäviä puutteita. Koska verkkosivuilla tila ei ole rajoitettua, ei kuitenkaan ole hyväksyttävää jättää kertomatta totuutta mahdollisimman yksityiskohtaisesti jossain. Jos tietokantaan on luetteloitu jotain, on tiedettävä, mikä ryhmä on kyseessä. Jos tietokantaan on viety poimintoja tietystä arkistoaineistosta, käyttäjän pitää ymmärtää, millä kattavuudella tämä on tehty. On myös merkityksellistä tietää, onko tietokantaan viety suoraan aineistosta tehtyjä havaintoja vai aiemmin toisessa muodossa julkaistu kokonaisuus.

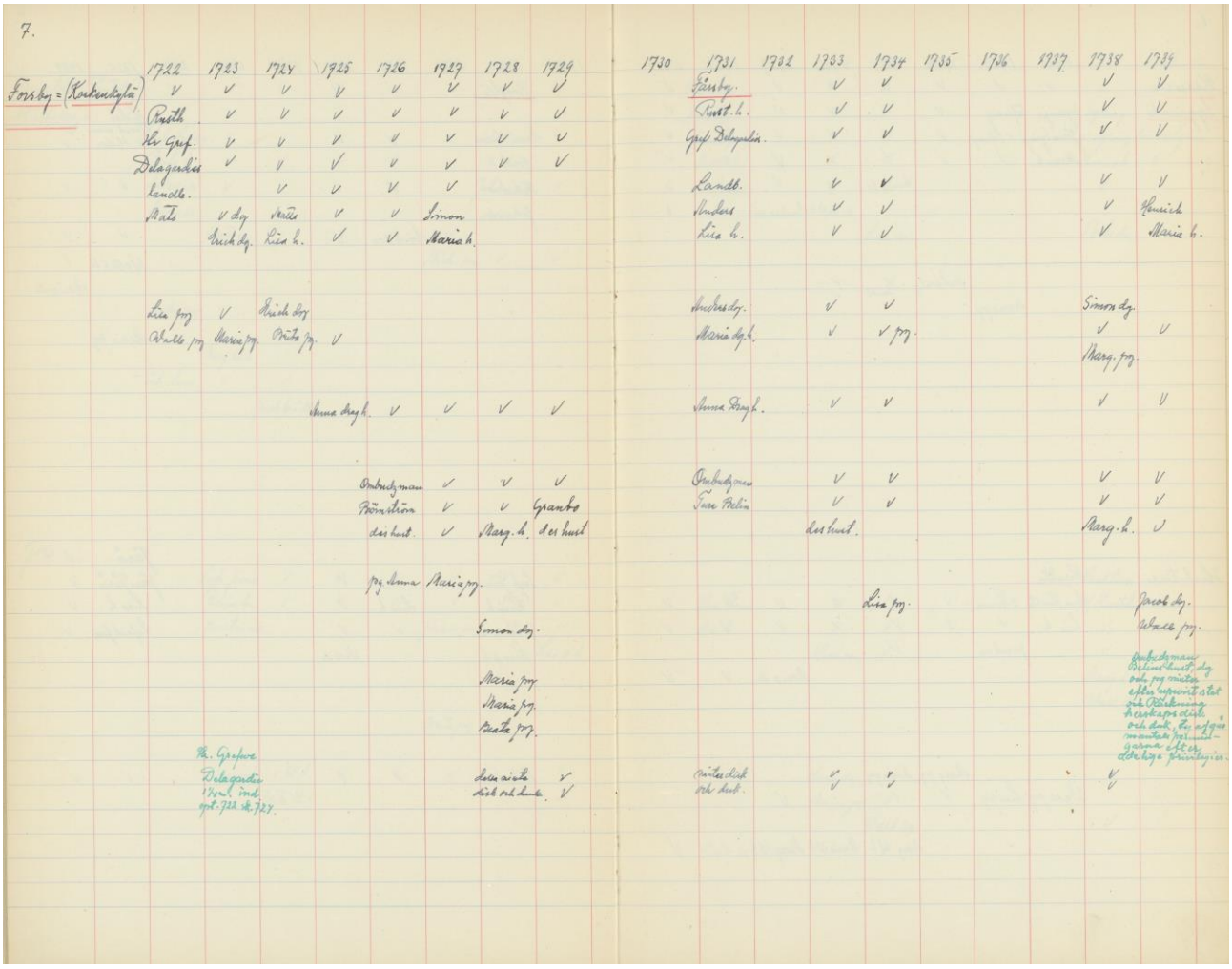
Tietokantojen hallinnoijat eivät tietenkään ole vastuussa kaikesta tutkijoiden kohtaamasta viestinnästä. Kuten edellä totesin, tyypillisesti tieto kulkee tutkijalta toiselle. Se, kuinka yksityiskohtaista, kattavaa ja oikeaa välitetty kuvaus on, riippuu kertojan omista tiedoista ja tiedonvälitystilanteen luonteesta. Erityisesti, jos kertoja on auktoriteetti ja kuulija aloittelija, kuvaus voidaan ottaa totuutena, jota ei ymmärretä tarkistaa.

Valitettavan usein tietokannan rajauksen tarkistus on mahdotonta. Tekijöille, aivan kuten yksittäisille historiantutkijoille taulukoidensa kanssa, tietokannan laajuus on ehkä itsestäänselvyys. Tai sitten sen alkuperäinen kuvaus on irronnut tietokannasta vuosien varrella. Omissa tutkimuksissani ikävin vastaan tullut esimerkki on ruotsalainen Linköpingin yliopiston hallussa oleva tietokanta *Medicinhistoriska databasen*, johon on puhtaaksikirjoitettu erilaisia lääketieteen historiaan liittyviä asiakirjoja.[42] Osalle asiakirjatyypeistä on annettu hakuliittymässä rajavuodet, mutta kaikille ei. Haku vaatii jonkun sanan eli rajavuosia ja tietueiden määrää ei voi selvittää tyhjällä haulla. En ole löytänyt käyttöliittymästä mitään selostusta siitä, oliko mukana kaikki säilyneet kappaleet kustakin asiakirjatyypistä vai esimerkiksi maantieteellinen otanta. Vaikka olisin saanut tyhjällä haulla asiakirjalistauksen, en olisi pystynyt ilman arkistotietokantojen kaivelua arvioimaan mahdollisten puutteiden tai vinoumien syytä eli sitä oliko kyse säilyvyydestä vai tietokannan aineistopoiminnasta. Sanahakujeni tulokset olivat näin ollen lähes käyttökelvottomia. Ne olivat esimerkkejä tutkimastani asiasta, mutta en voinut sanoa mitään ilmentymien yleisyydestä, mikä olisi ollut tutkimukselleni merkityksellistä.

Tietokannan sisällön hahmottaminen

Hakutulosten täysimittainen käyttö vaatii tietokannan rajojen tunnistamisen lisäksi tietoa sisällön rakenteesta ja muodostuksesta. Erityisesti nimekkään muistiorganisaation palvelussa tutkija kuvittelee aivan liian helposti, että tuotantoprosessit ovat "täydellisiä". Eli että oikeat asiat on yhdistetty keskenään, normalisoinnit tehty järjestelmällisesti ja asiasanoitus käyttökelpoinen. Koska koko tietokannan laatua ei pysty yksinkertaisesti tarkastamaan, mielikuva laadusta muodostuu usein ensimmäisistä havainnoista. Jos nämä ovat positiivisia, ongelmat voivat myöhemmin tulla esiin yllätyksinä.

Tietokoneaikaa edeltävä esimerkki listaa monimutkaisemmasta rakenteisesta tiedosta on Jalmari Finnen 1920-luvulla aloittama *Suomen asutuksen yleisluettelo* (SAY), jonka tarkoituksena oli tuoda yhteen verotustiedot kaikista Suomen maatiloista vuosilta 1539–1809. Yleisluettelon tekemiseksi kopioitiin maakirjoista, kymmenysluetteloista ja henkikirjoista tilojen veroluvut, henkilönimet ja toisinaan muita huomautuksia. Näitä ei jätetty yksinkertaisiksi luetteloiksi, vaan merkintöjä tulkittiin ja yhdistettiin kooten yhden tilan tiedot yhdelle aukeamalle. Ruutupaperille tehty rakenteistus on yhtä armoton kuin tietokanta: tiedot on asettava jonnekin. Erityisesti varhaisempien vuosien osalta tulkinnoissa on epävarmuutta, jota kaikki käyttäjät eivät välttämättä hahmota, vaikka data on kokonaisuudessaan näkyvissä. Jos tiedot olisivat tietokannassa, josta saisi eteensä vain yhden tilan valmiin isäntäluettelon, laadun ymmärtäminen olisi lähes mahdotonta.



Kuva 2. Esimerkki yleisluettelon aukeamasta (Kokemäen Forsby 1722–39). Kuvalähde: Kansallisarkisto. Suomen asutuksen yleisluettelo -arkisto. Kokemäen asutuksen yleisluettelo. Kokemäki (1719–1739), 7, https://astia.narc.fi/uusiastia/getFile.php?fileId=62298019.

SAY:n tekemisen ajan teknologiaa oli paperin ohella eriväriset kynät, joiden jäljestä on erotettavissa, mistä asiakirjatyypistä tiedot ovat peräisin. Yleisluetteloa on siis mahdollista käyttää lähdekriittisesti ja paluu alkuperäislähteeseen onnistuu osaajalta. Lähdekriittisyys on välttämätöntä, sillä työtä tehtiin vuosikymmeniä, linjauksia muutettiin ja tekijöitä oli monia.[43]

Laatuongelmia ja pitkän ajan epätasaisuutta on monissa tietokannoissakin, mutta niiden hakuliittymät tai käyttöohjeet eivät tätä helposti paljasta. Yksi esimerkki tästä on Suomalaisen kirjallisuuden seuran hallinnoima *Henrik-tietokanta*, johon on viety perukirjojen ja huutokauppakamarien mainintoja kirjaomistuksista Suomen kaupungeissa ennen vuotta 1810.[44] Olennaista tietosisältöä ovat kirjan omistajien, myyjien ja ostajien nimet, ajankohdat, paikat sekä kirjojen tiedot. Näitä ei ole jätetty aivan sikseen, sillä ohjesivulla sanotaan henkilöiden osalta, että "Nimiä on yhdenmukaistettu, kun todennäköisyys identtisyydestä on ollut suuri".[45]

Mitään puhetta ei ole kirjojen nimikkeiden tulkinnasta, mutta kun tietokantaan tekee tuloksia tuottavan haun käy ilmi, että tietueissa on kenttä "Lähdeteksti", johon on kopioitu alkuperäinen merkintä, sekä kentät "Tekijä" ja "Nimeke", jotka näyttävät siltä, että tallentaja on voinut esittää normalisoidun tulkintansa. Jos näin olisi tehty kattavasti ja samoja kirjoitusmuotoja käyttäen, tietokannasta saisi suoraan tietyn nimekkeen tai kirjailijan kaikkien tuotosten omistustiedot. Mutta tietueiden läpikäynti esimerkiksi vuosittain tai yleisellä sanalla kuten 'kung' paljastaa, että normalisointi on epätasaista ja toisinaan tulkinta

on sijoitettu hakasulkuihin osaksi nimekettä. Normalisoidun arvon kirjoittamista parempi ratkaisu olisi ollut luoda nimikkeistä oma taulunsa, johon linkitettäisiin kustakin asiakirjamaininnasta. Näin nimikkeisiin olisi voitu liittää helpommin ja mahdollisesti johdonmukaisemmin asiasanoitus sekä linkittää julkaisut kirjastotietokantoihin ja digitoituihin kappaleisiin.

Jos tuntee *Henrik-tietokannan* historiaa tai kurkistaa Tekijät-sivun tallentajien mittavaa nimilistaa, on helppo ymmärtää, että epäjohdonmukaisuus on peräisin pitkään tehdystä työstä. Satunnaiselta käyttäjältä tätä ei voi kuitenkaan odottaa ja kokemuksesta on helppo kuvitella, että yksittäisellä haulla esiin tullut hakutulos vakuuttaa ja tyydyttää, vaikka se todellisuudessa on todennäköisesti epätäydellinen. On aivan liian helppo ohittaa ohjeet-sivun loppukaneetti "HENRIK-tietokannassa saattaa esiintyä virheitä monista eri työvaiheista johtuen. Tämä on huomioitava, kun tietokantaa käytetään tieteelliseen tarkoitukseen. Virhemahdollisuus on otettava huomioon esimerkiksi koskien kirjojen lukumäärää tai kieltä."

> *"Jää semmoinen epävarmuustekijä, että olenko nyt osannu hakea oikeilla sanoilla ja olenko mä löytänyt edes tästä kyseisestä aineistosta kaiken sen, mikä olis relevanttia…"*
> Haastateltu historiantutkija[46]

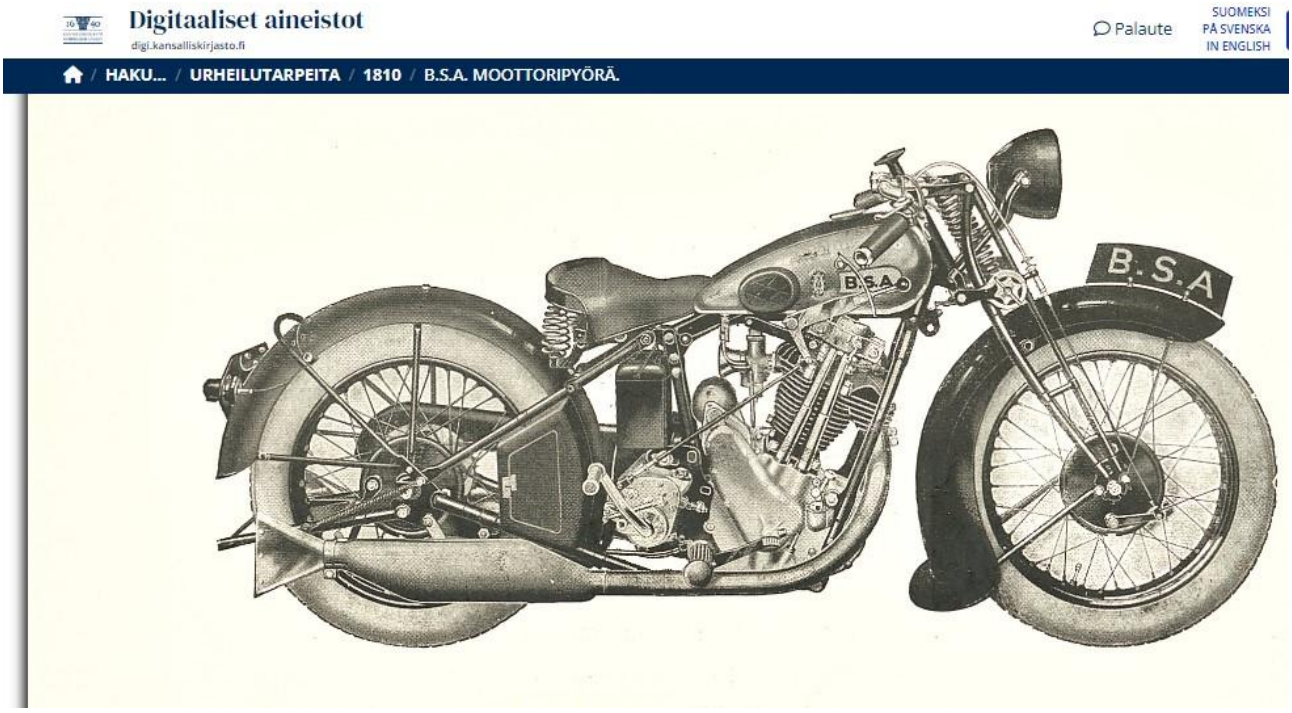
Ottaen huomioon, että *Henrik-tietokanta* on tarkoitettu kirjahistorian tutkimukseen, on hieman outoa, että suurin vaiva on nähty kirjojen haltijoiden identiteettien yhdistämiseksi ja normalisoimiseksi. On mahdollista, että taustalla on toinen aineistoryhmistä eli perukirjat, jotka koskevat tiettyä henkilöä. Henkilökeskeinen rakenne on säilynyt tietokannan tekoon käytetyssä Henrik Grönroosin aineistossa, jota on jo ennen tietokantaa julkaistu henkilöittäin ryhmiteltynä.[47] Mutta lähtökohtana voi myös olla historiantutkijoiden yleinen tapa hakea erityisesti henkilö- ja paikannimiä.[48]

Valitettavasti henkilö- ja paikannimien tunnistaminen ja käsittely hakukelpoisiksi ei ole kovin yksinkertaista. Käsin poiminta on hidasta ja tuottaa oman epävarmuutensa. Lisäksi siinä (eli datan muodostuksessa historiallisesta aineistosta) on noudatettava säännönmukaisuuksia, jotka (jälleen) eivät ole välttämättä käyttäjälle itsestäänselvyyksiä. Kirjoittaessani pientä tekstiä Topeliuksen tositapausta kuvaavasta runosta *Grafven i Karttula* mieleeni tuli tarkistaa, miten siinä esiintyvät henkilöt on esitetty Topeliuksen tuotannon digitaalisessa editiossa.[49] Hämmästyksekseni henkilönimiä ei ollut runoon merkitty ollenkaan. Tiedusteluuni vastattiin asiallisesti: "När det är fråga om fria, poetiska fiktionaliseringar, som i nämnda dikt, följer vi principen att inte personkoda, trots att vi kan identifiera de historiska personerna." Täysin ymmärrettävää, mutta ei ilmeistä. Verkkosivustolta löytyy metodeista ja tekstikoodauksesta dokumentaatio, jonka yhteydessä todetaan, että sitä ei päivitetty vuoden 2017 jälkeen, joten osa sisällöstä ei vastaa nykyistä julkaisua.[50]

Ohjelmallisesti tehtynä nimien tunnistamisen (Named entity recognition (NER)) tulos on nykytekniikalla väistämättä epätäydellinen, sillä teksteissä on esimerkiksi sanoja, jotka ovat sekä yleisnimiä että erisnimiä. Suomalaisten sukunimien erottelu paikannimistä vaatisi hyvin raskasta analytiikkaa. Täydellisessä maailmassa paikannimet myös eroteltaisiin kaimoistaan, paikannettaisiin ja niille olisi erilaisia historiallisia ontologioita, joiden tuella voisi hakea kaikkia mainintoja esimerkiksi menneisyyden hallintoalueelta.[51] Tämä on edelleen unelmaa. (Omat ja isommat haasteet liittyvät moniulotteisempien asioiden hakuun digitaalisesta aineistosta. [52])

Diplomatarium Fennicum -tietokannan kehityksessä yritettiin sekä ihmistyönä että ohjelmallisesti nimientiteettien tunnistusta, mutta työ jäi kesken. Kehittäjien päätös oli, että luotu aineisto kuitenkin julkaistiin hakujen tukemiseksi.[53] Riskinä (jälleen) on, että käyttäjä kuvittelee dataa täydelliseksi ja tulkitsee tuloslistauksensa virheettömäksi.

En kuitenkaan usko, että tämä on historiantutkijoiden osalta todennäköistä.[54] Ainakin itse pyrin tekemään erilaisia hakuja ja pitämään rajoitukset minimissään, eli mieluummin kahlaan läpi tarpeetonta, kuin otan riskin näkemättömyydestä. Toisinaan työekonomisista syistä on kuitenkin käytettävä esimerkiksi monien kirjasto- ja arkistotietokantojen tarjoamia "fasetteja" eli hakutuloksen rajaamista vuosilla, asiasanoilla, aineistotyypillä jne. Nämä ovat hyvin houkuttelevia, mutta on jaksettava pitää mielessä, että ne tarttuvat vain sellaisiin tietueisiin, joissa kyseinen asia on määritelty ja määritelty oikein. Olen parissa arkistotietokannan käyttäjätestauksessa hermostunut nähdessäni valikossa esimerkiksi mahdollisuuden rajata tulokset "karttoihin". Totta kai tämä olisi käytännöllistä ja ihanaa, mutta jos arkisto ei ole tunnistanut jokaista karttaansa kartaksi, hakurajauksen tarjoaminen on omasta mielestäni vastuutonta.



Kuva 3. Kansalliskirjasto on digitaalisten aineistojensa esityksessä valinnut puuttuvan painovuoden korvikkeeksi luvun 1810. Niinpä käyttöliittymässä tämäkin moottoripyöräesite näyttää olevan yli 200 vuotta vanha. Kuvalähde: Oy. Urheilutarpeita. 1810. https://digi.kansalliskirjasto.fi/pienpainate/binding/2145323.

Paradoksisesti fasettien ongelmat on helpompi kestää ympäristössä, jossa ne ovat vielä suurempia. "Kulttuurin ja tieteen aarteet yhdellä haulla" lupaavaa Finnaa ei pysty käyttämään eksploratiivisiin hakuihin hyödyntämättä metadataa ja on vain hyväksyttävä se, että "yhdellä haulla" ei tavoita kaikkea sitä, mitä ehdoillaan pyrkii hakemaan. Kuvailudata voidaan tuoda yhteen paikkaan haettavaksi, mutta kukin organisaatio on luonut sitä omilla säännöillään, jotka eivät ole pysyneet samoina vuosien läpi.[55] Luettelointiin on ollut käytettävissä erilaisia määriä ja erilaisin kompetenssein varustettuja ihmisiä. Dataa on tuotettu erilaisissa konteksteissa: päivittäisen työn ohessa tai isoissa projekteissa. On siis täysin ymmärrettävää, että Aalto-yliopiston arkiston tuhansissa arkkitehtuurin opetuskuvissa on ainoastaan asiasanat 'opetusaineisto' ja 'arkkitehtuuri' ja että resurssien säästämiseksi Kansalliskirjasto lopetti vuonna 2016 ulkomaisen tutkimuskirjallisuuden asiasanoituksen. Ei ole vaikea keksiä erilaisia syitä sille, miksi museoiden historiallisten valokuvien ja esineiden ajoitustiedot eivät välttämättä ole hyödyllisessä muodossa ja määriteltyjä järjellisellä tarkkuudella. Tai sille, että arkistotietokanta ei anna tuloksia ensimmäisellä mieleen tulevalla sanalla.

"Jos mä haen henkikirjaa, niin se antaa mulle 1800-luvun henkikirjat sillä hakusanalla. Se ei anna niitä läänintilien tositekirjoissa olevia henkikirjoja, koska ne on tositekirjoja. Ne henkikirjat niin kun tavallaan on siellä tosite-kirjoissa 1600- ja 1700-luvullakin, niin ne ei tuu esiin sillä henkikirja-sanalla silloin." Haastateltu historiantutkija[56]

Puutteiden merkitys on toisinaan pieni ja toisinaan hyvin merkittävä. Väitöskirjaprojektini alkuhetkillä kesällä 2019 minua neuvottiin aloittamaan Suomen ja Ruotsin historiallisista bibliografioista, joiden olemassaolon olin rehellisesti sanoen unohtanut kirjastotietokantojen vuosikymmeninä. Näistä *Svensk historisk bibliografi* oli vieraampi.[57] Koska olin tämän osalta saanut ohjeen "kaikki on nyt tietokannassa" ja digitaalisen version etusivulla lukee "SHBd är den mest kompletta bibliografin över svensk historia. [...] Posterna har hämtats från de tryckta SHB-volymerna (1771–1976) och från LIBRIS (1977–2010)", pienen hapuilun jälkeen kävin läpi kaikki relevanteilta näyttävät asiasanat. Läpikäytävän kirjallisuuden luetteloon kertyi rivejä, joten luulin kaiken olevan kunnossa. Näin jälkikäteen en osaa sanoa, olisiko julkaisuvuosien jakauma verrattuna tietokannan nimelliseen kattavuuteen voinut hälyttää ilmeisestä ongelmasta, joka selvisi minulle vasta myöhemmin. Tietokantaan oli nimittäin kyllä viety tietueet vanhoista painetuista bibliografioista, mutta niissä ei ollut muuta metatietoa kuin linkki julkaisuun. Asiasanahakuni eivät siis tarttuneet näihin ollenkaan. Jos joku lisäarvoa tuottava hakutapa on olemassa, en edelleenkään sitä tunnista.

Mitä laadukkaammalta sivusto näyttää ja mitä helpompi haku on, sitä todennäköisemmin erehtyy ajattelemaan, että sisältö on "täydellinen". Tämä on ainakin meriselitykseni omalle erehdykselleni ruotsalaisen *Litteraturbankenin* kanssa.[58] *Litteraturbanken* on komea käyttöliittymä kuratoituun kokoelmaan digitoitua ruotsalaista kirjallisuutta, jota voi etsiä kuvailutietojen perusteella, mutta siihen voi tehdä myös sanahakuja. Samaan aikaan kun aloin väitöskirjaani varten 1700-luvun tutkimisen, sivustolle vietiin saman vuosisadan digitoitua kirjallisuutta. Luonnollisestikin innostuin hakemaan mahdollisia aikalaismainintoja tutkimusaiheestani, mutta sain pettymyksekseni nollatuloksen. Vedin tästä johtopäätöksen, että aihe ei esiintynyt aineistossa. Virheeni paljastui vasta, kun avasin yksittäisen 1700-luvun painatteen ja huomasin, että optista tekstintunnistusta ei ollut käytetty fraktuuratekstiin tai se ei ollut siihen ollenkaan tehonnut. Käytännössä kaikki tai lähes kaikki 1700-luvun tekstit olivat siis sanahaun ulottumattomissa. Sivustolta ei löydy vastaavaa varoitusta kuin Kansalliskirjaston hakuohjeista: "Haun kattavuus tai tarkkuus eivät ole täydellisiä sillä sivujen tekstisisältö on luettu ohjelmallisesti ja se sisältää vaihtelevan määrän tunnistusvirheitä alkuperäisen aineiston kunnosta riippuen."[59] Tekstintunnistuksen puutteet voivat tosiaan olla hankalammin havaittavia ja rajattavia eli johtua digitoidun paperin ja musteen laadusta tai epätavallisesta ladonnasta. Tai siitä, että modernista julkaisusta muodostetut pdf-tiedostot ovat korruptoituneet tai muuten hävittäneet hakukelpoisen tekstin, jonka julkaisuarkistossa hakuja tehdessään odottaa olevan kokotekstihaun tavoitettavissa.

Entä sitten?

Osan edellä kuvatuista ongelmatilanteista olisin voinut välttää, jos olisin noudattanut "uuden lähdekritiikin" ohjeita ja ottanut kaikessa rauhassa selvää tietokannan sisällöstä sekä lukemalla tarjotun dokumentaation että tekemällä testaushakuja. Hakemalla itselleen tuttua asiaa tai yleistä sanaa saa jonkinlaisen käsityksen mahdollisuuksista ja laadusta. Jos aikarajauksien käyttö on tarpeen, kannattaa ensin yrittää selvittää, missä laajuudessa ja millä tarkkuudella ajoituksia on aineistossa haettavissa. Ja niin edelleen. Tämä olisi tietenkin helpompaa, jos tietokannasta on tarjolla dokumentaatio, joka on yhtä aikaa järjellisen mittainen, yksityiskohtainen ja ymmärrettävästi kirjoitettu. Mutta tämä on käytännössä yhtä mahdotonta,

kuin historiantutkimuksen täydellinen kuvaus tai tietokanta ja hakuliittymä, jotka vastaavat kaikkien tarpeisiin. Mikä ei tarkoita, etteikö tekijöiden kannata ohjeiden kehitystä yrittää.

"Uuden lähdekritiikin" kysymyslistojen opettelun ja käytön ohella pitää aina pitää mielessä se, että tietokanta tai sivusto on ihmisten tekemä. Mahdollisesti hakutuloksiin sisältyvää epävarmuutta ja epäselvyyksiä voisi oppia vertaamaan menneisyyden lähteisiin, joista ei ole koskaan käytettävissä "kaikkea"? Se, että oma haku ei tavoita kaikkea mahdollista tietokannasta, ei tee siitä hyödytöntä. Historiantutkija voi käyttää rakenteista muotoa eksploratiivisesti pyrkimättä johtopäätöksiin, todistusvoimaisuuteen tai hypoteesitestaukseen. Esimerkiksi aikasarjat sanojen esiintymisfrekvensseistä voivat synnyttää ajatuksia ja uusia kysymyksiä, vaikka pohjalla olevan tekstimassan koostumuksessa olisi epäselvyyksiä tai epävarmuuksia.[60] Hakukäyttöliittymien rakentajatkin voisivat miettiä ratkaisuja, jotka eivät tarjoa "oikeaa" tuloslistaa vaan mahdollisuuden selailuun ja serendipiteettisiin eli odottamattomiin ja onnekkaisiin löytöihin.[61]

Olisi kuitenkin toivottavaa, kuten muun muassa historiantutkija Mats Fridlund on korostanut, että historiantutkijat tekisivät hakunsa tietoisesti ja kehittäisivät näitä kuvaavia ilmaisuja. Näin voisi syntyä yhteinen käsitys hakutapojen riskeistä ja mahdollisuuksista.[62] Tai edes yhteistä kieltä ja ymmärrystä.

[1] Gunilla Jonsson, "Luckan" på tryck (?) eller trycket i "Luckan",' i *Framsynthets ihugkommelse. En vänbok till Sune Lindqvist på 80-årsdagen den 26 oktober 1991*. Acta Bibliothecae Regiae Stockholmiensis 51, red. Folke Sandgren och Ingrid Ström. (Stockholm: Kungl. Biblioteket 1991), 46–47.

[2] Vrt. Jesse Lindström ja Tiina Rossi, *"Vähän niin kuin menu ravintolassa". Arkistojen kuvailutiedot tukemassa historiantutkijoiden lähdemateriaalin etsintää*, (pro gradu, Tampere: Tampereen yliopisto, 2016), 55–56, https://urn.fi/URN:NBN:fi:uta-201602251256.

[3] Esim. Lara Putnam, "The Transnational and the Text-Searchable. Digitized Sources and the Shadows They Cast," *The American Historical Review* 121, no. 2 (2016): 377–402, https://doi.org/10.1093/ahr/121.2.377; David A. Bell, "The Bookless Future. What the Internet is doing to Scholarship," *The New Republic* 232, no. 16/17 (2005): 27–33; T. Mills Kelly, *Teaching History in the Digital Age* (Ann Arbor, MI, USA: University of Michigan Press 2013), 26–34, https://doi.org/10.3998/dh.12146032.0001.001; Patrick Leary, "Googling the Victorians," *Journal of Victorian Culture* 10, no. 1 (2005): 72–86.

[4] Tuomas Tepora, Mirkka Danielsbacka ja Matti O. Hannikainen, "Johdanto. Tutkimuksen työkalut," teoksessa *Avaimia menneisyyteen. Opas historiantutkimuksen menetelmiin*, toim. Mirkka Danielsbacka, Matti O. Hannikainen ja Tuomas Tepora ([Helsinki]: Gaudeamus, 2022), 12.

[5] Lindström ja Rossi, *Vähän niin kuin menu*, 59.

[6] Jorma Kalela, *Historiantutkimus ja historia* ([Helsinki]: Gaudeamus, 2020), 16.

[7] Jussi T. Lappalainen, *Haluatko historiankirjoittajaksi?*, Suomalaisen Kirjallisuuden Seuran Toimituksia 895 (Helsinki: Suomalaisen Kirjallisuuden Seura, 2002), 25–29, 38–44.

[8] Lappalainen, *Haluatko historiankirjoittajaksi?*, 22; Ville Vuolanto, "Tutkimusprosessi, metodit ja historiantutkimuksen ominaislaatu," *Historiallinen Aikakauskirja* 105, no. 3 (2007), 309, 314; Jorma Kalela, "Miksi ei pidä ajatella, että historiantutkija tuottaa kertomuksia?," teoksessa *Näkökulmia kertomuksen tutkimukseen*, Tietolipas 226, toim. Samuli Hägg, Markku Lehtimäki, Markku ja Liisa Steinby (Helsinki: Suomalaisen Kirjallisuuden Seura 2009), 301; Mirkka Danielsbacka, Matti O. Hannikainen, ja Tuomas Tepora, "Teoriaton historia?," teoksessa *Menneisyyden rakentajat. Teoriat historiantutkimuksessa*, toim. Mirkka Danielsbacka, Matti O. Hannikainen, ja Tuomas Tepora ([Helsinki]: Gaudeamus, 2018), 14.

[9] Vuolanto, *Tutkimusprosessi*, 306; vrt. Frederick W. Gibbs and Trevor J. Owens, "The Hermeneutics of Data and Historical Writing," in *Writing History in the Digital Age*, eds. Jack Dougherty and Kristen Nawrotzki (Ann Arbor, MI, USA: University of Michigan Press 2013), 163–164, https://doi.org/10.3998/dh.12230987.0001.001.

[10] Vrt. Mats Fridlund, "Chapter 4. Digital History 1.5. A Middle Way between Normal and Paradigmatic Digital Historical Research," in*Digital histories. Emergent approaches within the new digital history*, eds. Mats Fridlund, Mila Oiva and Petri Paju (Helsinki: Helsinki University Press, 2020), 75–76, 79–82, https://doi.org/10.33134/HUP-5-4.

[11] Esim. Inés Matres, "Digital historical materials for academics, educators, hobbyists, creatives and browsers: the visitors evaluate digi.kansalliskirjasto.fi," *Informaatiotutkimus* 35, no. 3 (2016): 48–49, https://journal.fi/inf/article/view/59436; Tuula Pääkkönen ja Kimmo Kettunen, "Kansalliskirjaston sanomalehtiaineistot: käyttäjät ja tutkijat kesällä 2018," *Informaatiotutkimus* 37, no. 3 (2018): 15–19, https://doi.org/10.23978/inf.76067; Tuula Pääkkönen, "Syvään päähän. Digi.kansalliskirjasto.fi-palvelun käyttäjäkyselyn tuloksia," *Tietolinja*, no.1, (2022), https://urn.fi/URN:NBN:fi-fe2022061546944; Sanna W. Kumpulainen and Heli Kautonen, "Accidentally Successful Searching. Users' Perceptions of a Digital Library," in *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)* (New York, NY, USA: Association for Computing Machinery, 2017), 257–260, https://doi.org/10.1145/3020165.3022124.

[12] *Kulttuuriperintöä käyttäjä edellä. Tuloksia Digitaalinen avoin muisti -hankkeesta*, toim. Miia Kosonen (Mikkeli: Kaakkois-Suomen ammattikorkeakoulu, 2021), https://urn.fi/URN:ISBN:978-952-344-357-0.

[13] Denis Mandrov, *Käyttäjälähtöinen digihumanismi. Tapaustutkimuksena Diplomatarium Fennicum -tietokannan kehittäminen tutkijayhteisön kanssa* (pro gradu, Turku: Turun yliopisto, 2018), https://urn.fi/URN:NBN:fi:uta-201811262892; Johanna Mieto, *Arkistotietojärjestelmän käyttäjälähtöinen kehittäminen. Case: Helsingin kaupunginarkiston Sinetti -arkistotietojärjestelmä* (YAMK-opinnäyte, [Mikkeli]: Kaakkois-Suomen ammattikorkeakoulu, 2019), https://urn.fi/URN:NBN:fi:amk-201903223640.

[14] Outi Hupaniittu, *Tutkijoiden ääni ja sähköiset aineistot. Selvitys muistiorganisaatioiden asiakkaitten digitoitujen aineistojen tarpeista ja saatavuudesta* ([Helsingfors]: Svenska litteratursällskapet i Finland 2012).

[15] Tiina Hölttä, *Digitoitujen kulttuuriperintöaineistojen tutkimuskäyttö ja tutkijat* (pro gradu, Tampere: Tampereen yliopisto, 2016), http://urn.fi/URN:NBN:fi:uta-201603171337.

[16] Lindström ja Rossi, *Vähän niin kuin menu*, 59; Janne Jokinen, *"Niin alkuperäisille kuin mahdollista" Arkistoaineiston käyttö tieteellisessä historiantutkimuksessa* (pro gradu, Turku: Turun yliopisto, 2016), https://urn.fi/URN:NBN:fi:uta-201602261278.

[17] Laura Korkeamäki, *Historiantutkijoiden tehtäväperusteinen informaatiovuorovaikutus digitaalisissa ympäristöissä* (pro gradu, Tampere: Tampereen yliopisto, 2018), https://urn.fi/URN:NBN:fi:uta-201809062467; Laura Korkeamäki and Sanna Kumpulainen, "Interacting with Digital Documents. A Real Life Study of Historians' Task Processes, Actions and Goals," In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*. (New York, NY, USA: Association for Computing Machinery, 2019), 35–43, https://doi.org/10.1145/3295750.3298931: Sanna Kumpulainen, Heikki Keskustalo, Boyang Zhang and Kostas Stefanidis, "Historical reasoning in authentic research tasks. Mapping cognitive and document spaces", *Journal of the Association for Information Science and Technology* 71, no. 2 (2020): 230–241, https://doi.org/10.1002/asi.24216; Heikki

Keskustalo, Elina Late, Laura Korkeamäki, Sanna Kumpulainen ja Kimmo Kettunen, "Informaatiovuorovaikutus historian tutkimuksessa. Tiedonlähteiden käytön ymmärtämisestä käytön tukemiseen", *Informaatiotutkimus* 40, no. 3 (2021): 245–268, https://doi.org/10.23978/inf.107890; Sanna Kumpulainen and Elina Late, "Struggling with digitized historical newspapers. Contextual barriers to information interaction in history research activities," *Journal of the Association for Information Science and Technology* 73, no. 7 (2022): 1012–1024. https://doi.org/10.1002/asi.24608; Elina Late and Sanna Kumpulainen, "Interacting with digitised historical newspapers. Understanding the use of digital surrogates as primary sources," *Journal of Documentation* 78, no. 7 (2022): 106–124, https://doi.org/10.1108/JD-04-2021-0078.

[18] Tutkimukseni ovat pääasiallisesti henkilöhistoriallisia, joten hyödynnän erityisesti sukututkijoille suunnattuja palveluita ja digitoitua kirjallisuutta ja lehdistöä. Parhaillaan teen väitöskirjatutkimusta 1700-luvun sanomalehdistä. Katsauksen huomioita olen julkaissut aiemmin blogiteksteinä: "Tietokanta, jossa on "kaikki"," *Sukututkijan loppuvuosi* 4.2.2019, https://sukututkijan-loppuvuosi.blogspot.com/2019/02/tietokanta-jossa-on-kaikki.html; "Kertausläksyjä tarkkaavaisuudesta tietokantahauissa," Sukututkijan loppuvuosi 5.11.2019, https://sukututkijanloppuvuosi.blogspot.com/2019/11/kertauslaksyja-tarkkaavaisuu-desta.html; "Muistutus ja kertaus: tietokannat eivät ole täydellisiä," Sukututkijan loppuvuosi 3.8.2020, https://sukututkijan-loppuvuosi.blogspot.com/2020/08/muistutus-ja-kertaus-tietokannat-eivat.html.

[19] Myös tekstit muodostavat tietokannan, mutta hahmottuvat käyttäjälle usein erilaisena kuin täysin rakenteisempi tieto.

[20] Esim. Jessica Parland-von Essen, "Granskning av digitala resurser," *Essetter* 9.1.2012 http://essetter.blogs-pot.com/2012/01/granskning-av-digitala-resurser.html; Jessica Parland-von Essen, "Källkritik och hänvisningar i material i digitala arkiv," *Historia i en digital värld* (2016), https://digihist.se/hdv2/3-material/#3_6; Kenneth Nyberg, "Källkritik och hänvisningar på webben," *Historia i en digital värld* (2016), https://digihist.se/hdv2/3-material/#3_5; Bo Pettersson, "Fjärrläs-ning och närläsning, stordata och smådata," *Historia i en digital värld* (2016), https://digihist.se/hdv2/fordjupningar/fjarrlasning-och-narlasning-stordata-och-smadata/; Jessica Parland-von Essen, "Chapter 5. Building Historical Knowledge Byte by Byte. Infrastructures and Data Management in Modern Scholarship," in *Digital histories. Emergent approaches within the new digital history*, eds. Mats Fridlund, Mila Oiva and Petri Paju (Helsinki: Helsinki University Press, 2020) https://doi.org/10.33134/HUP-5-5; Trevor Owens and Thomas Padilla, "Digital sources and digital archives: historical evidence in the digital age," *International Journal of Digital Humanities* 1 (2021), https://doi.org/10.1007/s42803-020-00028-7.

[21] Henkilökohtaisen kokemuksen ohella esim. Kumpulainen et.al., *Historical reasoning.*

[22] Lappalainen, *Haluatko historiankirjoittajaksi?*, 69–70. Kukin neulakortin reunojen reikä kertoi täyttikö kortin sisältö tietyn ehdon vai ei. Reikiin työnnetyillä sauvoilla kortteja voitiin poimia yhdistäen useampia kriteerejä.

[23] Anu Lahtinen, "Datanhallinta käytännössä: havaintoja historian graduseminaarista," *Think open* 14.6.2019. https://blogs.hel-sinki.fi/thinkopen/historian-graduseminaarilaiset-ja-datanhallinta/.

[24] Korkeamäki, *Historiantutkijoiden tehtäväperusteinen*, 33.

[25] Sharon Leon, "The Peril and Promise of Historians as Data Creators: Perspective, Structure, and the Problem of Represen-tation", *[bracket]* 24.11.2019 https://www.6floors.org/bracket/2019/11/24/the-peril-and-promise-of-historians-as-data-cre-ators-perspective-structure-and-the-problem-of-representation/ ; Ks. myös Jessica Parland-von Essen, "Att arbeta med data-baser," *Historia i en digital värld* (2016), https://digihist.se/hdv2/4-metoder/#4_3.

[26] Leon, "*The Peril* ".

[27] Leon, "*The Peril*".

[28] Leon, "*The Peril* "; Fridlund, "*Chapter 4*," 81–82; Trevor Owens, "Defining Data for Humanists: Text, Artifact, Information, or Evidence*?," Journal of Digital Humanities* 1, no. 1 (2012), http://journalofdigitalhumanities.org/1-1/defining-data-for-humanists-by-trevor-owens/; Mateusz Fafinski, "Historical data. A portrait", *History in Translation* 29.9. 2020, https://mfafinski.github.io/Historical_data/; Frederick W. Gibbs, "New Forms of History: Critiquing Data and Its Representations", *The American Historian* (2016), https://www.oah.org/tah/issues/2016/february/new-forms-of-history-critiquing-data-and-its-representations/.

[29] Andrew Prescott, The Imaging of Historical Documents, in *The virtual Representation of the Past*, eds. Mark Greengrass and Lorna Hughes. (Sheffield, England: Ashgate 2008), 13–14.

[30] Leon, "*The Peril* "; Fafinski, "*Historical data*".

[31] Finna.fi, https://finna.fi/.

[32] Janne Jokinen, *Niin alkuperäisille,* 33.

[33] Historiakirjat, Suomen Sukututkimusseura, https://hiski.genealogia.fi/hiski/.

[34] Svenska dagstidningar, Kungl. Biblioteket, https://tidningar.kb.se/.

[35] "Kokoelmia ja sisältöjä verkossa", Kansalliskirjasto, https://www.kansalliskirjasto.fi/fi/kokoelmia-ja-sisaltoja-verkossa.

[36] Leena Jansson, "Kansalliskirjasto digitoi Turun akatemian väitöskirjat. Suomalaisen tieteen varhaishistorian keskeistä lähde-aineistoa nyt verkossa", Kansalliskirjasto 52, no. 2 (2010): 17, https://urn.fi/URN:NBN:fi:ELE-1569856.

[37] Lainaus sähköpostista.

[38] " Fennica – Suomen kansallisbibliografia", Kansalliskirjasto, https://www.kansalliskirjasto.fi/fi/palvelut/fennica-suomen-kansallisbibliografia.

[39] Lainaus sähköpostista.

[40] Diplomatarium Fennicum, http://df.narc.fi/.

[41] Eljas Orrman, ”Näkökohtia Diplomatarium Fennicum -tutkimustietokantahankkeen toteutukseen,” *Historiallinen Aikakauskirja* 118, no. 1 (2020):106–110.

[42] Medicinhistoriska databasen (Linköping University Electronic Press) https://doi.org/10.3384/db.medhist.

[43] Eljas Orrman, *Suomen asutuksen yleisluettelon opas*. Valtionarkiston julkaisuja 7 (Helsinki: [Valtionarkisto] 1975).

[44] Henrik-tietokanta http://dbgw.finlit.fi/henrik/henrik_tietokanta.htm.

[45] HENRIK-hakuohjeet, Henrik-tietokanta, http://dbgw.finlit.fi/henrik/henrik_ohjeet.htm.

[46] Lindström ja Rossi, *Vähän niin kuin menu*, 53.

[47] Henrik Grönroos, Ann-Charlotte Nyman, *Boken i Finland: bokbeståndet hos borgerskap, hantverkare och lägre sociala grupper i Finlands städer enligt städernas bouppteckningar 1656–1809*. Skrifter utgivna av Svenska Litteratursällskapet i Finland 596 (Helsingfors: Svenska Litteratursällskapet i Finland 1996).

[48] Esim. Olli-Pekka Kasurinen, Lauri Leinonen ja Sanna Supponen, ”Modernien tutkimustietokantojen ja tiedon jalostusmenetelmien mahdollisuudet historiantutkimuksessa,” *Historiallinen Aikakauskirja* 117, no.3 (2019):181–183; Kumpulainen et. al., *Historical*, 236–237; Lindström ja Rossi, *Vähän niin kuin menu*, 52.

[49] Zacharias Topelius Skrifter, https://topelius.sls.fi/#/home.

[50] “En historik-kritisk utgåva”, Zacharias Topelius Skrifter, https://topelius.sls.fi/#/content/sv-03-03.

[51] Lindström ja Rossi, *Vähän niin kuin menu*, 55; Parland-von Essen, *Att arbeta*.

[52] Lindström ja Rossi, *Vähän niin kuin menu*, 52, 57; Tuula Juvonen, ”Aikamatkalle sateenkaarihistorian hakusanakirjan avulla,” *Sateenkaarihistorian ystävien kirjoituksia* 16.5.2021 https://sateenkaarihistoria.fi/aikamatkalle-sateenkaarihistorian-hakusanakirjan-avulla/; Heikki Keskustalo, Laura Korkeamäki, Selja Vanamo, Kimmo Kettunen and Sanna Kumpulainen, ”Analyzing gender clues in war-time letters,” *Digital Scholarship in the Humanities*, ennakkojulkaisu (2022), https://doi.org/10.1093/llc/fqac035.

[53] Kasurinen et al , ”*Modernien*,” 184–185.

[54] Vrt. Lindström ja Rossi, *Vähän niin kuin menu*, 60.

[55] Kuvailudatan merkityksestä laajemmin ks. Kimmo Elo, “Chapter 7. Big data, bad metadata. A methodological note on the importance of good metadata in the age of digital history,” in *Digital histories. Emergent approaches within the new digital history*, eds. Mats Fridlund, Mila Oiva and Petri Paju (Helsinki: Helsinki University Press, 2020), https://doi.org/10.33134/HUP-5-6.

[56] Lindström ja Rossi, *Vähän niin kuin menu*, 52.

[57] Svensk historisk bibliografi, Kungl. Biblioteket, https://shb.kb.se/F/?func=file&file_name=find-b&local_base=shb.

[58] Litteraturbanken, https://litteraturbanken.se/.

[59] Digitaaliset aineistot, Kansalliskirjasto, https://digi.kansalliskirjasto.fi/search.

[60] Gibbs and Owens, *Hermeneutics*, 161–163; Trevor Owens, “Discovery and Justification are Different: Notes on Science-ing the Humanities,” 19 November 2012. http://www.trevorowens.org/2012/11/discovery-and-justification-are-different-notes-on-sciencing-the-humanities/.

[61] Mitchell Whitelaw, “Generous Interfaces for Digital Cultural Collections,” *Digital Humanities Quarterly* 9, no. 1 (2015), http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html.

[62] Fridlund, *Chapter 4*, 80–81; Eva Andersen, “From search to digital search. An exploration through the transnational history of psychiatry,” in *Digital History and Hermeneutics: Between Theory and Practice*. Studies in Digital History and Hermeneutics 2, eds. Andreas Fickers and Juliane Tatarinov (Berlin: De Gruyter Oldenbourg, 2022), 131–137, 152–154, https://doi.org/10.1515/9783110723991.