þÿ D i g i t a a l i s e t i h m i s t i e t e e t (D i g i t a l H u m a n i t i e s) j

Tolonen, Mikko Sakari

Gaudeamus
2018

þÿ T o l o n e n , M S & L a h t i , L 2 0 1 8 , D i g i t a a l i s e t i h m i s t i e t e e t (D i g i t a l H u m
historiantutkimus . julkaisussa M Hannikainen , M Danielsbacka & T Tepora (toim) ,
Menneisyyden rakentajat : teoriat historiantutkimuksessa . Gaudeamus , Helsinki , Sivut
235-258 .

http://hdl.handle.net/10138/309808

unspecified
publishedVersion

DIGITAALISET IHMISTIETEET JA HISTORIANTUTKIMUS

Mikko Tolonen & Leo Lahti

> Tutkija astelee arvokkaasti arkistoon ja jättää asiaankuuluvasti berberinsä naulakkoon. Mutta pienen agenttisalkkunsa hän pitää mukanaan […] Yllätys yllätys! Agenttisalkku ei olekaan salkku! Sen kannen alta paljastuu näppäimistö, jonka takaa kääntyy ylös pieni näyttöruutu […] Olisiko kuvattu 'agenttisalkku' mahdollinen? Olisiko se tarpeellinen vai pelkkää hienostelua? Ja ennen kaikkea: olisiko siitä hyötyä esimerkiksi neulakortteihin verrattuna?[1]
>
> – Jussi T. Lappalainen, 1983

Digitaalisten aineistojen käytöstä historiantutkimuksessa on keskustelu viime aikoina paljon.[2] Itseään "digihistorioitsijoiksi" tituleeraavien henkilöiden määrä kasvaa maailmalla tasaisesti, ja historiantutkimusta tehdään myös Suomessa "ilman papereita", niin kuin 1980-luvulla visioitiin. Tästä huolimatta digitaalisten menetelmien käyttö on vielä vain harvoissa tapauksissa vaikuttanut historiantutkimuksen traditioiden ytimeen tai muokannut historia-alan tutkijakoulutusta.[3] Digitaalisten menetelmien hyödyntäminen historiantutkimuksessa ei kuitenkaan ole mikään uusi asia. Historioitsija Viljo Rasila kirjoitti jo 1960-luvulla tietokoneiden käytöstä historiantutkijoiden apuvälineenä *Historiallisessa Aikakauskirjassa*.[4] Tämän luvun alun sitaatissa kannettavasta tietokoneesta "agenttisalkkuna" puhunut historioitsija Jussi T. Lappalainen jatkoi keskustelua Suomessa 1970- ja 1980-luvuilla.[5]

Todellinen muutos syntyy kuitenkin vasta silloin, kun uusia datatieteen menetelmiä aletaan soveltaa aikaisempien tutkimustraditioiden varaan. Datatiede on yleistermi, joka viittaa modernien tietojenkäsittelytieteen menetelmien ja apuvälineiden kehittämiseen hallinnoitaessa ja analysoitaessa laajoja digitaalisia tietoaineistoja. Aineistot voivat sisältää esimerkiksi tekstiä, kuvaa, ääntä tai liikettä. Tätä datatieteen tarjoamien uusien välineiden ja laajojen digitaalisten tietoaineistojen hyödyntämistä humanistis-yhteiskuntatieteellisessä tutkimuksessa kutsutaan yleisnimellä digitaaliset ihmistieteet (*digital humanities*).[6] Suomen Akatemia päätyi keväällä 2015 tähän käännökseen, joka kattaa näin sekä humanistiset että yhteiskuntatieteet.[7]

Datatieteen mahdollisuuksia hyödynnettäessä ja nopeasti karttuvia digitaalisia tietoaineistoja käytettäessä myös lähdekritiikki saa uusia muotoja ja merkityksiä. Tulkintaongelmat ja virhelähteet muuttuvat yhä lukuisemmiksi ja monisyisemmiksi. Historiantutkimuksen perusmetodit eivät silti tulevaisuudessakaan katoa, vaan lähdekritiikin hallitsemisesta tulee yhä tärkeämpää teknisen osaamisen syventymisen rinnalla. Historiantutkijan koulutus alkaa toden teolla muuttua digitaalisten ihmistieteiden vaikutuksesta, kun saadaan aikaan tutkimustuloksia, joihin ”perinteisin” keinoin ei olisi päästy. Tämä paradigman muutoksen aika ei vielä ole kokonaisvaltaisesti koittanut, mutta sitä kohti edetään jo vauhdilla.

Tässä luvussa tarkastelemme digitaalisten ihmistieteiden merkitystä historiantutkimuksessa ja pohdimme niiden suhdetta erilaisiin teoreettisiin kysymyksenasetteluihin.[8] Käsittelemme mahdollisuutta muodostaa teoriaa digitaalisista ihmistieteistä historiantutkimuksessa työskentelymenetelmien, tutkimusaineistojen ja monitieteisen yhteistyön näkökulmasta. Samalla arvioimme tämän nuoren tutkimusalan ominaispiirteitä suhteessa muuhun tutkimukseen. Lisäksi pohdimme uudenlaisten menetelmien ja tutkimuksen piiriin saatujen lähdeaineistojen merkitystä. Heiluriliike näyttäisi vievän taas kohti aikaa, jossa historiantutkimuksen linjat uudistuvat mikrohistoriallisesta otteesta kohti ”Ison Historian” näkökulmaa. Massadata (*big data*) -tutkimus vaatii teknisiä erityistaitoja ja -välineitä, mutta samalla myös suurten aineistojen hallintaan ja tutkimukseen käytetyt kvantitatiiviset ja kvalitatiiviset menetelmät voivat yhtenäistyä.[9] Punnitsemme tämän kehityksen hyviä ja huonoja puolia.

Pohdimme lisäksi kulttuurista muutosta, jonka uudenlaisten lähestymistapojen käyttöönotto synnyttää. Esimerkiksi pelkän lähdeaineistojen digitoimisen ei sinänsä katsota vielä kuuluvan alan varsinaisen tutkimuksen piiriin, eikä humanistisessa tutkimustraditiossa tietotekniikan käytöllä ole ollut itseisarvoa. Digitaalisen tutkimuksen määritteleminen on haastavaa, sillä tutkimusaineistojen keruu ja järjestäminen on keskeinen – eli välttämätön mutta ei kuitenkaan riittävä – osa tieteellistä tutkimusprosessia. Digitaalisten ihmistieteiden tutkimusta määrittääkin monitieteisyys, jossa yhdistellään tutkimuskysymyksiä, menetelmiä ja lähdeaineistoja useiden tutkimusalojen piiristä.[10] Historiantutkimuksen näkökulmasta mielenkiintoista on erityisesti se, millaisia uusia vastauksia perinteisiin tutkimuskysymyksiin voidaan saada, kun käyttöön avautuu uusia aineistoja ja menetelmiä. Tutkimuksen merkitys syntyy kuitenkin tiedosta, jonka tutkija kykenee tarjoamaan ihmisille luotettavasti, ei tutkimuksen toteuttamisen tavoista.

Historiantutkimuksen tekemiseen liittyy monenlaisia vuosisatojen aikana vakiintuneita kulttuureja, ja kärjistäen voidaan todeta, että verrattaessa esimerkiksi biotieteisiin historia on vahvemmin yksittäisen tutkijan henkilön ympärille rakentuva tieteenala. Tämä näkyy selvästi esimerkiksi eri alojen julkaisumalleista. Historiankirjoituksessa yhteisjulkaisuja ei perinteisesti arvosteta ja ihanteena on yksittäisen kirjoittajan monografia. Tämä voi johtaa erilaisiin lieveilmiöihin, kuten epäsuhtiin tutkimustyötä tukevissa rahoitusmalleissa.[11] Samalla kun moderni luonnontieteellinen tutkimus rakentuu monialaisen yhteistyön varaan, historiantutkimuksessa yksittäisten tutkijoiden ja koulukuntien näkemyksiin voidaan tukeutua jopa objektiivisuuden ja näkökantoja perusteleviin tausta-aineistoihin perehtymisen kustannuksella. Toisaalta luonnontieteitä painottavilla tutkimusaloilla – vaikkapa luonnonsuojelun alalla – voitaisiin toki hyödyntää enemmän ja monipuolisemmin humanistien osaamista, esimerkiksi suojeltavien kohteiden kulttuuristen merkitysten selvittämisessä. Monitieteisessä yhteistyössä ymmärryseroja voi aiheuttaa se, että humanistisessa tutkimuksessa tärkeä rooli on hermeneuttisella, vähitellen tarkentuvalla otteella. Siinä missä luonnontieteet etsivät universaaleja lainalaisuuksia, humanistisessa tutkimuksessa ilmiöt ovat usein ainutkertaisia ja tulkinnat monimielisiä. Yksi historiantutkimuksen tulevaisuuden keskeisistä haasteista onkin: miten kyetään säilyttämään tämä humanismin ydin ja lisäämään samalla hedelmällistä vuoropuhelua eri tutkimussuuntausten välillä?

Miksi tulisi puhua erikseen digitaalisista ihmistieteistä tai digitaalisesta historiantutkimuksesta ja pyrkiä määrittelemään, mihin näillä termeillä viitataan?[12] Tilastollisten menetelmien hallinta ei voi korvata tutkimusalaan liittyvää perinteistä asiantuntemusta mutta voi täydentää sitä ja tuoda tutkimukseen uusia vaikutteita. Huomiota tulisi kiinnittää erityisesti sellaisiin humanistisiin ja yhteiskuntatieteellisiin tutkimustraditioihin, joilla on paras uudistumispotentiaali. Samalla digitaalisten ihmistieteiden tulee pystyä asemoimaan itsensä perinteisen humanistisen tutkimuksen kentällä. Digitaalisuuden painottaminen muuttaa paitsi tutkimusta myös tutkimuskohdetta: esimerkiksi digitaalinen sanomalehti lähteenä ei ole sama asia kuin painettu sanomalehti.[13]

Digitoidussa sanomalehdessä oleva tieto saattaa osittain säilyä uudessa muodossa, mutta se ei koskaan voi korvata materiaalista lähdettä. Aikaisemmasta alkuperäinen/kopio-keskustelusta onkin nykyään edetty siihen, että sekä ”perinteiset” että digitaaliset kulttuuriaineistot ymmärretään omanlaisina objekteinaan.[14] Perinteisiin analogisiin aineistoihin voi sisältyä monenlaisia yllättäviä piirteitä, vaikkapa digitoinnin ulkopuolelle jääviä marginaalimerkintöjä. Vastaavasti digitaalinen tieto sisältää elementtejä, joita alkuperäisessä kulttuuriperintöaineistossa ei ole. Näitä ovat esimerkiksi mahdollisuus yhdistää ja eritellä erilaisia laajoja aineisto-osia sekä digitaalisen aineiston yleinen muokattavuus. Cambridgen yliopiston piirissä toimiva Concept Lab pyrkii muuttamaan yleisesti käsitteiden tutkimusta korostamalla sitä, että digitaalista ja painettua lähdeaineistoa ei tutkita aatehistorian kannalta samoin tavoin.[15] Digitaalisten aineistojen avulla esimerkiksi aiemman tutkimuksen hypoteeseja ja tuloksia voidaan testata ja todentaa uudella tavalla. Datatieteen menetelmien soveltaminen myös mahdollistaa uudenlaiset kysymyksenasettelut mutta edellyttää sujuvaa vuorovaikutusta eri tieteenalojen asiantuntijoiden välillä. Lisäksi tutkimuksen uudet menetelmät ja muodot muokkaavat tutkimuskysymyksiä ja luovat uusia kysymyksiä. Lundin yliopiston projektissa on esimerkiksi tutkittu filosofisesti tiedon luonnetta digitaalisessa maailmassa.[16]

Tutkimuksen teoreettisessa näkökulmassa on otettava huomioon aineellisten ja aineettomien tutkimuskohteiden erot ja molempien vaikutukset itse tuloksiin. Aikaisempi keskustelu määrittää voimakkaasti uusienkin tutkimustulosten merkitystä. Jos tutkii vaikkapa David

Humen (1711–1776) teoksia aihemallintamisen (*topic modelling*)[17] avulla, on pystyttävä perustelemaan, millä tavoin se kytkeytyy muuhun Hume-tutkimukseen. Toisin sanoen, jos tutkii, miten David Hume käyttää vapauden käsitettä, ei riitä, että tutkimuksen kohteena ovat vain Humen digitoidut tekstit yleisesti. Uudet tulkinnat tulee pystyä asemoimaan suhteessa aikaisempaan Humen vapauden käsitettä tutkineeseen kirjallisuuteen. Tämä aatehistorioitsijoille itsestään selvä näkökulma ei aina ole sitä digitaalisissa ihmistieteissä, joiden tutkijat saattavat tulla toisenlaisten tutkimustraditioiden piiristä. Osa tutkijoista tuntuu ajattelevan, että uusia menetelmiä käytettäessä tutkimus voitaisiin irrottaa aiemmista viitekehyksistä. Digitaalisten ihmistieteiden tutkimus ei voi kuitenkaan täysin irrota aikaisemmasta, samankaltaista kulttuuriperintöaineistoa lähteenä käyttävästä tutkimustraditiosta. Uusien kysymystenasettelujen tulee siis kyetä operoimaan luontevasti perinteisen historiantutkimuksen kanssa.[18] Kun digitaalisia menetelmiä on toistaiseksi otettu käyttöön historiantutkimuksessa melko vaatimattomasti, tulisi huomio kiinnittää erityisesti niihin datatieteen elementteihin joilla on paras potentiaali edistää tutkimusta sen perinteisistä lähtökohdista.[19]

Digitaalisten ihmistieteiden määritelmä on tätä taustaa vasten suoraviivainen: kun tutkimusta tehdään esimerkiksi digitoiduilla sanomalehdillä ja siinä hyödynnetään datatieteen menetelmiä, tutkimus kuuluu digitaalisten ihmistieteiden alle. Pelkkä digitaalisten sanomalehtien käyttäminen luettavina lähteinä ei vielä kuulu digitaalisten ihmistieteiden piiriin. Toisaalta myöskään puhtaan teoreettinen menetelmäkehitys ei itsessään ole vielä historiantutkimusta vaan lukeutuu esimerkiksi tietojenkäsittelytieteen tai kieliteknologian piiriin. Soveltavien menetelmien kehittäminen voidaan sen sijaan katsoa digitaalisiin ihmistieteisiin kuuluvaksi tutkimukseksi.

Datatieteen menetelmien kirjo on laaja, mutta moninaisia ovat myös käyttäjien tarpeet. Esimerkiksi aineistojen hakeminen ja selaaminen tähän tarkoitukseen suunniteltujen käyttöliittymien avulla vaativat erilaisia teknisiä ratkaisuja ja palvelevat erilaisia käyttäjiä kuin samoihin tietoaineistoihin kohdistuva laskennallinen analyysi. Tällaisessa analyysissa tutkitaan tilastollisten ohjelmointikielten avulla aineistossa esiintyviä laajoja säännönmukaisuuksia ja niiden tilastollista merkitsevyyttä, vaikkapa kirjapainotoiminnan leviämisvauhdin vertailua eri kielialueilla ja tieteenaloilla. Sama tutkimusperinne voi kuitenkin

yhdistää sekä perinteisempää sanomalehtitutkimusta, jossa digitoitua sanomalehteä käytetään esimerkiksi tavanomaisen sanahaun avulla, että uusien laskennallisten menetelmien hyödyntämistä. Pelkkä aikaisemmasta fyysisestä lähteestä muodostettu digitaalinen ”kuva” ei silti vielä itsessään edusta digitaalista tutkimusta.

Digitaaliset tutkimusaineistot voidaan jaotella digitaalisena syntyneisiin (esim. sosiaalinen media) tai muuten syntyneisiin (esim. digitoidut sanomalehdet). Vaikka tämän luvun painopiste on varhaisemman ajan tutkimuksessa ja digitoiduissa historiallisissa aineistoissa, on huomioitava, että historiantutkimusta voidaan tehdä myös digitaalisina syntyvistä aineistoista. Yksi esimerkki tästä on Kansalaisten mielenliikkeet -hanke, jossa on tutkittu Suomi24-keskustelupalstan aineistoa. Historiantutkimusta hankkeessa ovat edustaneet Jaakko Suominen ja Anna Sivula, ja he ovat tutkineet keskustelupalvelua muuttuvana alustana – siis historiallisena ilmiönä. Tämä esimerkki osoittaa, miten on mahdollista uudistaa historiantutkimusta uusilla menetelmillä ja samalla siirtää historiantutkimuksen traditiota kohti modernin maailman ilmiöitä. Kyseisessä hankkeessa uusi ja vanha kohtaavat.[20]

Digitaalisuuteen ja sen hyödyntämiseen liittyy vahvasti massadatan käsite.[21] Tällä tarkoitetaan erityisen laajoja tietoaineistoja, joiden käsittelyyn tarvitaan uudenlaista tietojenkäsittelyinfrastruktuuria ja tilastollisia menetelmiä. Yksi määrittely massadatalle on, ettei aineisto mahdu yksittäiseen tietokoneeseen, vaan aineiston hallintaan, käsittelyyn ja analysointiin tarvitaan järeitä infrastruktuuriratkaisuja ja skaalautuvia eli myös suurten tietoaineistojen tutkimukseen riittävän kevyitä ja nopeita menetelmiä. Suomessa näitä välineitä ja palveluja tarjoaa etenkin Tieteen Tietotekniikan Keskus (CSC). Tällaisessa massadatan tutkimuksessa ja pienemmänkin mittakaavan data-analyysissa aineistoista pyritään tyypillisesti etsimään yleisiä tilastollisia trendejä. Lähestymistavan etuna on, että havaintoja voidaan tehdä erittäin suuressa mittakaavassa, mikä ei ole aikaisemmin ole ollut mahdollista. Karkeakin aineisto voi osoittautua arvokkaaksi, kun etsitään ja kuvataan laajoja tilastollisia säännönmukaisuuksia. Laajaan kvantitatiiviseen analyysiin perustuvassa tutkimuksessa datasta voidaan tehdä myös odottamattomia havaintoja, ja ne voivat suunnata tutkimusta ennakoimattomalla tavalla. Molemmilla lähestymistavoilla – huolella laadituilla pienemmän mittakaavan täsmä-analyyseilla ja laajempien karkeiden aineistojen tutkimuksella – on paikkansa. Lähestymistavasta

riippumatta tutkimusaineiston siistiminen ja järjestäminen käyttökelpoiseen muotoon ja tulosten ymmärrettävä esittäminen on kuitenkin tutkimuksessa usein yksi työläimmistä ja eniten aikaa vievistä vaiheista.

DIGITAALINEN HISTORIANTUTKIMUS KÄYTÄNNÖSSÄ

Digitaalisten ihmistieteiden piirissä on käyty mittavaa keskustelua metodeista ja niiden merkityksestä.[22] Mitä datatieteen menetelmät voivat todella tarjota historiantutkimukselle? Digitaalisissa ihmistieteissä on muodikasta alleviivata, että suuren lähdeaineiston kuvaaminen visuaalisesti uudella tavalla voi tarjota uusia näkökulmia tutkittavaan aiheeseen. Datan visualisoinnista onkin muodostunut oma metodinen suuntauksensa, joka ulottuu perinteisten tieteenalojen yli. Se on sekä tiedeviestinnän väline että kokeileva tutkimusmetodi, jolla aineistosta voidaan seuloa parhaimmillaan uudenlaisiin tutkimuskysymyksiin johtavia havaintoja.[23] Värikkäät visualisoinnit eivät voi kuitenkaan korvata syvällisempää tilastollista mallinnusosaamista.

Suurten aineistojen hyödyntämiseen liittyvien ongelmien ratketessa tutkimuksen painopiste siirtyy kohti soveltavaa tutkimusta, jota edustaa esimerkiksi tiedontuotannon tarkasteleminen sanomalehdissä.[24] Tällaisessa tutkimuksessa pyritään lähestymään kohteita monimuotoisen digitaalisen aineiston kautta ja erilaisista näkökulmista. Usein tutkimus kuuluu osaksi laajempaa kokonaisuutta, jossa esimerkiksi sosiolingvistinen tai sosiologinen tutkimus yhdistyy uusiin datatieteen menetelmiin. Sanomalehdet tarjoavat oivan esimerkin digitaalisen humanismin monimuotoisesta tutkimuskohteesta, sillä perinteisen sisältöanalyysin tekemisen ohella myös lehtien materiaalisuuden ja fyysisten ominaisuuksien tutkiminen voi tuoda kiinnostavia johtopäätöksiä. Niiden tarkasteleminen voi olennaisesti auttaa tekemään entistä luotettavampia laadullisia tulkintoja Suomessa tapahtuneesta julkisesta keskustelusta, kuten myöhemmin osoitamme kirjapainojen tuotannosta tehtyjen analyysien avulla.

Toinen esimerkki ilmiöpohjaisesta lähestymisestä on kaupunkitutkimus. Digitaalisin menetelmin voidaan tutkia, miten kaupunkien toimintaan liittyvä data näkyy kaupunkien kehityksen kontekstissa. Tällainen tieto voi olla hyödyllistä kaupunkien toiminnan suunnittelussa. Kaupunkeihin ja yhteiskuntaan liittyvää dataa ja erilaisia laskennallisia

menetelmiä on saatavilla yhä enemmän perinteisen laadullisen tutkimuksen tueksi. Humanistien panos voi tuottaa näkökulmia kaupunkien toimintaan kuuluvien ilmiöiden kehittymisestä ja niiden yhteyksistä ihmisten käyttäytymiseen, hyvinvointiin, vuorovaikutukseen ja maailmankuvaan. Historiallinen analysointi liittyy juuri tähän tehtävään, mutta tutkimuksen tavoitteena voi olla löytää myös sellaisia näkökulmia, joilla on merkitystä esimerkiksi poliittisessa päätöksenteossa.[25]

Eri tieteenalojen kohdatessa ilmiöitä voidaan tutkia monista täydentävistä näkökulmista. Tämä pakottaa tutkijat ulos asiantuntijarooleistaan ja perinteiseltä mukavuusalueeltaan. Kun tarkoituksena on ymmärtää laajoja ja monimutkaisia kokonaisuuksia, kasvaa tarve yhteistyölle eri alojen asiantuntijoiden kesken. Digitaalisissa ihmistieteissä korostuvatkin monitieteinen tutkimusote, yhteistyö ja pragmaattisuus.

Alan osaamisella olisi tarkoitus uudistaa ja rikastuttaa muitakin tutkimusaloja kuin historiantutkimusta: oma oppiala tulee asettaa osaksi laajempaa humanististen ja sosiaalitieteiden perinnettä, jossa ihmisen toimintaa on mallinnettu eri tavoin kuin historiantutkimuksessa. Digitaalisten menetelmien mahdollisuuksia on kuitenkin pystyttävä hyödyntämään menettämättä humanistisen tutkimuksen hermeneuttista erityispiirrettä. Toisaalta on kiinnitettävä huomiota tutkimusaineistojen edustavuuteen ja esikäsittelyyn sekä tilastollisten tulkintojen merkitsevyyteen ja toistettavuuteen. Digitaalisten ihmistieteiden tutkimusta onkin perusteltua tehdä hankkeissa, joihin kootaan eri tieteenalojen osaamista.

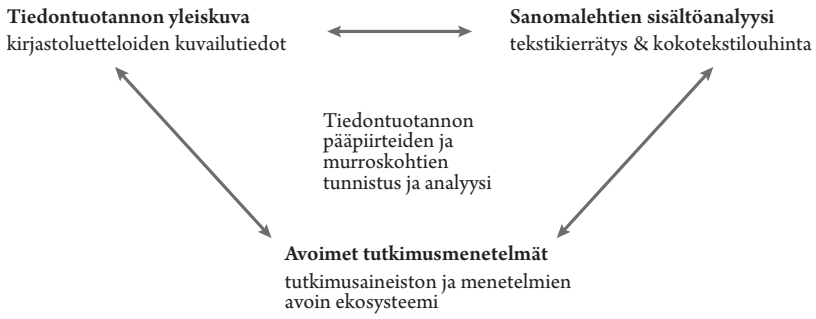
Digitaalinen historiantutkimus keskittyy usein tekstiaineistoihin. Tärkeimpiä näistä ovat aineiston kuvailutiedot (*metadata*) sekä kokotekstit eli digitoidut kirjalliset aineistot, joista voi tehdä avoimia tekstihakuja (Kuva 1).[26] Näitä on hyödyntänyt muun muassa Suomen Akatemian COMHIS-yhteistyöhanke (2016–2019),[27] jossa on tutkittu suomalaista julkisen keskustelun historiallista kehitystä eurooppalaisessa kontekstissa (Kuva 2). Hankkeessa kiinnitetään erityistä huomiota aineistonhallintaan. Tätä tukemaan on kehitetty Octavo-niminen avoimen tieteen alusta, jolla tutkimusaineistoja voidaan yhdistää ja analysoida tehokkaasti ja läpinäkyvästi.[28] Automatisoidut ratkaisut mahdollistavat sujuvasti uusien lähdeaineistojen mukaan tuomisen, aineistoista löytyvien virheiden korjaamisen ja tilastollisten yhteenvetojen täydentämisen.

Tiedonlouhinnan aineistoja aatehistorian näkökulmasta

Kuvailutiedot & tilastollinen analyysi	Kokotekstit & tekstilouhinta
• **Tavoite:** laajojen muutosten kuvailu	• **Tavoite:** yksityiskohtaisempi käsitteellisen muutoksen ja kielen käytön analyysi
• **Aineistot:** kirjastot ja arkistot tulvivat erilaisia kuvailutietokantoja (metadata)	• **Aineistot:** tekstitietokannat (ECCO, EEBO, digitoidut sanomalehdet ym.)
• **Potentiaali:** tutkimuskäyttö aliarvioitu	• **Potentiaali:** kvantitatiivinen evidenssi laadullisen tutkimuksen tukena; uudet havainnot ja etäluenta
• **Saatavuus:** saatavuudessa ongelmia mutta vähemmän kuin kokotekstien kohdalla	• **Saatavuus:** alkuperäisaineistot harvoin avoimesti saatavilla tai sidottu rajoitettuihin käyttöliittymiin
• **Skaalautuvuus:** erinomainen; data on rakenteista ja sen koko rajattu	• **Skaalautuvuus:** rajattu; data on massiivista ja voi vaatia huomattavaa jalostamista; aineistoista ja menetelmistä riippuvainen
• **Metodologinen näkökulma:** tarjoaa selkeän lähtökohdan uusien käytäntöjen soveltamisessa humanistiseen tutkimukseen	• **Metodologinen näkökulma:** aatehistorian sovelluksia toistaiseksi hyvin vähän

Kuva 1. Kuvailutietojen ja kokotekstien sovelluspotentiaali laskennallisessa historiantutkimuksessa.

Julkisen keskustelun muutos Suomessa vuosina 1640–1910

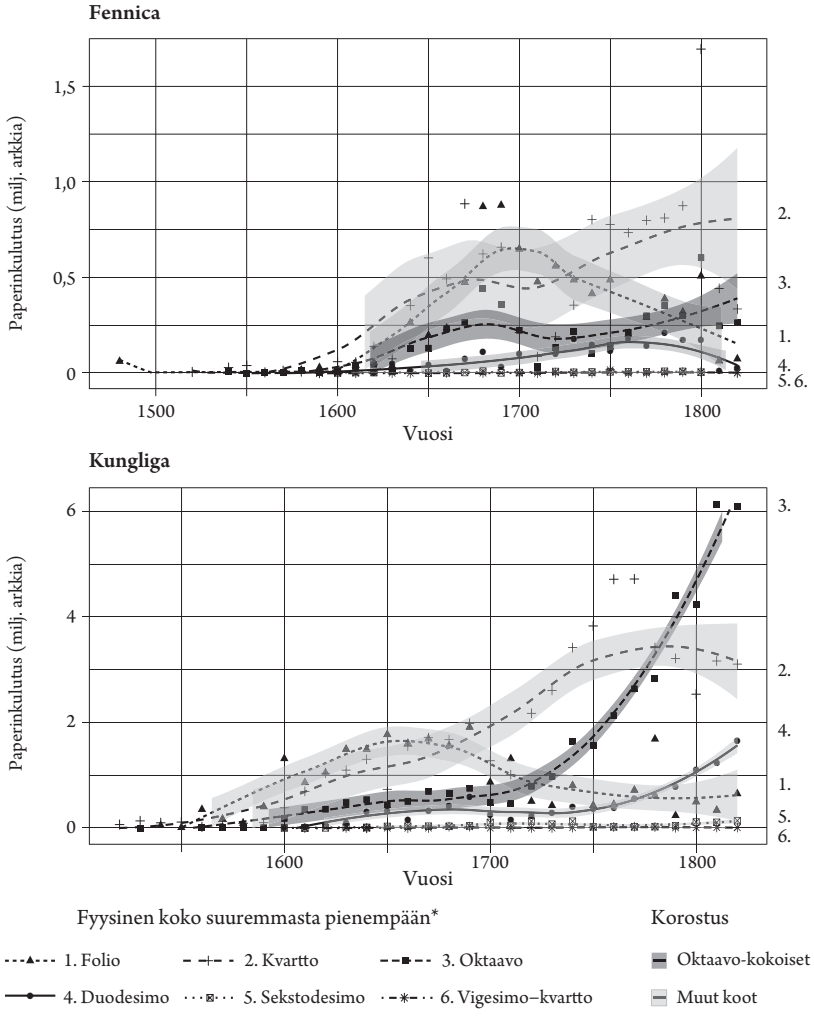


Kuva 2. Julkisen keskustelun muutos Suomessa vuosina 1640–1910. Kotimaisen tiedontuotannon kehitystä on tutkittu yhdistämällä painetun materiaalin yleisiä kuvailutietoja tekstisisältöihin. Avoimen tieteen menetelmät ovat keskeisiä suurten tietomäärien laadukkaassa tutkimuskäytössä.

Kirjastoluetteloiden sisältämien kuvailutietojen avulla pystytään kartoittamaan tiedontuotannon yleinen kehitys huomattavasti kattavammin ja objektiivisemmin kuin aikaisemmin analogisessa

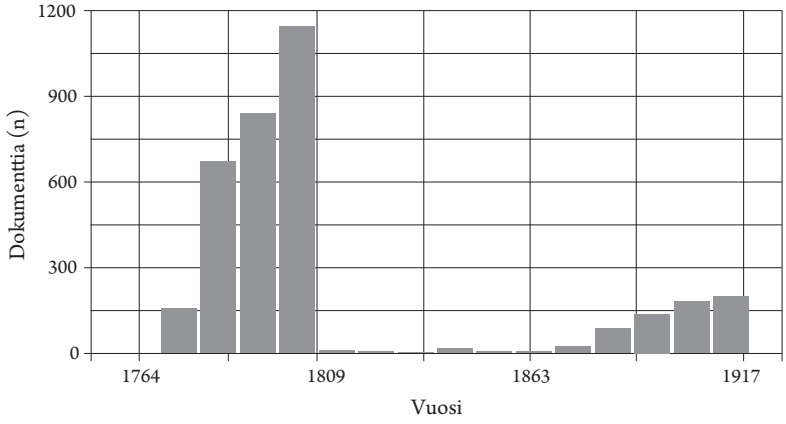
tutkimuksessa. Yhdistelemällä taustatietoa kirjojen aiheista, kirjoittajista sekä painovuosista, -paikoista ja -määristä voidaan tarkastella esimerkiksi kirjapainotoiminnan ja muun tiedontuotannon kehittymistä. Lähteiksi käyvät monenlaiset aineistot arkistomateriaaleista rekisteriaineistoihin ja kirjastoluetteloihin. Automatisoitu jäsenteleminen on usein helpompaa rakenteeltaan yhtenäisille kuvailutiedoille kuin laajemmille kokotekstitietokannoille, koska kuvailuaineistojen koko on rajatumpi ja rakenne selkeä. Niitä voidaan myös tarkastella suoraviivaisemmilla menetelmillä. Kuvailutietojen olemassaolo täydentää näin merkittävästi tekstipohjaisen tutkimuksen mahdollisuuksia. Koneluettavien kuvailutietokantojen hyödyntämistä alan tutkimuksessa onkin merkittävästi aliarvioitu.

Olemme hyödyntäneet kuvailutietoja muun muassa vertailemalla ruotsalaisen Kungliga-kirjaston luettelointitietoja Suomen kansallisbibliografia Fennican vastaaviin määrityksiin. Tämän pohjalta olemme voineet tehdä suoraviivaista analyysia kirjojen eri fyysisten julkaisumuotojen vaikutuksesta tiedontuotantoon varhaismodernina aikana (Kuva 3). Kotimaisen historiantutkimuksen kannalta kiintoisan esimerkin tarjoaa Fennican tutkiminen tieteellisten laskentakirjastojen avulla. Avoimet tieteelliset laskentakirjastot ovat hyvin dokumentoituja ohjelmistoja tai algoritmikokoelmia, jotka antavat tutkijalle sujuvia mahdollisuuksia soveltaa laskennallisen tieteen ja datatieteen uusimpia menetelmiä tutkimusaineistojen käsittelyyn ja analysointiin. Vaikka Fennican alkuperäisaineistosta vielä löytyvät puutteet tuleekin tiedostaa ja korjata, alustava analyysi osoittaa esimerkiksi suoraan historiallisiin tapahtumiin kytkeytyvät tiedontuotannon huiput ja aallonpohjat Vaasassa, Helsingissä, Turussa ja yleisesti Suomen alueella vuosina 1764–1917 (Kuvat 4–8). Kartoittamalla vaikkapa ruotsalaisiin tai venäläisiin paikannimiin tehtävien viittausten määrää ja laatua suomalaisissa sanomalehdissä voidaan saada uutta tietoa suomalaisen julkisen keskustelun rakenteesta ja kehityksestä historian eri vaiheissa. Paikallisemmassa mittakaavassa vastaavaa tietoa voidaan puolestaan saada tutkimalla viittauksia naapurikuntien ja suurempien kaupunkien välillä. Merkittävien suomalaisten elämäkertatietoja sisältävän Kansallisbiografian historiallisten henkilöiden nimien paikantaminen ja kartoitus voi sekin kehittää suomalaisen julkisen toiminnan tutkimusta. Näin uudet menetelmät voivat antaa historiantutkimuksen käyttöön sekä uudenlaista materiaalia että uusia näkökulmia.

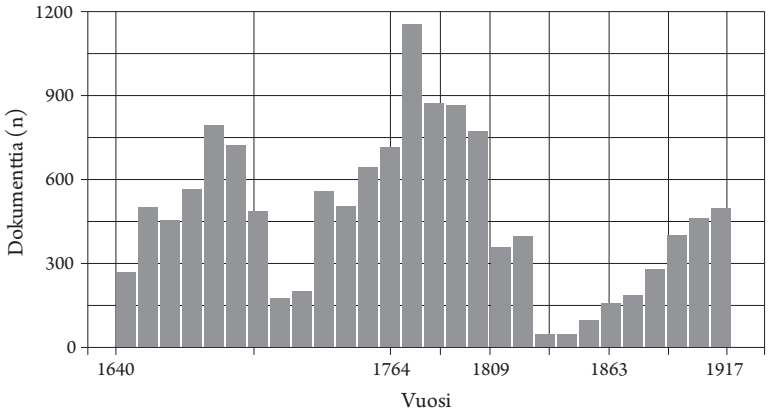


* Mm. kokoluokituksista ks. Laine 2018.

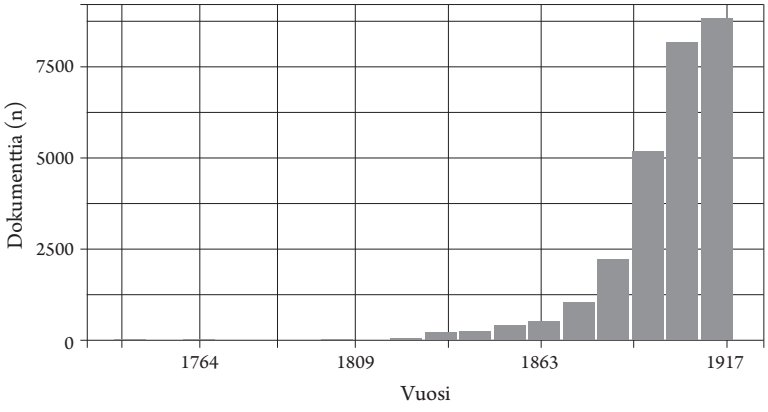
Kuva 3. Paperinkulutus eri kirjaformaattien mukaan vuosina 1600–1828 Fennicassa ja Kungligassa. Kuvasta näkyy erityisesti Kungligan ja Ruotsin kohdalla yleiseurooppalainen trendi, jossa helpommin kannettava oktaavo-kokoinen kirja (tumma korostus, vastaa nykyistä noin A5–B5-kokoa) nousee suurempien kirjaformaattien sijaan keskeisimmäksi tiedontuotannon muodoksi 1700-luvun aikana. Tämä kertoo paperinkulutuksen kasvusta, joka puolestaan kertoo lukemisen leviämisestä laajempien kansanosien keskuuteen valistuksen aikakaudella. Fennicassa havaittavat lievemmät muutokset heijastelevat sitä, että painotuotanto Suomessa ei kohdistunut niinkään kirjoihin vaan virallisiin asiakirjoihin. Kirjoja tuotiin Suomeen esimerkiksi Ruotsin alueelta.



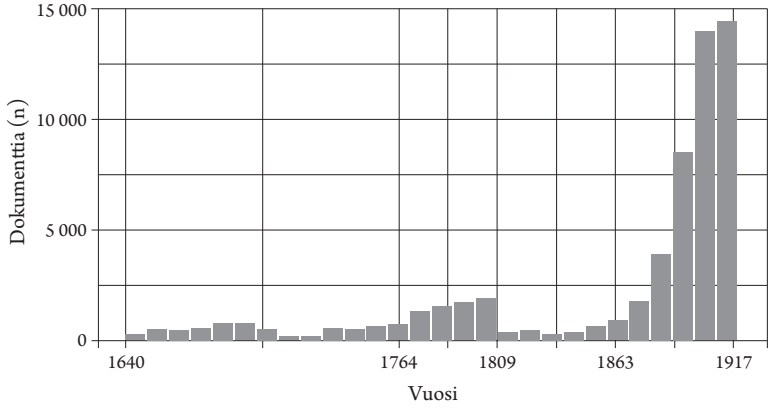
Kuva 4. Julkaisutoiminta Vaasassa vuosikymmenittäin vuosina 1760–1917. Aineisto päättyy vuoteen 1917, joten vertailussa on huomioitava, että kaikissa kuvissa 4–8 viimeinen palkki kattaa muihin palkkeihin verrattuna kaksi vuotta lyhyemmän aikajakson. Oheinen kuva viittaa tiedontuotannon romahdukseen Vaasassa välittömästi Venäjän vallan alussa. Tämä saattaa osittain selittyä sillä, että painotoimintaa harjoitti Vaasassa vuosina 1776–1804 yksi suurempi toimija (Georg Wilhelm Londicer) ja ilmeisesti myös hovioikeuden asiakirjojen julkisuusperiaate ja painaminen muuttuivat Venäjän vallan myötä. Kuvio havainnollistaa yksinkertaisella tavalla Suomen historiassa tapahtuneita julkisuuteen ja julkiseen keskusteluun liittyviä käänteitä. Tämän Suomen kansallisbibliografia Fennicasta tehdyn analyysin perusteella julkisuuden kehitys painotuotannon muodossa näyttäisi olevan Venäjän vallan aikana hitaampaa kuin mitä yliopiston siirron ja Helsingin kehittämisen merkitystä korostanut historiantutkimus on aikaisemmin esittänyt.[29] Mukana on muitakin selittäviä tekijöitä, kuten systemaattisia keisarikunnan aikaiseen painotuotantoon liittyviä puutteita kirjastoluettelossa (on tunnettua, että Fennican luetteloinnista puuttuu muun muassa 1800-luvun ruotsinkielistä kaunokirjallisuutta). On silti selvää, että aineistojen kvantitatiivinen analyysi tarjoaa tehokkaan tavan tunnistaa tiedontuotannon historiaan liittyviä trendejä. Lopulliset johtopäätökset tulee varmistaa myös muilla keinoin. Yksi syy siihen, miksi tahdomme tuoda nämä vahvistamattomat tilastolliset näkökulmat esiin, on se, että emme yksinkertaisesti voi jäädä odottelemaan kansallisbibliografian hidasta täydentymistä seuraavina vuosikymmeninä, vaan tämä tärkeä luettelointityö on saatava viimeisteltyä mahdollisimman pikaisesti.[30]



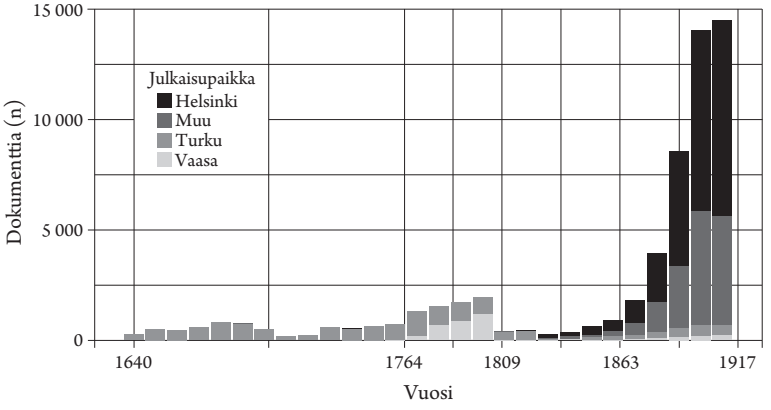
Kuva 5. Julkaisutoiminta Turussa vuosina 1640–1917. 1700-luvun kehitys näyttäisi seuraavan eurooppalaista toiminnan kiihtymisen trendiä. Nykyisten luettelointitietojen valossa kirjapainotoiminnan kehitys 1820-luvulta lähtien on taas erittäin hidasta.



Kuva 6. Julkaisutoiminta Helsingissä vuosina 1740–1917. Helsingin kirjapainotoiminnan kehitys Fennican sisältämän aineiston perusteella viittaa siihen, että ennakkosensuurin osittaisen poistumisen vaikutukset 1850-luvulla ovat yllättävän vähäiset ja varsinainen painotoiminnan kehitys lähtee käyntiin vasta huomattavasti myöhemmin. Yliopiston siirtäminen Helsinkiin vuoden 1828 jälkeen näyttää olleen julkaisutoiminnan kannalta jos ei suorastaan lamauttava niin ainakin hidastava toimenpide, vaikka painotoiminnan romahdus Fennican mukaan alkaakin Turussa heti vuonna 1809.



Kuva 7. Julkaisutoiminta Suomen alueella vuosina 1640–1917. Toiminnan vilkastuminen 1700-luvun lopun Ruotsissa näyttää suunnilleen samalta kuin yleisesti Euroopassa (ks. Kuva 3). Vuoden 1809 jälkeen toiminta Suomen alueella taantuu, ja kestää 1890-luvulle saakka, ennen kuin painotuotteiden lukumäärissä noustaan samalle tasolle kuin Ruotsin vallan aikana. Tämä kuva tulee varmasti tarkentumaan kansallisbibliografian täydentymisen myötä.



Kuva 8. Julkaisutoiminta koko Suomessa sekä erikseen Helsingissä, Turussa ja Vaasassa vuosina 1640–1917. Tässä esitettävä yhteenveto korostaa Suomessa tapahtuneita julkaisutoiminnan yleisiä muutoksia. Suomalainen painotoiminta keskittyy vahvasti Helsinkiin 1800-luvun loppupuolella. Turun painotoiminta on laajassa mittakaavassa stabiilia 1800-luvun alkuun saakka, jolloin myös Vaasan julkaisutoiminta notkahtaa.

Alkuperäiset kuvailutietueet eivät sellaisenaan sovellu tilastolliseen analyysiin, vaan ne pitää ensin jalostaa sopivaan muotoon. Tämä työ sisältää muiden muassa maantieteellisten ja henkilönimien erilaisten kirjoitusmuotojen yhtenäistämistä, samannimisten kirjoittajien erottelua, kirjoitusvirheiden käsittelyä ja kirjojen sivumäärien arviointia bibliografisten standardimerkintöjen pohjalta. Tietoja voidaan lisäksi täydentää ja rikastaa muiden lähdeaineistojen avulla. Kirjoittajista voidaan selvittää puuttuvat elinvuodet ja sukupuoli sekä muuta taustatietoa, esimerkiksi painatusmaa painopaikkojen perusteella. Valtaosa jalostamistyöstä voidaan ja kannattaa automatisoida, kun tutkittavien aineistojen määrät kasvavat kymmeniin tai satoihin tuhansiin dokumentteihin. Samalla analyysien tilastollinen voima kasvaa ja satunnaisten virheiden vaikutus johtopäätöksiin hälvenee. Kaikkea ei voida kuitenkaan automatisoida. Juuri tästä syystä joustavaan lähdekoodiin perustuvat räätälöidyt laskentakirjastot tarjoavat huomattavasti parempia mahdollisuuksia alan tutkimukselle verrattuna valmiisiin ohjelmistoihin. Nämä kun olettavat, että aineisto on valmiiksi saatettu kulloinkin käytössä olevan ohjelmiston vaatimaan muotoon ja että aineistoon sisältyvät virheet on korjattu. Tämä on merkittävä ongelma, sillä käytännössä jopa 80 prosenttia massadataa käyttävän tutkijan työajasta on arvioitu kuluvan aineiston esikäsittelyyn.[31] Ohjelmointitaitoa tarvitaan, jotta raakadatan jalostaminen ja harmonisointi voidaan automatisoida mahdollisimman kattavasti. Samalla tutkijalle täytyy jäädä vapaat kädet joustavien, aineistokohtaisesti räätälöityjen luokittelujen laadintaan. Avoimen lähdekoodin laskentakirjastot ovat avainasemassa, koska niiden avulla tutkija saa tehokkaita ja joustavia työkaluja niin aineistojen jalostamiseen kuin tilastolliseen analyysiin ja tulosten esittämiseen. Samalla kaikki työvaiheet voidaan toteuttaa läpinäkyvällä ja toistettavalla tavalla, tieteellisen avoimuuden periaatteita kunnioittaen.

Laajoja digitoituja tekstiaineistoja voidaan tutkia kahdella toisiaan täydentävällä tavalla: 1) louhimalla rakenteetonta tekstiä ja 2) hyödyntämällä näiden aineistojen kuvailutietoja (sisältäen kirjasto-, arkisto- ja museoluetteloinnit). Tiedonlouhinta viittaa menetelmiin, joilla suurista tietomassoista voidaan etsiä tutkimuksen kannalta kiinnostavia rakenteita ja säännönmukaisuuksia myös silloin, kun tutkittavasta aiheesta on saatavissa vain vähän ennakkotietoa ja tutkimushypoteesien tarkka määrittely on haastavaa. Nämä menetelmät eivät korvaa perinteistä lähteiden lähiluentaa vaan täydentävät sitä. Tekstilouhinnan

tutkimuspotentiaali on silti valtava. Jos historioitsija on esimerkiksi kiinnostunut oikeudenmukaisuuden käsitteen muuttumisesta brittiläisessä julkisessa keskustelussa, tätä voidaan tutkia vaikkapa louhimalla kansainvälisiä aineistoja. Näitä ovat muun muassa *Early Modern English Books Online* (EEBO) ja *Eighteenth-Century Collections Online* (ECCO), jotka yhdessä sisältävät käytännössä kaikki varhaisella uudella ajalla (1470–1800) Britanniassa painetut kirjat. Muita vastaavia aineistoja tarjoaa esimerkiksi ranskalainen kansallinen Gallica-projekti.[32]

Samalla louhintaan voidaan yhdistellä käsitehistorian metodeja. Kun käytössä on automatisoituja, laajojen tietoaineistojen käsittelyyn skaalautuvia mallintamismenetelmiä, käsitteen muutosta voidaan ryhtyä tutkimaan aivan eri tavoin verrattuna perinteiseen lähiluentaan. Kotimaisena esimerkkinä on kaikkien historiallisten suomalaisten sanomalehtien tekstilouhinta, jossa aineistoista haetaan paikannimiä ja henkilöitä.[33] OCR-menetelmä (*optical character recognition*), jonka avulla tekstiaineisto digitoidaan koneluettavaan muotoon, ei ole täydellinen; parhaimmassakin tapauksessa sanojen tunnistustarkkuus on vain noin 80 prosenttia. Tunnistusmenetelmiä pyritäänkin kehittämään siihen suuntaan, että ne voivat toimia vakaasti, vaikka aineisto olisikin käsinkirjoitettua tai muuten epästandardimaista.[34] Alan tutkimuskäytännöt muuttuvat nopeasti digitointimenetelmien kehittyessä ja yleistyessä.

AVOIMEN TIETEEN MERKITYS

Digitaalisten aineistojen suurin merkitys on nähty ensisijaisesti saatavuuden parantumisessa. Silti aineistojen avoimuuteen ja käsittelyyn liittyy myös huomattavia periaatteellisia ongelmia. Viime aikoina avoimen tiedon (*open knowledge*) tai avoimen tieteen (*open science*) periaatteet on alettu nähdä laajempana kokonaisuutena, joka tarjoaa runsaasti uusia välineitä, lähestymistapoja ja mahdollisuuksia tehdä historiantutkimusta.

Avoimessa tieteessä korostetaan läpinäkyvien ja yhteisöllisten toimintamallien merkitystä. Käsitteen katsotaan viittaavaan esimerkiksi lähdeaineistojen, tutkimusmenetelmien ja julkaisemisen avoimuuteen. Avoimuus sisältää vapauden tallentaa, muokata ja jakaa edelleen tutkimusmateriaalia tavanomaisten tieteellisten viittauskäytäntöjen mukaisesti ilman muita rajoituksia. Tutkimusaineistojen koon kasvaessa

ja niiden käsittelyyn tarvittavien informaatiotekniikan menetelmien monimutkaistuessa tutkimuksen avoimuus on kuitenkin kohdannut uudenlaisia haasteita. Pelkkä tutkimusprosessin yleinen kuvaus ei vielä takaa, että tutkimus olisi toistettavaa tai että muu tutkimusyhteisö pystyisi rakentamaan sen varaan tehokkaasti uutta tutkimusta. Ihmistieteissä kvantitatiivisen tutkimuksen haasteisiin ei ole vielä kiinnitetty samanlaista huomiota kuin luonnontieteissä, kuten fysiikassa tai biolääketieteessä, joissa avoimuuden puutteeseen liittyvät ongelmat ovat tulleet esiin. Luonnontieteen aloilla tutkimuksen avoimuuden edistämiseen on kehittynyt viime vuosikymmeninä lukuisia uusia tapoja, joista mainittakoon esimerkkeinä avoimet julkaisuarkistot, avoimet tieteelliset laskentakirjastot sekä tutkimusaineistojen keskitetty kokoaminen ja jakelu.[35]

Kun luonnontieteissä eri maiden tutkimuslaitokset jopa kilpailevat avointen tietokantojen laadulla,[36] humanistis-yhteiskuntatieteelliselle tutkimukselle merkittäviä aineistoja ylläpitävät laitokset vaikuttavat jopa olevan välinpitämättömiä edistämään aineistojen avointa käyttöä. Kansainvälisessä tarkastelussa tieteenala onkin avointen käytäntöjen osalta jäänyt jälkeen luonnontieteistä. Historiantutkimuksessa voi törmätä kirjaviin ja jopa virheellisiin käsityksiin tutkimusaineistojen ja julkaisemisen avoimuudesta. Muutosta on haettu muun muassa Opetus- ja kulttuuriministeriön vuosina 2014–2017 toimineen Avoin tiede ja tutkimus (ATT) -hankkeen myötä.[37] Yksi hankkeen tärkeimmistä aikaansaannoksista on huomion kiinnittäminen tekijänoikeuksien ja yksilösuojan merkitykseen ja niistä huolehtimiseen mutta toisaalta myös vaikeuksiin saada aineistoja tutkimuksellisen tiedonlouhinnan käyttöön.[38] Eurooppalainen lainsäädäntö muuttuu tiedonlouhinnan osalta vuoden 2018 aikana.[39]

Avoin data (*open data*) tarkoittaa sitä, että tutkimuksen alkuperäiset tietoaineistot ovat saatavilla joko maksutta tai minimaalisin tuotantokustannuksin ja koneluettavassa muodossa eikä aineistojen jatkokäytölle ole asetettu teknisiä, juridisia, kaupallisia tai muita rajoitteita.[40] Tällaista aineistoa voi siis käyttää, jalostaa, yhdistellä muihin aineistoihin ja jakaa edelleen haluamallaan tavalla. Ainoa ehto avoimen aineiston käytölle on viittaaminen alkuperäiseen lähteeseen tavanomaisen tieteellisen käytännön mukaisesti. Tämä avoimen datan määritelmä on sopusoinnussa tieteen avoimuuden määritelmän ja vakiintuneiden viittauskäytäntöjen kanssa.

Luonnontieteellisen tutkimuksen parissa on toimittu edelläkävijöinä myös aineistojen avoimessa jakelussa. Luonnontieteissä on tiedostettu alkuperäisaineistojen saatavuuden merkitys tutkimuksen läpinäkyvyydelle ja avoimuudelle, joka on yksi tieteellisen toiminnan kulmakivistä. Aineistojen jakamiseen esimerkiksi fysiikassa ja biotieteissä on perustettu keskitettyjä julkisia tietokantoja, joiden pitkän aikavälin rahoitus on turvattu. Humanistis-yhteiskuntatieteellisen tutkimuksen lähdeaineistot on puolestaan koottu erityisesti maailmalla kaupallisen mallin mukaisesti tai esimerkiksi kirjastojen toimesta sillä ajatuksella, että niiden käyttöoikeuksia voidaan myöhemmin myydä. Avoimen julkaisemisen ei ole perinteisesti nähty tuovan merkittävää lisäarvoa. Tulevaisuudessa esimerkiksi perustavanlaatuiset aatehistorian metodologiset lähestymistavat, kuten skinneriläinen kontekstualismi ja koselleckilainen käsitehistoria, voivat kuitenkin näyttäytyä uudenlaisessa perspektiivissä digitoitujen aineistojen hyödyntämisen myötä. Tämä edellyttää, että tutkimuskysymysten ehdoilla etenevä historiantutkimus voi ammentaa kieliteknologiasta, koneoppimisesta ja muista informaatiotieteiden menetelmistä ja suuntauksista. Samalla olisi tärkeää, että yhteistyötä tutkimuksen tietoaineistoja hallinnoivien kirjastojen ja muiden muistiorganisaatioiden kesken syvennetään huomattavasti.

Käytännössä pääseminen historiantutkimukselle keskeisten alkuperäisaineistojen äärelle on osoittautunut vaikeaksi jopa luottamuksellisen tutkimusyhteistyön kautta – puhumattakaan siitä, että samoja aineistoja julkaistaisiin avoimesti verkossa.[41] On hämmentävää, ettei historiallisia aineistoja usein pystytä saamaan tutkimuskäyttöön ilman monimutkaisia sopimusneuvotteluja, vaikka käyttöön ei välttämättä liity lainkaan tekijänoikeus- tai yksilönsuojaongelmia. Tämä on esimerkki siitä, kuinka aineistojen käyttö ja hallinta eivät ole tutkijakeskeisiä, vaan toimintaa säätelee aggressiivinen tekijänoikeuksien haltijoiden edunvalvonta.[42] Keinotekoisten kaupallisten ja teknisten rajoitteiden yleisyys digitoitujen humanistis-yhteiskuntatieteellisten aineistojen yhteydessä poikkeaa merkittävästi luonnontieteellisen tutkimuksen yleisestä käytännöstä, jossa alkuperäiset mittausaineistot tallennetaan julkaisun yhteydessä avoimiin tietokantoihin. Tieteenalojen erilaisia toimintatapoja selittävät erot tutkimuskulttuureissa, aineistojen jakelun järjestämisessä ja rahoitusmalleissa. Toivottavasti esimerkiksi Kansalliskirjasto onnistuisi tekemään digitaalisten aineistojen käytöstä ja tätä

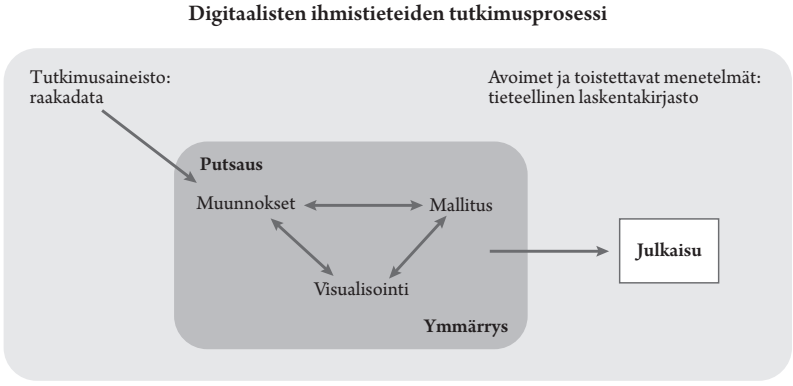
varten rakennettavien työkalujen hyödyntämisestä nykyistä ketterämpää. On tärkeää, että tutkimuksen mahdollistaminen etenee myös Suomessa mahdollisimman suoraviivaisesti.

Avoimuuden puutteet muodostavatkin tällä hetkellä huomattavan pullonkaulan humanistis-yhteiskuntatieteellisessä tutkimuksessa, ja ne voidaan kiteyttää kolmeen keskeiseen ongelmaan. Ensinnäkin digitoidut lähdeaineistot ovat usein saatavilla vain kaupallisten lisenssien tai luottamuksellisten yhteistyösopimusten kautta. Toiseksi monet aineistot ovat saatavilla ainoastaan rajatun verkkoliittymän kautta, eikä tutkija välttämättä pääse käsiksi koko aineistoon ja alkuperäisiin lähdeaineistoihin, joita tarvittaisiin kattavan kokonaiskuvan muodostamiseen. Yksityiskohtaisempi ja alaa uudistava tekstilouhinta edellyttää pääsyä koko alkuperäisaineistoon ja mahdollisuuksia sen vapaaseen jatkokäsittelyyn. Tähän liittyy kolmas ongelma eli aineistojen käsittelyyn räätälöityjen joustavien tieteellisten laskentakirjastojen puute. Valmiit graafiset käyttöliittymät eivät voi koskaan tarjota samaa joustavuutta aineistojen analysointiin kuin tutkijan itse kirjoittamat ohjelmakoodit, joiden laadinnassa on mahdollista hyödyntää laskentakirjastojen kautta saatavia valmiita rakennuspalikoita.

Datan avoimuuden lisäksi keskeistä tutkimuksen läpinäkyvyydelle ja uudelleenkäytölle onkin tutkimusmenetelmien avoin julkaiseminen.[43] Avoimesta lähdekoodista keskusteleminen saattaa tuntua historiantutkijoille vieraalta, mutta digitaalisten menetelmien käytön lisääntyessä avointen toimintamallien edistäminen on tärkeää. Käytäntö on nopeasti yleistynyt erityisesti luonnontieteissä, joissa vertaisarvioitujen tutkimusten laadintaan liittyviä lähdekoodeja julkaistaan hajautettujen versionhallintapalveluiden kautta.[44]

Kvantitatiivisen tutkimuksen ohessa syntyy aineiston jalostamiseen ja tilastolliseen analysointiin soveltuvaa lähdekoodia, joka voidaan järjestää tieteelliseksi laskentakirjastoksi ja jota on mahdollista hyödyntää myöhemmin muiden vastaavien aineistojen tutkimuksessa. Avoin tieteellinen laskentakirjasto (Kuva 9) pitää sisällään yhteisöllisesti kehitettäviä avoimen lähdekoodin välineitä aineistojen tutkimuskäyttöön. Jokainen analyysiprosessin vaihe alkuperäisen avoimen raakadatan putsaamisesta lopullisiin yhteenvetoihin on läpinäkyvä. Tämä vähentää mahdollisuuksia vääristymien piilotteluun ja parantaa tutkimuksen laatua.[45] Kun alan tutkija voi rakentaa valmiista laskentakirjastoista omiin tarpeisiinsa sopivan kokonaisuuden, rutiininomaiset työvaiheet

vähenevät. Historiallisissa lähteissä esimerkiksi kirjastoluetteloiden kuvailutiedot ovat varsin yhtenäisiä ylläpitäjästä riippumatta. Tämä tarjoaa mahdollisuuksia eri maiden kirjastoluetteloiden yhdistämiseen ja tiedontuotannon tutkimiseen kansainvälisesti.



Kuva 9. Datasta tietoon. Alkuperäisten tutkimusaineistojen jalostaminen tiedoksi ja lopulliseksi julkaisuksi on monivaiheinen prosessi, jossa tarvitaan välineitä aineistojen järjestämiseen, yhdistelyyn ja tilastolliseen analysointiin.[46] Merkittäviä osia työstä voidaan automatisoida laatimalla tähän tarkoitukseen suunnattuja tieteellisiä laskentakirjastoja. Käytettyjen lähdekoodien avoin julkaiseminen tekee tutkimusprosessista toistettavan ja läpinäkyvän, mahdollistaa datan ja menetelmien jatkokäytön ja edistää laajemman tutkijaverkoston osallistumista kehitystyöhön. Lopullinen julkaisu voikin sisältää tieteellisen analyysin lisäksi myös tutkimusaineistoja ja tilastollisia tutkimusmenetelmiä.

Avoimuus parantaa myös tutkimuksen toistettavuutta ja läpinäkyvyyttä. Nykytilanteessa tutkija saattaa tehdä suljettujen käyttöliittymien avulla monimutkaisiakin tietokantahakuja ja muodostaa niiden perusteella heuristisia päättelyketjuja tiettyjen termien käytön historiallisesti kehityksestä, mutta tehdyistä hauista ja niiden sisällöstä tai päättelyketjusta ei välttämättä jää edes tutkijalle itselleen selkeää kirjanpitoa. Automatisointi tarjoaa tähän ratkaisun. Avoimuuden ansiosta muutkin voivat hyödyntää samoja ratkaisuja omassa työssään, jatkokehittää niitä omiin tarpeisiinsa ja antaa tulokset takaisin yhteisön käyttöön. Tiedeyhteisön näkökulmasta tällainen malli voi vähentää merkittävästi päällekkäistä työtä ja tarjota mahdollisuuksia verkostomaiseen, perinteiset

institutionaaliset ja maantieteelliset rajat ylittävään yhteistyöhön. Tämä hyödyttää kaikkia alan toimijoita ja vauhdittaa tutkimusta. Onkin tärkeää, että historiantutkimuksessa voidaan ennakkoluulottomasti kokeilla myös datatieteen kaikkia parhaita käytäntöjä.

Avoimuus voi uudistaa tieteenalan toimintakulttuuria ja suuntaa. Kansainvälisessä kilpailussa on pystyttävä tekemään hajautettua yhteistyötä, jossa yhteisten pelisääntöjen kautta pyritään avaamaan koko tutkimusprosessi niin, että samoista tai samankaltaisista kysymyksistä kiinnostuneiden tutkijoiden ei tarvitse keksiä pyörää uudestaan vaan yhteisvoimin voidaan kehittää rikkaampia ja laadukkaampia kokonaisuuksia. Luonnontieteen alojen tutkimuksessa tämä on usein jo arkipäivää avointen tietokantojen ja tutkimusyhteisön kehittämien ohjelmistokirjastojen myötä. Digitaalisten ihmistieteiden alalla kotimaisten tutkimusryhmien kannattaisikin tehdä yhteistyötä esimerkiksi suomalaisten sanomalehtiaineistojen analysoinnissa, jotta niukat tutkimusresurssit eivät valuisi hukkaan tutkimusryhmien tehdessä päällekkäistä työtä.

Tinkiminen mistä tahansa avoimuuden osa-alueesta heikentää koko tutkimusprosessin avoimuutta. Humanistisilla aloilla puhutaan paljon avoimesta julkaisemisesta mutta vähemmän tutkimusaineistojen ja tutkimusprosessien avoimuudesta, vaikka jälkimmäinen avoimuus saattaa olla tutkimuksen kehittymisen kannalta jopa edellistä olennaisempaa. Paradoksaalisesti tätä merkitystä ei aina havaita, kun perinteisesti ajatellaan, että tiedeyhteisön kannalta hyödyllisin julkaisuyksikkö on lopullinen, puntaroitu ja viilattu tutkimusjulkaisu. Kuitenkin myös tutkimusmenetelmät ja tutkimusprosessin aikana käytävä ajatustenvaihto ovat keskeisiä tutkimusprosessin osia. Näiden avoin jakaminen tiedeyhteisön kesken tarjoaa laadullisten tulosten lisäksi arvokasta materiaalia ja välineitä, joiden pohjalta uutta tutkimusta voidaan rakentaa tehokkaasti. Avoimia käytäntöjä omaksuneilla tieteenaloilla on hyödynnetty Linux-maailmasta tuttua ”julkaise aikaisin ja usein” -periaatetta, jonka mukaan koko kvantitatiivinen tutkimusprosessi ja sen eteneminen vertaisarviointeineen avataan yleisölle.[47] Avoimuuden merkitystä tutkimusetiikan ja metodiikan ytimessä tuleekin korostaa. Näin tutkimusprosessi ja -tulokset liikkuvat sujuvammin tutkijoiden ja tutkimusalojen välillä, jolloin niitä voidaan paremmin hyödyntää opetuksessa, kansalaisyhteiskunnan toiminnassa sekä muussakin innovaatiotoiminnassa. Avoimuuden tutkimukselle ja yhteiskunnalle tuottama lisäarvo onkin arvioitu huomattavasti avoimuuden edistämisen kustannuksia suuremmaksi.[48]

DIGITAALISUUDEN ASEMASTA HISTORIAN TUTKIMUSTRADITIOSSA

Historiantutkimuksessa on oltu pidättyväisiä digitaalisten aineistojen käytölle ja tiedonlouhinnalle, vaikka esimerkiksi kieliteknologian menetelmät ovat kehittyneet jo ainakin kolmenkymmenen vuoden ajan humanistisen tutkimustradition parissa.[49] Kärjistäen voidaan todeta, että perinteisessä historiantutkimuksessa yksittäinen sankaritutkija lukee ja kirjoittaa, eikä tilastopuuhastelun nähdä välttämättä tuovan mainittavaa lisäarvoa. Itsenäistä asiantuntemukseen pohjautuvaa luovaa ajattelua mikään kone ei voikaan tutkijan puolesta tehdä. Toisaalta tutkijat eivät ole välttämättä tietoisia datatieteen avaamista uusista tutkimusmahdollisuuksista. Historian alalla digitaalisten ihmistieteiden sovellusten puute selittyy osin sillä, että historiantutkimuksen ydinosaajat eivät ole lähteneet kehittämään menetelmiä. Näin ei ole päässyt edes syntymään historian tarpeisiin vastaavia, tietojenkäsittelyä hyödyntäviä menetelmiä. Laskennallisten tieteiden edustajat eivät välttämättä edes osaa ennakoida historioitsijoiden tarpeita. Vaikka joitakin menetelmiä on jo kehitelty, eri tutkimuskulttuurien kohtaamisessa on vielä paljon työsarkaa.[50] Uusi tutkimusala tarjoaakin lupaavia mahdollisuuksia ja tuoreita tutkimuskohteita myös menetelmätutkijoille.

Tutkimusta voitaisiin monissa tapauksissa tehostaa ottamalla käyttöön tilastollinen näkökulma laajempien aineistojen tarkasteluun. Digitaalisten aineistojen käytöllä voidaan ainakin laajentaa tutkimuksen ja muutosten havaitsemisen aikajännettä sujuvasti vuosikymmenistä vuosisatoihin.[51] Tämänsuuntainen ajattelu ei ole uutta, ja ainakin käsitehistorian piirissä tätä mahdollisuutta on myös Suomessa pohdittu jo pidempään.[52] Mahdollisuuksien realisoituminen käytännön historiantutkimuksessa vaatii kuitenkin monitieteistä yhteistyötä ja muutoksia tutkimuskulttuuriin.

Uusien lähestymistapojen ilmaantuminen nostaa esille tarpeen korostaa humanistisen tradition omien tutkimuskysymysten merkitystä ja hahmottaa, mikä on alan perustutkimuksen ydin. Tutkimuksen lopullisen arvon sanelevat käytännön tulokset ja sen välittämä ymmärrys menneisyydestä. Mikään tutkimusmenetelmä tai -aineisto ei silti korvaa hyviä kysymyksiä ja hypoteeseja, joiden onnistunut määritteleminen ja niihin vastaaminen saatavilla olevien aineistojen perusteella ratkaisee viime kädessä onnistumisemme tutkijoina.

Kuvailutietojen käyttämisen, tekstilouhinnan ja datatieteiden välineiden hyödyntämisen yhdistäminen tarjoaa uusia tapoja ymmärtää esimerkiksi painotuotteita fyysisinä objekteina, niiden tuottamiseen liittyviä näkökulmia ja laajempia käsitteiden muutoksia. Digitaalisten ihmistieteiden käsite on tarkoitettu väliaikaiseksi; tällä hetkellä se luonnehtii tietyntyyppistä digitaalisilla aineistoilla tehtävää tutkimusta, jonka varsinaisena päämääränä on kuitenkin uudistaa humanistista tutkimusperinnettä. Todennäköistä on, että digitaalisten aineistojen käyttö leviää suhteellisen nopeasti tutkimuksen valtavirtaan myös historia-alalla. Tällöin voidaan palata puhumaan pelkästään historiantutkimuksesta ja humanismista ilman digitaalisuutta tai muita etuliitteitä. Yksittäisten tutkimusperinteiden jopa vuosisataiset jatkumot eivät katoa vaan vain ammentavat menetelmäkehityksen uusista mahdollisuuksista tuoreita näkökulmia.

Historiantutkimukselle käyttökelpoisia digitaalisia aineistoja on jo runsaasti saatavilla. Yhden kiintoisan tutkimuskohteen tarjoavat kotimaiset aineistot, esimerkiksi Kansalliskirjaston kokoelmat. Haasteena onkin kehittää menetelmiä, joilla tekstiaineistoja yhdistäviä ja erottavia piirteitä voitaisiin sujuvasti analysoida yli kielirajojen. Suomen kaksikielisyys tarjoaa erinomaisia mahdollisuuksia pilotoida rajatussa mittakaavassa eri kielten käyttöä ja yhdistelyä tekstilouhinnassa. Suomalaisten sanomalehtien tutkimuksessa voidaan yhdistää monia kotimaisen historian osaamisen parhaita puolia, kuten käsitehistorian ja lingvistiikan traditioita. Lisäksi operoiminen sisällöllisesti tutun aineiston parissa tarjoaa erinomaisia mahdollisuuksia kehittää yhteistyötä tieteellisen laskennan asiantuntijoiden kanssa. Tämä voi tuoda uusia ideoita sekä ihmistieteiden tutkimukseen että datatieteen menetelmäkehitykseen.

Ihmistieteissä seuraavat tavoitteet voisivat liittyä kvantitatiivisten tutkimusmenetelmien yhtenäistämiseen ja yhteistyön edistämiseen luonnontieteissä jo pidemmälle vakiintuneen avoimen toimintamallin keinoin. Samalla tulee kehittää alan opetusta, jotta uudet tutkijasukupolvet kykenevät haastamaan aikaisempia historiantulkintoja tutkimusalan omista lähtökohdista ja osaavat ammentaa lisäarvoa monitieteisestä yhteistyöstä. Olisi hyvä tehdä historiantutkimuksen ja digitaalisten ihmistieteiden yhteisiä linjauksia alan kotimaisen opetuksen toteuttamisesta. Humanistisen tutkimustradition sisällä suurin

muutos liittyy uusien tutkimusmenetelmien ja aineistojen käytettävyyden edistämiseen. Tässä avoimen tieteen puolestapuhujien korostamat ajatukset ovat keskeisiä. Tarvitaan konkreettisia ja pitkäjänteisiä toimia alan toimintakulttuurin uudistamiseksi. Tieteellisen laskennan asiantuntijoita on saatava mukaan ihmistieteiden pariin sellaisella tavalla, joka parhaiten palvelee ihmistieteiden perinteisten ydinteemojen tutkimusta.*

* Kirjoitettu osana Suomen Akatemian hankkeita *Digitaalinen historiantutkimus ja julkisuuden muutos Suomessa 1640–1910* (293316) ja *Ihmisen suolistomikrobiston ekologinen mallitus: yksilöllisyys, dynamiikka ja toiminnallisuus laajoissa väestötutkimuksissa* (295741).

1. Lappalainen & Lappalainen 1983, 75–78.
2. Luku on muokattu aikaisemmin julkaistusta artikkelista ”Aatehistoria ja digitaalisten aineistojen mahdollisuudet”, Tolonen & Lahti 2015. Tekstiä, esimerkkejä ja viitteitä on päivitetty ja sovitettu historian metodiopetuksen tarpeisiin. Kiitämme Teemu Ojannetta luvun erinomaisesta toimittamisesta.
3. Digitaalisten metodien kirjosta historiantutkijoille ks. Graham ym. 2014. Tämän kritiikistä vastaavasti ks. Hitchcock 2014.
4. Rasila 1967, 140–146. Rasila oli kiinnostunut myös tilastollisten menetelmien käytöstä historiankirjoituksessa, ks. Rasila 1977. Digitaalisesta historiantutkimuksesta tehtiin keväällä 2016 Petri Pajun johdolla kartoitus, johon osallistui 17 vastaajaa, eli otos on

melko suppea ja tuloksista on hankala tehdä pidemmälle meneviä johtopäätöksiä, Paju 2016. Syksyllä 2016 kartoitettiin Inés Matresin toimesta ja Mikko Tolosen johdolla yleisemmin humanistisen alan digitaalisten metodien käyttöä Suomessa, vastaajia oli 239, ja nämä tulokset kannattaa yhdistää edelliseen, Matres 2016.

5. Kiitämme anonyymia arvioitsijaa Jussi T. Lappalaisen mainitsemisesta.
6. Aiheesta *digital humanities* ks. esim. Gold 2012.
7. Mielenkiintoinen kehitys on aiemmin suositun *humanities computing* -termin suhde *digital humanities* -termin käyttöön. Ensin mainittua termiä on määritelty 1998–2006, ks. esim. McCarty 1998; Unsworth 2000; Orlandi 2002; Hockey 2004; McCarty 2005 & Short 2006; University College of Londonin sivustolla listattuja *digital humanities* -alaa määritteleviä blogikirjoituksia vuosilta 2014–16, UCL 2016; laajempana katsauksena ks. Haverinen & Suominen 2015.
8. Samasta aiheesta ks. Elo 2016.
9. Tästä laajempien aikajaksojen tutkimuksesta ks. Guldi & Armitage 2014. On hyvä huomioida, että kyseinen teksti on saanut osakseen myös kritiikkiä.
10. Monitieteisyyden merkitys korostuu digitaalisissa ihmistieteissä. Esimerkki tästä on digitaalisten ihmistieteiden opintokokonaisuus Helsingin yliopistossa, jossa koulutuksen tähtäimenä on pikemminkin monitieteinen yhteistyö kuin yksittäisten tutkijoiden varustaminen tutkimusprosessin kaikkien osien hallintaan, HELDIG 2017.
11. UTA 2016.
12. Digitaalisten ihmistieteiden merkityksestä ks. esim. Svensson 2010; 2012 ja 2016.
13. Tästä aiheesta ks. Häyrinen 2012. Kiitämme Jaakko Suomista vinkistä.
14. Modaliteettien ja materiaalisuuden merkitys analoginen/digitaalinen-debatissa on ollut merkittävä erityisesti museoiden kontekstissa, ks. esim. Cameron 2007.
15. Concept Lab 2016; hankkeen teoreettisesta asetelmasta ks. de Bolla 2013.
16. Lund University 2016.
17. Brett 2012.
18. ADHO 2016. Tutkimalla vuosittaisen, kansainvälisesti suurimman ADHO-yhteisön organisoiman *digital humanities* -konferenssin abstrakteja vuodelta 2016 käy esimerkiksi ilmi, että alalla kytkentä historiallisten aiheiden kohdalta varsinaiseen historialliseen tutkimustraditioon ei ole mitenkään itsestään selvää, vaikka useat aiheet itsessään olisivat historiallisia.
19. Kansainvälisinä keihäänkärkinä historiantutkimuksen kannalta digitaalisissa ihmistieteissä tulee mainita ainakin Frédéric Kaplanin johtama Digital Humanities Laboratory (DHLAB 2016) ja Stanfordin yliopiston CESTA, jossa on viety läpi useita historiantutkimusta muokkaavia projekteja, CESTA 2016.
20. Suomi24-aineistosta historiantutkimuksen kannalta ja digisyntyisten ilmiöiden historiantutkimuksesta perusteellisesti ks. Suominen & Sivula 2016.
21. Tolonen & Lahti, 2015; ks. myös Purhonen & Toikka 2016.
22. Lukuisista aihetta käsittelevistä kokoelmista ks. esim. Deegan & McCarty 2012 ja Crompton, Lane & Siemens 2017.
23. Hyvä yksinkertainen esimerkki Berhane 2016.
24. Samalla metodit kehittyvät kiihtyvää vauhtia, ks. AHDO 2016.
25. Pohjoismaisista SmartCity-hankkeista ainakin Aarhusissa humanisteilla on Dariah-verkoston kautta ollut selvä rooli, ks. Smart Aarhus 2016. Hankkeesta enemmän, ks. Smart Culture 2014.
26. Ks. Tolonen & Lahti 2015.

27. Suomen Akatemian hanke ”Digitaalinen historiantutkimus ja julkisuuden muutos Suomessa 1640–1910” (hankenro 293316).
28. Octavo 2017.
29. Klinge 1989; 2012.
30. Ks. esim. Hakapää 2008.
31. Tämä arvio perustuu COMHIS-akatemiahankkeen yhteydessä tehtyyn työhön kirjastoluettelojen siivoamisessa, jotta niillä on päästy tekemään varsinaista tilastollista analyysia ja tutkimusta. Varoittava esimerkki vastaavien aineistojen käytöstä ilman esikäsittelyä ks. GDELT 2016. Kyseisessä tapauksessa dataa on ollut riittävästi ja visualisoinnit ovat olleet vastaavia kuin muissakin projekteissa, mutta jos lähdeaineistoon ei voida luottaa, sama pätee myös johtopäätöksiin. Ks. myös Lahti, Ilomäki & Tolonen 2015 ja Tolonen ym. 2016.
32. Latinankielisestä aineistosta yleisemmin ks. Brepolis 2016.
33. DIGI 2016.
34. Kettunen, Pääkkönen & Koistinen 2016.
35. Ioannidis 2005.
36. Gardiner-Garden & Littlejohn 2001.
37. Avoin tiede 2016.
38. Still ym. 2016.
39. European Commission 2016.
40. Molloy 2011.
41. Esimerkiksi Dewey-järjestelmän (kirjastoaineiston klassinen luokittelujärjestelmä) saaminen tutkimuskäyttöön on osoittautunut erittäin hankalaksi. Samanlaisia kokemuksia liittyy niin ECCO-aineiston raakadatan kuin myös eri kirjastoluetteloiden marc-tiedostojen saamiseen tutkimuskäyttöön.
42. Tästä keskustelusta ks. esim. Salmi & Tolonen 2017.
43. Morin ym. 2012; Ioannidis 2014.
44. Esim. GITHUB 2017, GITLAB 2017 ja Bitbucket 2017.
45. Head ym. 2015.
46. Hadley & Grolemund 2017; avoimen tieteen tutkimusprosessin kehityksestä ks. myös Lahti ym. 2017.
47. Ks. Ioannidis 2005.
48. Manyika ym. 2011; The World Bank 2014.
49. Helsingin yliopistossa esim. Kimmo Koskenniemi, Koskenniemi 1984.
50. Mielenkiintoista kehitystä ja seurattavaa toki riittää, kuten esim. Schmidt 2016.
51. *The History Manifesto* 2016; Goethe Universität 2016; ks. myös Armitage 2012.
52. Ks. esim. Ihalainen 2005; 2010 sekä Marjanen 2013.