KATSAUS

Kimmo Elo

Pohdintoja laskennallisten ja digitaalisten ihmistieteiden suhteesta

ABSTRAKTI / ABSTRACT

Kokoomateos "Information flows across the Baltic Sea" herättää lukijassa kysymyksen digitaalisen ja laskennallisen tutkimusorientaation välisestä suhteesta historiantutkimuksen piirissä. Vaikka näitä käsitteitä tunnutaan käytettävän usein synonyymeinä, niiden välillä on sekä paradigmaattinen että metodologinen ero, jolla on olennainen vaikutus tutkimusprosessiin. Siinä missä digitaalisuus on eräälainen sateenvarjokäsite, laskennallisuus paradigmana pyrkii tuottamaan parempaa ymmärrystä monimutkaisesta maailmasta informaatiota mallintamalla ja prosessoimalla. Digitaalisen ja laskennallisen välisen eron korostaminen ei siten ole akateemista hiustenhalkomista, koska tällä jaottelulla on perustavanlaatuinen merkitys sille, millaista tutkimusprosessia tutkija tutkimansa aiheen ympärille aikoo rakentaa.

Digitaaliset ihmistieteet, laskennalliset ihmistieteet, mediahistoria, historiantutkimus, metodi, paradigma

Kimmo Elo on Turun yliopiston eduskuntatutkimuksen keskuksen vanhempi erikoistutkija. Hänen tutkimusalojaan ovat Saksan ja Euroopan historia, parlamenttitutkimus sekä laskennalliset ihmistieteet. Hän toimii Suomen Akatemian rahoittaman LAWPOL-tutkimusinfrastruktuurihankkeen (2023–2025) varajohtajana ja eduskuntatutkimuksen keskuksen LAWPOL-osahankkeen PI:na. Sposti: kimmo.elo@utu.fi

Johdanto

Sanomalehdet ovat jo vuosisatojen ajan olleet yksi merkittävimmistä ajankohtaisten yhteiskunnallisten tapahtumien peilaajista, mikä selittää myös niiden keskeistä asemaa yhtenä yhteiskuntahistorian tutkimuksen keskeisimmistä aineistoista. Sanomalehtien kautta on ollut mahdollista myös jälkikäteen kurkistaa tietyn aikakauden yhteiskunnalliseen moninaisuuteen ja saada tietoa asioista, jotka ovat tavalla tai toisella olleet merkityksellisiä laajoille yhteiskuntaryhmille. Sanomalehtien merkitys menneisyyden dokumentoijana selittänee osaltaan myös niitä laajoja digitointihankkeita, kiitos joiden menneisyyden

arkipäivään on tänään mahdollista kurkistaa myös oman tietokoneen näytöltä ja hyödyntämällä digitaalisia tutkimustyökaluja.

Kokoomateos *Information Flows across the Baltic Sea. Towards a Computational Approach to Media History*[1] pureutuu ruotsinkielisten sanomalehtien historiaan digitaalisen mediahistorian näkökulmasta. Teoksen kaikki tutkimusartikkelit ovat syntyneet osana samannimistä tutkimusprojektia, johon osallistui tutkimusryhmiä Turun ja Helsingin sekä Örebron että Uumajan yliopistoista. Teoksen aloittava, teoksen toimittajien yhdessä kirjoittama johdantoartikkeli yhdistää tiiviin katsauksen ruotsalaiseen lehdistöhistoriaan *text reuse* -tietokannan perusteiden ja peruskäytön esittelyyn. Varsin tiiviistä – ja etenkin tietokannan osalta paikoin teknisestä – esitystavasta huolimatta artikkeli onnistuu paketoimaan molemmat teemat selkeälukuiseksi kokonaisuudeksi. Ratkaisua voi pitää onnistuneena myös siksi, että käsittelemällä kaikille artikkeleille yhteiset tausta-asiat johdantoluvussa, teoksen tutkimusartikkeleille jää kullekin enemmän tilaa varsinaiselle empiiriselle analyysille. Näin vältetään monia kokoomateoksia vaivaava toisto ja päällekkäisyys juuri aineisto- ja menetelmäkuvailujen osalta.

Teoksen tutkimusartikkelit liikkuvat laajalla temaattisella alueella, kun teoksessa käsitellään yleisen uutisvälityksen lisäksi muun muassa sellaisia kapeampi teemoja kuten raittius, visuaalinen viihde tai ihmelääkkeitä esitelleet ”puoskarit”. Jokainen artikkeli avaa kiinnostavia historiallisia kurkistusikkunoita sanomalehtien muodostamaan mediahistorialliseen ekosysteemiin ja alleviivaa sanomalehtien keskeistä merkitystä paitsi tiedonvälityksen, myös eri aikakausina vallinneiden yhteiskunnallisten virtausten peilinä. Toisaalta, vaikka kaikki artikkelit osaltaan valottavat *text reuse* -tietokannan mahdollisuuksia ja rajoitteita, ovat artikkelit sen verran heterogeenisiä sekä teemoiltaan että kontekstiltaan, että niitä yhdistää suurelta osin vain tekstien uudelleenkäyttö ajassa ja paikassa.

Vaikka olen pitkälti samaa mieltä *Information flows* -kokoomateoksen kirjoittajien kanssa siitä, että ilman modernia tietojenkäsittelytiedettä tekstien uudelleenkäytön tutkiminen laajassa mittakaavassa ei olisi ollut mahdollista, teos nostaa esille monissa muissakin yhteyksissä laajaa keskustelua herättäneen kysymyksen laskennallisten menetelmien suhteesta eri tieteenalojen omaleimaisuuteen. Samalla kyse on laskennallisen ja digitaalisen välisestä suhteesta eli siitä, käytetäänkö näitä kahta synonyymeina vai vallitseeko niiden välillä paradigmaattinen ero. Siinä missä digitaalinen on vakiinnuttanut asemansa myös historiatieteiden sisäisissä keskusteluissa jo 2000-luvun ensimmäisellä vuosikymmenellä,[2] suhde laskennallisuuteen on ollut kompleksimpi. Laskennallisuus on paradigmana digitaalista kapeampi ja voimakkaammin datatieteisiin orientoituva, mikä on herättänyt vahvoja, tieteenalaidentiteettiin liittyviä kysymyksiä. Kysymys tämän jännitteisen suhteen vaikutuksista tutkimukseen ei ole mielestäni merkityksetön tilanteessa, jossa digitaalisuus yhä enemmän työntyy myös ihmistieteisen tutkimuksen kenttään tutkimusparadigman muodossa. Nyt käsillä oleva teos tarjoaa oivallisen lähtökohdan tämän suhteen pohtimiselle.

Laskennallisuus vs. digitaalisuus

Information Flows across the Baltic Sea -teos koostuu yhteensä yhdeksästä (9) artikkelista, jotka kaikki perustuvat *Svenska Litteratursällskapet i Finland* -seuran ylläpitämään avoimeen tietokantaan (https://textreuse.sls.fi/). Tietokanta on rakennettu digitoidun, ruotsinkielisen sanomalehtikorpuksen pohjalta. Korpus kattaa usean sadan vuoden aikajänteen vuodesta 1645 vuoteen 1918. Tietokannan ja sen varaan rakentuvien hakutoimintojen toteuttamisen haasteena on ollut automaattisen tekstintunnistuksen (OCR, *optical character recognition*) vaihteleva laatu, joka näkyy – omien testihakujeni perusteella arvioituna – loppukäyttäjälle aineiston ”kohinana” eli virheellisesti tunnistettuina merkkeinä,

jotka tekevät alkuperäisten artikkelien digitoiduista kokoteksteistä paikoin mahdottomia ymmärtää. BLAST-menetelmän sovelluksena toteutettu hakualgoritmi näyttäisi kuitenkin toimivan varsin hyvin myös ”kohinaisella” datalla, minkä saavuttaminen on varmasti vaatinut innovatiivisuutta algoritmien kehittäjiltä ja on siten kiistämättä sulka turkulaisten tietokonelingvistien hattuun.

Vaikka BLAST-menetelmää voidaan kiistatta pitää laskennallisena menetelmänä, tämän artikkelin kannalta olennaisempi kysymys liittyy siihen, millainen rooli laskennallisilla menetelmillä on tutkimusartikkeleissa. Jos kysymystä lähestytään hyödyntämällä Anna Haverisen ja Jaakko Suomisen jo vuonna 2015 hahmottelemaa digitaalisen tutkimuksen nelikenttää[3], *Text reuse* -tietokantatyökalu on ensisijaisesti teknologiaa hyödyntävä digitaalinen työkalu, instrumentti, jonka kehittäminen on tullut mahdolliseksi – ja toki myös tarpeelliseksi – sanomalehtien digitoinnin edetessä sellaiselle tasolle, ettei aineistomassaa ole enää mielekästä tai mahdollista hallita manuaalisesti. *Text reuse* -tietokannan kehitystyön tapahtuminen selkeästi fokusoituneen tutkimushankkeen ja valmiiksi digitoidun aineistokorpuksen asettamissa raameissa näkyy myös siinä, että työkalua on voitu kehittää tiettyjen, selkeästi määriteltyjen tavoitteiden pohjalta, jotka tavoitteet kuitenkin ammentavat ensisijaisesti taustalla olevan tutkimushankkeen tarpeista. Hakuportaali mahdollistaa tekstilohkojen uudelleenkäytön tutkimisen, jolloin tutkija pystyy tarkastelemaan tietyn termin, käsitteen tai fraasin leviämistä sekä ajassa että maantieteellisesti. Aihepiirin tutkijoille hakuportaalista on eittämättä suuri apu, vaikka kiinnostus kohdistui vain tietyn tekstin esiintymiseen tietokantaan kuuluvissa sanomalehdissä. Painotus on siis ollut, Haverisen ja Suomisen terminologiaa lainaten, teknologian kehittämisessä välineenä.

Claudio Cioffi-Revillan määritelmään nojautuen laskennallisuus - siis *computational* – tarkoittaa paradigmaattista näkökulma, joka koostuu sisällöllisestä (substantiaalisesta) ja metodologisesta ulottuvuudesta ja korostaa informaation prosessoinnin keskeisyyttä laskennallisille ihmistieteille.[4] Substantiaalisuus viittaa informaation prosessoinnin keskeiseen merkitykseen, kun haluamme selittää ja ymmärtää monimutkaista sosiaalista todellisuutta. Metodologinen ulottuvuus puolestaan alleviivaa tietokonepohjaista laskentaa (*computing*) keskeisenä välineenä mallintaa monimutkaista sosiaalista todellisuutta, jotta ymmärtäisimme sitä paremmin. Keskeistä tässä on, että laskennallisuuden odotetaan tuottavan ontologista lisäarvoa suhteessa muihin lähestymistapoihin ja siten parantavan ymmärrystämme kohteena olevista ilmiöistä. Mutta yhtä olennaista on myös analyysin perustuminen data-analytiikan ja/tai tietojenkäsittelytieteiden piirissä kehitettyjen menetelmien soveltamiseen ihmistieteiden kentässä.[5] Esimerkkeinä laskennallisten menetelmien kautta saavutetusta uudesta historiallisesta näkökulmasta voi nostaa esille Suomen sisällissotaan osallistuneiden punaisten naisten tarkastelun verkostoanalyysin välinein[6] sekä tekstinlouhinnan avulla saavutetut uudet näkökulmat parlamenttipuheeseen.[7]

Laskennallinen näkökulma on kuitenkin läsnä vain teoksen avaavassa johdantoartikkelissa, mutta siinäkin vain BLAST-työkalua käsittelevässä osassa. Teoksen muissa artikkeleissa tietokanta on lähinnä käyttöliittymä aineistoon, jonka avulla sekä rajataan lähiluettavaa aineistomassaa että tuotetaan kuvailevia analyyseja tutkijan tulkittavaksi. Haverisen ja Suomisen nelikenttään projisoitaessa näissäkin suhde teknologiaan säilyy välineellisenä, ontologisen intressien siirtyessä pyrkimykseen luoda työkalun tuottamien havaintojen avulla ymmärrystä sanomalehdistön roolista ja merkityksestä informaation levittäjänä. Vaikka kaikissa artikkeleissa analyysien perustana ovat *Text reuse* -tietokantahaun tulokset, yhdenkään artikkelin kirjoittaja tai kirjoittajat eivät prosessoi tätä informaatiota laskennallisin tai data-analyyttisin menetelmin, vaan kaikki jatkoanalyysit tapahtuvat laadullisen lähiluennan ja aiemman tutkimuksen kanssa käytävän reflektoinnin muodossa.[8] Toisin sanoen, vaikka aineistojen keruu tapahtuu varsin oivaltavaan algoritmiin perustuvaa digitaalista työvälinettä hyödyntäen, varsinainen analyysi noudattaa historiantutkimuksen perinteistä metodiikkaa.

Juuri tässä kohdassa laskennallinen yhteiskuntatieteilijä minussa nosti kriittisen sormensa väittäen, että tiukasti tarkastellen tämä teos kyllä edustaa *digitaalisia* ihmistieteitä, mutta ei – jos asiaa tarkastellaan edellä esitellyn määritelmän valossa – *laskennallisia* ihmistieteitä. Hieman yksinkertaistaen ilmaistuna kyse on siitä, käytetäänkö laskennallisia menetelmiä tieteenalaspesifien kysymysten ja tutkimusongelmien ohjaamana vai edellyttääkö näiden menetelmien käyttö tutkimusongelmien sovittamista menetelmien asettamiin reunaehtoihin. Nyt käsillä ollut teos jättää tässä suhteessa ristiriitaisen olon. Yhtäältä vaikuttaa ilmeiseltä, että tekstien uudelleenkäyttö ja kierrätys on mediahistorian tutkimuksessa tärkeä osa-alue, johon mahdollisuus hyödyntää suuria, pitkän aikajänteen kattavia media-aineistoja tuo merkittävää lisäarvoa suhteessa kapeamman fokuksen lähiluentaan perustuvalle tutkimusmenetelmälle. Tästä näkökulmasta tarkasteltuna *Text reuse* -työkalut näyttäisivät tulleen kehitetyiksi juuri mediahistorialle tieteenalana ominaisten kysymysten tutkimisen tueksi.

Toisaalta, kaikissa artikkelissa *text reuse* -tietokannasta saadut tulokset esitetään sellaisinaan ja pääosin hakuportaalin tarjoamin esitystavoin. Toki artikkeleissa tuloksia tulkitaan hyvinkin monipuolisesti aikalaiskontekstissa, mutta lähtökohtaisesti hakutuloksia näytetään pidettävän sekä relevantteina että valideina, mutta myös edustavina valitun hakufraasin osalta. Kriittisempää pohdintaa olisin toivonut etenkin sen osalta, miten paljon ”kohinaisesta” datasta jää löytämättä asioita, vaikka algoritmin kohinansieto vaikuttaakin kohtuullisen hyvältä. Ylipäätään työkalun avulla tehtyjen havaintojen luotettavuuden pohdinta jää varsin ohueksi, osasta artikkeleita tällainen pohdinta puuttuu kokonaan. Historiantutkijoille suunnatussa teoksessa ei myöskään pitäisi sortua sellaiseen anakronismiin, jossa työkalun geospatiaalisissa visualisoinneissa taustakartat perustuvat nykyisiin valtiorajoihin, tieverkostoihin tms. Vaikutelmaksi jää, että teknologian rooli painottuu aineiston hallintaan – siis digitaalisuuteen – ei niinkään pyrkimykseen mallintaa monimutkaista sosiaalista todellisuutta, jotta ymmärtäisimme sitä aiempaa paremmin, mikä taas on keskeistä laskennallisuudelle.

Miksi semantiikalla on väliä?

Information flows across the Baltic Sea -teoksen toimittajat sijoittavat teoksen jo otsikkotasolla osaksi laskennallisen mediahistorian (*computational media history*) kenttää. Kuten edellä on osoitettu, laskennallisuus on paitsi täsmällisempi, myös kapeampi lähestymistapa verrattuna jonkinlaisena metakäsitteenä asemansa vakiinnuttaneeseen digitaaliset ihmistieteet(*Digital Humanities*) -paradigmaan.

Onko sitten oikeasti merkitystä sillä, onko jokin tutkimus digitaalista vai laskennallista? Omasta mielestäni kysymys on aivan keskeinen, koska tällä jaottelulla on – kuten edellä olen pyrkinyt osoittamaan – perustavanlaatuinen merkitys sille, millaista tutkimusprosessia tutkija tutkimansa aiheen ympärille aikoo rakentaa. Digitaaliset ihmistieteet myös kärsivät jo nyt, pääosin yhteiskunnan digitoitumisen oheisvaikutuksena, määritelmällisestä laajentumisesta. Erityisen voimakkaasti tähän näyttää vaikuttavan sähköisesti saatavilla olevien aineistojen kiihtyvä kasvu, mikä näkyy suoraan erilaisten aineistojen hyödyntämiseen tarkoitettujen www-pohjaisten portaalien räjähdysmäisenä kasvuna. Moni tutkimus, joka hyödyntää tällaisia aineistoportaaleja, sijoittaa itsensä digitaalisten ihmistieteiden kenttään, mikä toki sopii tässäkin artikkelissa hyödyntämäni Haverisen ja Suomisen kehittelemän nelikentän perusajatukseen. Riskinä on kuitenkin se, että tämän seurauksena Fitzpatrickin esittämä ”humanities, done digitally”-rajaus liudentuu – ja pahimmassa tapauksessa pitkälle digitoituneessa yhteiskunnassa kaikki ihmistieteellinen tutkimus on digitaalista ihmistiedettä.

Laskennallisen samaistaminen digitaaliseen voisi myös johtaa ongelmiin erottaa se, mitä tutkitaan siitä, miten ja millä välineillä tutkitaan. Sähköisten aineistojen käyttö tekee tutkimuksesta yhtä

vähän laskennallista kuin tilastoaineistojen käyttö tekee tutkimuksesta määrällistä tai tilastollista tutkimusta. Laskennallisuus tulisi siis jatkossakin varata kuvaamaan sellaisia tutkimuksia, joissa tutkittavan ilmiön monimutkaisuutta pyritään ymmärtämään analysoimalla informaatiota tietojenkäsittelytieteiden ja data-analytiikan piirissä kehitettyjä työkaluja ja menetelmiä hyödyntämällä. Tämä ei kuitenkaan tarkoita sitä, että laskennalliset ihmistieteet tutkisivat tietojenkäsittelytieteille ominaisia kysymyksiä. Vaan sitä, että me voimme hyödyntää ihmistieteille relevanttien ilmiöiden tutkimuksessa sellaisia informaatiolähteitä, joiden tutkiminen perinteisemmin tutkimusmenetelmien olisi hankalaa, ellei peräti kokonaan mahdotonta.

Edellä esittämästäni paradigmaattis-metodologisesti kritiikistä huolimatta totean, että lukukokemuksena teos oli myös empiiristä aihepiiriä vähemmän tuntevalle kiinnostava – ja kauniin aikalaiskuvituksen ansiosta myös visuaalisesti fasinoiva – ja katson sen aidosti tuottavan uutta ymmärrystä sanomalehtien roolista informaation levittäjänä. Kun kokoomateoksen artikkeleita yhdistää ei niinkään jokin metakysymys, vaan yhteinen aineistopohja, tulos on väistämättä mosaiikkimainen. Lukijana jäin kaipaamaan johdantoluvun kaltaista päätäntölukua, johon olisi suodatettu tutkimusartikkelien keskeiset havainnot tekstien uudelleenkäytön näkökulmasta, mutta myös tuotu teoksen havaintoja ja johtopäätöksiä nykypäivään. Teoksessa historiallisina ilmiöinä esille nostetut tekstien kierrätykset eri tavoin kehystettyinä tai sisältöjen nopea maantieteellinen leviäminen ovat ilmiöinä relevantteja myös 21. vuosisadan mediamaailmalle. Tämän pitkän ajallisen kaaren pohdiskelua olisi mielestäni voinut tehdä ilmiön tasolla rohkeamminkin.

[1]Partik Lundell, Hannu Salmi, Erik Edoff, Jani Marjanen, Petri Paju & Heli Rantala (toim.), *Information Flows across the Baltic Sea. Towards a Computational Approach to Media History* (Föreningen Medihistoriskt Arkiv: Lund, 2023).
[2]Kimmo Elo, "Digitaalisen historiantutkimuksen kenttää louhimassa", teoksessa Kimmo Elo (toim.) *Digitaalinen humanismi ja historiatieteet* (Turku: Turun Historiallinen yhdistys, Historia Mirabilis 12, 2016), 11–35.
[3]Anna Haverinen & Jaakko Suominen, "Koodaamisen ja kirjoittamisen vuoropuhelu? Mitä on digitaalinen humanistinen tutkimus," *Ennen & Nyt* (2015), https://journal.fi/ennenjanyt/article/view/108634/63637.
[4]Cluadio Cioffi-Revilla, *Introduction to Computational Social Science: Principles and Applications,* (Cham: Springer, 2. laitos, 2017), 2–3.
[5]Ks. myös Kathleen Fitzpatrick, "The Humanities, Done Digitally," teoksessa: Matthew K. Gold (toim.) *Debates in Digital Humanities* (Minnesota: Minnesota University Press, 2012), http://dhdebates.gc.cuny.edu/debates/text/30; Pertti Ahonen, "Laskennalliset koneoppimisen menetelmät politiikan tutkimuksen kannalta: institutionaalinen tarkastelu ja tieteenfilosofisen ja teoreettisen syventämisen tarve," *Politiikka* 60, no. 2 (2018), 157–163.
[6] Tiina Lintunen & Kimmo Elo, "Valtiorikosoikeuteen joutuneiden punaisten naisten verkostot Porin alueella,", *Historiallinen Aikakauskirja* 116, no. 2 (2018), 139–152.
[7]Jens Edlund, Daniel Brodén, Mats Fridlund, Cecilia Lindhé, Leif-Jöran Olsson, Magnus P. Ängsal & Patrik Öhberg, "Multimodal Digital Humanities Study of Terrorism in Swedish Politics: An Interdisciplinary Mixed Methods Project on the Configuration of Terrorism in Parliamentary Debates, Legislation, and Policy Networks 1968–2018," teoksessa: Kohei Arai (toim.) *Intelligent Systems and Applications 2021*, (Cham: Springer, Lecture Notes in Networks and Systems 295),. https://doi.org/10.1007/978-3-030-82196-8_32; Derek Greene & James P. Cross, "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach," *Political Analysis* 25, no. 1 (2017), 77–94, https://www.sciencegate.app/app/document/download#10.1017/pan.2016.7; Deborah Kilroy, "All the king's men? A demographic study of opinion in the first English Parliament of James I, 1604–10," *Parliaments, Estates and Representation* 41, no. 1 (2021), 1–23, https://doi.org/10.1080/02606755.2020.1857546; Anna Ristilä & Kimmo Elo, "Observing political and societal changes in Finnish parliamentary speech data, 1980–2010, with topic modelling," *Parliaments, Estates and Representation* 43, no. 2 (2023), 149-176, https://doi.org/10.1080/02606755.2023.2213550.
[8]Laajemmin vastaavasta kritiikistä, ks. esim. Gary Hall, "Toward a Postdigital Humanities: Cultural Analytics and the Computational Turn to Data-driven Scholarship," *American Literature* 85, no. 4 (2013), 781–809.