Digitaaliset menetelmät rantautuvat historiatieteeseen

Kimmo Elo (toim.): Digitaalinen humanismi ja historiatieteet. Historia mirabilis 12, THY 2016. 258 s. ISBN 978-952-7045-05-3.

Muun yhteiskunnan mukana myös humanistinen tutkimus digitalisoituu. Kimmo Elon toimittama kokoelma *Digitaalinen humanismi ja historiatieteet* on tärkeä pelinavaus historiatieteen alueella. Erilaisia laskennallisia menetelmiä ja pitkiä lähdesarjoja on tietenkin käytetty historiatieteessä aiemminkin, mutta monet tekstilouhinnan ja klusteroinnin menetelmät on kehitetty vasta 2000-luvulla tietokoneiden kapasiteetin kasvun myötä. Elon toimittama teos toimii suomenkielisenä johdatuksena digitaalisten menetelmien tarjoamiin mahdollisuuksiin historiantutkimuksessa ja se tulee todennäköisesti löytämään tiensä myös yliopistojen tenttikirjaksi. Kokoelma jakaantuu johdannon jälkeen kahteen osaan: neljään yleisempään metodologiseen katsausartikkeliin sekä kolmeen tapaustutkimukseen. Tapaustutkimukset käyttävät kaikki kotimaisia aineistoja.

Kimmo Elon kirjoittama johdanto on erinomainen katsaus aiheeseen ja kontekstualisoi digitaalista humanismia yhteiskuntatieteiden ja tietojenkäsittelytieteen tutkimuskäytäntöihin. Elo painottaa aivan oikein sitä, että digitaaliset menetelmät vaativat tavallisesti aidosti monitieteistä tutkimusryhmää, jonka jäsenet tulevat historiantutkimuksen ohella kielitieteestä, tietojenkäsittelytieteestä tai jopa bioinformatiikasta. Elon mukaan tällainen ryhmätyö edellyttää humanistilta oppimista pois yksittäisen sankaritutkijan roolista. Erityisesti laajoja aineistoja käyttävissä hankkeissa yksittäinen tutkija ei voi tehdä kaikkea itse aineiston keräämisestä, analysoinnista ja tulkinnasta tutkimustulosten julkaisemiseen, sillä digitaaliset menetelmät edellyttävät syvällistä ymmärrystä esimerkiksi tilastotieteen perusteista, tietokannoista kuten myös ohjelmointitaitoa.

Katja Fältin artikkeli on ansiokas esitys digitaalisten tutkimusaineistojen konservoinnista ja jakamisesta. Fältin tekstin abstraktiotaso on tosin varsin korkealla: hän peräänkuuluttaa aineistojen huoltamista niin, että ne säilyisivät käyttökelpoisina vuosikymmeniä ohjelmistojen ja tiedostoformaattien muuttuessa kertomatta kuitenkaan, miten tämä tapahtuu. Fält olisi voinut esimerkiksi todeta sen yksinkertaisen seikan, että data siirtyy parhaiten järjestelmästä toiseen, kun se on talletettu universaalien UTF-8 tai ASCII-standardien mukaisena muotoilemattomana tekstitiedostona eikä valmistajakohtaista binäärikoodia edellyttävässä tiedostomuodossa. Esimerkiksi Microsoft Excelin kaltaiset taulukkolaskentaohjelmat voivat tallentaa laajojakin tietokantoja puhtaassa tekstimuodossa käyttäen yksinkertaisesti pilkkua tietokannan eri kenttien erottamiseen (ns. CSV-standardi eli *Comma Separated Values*). Mikäli tietokanta on hierarkisoitu, tekstitiedostoissa voidaan käyttää XML- (*Extensible Markup Language*) tai JSON- (*JavaScript Object Notation*) merkintäkieltä, jotka muistuttavat verkkosivujen kirjoittamiseen käytettävää HTML-kieltä. Erona koneluettaviin binääritiedostoihin tällaiset merkintäkielet ovat lähtökohtaisesti ihmisluettavassa muodossa ja ne rinnastuvat läpinäkyvyydeltään ohjelmistojen avoimeen lähdekoodiin (*open source*). Kaikki kuviteltavissa olevat menneisyyden ja tulevaisuuden ohjelmat osaavat avata tiedostoja, jotka on tallennettu tällaisessa muodossa, kunhan käytetty merkistökoodaus tukee kyseisen kielen kirjaimistoa. Esimerkiksi Microsoft Wordin tai Excelin omat binääritiedostot voidaan avata vain ohjelmalla, joka tukee juuri näitä valmistajakohtaisia muotoja eivätkä ne siksi sovellu lainkaan tiedon arkistointiin, mikäli halutaan välttää tulevaisuuden yhteensopivuusongelmat.

Fältin artikkelia seuraa käännökset kahdesta Frédérick Clavertin alkujaan ranskaksi julkaistusta blogikirjoituksesta. Ainakin arvostelijalle jäi epäselväksi, ovatko myös Clavertin kirjoitukset käyneet muiden tekstien tavoin läpi vertaisarvioinnin vai onko blogikirjoitukset uudelleenjulkaistu alkuperäisessä muodossa. Clavertin tekstit ovat sinällään tärkeitä, koska ne esittelevät suomenkieliselle yleisölle esimerkiksi Franco Morettin kuuluisan erottelun lähi- ja etälukemiseen (*close and distant reading*). Clavertin valaisevana esimerkkinä on holokaustin historian uudelleenkirjoittaminen tarkastelemalla kaikkia saatavilla olevia dokumentteja samanaikaisesti. Clavert huomauttaa monien muiden tavoin, että etälukeminen ei missään tapauksessa korvaa perinteistä lähilukemista, vaan toimii sitä täydentävänä menetelmänä.

Jaakko Suominen ja Anna Sivula ottavat kantaa digisyntyisten aineistojen tutkimukseen ja säilyttämiseen. Facebookin, Twitterin sekä Suomi24-keskustelufoorumin kaltaiset sosiaaliset sivustot tarjoavat lähihistoriasta kiinnostuneelle jättimäisen tutkimusaineiston. Kun mukaan otetaan esimerkiksi pelien tutkimus, nousee esiin tarve kehittää täysin uusia ei-tekstuaalisten lähteiden tutkimusmetodeja. Vaikka pelit ovat audiovisuaalisia, ne poikkeavat esimerkiksi elokuvatutkimuk-

sesta, koska pelin juoni on usein täysin avoin. Toisaalta vaikka esimerkiksi Suomi24-foorumi koostuu tekstistä, se on rakenteeltaan avoimempi ja dynaamisempi kuin painetut julkaisut sekä sisältää yksityisyyden suojaan liittyviä eettisiä hankaluuksia tuottaen tutkijoille aivan uudenlaisia lähdekriittisiä ongelmia.

Teoksen käännösartikkeleihin lukeutuu myös Marten Düringin teksti agenttiperustaisesta mallintamisesta. Düring käsittelee tietokonemallien mahdollisuuksia esimerkiksi kontrafaktuaalisen historiantutkimuksen alueella. Historiallisten tapahtumien laskennallinen mallintaminen on ollut yleistä arkeologiassa, mutta historiassa sitä on sovellettu vähän. Düringin mukaan mallit eivät kuitenkaan sovellu kovin monimutkaisten tapahtumien tutkimiseen, sillä ne voivat käsitellä vain vähän erilaisia kausaalisia tekijöitä. Ne ovat parhaimmillaan yhden selkeän hypoteesin vahvistamisessa tai kumoamisessa.

Varsinaiset tapaustutkimukset aloittava Kimmo Elon ja Olli Kleemolan artikkeli käsittelee puolustusvoimien 2. maailmansodan SA-kuvaportaalia (http://www.sakuva.fi) ja sen louhintaa. Elo ja Kleemola ovat erityisesti kiinnostuneita saksalaisten mukana olosta kuvissa. Samoilla menetelmillä olisi mahdollista tutkia myös monia muita sotaan liittyviä teemoja kuten vaikkapa sitä, millaisia kuvia ja teemoja Neuvostoliittoon yhdistettiin. Elon ja Kleemolan artikkelin vahvuutena on tutkimusmenetelmän läpinäkyvyys, mutta tulosten esittäminen kirjan muodossa hankaloittaa niiden lukemista. Olisin toivonut erityisesti Gephillä luotujen kollokaatioverkostojen jakamista esimerkiksi jollakin verkkosivulla isolla resoluutiolla niin, että myös kuvion pienimmistäkin teksteistä saisi selvää ilman suurennuslasia. Vielä tärkeämpää tutkimuksen avoimuuden ja uusinnettavuuden kannalta olisi ollut siinä käytettyjen komentotiedostojen avoin jakaminen esimerkiksi GitHub-palvelussa, joka mahdollistaisi paitsi tutkimuksen toistamisen myös toisten tutkijoiden jatkotutkimukset aiheesta. Vaikka artikkelin aineistona ovat visuaaliset kuvat, itse louhinta on käytännössä toteutettu kuvien metatiedoille. Kirjoittajat perustelevat ratkaisuaan sillä, että kuvantunnistusteknologia ei kykene tunnistamaan kuvista saksalaisia elementtejä. Kirjoittajat eivät kuitenkaan mainitse esimerkiksi Lev Manovichin aiempia tutkimuksia, joissa vertaillaan suoraan kuvatiedostoja tai koulutettujen neuroverkkojen tarjoamia uusia mahdollisuuksia kuvantunnistuksessa. Elo ja Kleemola ovat luultavasti kuitenkin oikeassa siinä, että neuroverkkoihin liittyvä teknologia ei ole vielä riittävällä tasolla, jotta kuvatiedostojen saksalaisia elementtejä voitaisiin louhia luotettavasti.

Visuaalisen historian teemaa jatkaa tavallaan myös Antti Härkösen artikkeli geospatiaalisesta analyysistä. Tilallinen ja maantieteellinen historia on ollut yksi digitaalisen historiantutkimuksen perussovellus, joten tätä koskeva artikkeli on tärkeä lisä kirjassa. Härkönen selittää myös määrällisen tutkimuksen perusteita kuten autokorrelaatiota ja MAU-ongelmaa. Pitkä artikkeli tuo hyvin esiin tilallisen tutkimuksen mahdollisuuksia ja sudenkuoppia havainnollistaen niitä Käkisalmen lautamiehien käräjämatkojen tutkimisella.

Kirjan päättää Lauri Viinikkalan teksti yhdistetyn todellisuuden hyödyntämisestä erityisesti Louhisaaren kartanon 1820- ja 30-lukujen historian elävöittämisessä museovieraille. Viinikkala pohtii digitaalisten menetelmien suhdetta historiantutkimuksen narratiiviseen käänteeseen, sillä Louhisaaren historiaa on yritetty elävöittää projektissa nimenomaan siitä kertovilla tarinoilla. Vaikka yhdistetty todellisuus tuo uusia mahdollisuuksia esittää museovieraille esimerkiksi kadonneita maiseman tai arkkitehtuurin elementtejä, Viinikkala muistuttaa digitaalisenkin rekonstruktion olevan aina vain yksi tulkinta menneestä.

Teoksen tapaustutkimusten sekä hyvänä että huonona puolena on visuaalisuus. On tietenkin totta, että tekstuaalisuus on hallinnut liikaakin historiantutkimusta. Toisaalta nyt esimerkiksi luonnollisten kielten tutkimuksen parissa kehitetyt menetelmät jäävät kirjassa vähäiselle käsittelylle, mikä on teoksen pahin heikkous. Jotta teoksen voisi sanoa tarjoavan kattavan johdannon aihepiiriin, sen tulisi esitellä ainakin jollakin tavalla korpuslingvistiikan perustyökalut (esimerkiksi *N*-grammit) sekä näiden ohella myös *topic modeling* (aiheen mallinnus), klusterointi, tekstien vektorimallinnus, tekijyyden tunnistus sekä tekstin uudelleenkäytön tunnistus. Jään odottamaan vielä toista suomenkielistä kokoelmaa aiheesta, jossa myös nämä alueet saisivat yhtä laadukkaan esittelyn.

Asko Nivala, FT
Turun yliopisto