

Historiantutkimus digitaalisella aikakaudella

Kaksi blogikirjoitusta Frédéric Clavertin blogista
*L'Histoire contemporaine à l'ère numérique**

FRÉDÉRIC CLAVERT

Frédéric Clavert väitteli Strasbourgin yliopistosta lähihistoriaan liittyvästä aiheesta,¹ minkä jälkeen hän on tutkimuksissaan keskittynyt pohtimaan historiantutkimuksen lähteiden käyttöä digitaalisella aikakaudella. Hän toimi viisi vuotta Euroopan integraatiohistorian dokumentaatiokeskuksessa CVCE:ssä (www.cvce.com) Luxemburgissa sekä kaksi vuotta tutkimusinsinöörinä Pariisin Sorbonnen yliopistossa. Nykyisin Clavert toimii lähihistorian yliopistonlehtorina Sveitsissä Lausannen yliopistossa. Vuonna 2013 ilmestyi Clavertin yhdessä Serge Noiret'n kanssa toimittama teos *L'histoire contemporaine à l'ère numérique / Contemporary history in the digital age* (PIE Peter Lang, 2013).

Clavert, samoin kuin monet digitaalisen historiantutkimuksen puolestapuhujat, on alusta lähtien hyödyntänyt blogiaan sekä keskustelufoorumina että eräänlaisena koeareenana, jolla on mahdollista esittää, testata ja pohtia erilaisia ideoita. Digitaaliselle historiantutkimukselle ovatkin tyypillisiä uudet julkaisu- ja kommunikaatiokeinot, erityisesti internetin tarjoamat sosiaalisen median foorumit, joiden kautta tutkijat luovat uusia muotoja tieteelliselle julkaisemiselle. Blogit ja muut vuorovaikutteiset sosiaalisen median kanavat luovat tutkijoille uusia kontaktipintoja sekä uusien ajatusten kehittelemiseksi että tieteellistä keskustelun käymiseksi. Yksinkertaisen some-viestinnän lisäksi tutkijat voivat näiden työkalujen avulla avata uusia dialogiympäristöjä.

Tulevaisuudessa digitaalisten keskustelu- ja julkaisutilojen merkitys tulee entisestään korostumaan myös tiedeviestinnässä. Vaikka historian-

* Alkuperäiset tekstit on suojattu Creative Commons 3.0, CC-BY-SA-lisenssillä. Nyt julkaistavat tekstin julkaistaan alkuperäisen kirjoittajan suostumuksella lisenssiehtojen mukaisena suomennoksena.

1 Clavertin väitöskirja (2009) käsitteli 1920- ja 1930-lukujen Saksan Reichsbankin johtajaa Hjalmar Schachtia.

tutkijoita pidetään edelleen yleisesti yksinäisinä susina, on historiantutkimus myös joukkuepeleä, jossa tutkijoiden välisellä vuorovaikutuksella on tärkeä rooli. Digitaaliset ympäristöt vastaavat osaltaan tähän tarpeeseen. Mutta samalla digitaalisilla ympäristöillä on potentiaalia haastaa perinteiset tieteellisen julkaisemisen järjestelmät, joissa tieteen avoimuusperiaate törmää kustantajien kaupallisiin intresseihin. Perinteisen julkaisemisen rinnalle näyttäisikin olevan muodostumassa digitaalisessa ympäristössä tapahtuva orgaaninen tieteellinen keskustelu, joka monipuolisesti hyödyntää digitaalisia menetelmiä ja välineitä. Clavertin oman blogin alustana toimivaa *OpenEdition*-ympäristöä voidaan pitää yhtenä hyvänä esimerkkinä tällaisista digitalisaation avaamista mahdollisuuksista.² Clavert myös muokkaa blogikirjoituksiaan niihin tulneiden kommenttien perusteella, mikä korostaa blogikirjoittamisen vuorovaikutteisuutta ja keskustelevuutta.

Seuraavassa kahdessa blogikirjoituksessa Clavert pohtii digitaalisen historiantutkimuksen peruskäsitteitä ja yrittää soveltaa niitä omaan työhönsä. Ensimmäisessä kirjoituksessa Clavert käsittelee historiallisen luennan kahta eri ulottuvuutta (*close/distant reading* ja *human/machine reading*) ja päättyy siihen johtopäätökseen, että arkistoaineistojen tietokoneavusteinen lukeminen (*machine reading*) tulee olemaan historiallisten aineistojen etäluennan (*distant reading*) yksi tärkeimmistä muodoista, koska sen avulla voidaan suuria arkistoaineistomääriä tarkastella ikään kuin lintuperspektiivistä.³ Toisessa tähän teokseen valitussa kirjoituksessaan Clavert pohtii Viktor Mayer-Schönbergerin ja Kenneth Cukierin teosta *Big Data*, erityisesti siinä esiteltyä *datafication*-käsitettä. Hän erottaa *digitalisoinnin* (*numérisation* eli analogisten tietojen muuttamisen digitaalseksi) siitä prosessista, jolla näistä tiedoista tehdään varsinaista analysoitavaa dataa (*datafication*, ranskaksi *mise en données*).

Käännös ja saatesanat Louis Clerc

2 "Open Edition".

3 Clavertin käyttämät ranskankieliset termit ovat *lecture humaine* (*human reading*) ja *lecture computationnelle* (*machine reading*).

Historiantutkimuksen lähteiden lukeminen digitaalisella aikakaudella⁴

Jokaisella historioitsijalla on omanlaisensa suhde primäärilähteisiin. Kun 2000-luvun alussa kirjoitin väitöskirjaani, tekivät toiset arkistodokumentit minuun suuremman vaikutuksen kuin toiset. Muistan esimerkiksi edelleen ulkoa kokonaisia osia marraskuun 1938 juutalaisvainon (*Reichskristallnacht*) jälkeisten natsijohtajien kokousten pöytäkirjoista. Niissä saatettiin pohtia tuntikaupalla ”julkisen penkin” määritelmää, tavoitteena määritellä penkit, joilla juutalaisilla ei olisi enää oikeutta istua. On vaikea ymmärtää natsismin aikaa ilman perehtymistä tällaisiin, usein häiritsevän yksityiskohtaisiin dokumentteihin.

Yksittäisen dokumentin läheisyys voi kuitenkin olla myös ongelma. Väitöskirjatyöni pyrki osoittamaan, että – toisin kuin natsien kanssa liittoutuneet konservatiivit ja osa historioitsijoistakin väittivät⁵ – antisemitismillä oli merkittävä rooli saksalaisten konservatiivien, natsien ja armeijan välisen allianssin eräänlaisena ideologisena kittinä. Tämä allianssi mahdollisti kolmannen valtakunnan synnyn tammikuussa 1933. Yhtenä todisteena käytin Hjalmar Schachtin samana vuonna pitämää puhetta, joka ei ollut tilattu puheenvuoro ja jota propagandaministeri ei etukäteen tarkastanut.⁶ Schacht piti puheensa naisjärjestölle, jossa myös hänen vaimonsa oli aktiivinen. Puhe vetää suoran linjan 1500-luvun reformaattorista Martin Lutherista (katoliseen ja itävaltalaiseen) Adolf Hitleriin. Schacht rakentaa historiallinen rinnastuksen, ei talouden ja finanssipolitiikan, vaan Lutherin kahden antisemitistisen kirjan kautta.

4 ”Lecture des sources historiques à l'ère numérique”.

5 Ks. v. 2006 Hans Mommsenin ja Peter Hoffmannin välinen debatti Mommsenin ja Sabine Gillmannin Goerdeler-dokumenttien toimittamisesta: Hoffmannin kritiikki (toukokuu 2006): ”Hoffman 5-2006”. Mommsenin vastaus (kesäkuu 2006): ”Mommsen 6-2006”. Hoffmannin lisäys (heinäkuu 2006): ”Hoffmann 7-2006”. Mommsen kuvailee Goerdeleria ”piiloantisemitistiksi”, joka pyrki erottamaan juutalaiset kansankunnasta. Hoffmann korosti vastauksessaan, ettei näitä asioita voi tutkia ottamatta huomioon ajan henkeä, joka vaikutti Goerdelerin ajatuksiin. Ks. Gillmann ja Mommsen 2003.

6 BArch Deutsche Reichsbank R2501/6985, 176–197 – ”Luther als Deutscher”, Rede vom Dr. Schacht.

Mikäli olisin tuolloin osannut soveltaa digitaalisia menetelmiä tutkimustyössäni, olisin pystynyt laajentamaan tarkastelukulmani tästä yksittäisestä dokumentista suurempaan kokonaisuuteen, joka olisi voinut koostua Schachtin kaikista puheista. Olisin voinut vertailla Goebbelsin johtaman ministeriön tarkistamien puheiden semanttista kenttää siihen sanavarastoon, jota Schacht käytti sensuroimattomissa puheissaan. Tätä kautta olisin voinut nähdä paremmin, oliko antisemitismi konservatiivien ja natsien välisen liiton keskeinen elementti, ja samalla myös laajemmin arvioida sen painoarvoa. Edelleen olisin voinut analysoida, miten tämä antisemitistinen puhunta kehittyi natsivallan aikana, erityisesti sen jälkeen kun konservatiivit oli syrjäytetty.

Tämän ensimmäisen kokemuksen jälkeen pohdin historiantutkimusta digitalisaation aikakaudella erityisesti Franco Morettin julkaisujen pohjalta.⁷ Tutkimusteni kautta jouduin pohtimaan erityisesti sitä, miten Euroopan integraatiosta kiinnostuneet historiantutkijat tulevat pärjäämään EU:n valtavien digitalisoitujen aineistojen kanssa. Onko lähteiden *konelukeminen* eli tietokoneavusteinen louhinta, analysointi ja visualisointi mahdollista? Voiko tämä *konelukeminen* tuoda lisäarvoa suhteessa perinteiseen lähiluentaan, jolla on edelleen vahva asema historioitsijoiden lähdetyöskentelyssä?

Pohdintani jatkui tästä blogikirjoituksesta alkaneen keskustelun aikana.⁸ Päädyin puhumaan kahdesta akselistä, joiden avulla voimme lukea historiantutkimuksen lähteitä. Ensimmäinen akseli muodostuu lähi- ja kaukoluennan (*close reading/distant reading*), toinen ihmis- ja koneluennan (*human/machine reading*) välille.

7 Moretti 2007. Franco Moretti on Stanfordin yliopiston professori, jonka erityisalanana on romaanikirjallisuuden historia. Hän on tunnettu myös sympatioistaan marxistista koulukuntaa kohtaan – ja kyllä, hän on elokuvaohjaajan veli.

8 Kiitän Sébastien Barret'ta, Louis Clerciä ja Claire Clivazia kommenteista – mahdolliset virheet tai väärintulkinnat ovat kuitenkin omalla vastuullani.

Ensimmäinen akseli: *close/distant reading*

Jos lähdetään jo esitetyistä kahdesta esimerkistä, oma työni Schachtin yksittäisen puheen parissa tarkoitti vahvaa läheisyyttä yksittäiseen arkistodokumenttiin. Olisin kuitenkin voinut ottaa enemmän etäisyyttä tähän yksittäiseen arkistolähteeseen. Tällainen etäisyydenotto antaisi historioitsijalle mahdollisuuden nähdä elementtejä, jotka eivät tule esille, jos lähdettä luetaan liian läheltä. Samalla etäisyyden kasvattaminen antaisi myös mahdollisuuden asettaa dokumentti muiden asiakirjojen muodostamaan kokonaisuuteen, jonka elementit ovat suhteessa toisiinsa. Erityisesti lähihistoriassa (*histoire contemporaine*) yksittäinen arkistodokumentti on harvoin yksinään kovin tärkeä. Pikemminkin työmme on yhtä paljon yksittäisten dokumenttien tulkintaa kuin dokumenttisarjojen muodostamien suurten kokonaisuuksien tulkintaa. On myös tärkeää huomioida, millä tavoin arkistojen hoitajat ovat luokitelleet dokumentit ja sarjat.⁹ Miksi tietyn puheen luonnos esiintyy vain tässä yhdessä sarjassa eikä jossakin toisessa? Tämä kysymys on joskus tutkijalle yhtä tärkeä kuin itse dokumentin sisältö.

Esimerkiksi kolmannen valtakunnan valtakunnankanslerin arkistoissa on Kiinan valtiovarainministeri Kungin ja salkuttoman ministerin, *Reichsbankin* (joka toki oli huomattavasti heikentynyt) johtaja Schachtin väliseen kirjeenvvaihtoon liittyviä dokumentteja.¹⁰ Nämä kirjeet eivät siis olleet *Reichsbankin* arkistoissa vaan Hitlerin valtakunnankanslian arkistoissa. Kirjeiden arkistointipaikka todistaa, että ainakin valtakunnankanslian päällikkö Hans Heinrich Lammers oli tietoinen kirjeenvaihdosta. Mutta jos Kungin ja Schachtin kirjeenvaihto alkoi jo vuonna 1933, niin miksi vain muutamia kirjeitä löytyy valtakunnankanslerin arkistosarjasta R43?¹¹

Vastataksemme tähän kysymykseen meidän tulee tarkastella dokumentteja laajemmasta näkökulmasta, mutta myös ottaa huomioon

9 Minulla on erityisesti kokemusta valtiollisista arkistoista.

10 Kirjeenvaihto koskee saksalaisten aseiden kuljettamista Kiinaan. Kung valitsee aseiden laadusta ja natsi-Saksan tuesta Japanille, Schacht puolestaan pyrkii lepyttelemään häntä. Käänt. huom.

11 BArch Reichskanzlei R43/II/91428c, 70.

Kolmannen valtakunnan ulkopoliitiikkaan liittyviä tekijöitä. 1900-luvun alussa Saksa oli ollut perinteisesti Kiinan tukena paitsi Japania myös Iso-Britanniaa, Ranskaa ja Yhdysvaltoja vastaan. Vuodesta 1936 eteenpäin Saksan ideologinen ja strateginen läheisyys Japaniin aiheutti kuitenkin poliittista lähestymistä samaan aikaan, kun Japani ryhtyi valtaamaan osia Kiinasta. Saksalaisia aseita kuitenkin kuljetettiin edelleen Kuomintangin hallitsemaan osaan Kiinasta samalla, kun muutamat saksalaiset valtuuskunnat (joiden tavoitteet olivat enemmän tai vähemmän selkeitä) pyrkivät yhteyteen Tšiing Kai-šekin kanssa. Kyseinen kirjeenvaihto on valtakunnankanslerin arkistoissa juuri näiden tapahtumien takia: vuonna 1937 natsi-Saksa oli lähes kokonaan kääntynyt Japaniin päin, eikä Schacht enää saanut tukea Kiinaa.¹²

On olemassa toinenkin syy siihen, miksi nämä kirjeet ovat valtakunnankanslerin arkistoissa eivätkä Reichsbankin arkistoissa. Vuonna 1937 Hjalmar Schacht oli vaikeassa asemassa. Hän menetti talousministeriön entiselle taloustoimittajalle ja propagandistille Walter Funkille, joka toimi nelivuotissuunnitelmasta vastanneen Hermann Göringin alaisena. Schacht oli menettänyt vaikutusvaltansa myös ulkosuhteissa, eikä hän enää pystynyt – toisin kuin esimerkiksi van Zeelandin valtuuskunnan kanssa – käyttämään kansainvälistä mainettaan saadakseen poliittista valtaa Saksan sisällä.¹³ Samassa tilanteessa olivat myös muut konservatiivit, jotka olivat vielä jäljellä Hitlerin lähipiirissä. Kiinan asiat kulkivat nyt Hitlerin valtakunnankanslian kautta, eikä Schacht enää kokenut voitonsa toimia yksin.

On siis tärkeää ottaa etäisyyttä yksittäiseen dokumenttiin ja tarkastella laajasti kokonaista periodia ja kontekstia. Tutkijoiden käytössähän on esimerkiksi melko vähän yksittäisiä dokumentteja juutalaisten kansanmurhan suunnitteluprosessista kesäkuun 1941 ja tammikuun 1942 Wannseen konferenssin väliin jäävältä ajanjaksolta. Kuten tunnettua, Wannseen konferenssissa esitettiin ylimmälle natsijohdolle nk. ”lopullinen ratkaisu” eli Hitlerin hanke Saksan hallussa olevilla alueilla eläneiden

12 Vuoden 1937 kirjeissä Kung valittaa hänelle välitettyjen aseiden kehnosta laadusta.

13 Clavert 2009.

juutalaisten tuhoamiseksi.¹⁴ Tilanteessa, jossa yksittäiset dokumentit eivät yksinään riitä, ainoastaan kaikkien dokumenttien käsittely kokonaisuutena antaa mahdollisuuden nähdä, kuinka juutalaisten kansanmurhan suunnittelu ja toteutus prosessina eteni. On siis otettava etäisyyttä yksittäisiin dokumentteihin, jotta prosessia voitaisiin ymmärtää kokonaisuutena.

Tämä lähteiden etäluenta (*distant reading*), jossa yksittäiset dokumentit asetetaan osaksi laajempaa kokonaisuutta, voi tietokoneavusteisesti mennä vieläkin pidemmälle. Digitaaliset menetelmät antavat mahdollisuuden käsitellä suuria aineistokokonaisuuksia, asettaa yksittäisiä dokumentteja kontekstiinsa ja nostaa esille elementtejä, joita muun muassa aineistojen valtavan koon takia ei kyettäisi havainnoimaan perinteisen lähiluennan keinoin. Tällä tavoin voidaan hyödyntää ensimmäisen akselin potentiaali.

Toinen akseli: *human/machine reading*

Tämän päivän teknologioiden avulla on mahdollista löytää suurista arkistoaineistoista sellaisia yksityiskohtia tai toistuvia piirteitä, joita ihmisen suorittama luenta ei olisi kyennyt havainnoimaan. Samalla nämä uudet työkalut antavat mahdollisuuden todistaa ja vahvistaa aiempia, ihmisten tekemiä havaintoja. Menetelmiä voidaan soveltaa sekä yksittäisiin dokumentteihin¹⁵ että arkistokokonaisuuksiin¹⁶, jotka voivat sisältää hyvinkin suuria määriä dokumentteja.¹⁷

Esimerkkinä voidaan tarkastella Lauren F. Kleinin esitelmää vuodelta 2012 Hampurin *Digital Humanities* -konferenssissa, jossa Klein havainnollisti, kuinka hän oli kyennyt tutkimaan Yhdysvaltojen orjuu-

14 Ks. Édouard Hussonin ja Florent Brayardin välinen debatti arkistodokumenteista: Brayard 2009; Husson 2009.

15 ”Clavert”.

16 ”DH2012”.

17 Internetissä olevat sivut ovat esimerkki näistä valtavista datakokonaisuuksista. Internet Archive -organisaatio esimerkiksi lupaa 80 terabittia internetistä kerättyä dataa tutkimuskäyttöön: ”80 terabytes of archived web crawl data available for research” (ks. ”Internet Archive”).

den historiaa soveltamalla verkostanalyysia ja visualisointitekniikoita.¹⁸ Kuten monet yhdysvaltalaiset porvarit, myös Thomas Jefferson puhui orjistaan käyttämällä vain heidän etunimiään. Soveltamalla digitaalisia menetelmiä Jeffersonin papereista laadittuun digitaaliseen kokoelmaan Klein pystyi linkittämään useita eri tietolähteitä ja siten löytämään myös orjien sukunimet. Digitaalisten menetelmien käyttö vaati kuitenkin myös manuaalista työtä, koska Klein joutui usein tarkistamaan tietojärjestelmän antamat tulokset.¹⁹

Tämä dialogi ihmisen suorittaman lähiluennan ja konelukemisen välillä ei ole uutta, sillä se on jo pitkään kuulunut esimerkiksi kvantitatiivisen taloushistorian perusmetodologiaan. Franco Moretti aloittaa ensimmäisen graafeihin liittyvän artikkelinsa viittaamalla Annales-koulukunnan tapaan käyttää kvantitatiivisia tietoja kirjoittaessaan taloushistoriaa. Moretti ehdottaa samojen menetelmien soveltamista kirjallisuushistoriaan, joka yleensä keskittyy vain muutama kymmenen, korkeintaan muutama sataan kirjoittajaan ja muutamiin satoihin teoksiin. Mutta jo 1700-luvulta alkaen sekä kirjailijoita että julkaistuja teoksia on ollut valtavasti enemmän verrattuna siihen, mihin kirjallisuuden historia yleensä keskittyy.²⁰ Moretti väittää, että kirjallisuuden historia (sellaisena kuin sitä kirjoitetaan) on liian suppeaa. Hänen ratkaisunsa on käyttää menetelmiä kvantitatiivisesta taloushistoriasta (*graphs*), maantieteestä (*maps*) ja evoluutiotieteestä (*trees*). Moretti korostaa pragmaattista lähestymistapaa, joka lähtee tutkijan tarpeista.

Kvantitatiivisia menetelmiä kritisoitiin jo 1970-luvulla, jolloin muun muassa Paul Veyne korosti ilmiöiden ainutlaatuisuutta ja hylkäsi liian yksinkertaisina pitämänsä yleistävät käsitteet. Michel de Certeau puhui laajemmin realistisesta illuusiosta, kun historioitsijat uskovat tutkivansa

18 Lauren Frederica Klein, "Social Networks analysis and visualization of the papers of Thomas Jefferson", *Digital Humanities* 2012.

19 Kommenteissa Clavert ja muut korostavat, ettei konelukeminen ole mikään oikotie onneen, koska sen antamat tulokset joudutaan usein tarkistamaan manuaalisesti. Käänt. huom.

20 Annalesin lisäksi Yhdysvaltojen New Economic History käytti samoja menetelmiä. Peter Haber kuvailee näitä koulukuntia ja kutsuu niitä Digital Historyn esi-isiksi (Haber 2011).

”todellista historiaa”. Italialainen mikrohistoria, joka lähtee yksilöistä tutkiaksensa heitä ympäröivää yhteiskuntaa, kritisoiakin kvantitatiivista lähestymistapaa historian tutkimukseen. Eräät kvantitatiivisen historian puolestapuhujat, kuten Emmanuel Le Roy Ladurie, keskittyivätkin tutkimuksissaan paikallishistoriaan.

Nykyinen digitaalinen aikakausi, joka läpäisee yhteiskunnan eri kerrokset,²¹ luo kvantitatiiviselle historian tutkimukselle uusia mahdollisuuksia. Paperisten tai filmiarkistojen digitalisointi tai natiivisti digitaaliset arkistolähteet (tietokonelevykkeistä internetissä oleviin sisältöihin) tuovat uudelleen esille kvantitatiiviset menetelmät. Samalla syntyy tarve kehittää menetelmiä, joilla voidaan saada suurista aineistokokonaisuuksista esille myös laadullisia elementtejä.

Lähtökohtani on näkemys, jonka mukaan meidän täytyy yhdistää arkistojen ihmislukeminen – siis perinteinen arkistoissa istuminen ja lähdekriittinen lähiluenta – ja arkistojen tietokoneavusteinen tulkitseminen eli koneluenta. Koneluenta antaa mahdollisuuden tutkia suuria, jopa valtavia digitalisoituja tai natiivisti digitaalisia aineistomassoja tehokkaammin kuin pelkkä ihmislukeminen.

Paras keino tämän tavoitteen toteuttamiseksi on pyrkiä samanaikaisesti sekä koneluentaan että ihmislukemiseen. Ihmislukemisella tutkija säilyttää kontaktin toimijoihin ja heidän ympäristöönsä, koneluenta puolestaan mahdollistaa etäisyyden ottamisen kohteeseen. Etäisyys auttaa näkemään ja hahmottamaan erilaisten lähteiden välillä vallitsevia yhteyksiä – se siis etäännyttää tutkijan ”yhden dokumentin” tutkimuksen vaaroista ja avaa laajemman näkökulman.²²

21 Ks. Clivaz 2012.

22 Aiheesta lisää ks. Stéfán Sinclair, ”Distant reading”, *Observatoire de l’imaginaire contemporain*, MacGill Université du Québec à Montréal, 24.1.2012.

Mailman data, historian data?²³

Lukiessani Guido Kollerin blogia²⁴ *We think history* sain tietää Viktor Mayer-Schönbergerin ja Kenneth Cukierin *Big Data* -ilmiöön liittyvästä kirjasta.²⁵

Mailma datana?

Teos on erittäin monipuolinen, mutta haluaisin korostaa sen osaa, jossa kirjoittajat puhuvat *dataficationista*.²⁶ *Le Monde Diplomatique* käänsi termin ranskaksi muotoon "mise en données du monde" eli mailman tietojen muuttaminen dataksi. Kirjoittajat korostavat sitä, että mailmaa on "datafoitu" jo ennen digitaalista aikaamme: esimerkiksi Yhdysvaltojen merivoimat on 1800-luvulla systemaattisesti käyttänyt kirjallisia loki-tietoja parantaakseen alustensa reittejä tekemällä niistä lyhyempiä ja tehokkaampia.

Se, että ihminen on systemaattisesti kerännyt tietoja omasta ympäristöstään, on ilmiönä vanhempi kuin tietokoneet ja digitalisointi. Tavoitteena on ollut mitata ympäristöä ja sitä kautta ymmärtää paremmin omaa elinympäristöä. *Datafication* perustuu myös sille ajatukselle, että massiiviset datakokonaisuudet tuovat jotakin sellaista, jota ei voida löytää pienistä kokonaisuuksista. Tämä ajatus on *Big Data* n ytimessä: määrällinen muutos merkitsee myös laadullista muutosta.

Kirjassa *datafication*-termi määritellään seuraavasti: "To datafy a phenomenon is to put it in a quantified format so it can be tabulated and analysed" (s. 78). Kyse on siis eri asiasta kuin digitalisointi, joka määritelmällisesti tarkoittaa analogisen tiedon muuntamista digitaaliseksi, sarjaksi nollia ja ykkösiä. Ilmiöiden muuttaminen dataksi voi sen sijaan tapahtua muuallakin kuin digitaalisessa ympäristössä, joskin nyky-

23 "Mise en données du monde, mise en données de l'histoire?"

24 "Guido Koller".

25 Mayer-Schönberger ja Cukier 2013.

26 Käännetty *Le Monde diplomatique*ssa: Cukier ja Mayer-Schönberger 2013. Hyvä tiivistelmä kirjan pääargumenteista ks. "The New Republic". Käänt. huom.

päivänä ilmiöt (analogisuus ja digitaalisuus) ovat erottamattomasti linkittyneet toisiinsa.

Kirjoittajat kuvailevat ihmiskunnan pitkäaikaista yritystä mitata ja kvantifoida ympäristöään, mikä juontaa juurensa jo antiikin maailmasta. Mittaaminen ja arkistointi voidaankin nähdä *datafication*-prosessin ensimmäisinä etappeina (s. 79). Ne antavat mahdollisuuden toistaa ihmisten toimintaa, hyödyntää aiempaa kokemusta ja suunnitella. Mittaaminen (määrät ja pituudet, paikat, volyymit ja aika) on kehittynyt hitaasti. Arabialaisten numeroiden keksiminen oli tärkeä etappi, joka loi edellytykset matematiikan kehittymiselle. Matematiikka puolestaan antoi mahdollisuuden analysoida abstrakteja mittaauksia, siis dataa.

Edellä mainitut etapit saavutettiin pitkällä aikavälillä: arabialaiset numerot, matematiikka ja esimerkiksi kirjanpito kehittivät länsimaissa vuosisatojen aikana. 1800-luvulla kehitys kiihtyi, kun yhä uusia ilmiöitä pystyttiin mittaamaan – esimerkiksi vaikkapa sähkö, lämpötila, ilmanpaine tai ääni (s. 81–82). Toinen nopeasti kehittynyt väline on ollut tietokone. Tietokoneiden kehitys ei ainoastaan tehnyt kvantifioinnista halpaa ja luotettavaa vaan myös mahdollisti datan nopean matemaattisen analysoinnin. Tämä puolestaan toi tietomassasta esille uusia asioita, jotka eivät olleet paljastuneet ilman tietokoneanalyysia. Yhdessä datan digitalisoinnin kanssa tietokoneet ovatkin vieneet maailman mittaamisen ja sen tietojen analyysin aivan uudelle tasolle.

Kirjoittajat haluavat osoittaa *dataficationin* ja digitalisoinnin välisen eron käyttämällä esimerkkinä Google Booksia, tietokantaa, johon on digitalisoitu miljoonia kirjoja ja jonka kautta on mahdollista lukea niitä suoraan tietokoneen ruudulta. Tietokannan kirjoista on mahdollista etsiä tietoja esimerkiksi Google Ngram Viewer -työkalun avulla.²⁷ Ngram Viewer muuttaa Google Booksiin tallennettujen kirjojen tietoja datakokonaisuudeksi, josta on mahdollista etsiä tietoja ”massa-analyysilla” (s. 83–86). Kirjoittajat puhuvat myös paikkatiedoista, joista voidaan muodostaa dataa esimerkiksi GPS-järjestelmän kaupallisen mutta ennen kaikkea ei-kaupallisen soveltamisen kautta.²⁸

27 ”Ngrams”.

28 Erityisesti GPS-tietojen soveltaminen julkiseen terveysalaan.

Teoksen loppuosassa kirjoittajat pohtivat *dataficationin* ”uusia rajoja” eli sosiaalista kanssakäymistä.²⁹ He muistuttavat, että vaikka esimerkiksi Twitter rajoittaa yksittäisen viestin pituuden 140 merkkiin, jokaisen viestiin liittyy tarkkaa ja kiinnostavaa metadataa (eli tietoa tiedosta):

Like Google, a gaggle of social media networks such as Facebook, Twitter, LinkedIn, FourSquare, and others sit on an enormous treasure chest of datafied information that, once analyzed, will shed light on social dynamics at all levels, from the individual to society at large (s. 94).

Samaan hengenvetoon kirjoittajat kuitenkin korostavat, että datakokonaisuuksien hyödyntämistä rajoittaa pelkästään oma mielikuvitusemme. Mutta mitä *datafication* merkitsee historiantutkimukselle?

Historian *datafication*?

Cukier ja Mayer-Schönberger ilmaisevat merkityksen suoraan: kaikkea voidaan mitata ja kvantifioida, eli kaikesta voidaan tehdä dataa. Tämä vaikuttaa välttämättä myös historiantutkimukseen.

Käyttäkäämme esimerkkinä sosiaalista mediaa. Meillä ei ole ikinä ollut näin runsaasti informaatiota, jonka avulla kuvailla sosiaalisia käytäntöjä ja ilmiöitä – eikä näin paljon informaatiota, kuvineen ja teksteineen, miljoonien ihmisten arkielämästä, heidän päivänrutiineistaan, ruokatottumuksistaan³⁰, rakkaudestaan kissaansa kohtaan, lastensa kasvatustavoista, työsuhteista, suhteesta politiikkaan, perheeseen, ja niin edelleen. Nämä tiedot, joita sosiologit käsittelevät³¹, aiheuttavat kuitenkin mittavia metodologisia ongelmia. Edustavatko ne väestöä kokonaisuudessaan?³² Miten niitä voidaan käyttää? Millä tavalla niitä säilytetään? *Just in Time Sociology* on yksi tieteenaloista, joka yrittää vastata muun muassa näihin kysymyksiin. Tutkimuksissa huomioidaan, ettei esimerkiksi Twitter ole mikään ”yhteisö” vaan miljoonista yksilöistä

29 Muun muassa sosiaalinen media.

30 Tässä suhteessa Instagram on aarearkku.

31 Ks. Pailler 2013a.

32 Ks. Pailler 2013b.

koostuva massa, jonka sisälle muodostuvat *ad hoc* -yhteisöt voivat olla hyvinkin heterogeenisiä.³³

Jossain vaiheessa tämä data muuttuu myös historian tutkimuksen lähdeaineistoksi. Voidaan olettaa, että muutamien vuosien kuluttua lähihistorian väitöskirjat tehdään ei pelkästään digitaalisiin lähteisiin perustuen vaan myös sellaisilla työkaluilla, jotka antavat mahdollisuuden hyödyntää sekä *close* että *distant readingia*. Historioitsijat ja arkistotyöntekijät joutuvatkin laajasti pohtimaan datatulvan eri ulottuvuuksia. Lisäksi me digitalisoimme nopealla tahdilla jo olemassa olevia arkistoja, minkä seurauksena historioitsijat käsittelevät myös klassisissa arkistokeskuksissakin digitaalisia arkistoja. *Datafication* koskee kuitenkin myös historioitsijoita itseään ja heidän käytäntöjään.

Historioitsijoiden data?

Seuraava tarkasteluni keskittyy kolmeen eri näkökulmaan:

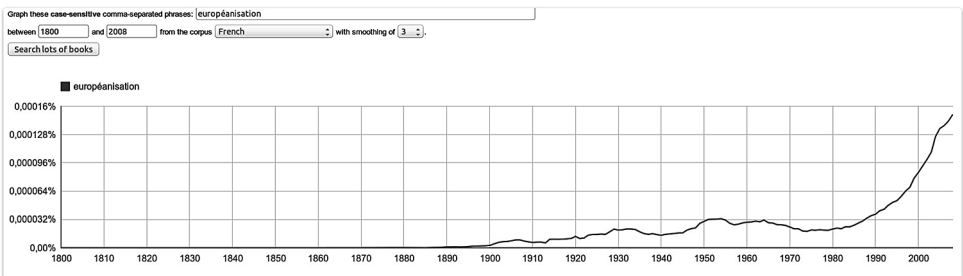
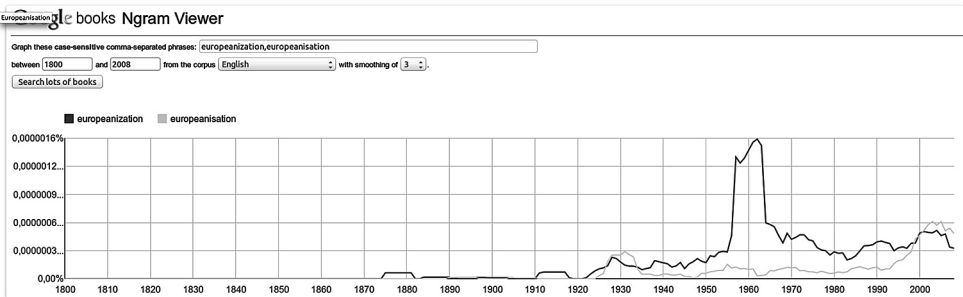
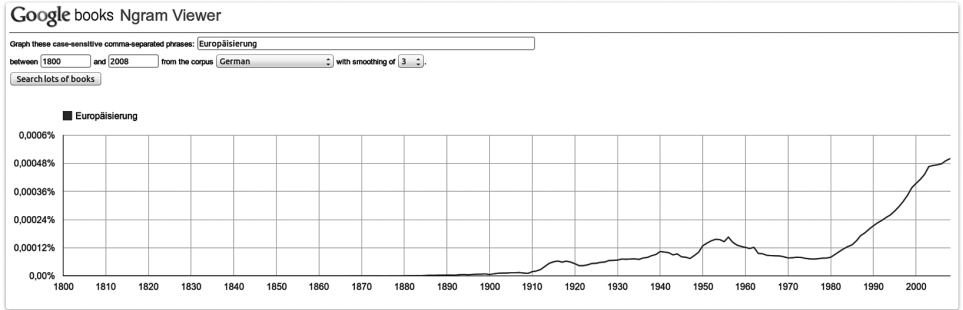
1. Mahdollisuuteen ”lukea” historiografiaa Big Data -työkaluilla
2. Mahdollisuuteen tarkastella reaaliaikaisesti kehittyvää historiografiaa sosiaalisessa mediassa tai vastaavassa ympäristössä
3. Mahdollisuuteen tarkastella historioitsijoiden käytäntöjä esimerkiksi ”avoimissa konferensseissa” (*Conférences ubiquitaires*).

Google Books, Google Ngram ja historiografia

Kun tutkija aloittaa tieteellisen tekstin kirjoittamisen, ensimmäinen askel on tarkistaa, mitä aiheesta on jo kirjoitettu – siis perehtyä aiheen historiografiaan. Nykyisin nämä historiografiset osat ovat usein rajallisia:

1. Julkaisuja on niin runsaasti, ettei ole mahdollista saada tarkkaa kuvaa koko kirjallisuudesta.
2. Kielimuurin vuoksi tutkija ei voi tuntea kaikkea historiografiaa kaikilla maailman kielillä.
3. Aika ei useinkaan riitä kaiken kattavan historiografian muodostamiseen.

33 Ks. Mitchell ja Hitlin 2013.



Kuva 1. ”Eurooppalaisuus”-käsitteen esiintyminen kirjallisissa lähteissä eri aikakausina Googlen books Ngram Viewer -sovelluksen mukaan.

Big Data -kirjan kirjoittajat käyttävät NGramia esimerkkinä tavoista, joilla voidaan käsitellä isoa datakokonaisuutta ja saada siitä esiin elementtejä, jotka olisi vaikea löytää vain ihmislukemisella. Google Ngram Viewer on yksinkertainen työkalu, joka mittaa kuinka usein tietty sana tai sanaryhmä esiintyy tietyssä korpuksessa ja erilaisilla kielillä. Korpus

on tässä tapauksessa Google Books, toisin sanoen useita miljoonia kirjoja.

Syyskuussa 2012 RICHIE (Réseau International de jeunes Chercheurs en Histoire de l'Intégration Européenne) -verkosto järjesti kesäkouluksensa teemasta ”*Europeanisation*”. Voidaan kysyä, milloin tämä käsite tuli esille ja kuinka usein sitä käytetään. Nopea NGram-tarkastelu antoi seuraavat tulokset (ks. s. 89):

- Saatiin tietoa siitä, milloin termi esiintyi ensimmäisen kerran englanniksi, ranskaksi ja saksaksi.
- Saatiin tietää, että (loogisesti) termiä käytettiin ensin puhuttaessa kolonialisaatiosta.
- Lopuksi saatiin tietää, että käsitettä käytettiin erityisen paljon 1980-luvun jälkeen Euroopan integraation kontekstissa.

Tulokset eivät ole yllättäviä, mutta niiden saaminen kesti vain muutamia minuutteja. On toki selvää, ettei NGram voi kokonaan sivuttaa klassista historiografista työtä, mutta se pystyy ohjaamaan tutkimusta, nopeuttamaan sitä ja löytämään myös tutkijalle ennestään tuntemattomia julkaisuja tai tuomaan esille uusia oivalluksia. Menetelmällä voidaan kyseenalaistaa monia historiografisia narratiiveja, esimerkkinä vaikkapa *Digital Humanities* -keskustelun kultainen legenda, jonka mukaan ensin tuli Roberto Busa, sitten ensimmäiset tienraivaajat, *Linguistic Computing*, *Humanities Computing*, ja lopulta vuonna 2005 termin *Digital Humanities* synty. Käyttämällä erilaisia työkaluja (NGram, joka asetetaan tutkimaan Google Scholar -palvelun tietoja sellaisina, kuin ne tulevat esille ohjelmassa *Harzing's Publish or Perish*) paljastuu kuitenkin toisenlainen narratiivi:

- Roberto Busa on toki vaikuttava. Hänen julkaisuihinsa viitataan jatkuvasti paljon, mutta erityisesti 1980-luvun jälkeen ja vielä enemmän 2000-luvulla.
- NGram-tulokset osoittavat, että ensimmäiset tutkijat, jotka käyttivät termiä *Linguistic Computing*, viittaavat kvantitatiiviseen taloustieteeseen eivätkä Busan kirjoituksiin. Samaan tapaan kuin *Annales* Ranskassa ja *New Economic History* Yhdysvalloissa, *Linguistic Computing* oli seurausta talous- ja tilastotieteen vaikutuksesta.

Tämän tyyppinen kartoittava tutkimus tulee olemaan tulevaisuudessa yhä yksinkertaisempaa, kun datakokonaisuudet tulevat isommiksi ja käyttäjäystävällisemmiksi. Mutta jos historioitsijat käyttävät jo nykyään paljon internetiä, voisiko olla mahdollista mennä vielä pidemmälle ja tutkia, kuinka historiografia kehittyy reaaliaikaisesti?

Historiografian reaaliaikainen kehitys?

Sosiaalinen media, esimerkiksi Twitter, antaa hyvän kuvan tästä potentiaalista. Siellä voidaan seurata keskusteluja niin kutsuttujen hashtagien kautta (esimerkiksi *#histoire*). Katrina Gulliver³⁴ otti muutamia vuosia sitten käyttöön hashtagin *#twitterstorians*. Jos Twitterin tiedoista poimitaan vain ne viestit, joissa tämä hashtag esiintyy, on mahdollista muodostaa kuvaa nuorten (kyllä, suurin osa heistä on nuoria) historioitsijoiden huolista ja keskustelunaiheista. Usein keskustelu pyörii työhakemusten ja niiden jakamisen ympärillä. Samalla tavoin esimerkiksi hashtagilla *#envhist* voidaan tutkia ympäristöhistoriaan liittyvää nykykeskustelua.

Tutkijat myös keskustelevat enenevässä määrin blogien kautta spesifeistä teemoista tai yleisesti omasta tutkimustyöstään. Nämäkin tiedot voivat olla hyödyllisiä, esimerkiksi tiettyyn sivustoon tai datakokonaisuuteen kohdistuvien niin kutsuttujen RSS-verkkosyötteiden kautta. *PressForward*-aloite³⁵ toimii juuri tällä periaatteella: tavoitteena on tuoda esille sellaisia tieteellisiä kirjoituksia, jotka julkaistaan suoraan internetissä. Historian alalla esimerkiksi *Global perspectives on digital history* -portaali³⁶ (jonka aktiiveihin lukeutui muun muassa edesmennyt Peter Haber³⁷) antaa nopeasti laajan näkemyksen digitaalisen historian tutkimuksen historiografiasta.

34 "Katarina Gulliver".

35 "PressForward".

36 "GPDH".

37 "Peter Haber".

Historioitsijoiden käytännöt: konferenssit

Toinen tapa muuttaa historioitsijoiden käytännöt dataksi on tieteellisen konferenssin järjestäminen. Konferenssit ovat monin tavoin tärkeitä, sillä niissä tavataan, esitellään tutkimustuloksia ja keskustellaan niistä sekä kehitellään uusia projekteja. Tähän asti konferensseista on useimmiten jäänyt vain yksi varsinainen tieteellinen jälki, konferenssijulkaisu. Mutta yhä useammin konferensseista jää monia erilaisia digitaalisia jälkiä: kotisivut, netissä olevat esitykset, blogikirjoitukset, sosiaalisen median päivitykset ja niin edelleen.

Näiden digitaalisten ”avointen konferenssien” jälkiä voidaan kerätä ja analysoida samalla tavoin kuin muutakin digitaalista dataa. Näin on mahdollista saada reaaliaikaisesti tietoa siitä, mitä ajatellaan ja mistä keskustellaan. Kyse on siis toisenlaisesta aineistosta kuin konferenssijulkaisu, joka koostuu useimmiten vain esitelmistä ja joka julkaistaan konferenssin jälkeen.

Big Datan ongelmat ja maailman *datafication*

Kriittinen, kyseenalaistava ote on kuitenkin tarpeen myös Big Datan kohdalla, kuten Kate Cawford ja Danah Boyd³⁸ osoittavat *Foreign Policy*-lehden artikkelissaan.³⁹ Big Data vaatii isoja korpuksia, jotka ovat mahdollisimman lähellä täydellisyyttä – eli niihin tulisi sisältyä kaikki data, joka aiheesta on mahdollista kerätä. Mutta jopa täydelliset korpukset ovat sosiaalisia rakennelmia, jotka perustuvat tiettyyn hypoteesiin. *Big Datan* kirjoittajat muistuttavat Chris Andersonin artikkelista⁴⁰ käydystä debatista ja korostavat erityisesti sitä, että digitaaliset korpukset perustuvat tilastotieteellisiin ja matemaattisiin teorioihin. Erehtymisen mahdollisuus on aina olemassa, siitä huolimatta (tai sen takia), että tutkija ajattelee saavansa täydelliset tiedot tietystä ilmiöstä.

38 Boyd ja Crawford 2012.

39 ”Think again big data”

40 Anderson 2008.

Digitaalisia työkaluja tulee siis käyttää huolellisesti. On myös hyvä tiedostaa, että tiedon ”muuttaminen dataksi” on yksi potentiaalinen virheiden lähde. Tim Hitchcock muistuttaakin, että epätarkasti syötetyt tiedot – esimerkiksi huonosti toimivan tekstintunnistusohjelman takia – voivat aiheuttaa useita ongelmia.⁴¹ On myös mahdollista, että mikäli virheellisiä digitaalisia korpuksia käytetään enemmän kuin ei-digitaalisia, kokonaisen aihepiirin historiografia voi vinoutua.⁴²

Lähteet

- Anderson, Chris: The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *WIRED* 23.6.2008.
- Boyd, Danah; Crawford, Kate: Critical Questions for Big Data. *Information, Communication & Society* 5/2012, 662–679.
- Brayard, Florent: Shoah: l’intuition et la preuve. Retour sur le processus décisionnel. *La vie des idées* 12.2.2009.
- ”Clavert”. Verkkosivusto. <<http://www.clavert.net/wordpress/?p=1071>>. 10.8.2016.
- Clavert, Frédéric: La mission van Zeeland, une tentative de clearing multilatéral. Teoksessa *L’Europe au coeur. Livre offert à Marie-Thérèse Bitsch*. Toim. Birte Wassenberg ja Martial Libéra. PIE Peter Lang, Brussels 2009.
- Clavert, Frédéric: *Hjalmar Schacht, financier et diplomate (1930–1950)*. PIE Peter Lang, Brussels 2009.
- ”Clavertin blogi”. Frédéric Clavertin internet-blogi. <<http://histnum.hypotheses.org/>>. 10.8.2016.
- Clivaz, Claire (toim.): *Lire demain*. Lausanne, Presses Polytechniques et Universitaires Romandes 2012.
- Cukier, Kenneth; Mayer-Schönberger, Viktor: Mise en données du monde, le déluge numérique. *Le Monde diplomatique* 7/2013.

41 Hitchcock 2011.

42 Esimerkiksi Maailman terveysjärjestö (World Health Organization, WHO) on digitalisoinut kaikki YK:n edeltäjän Kansainliiton dokumentit, jotka liittyvät malarian hoitoon. Tämän seurauksena Kansainliiton toimenpiteitä malariaa vastaan on tutkittu paremmin kuin järjestön toiminnan mitään muuta osa-aluetta. Toisena esimerkkinä vinoumasta voidaan mainita vähemmistökielet: Gerben Zaagsma on nostanut esille jiddišinkielisten lähteiden digitalisointiin ja analyysiin liittyvät ongelmat.

- "DH2012". Digital Humanities 2012 -konferenssin verkkosivusto. <<http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/introduction-to-distant-reading-techniques-with-voyant-tools-multilingual-edition/>>. 10.8.2016.
- "GPDH". Verkkosivusto. <<http://www.gpdh.org/>>. 10.8.2016.
- Gillmann, Sabine; Mommsen, Hans (toim.): *Politische Schriften und Briefe Carl Friedrich Goerdelers*. KG Saur, Berlin 2003.
- "Guido Koller". Kotisivu ResearchGate-verkkopalvelussa. <https://www.researchgate.net/profile/Guido_Koller2>. 10.8.2016.
- Haber, Peter: *Digital Past. Gescgichtswissenschaften im digitalen Zeitalter*. München, Oldenbourg 2011.
- Hitchcock, Tim: Academic History Writing and Its Disconnects. *Journal of Digital Humanities* vol. 1(1) 2011, 1–12.
- "Hoffmann 5-2006". H-NET-verkkosivusto 5/2006. <<http://h-net.msu.edu/cgi-bin/logbrowse.pl?trx=vx&list=H-German&month=0605&week=d&msg=sil6fzghu1HGx238vWd4ZQ&user=&pw>>. 10.8.2016.
- "Hoffmann 7-2006". H-NET-verkkosivusto 7/2006. <<http://h-net.msu.edu/cgi-bin/logbrowse.pl?trx=vx&list=H-German&month=0607&week=a&msg=6Fv3qRGX6inboIbYrtihLw&user=&pw>>. 10.8.2016.
- Husson, Édouard: Comprendre les origiens de la Shoah. Réponse à Florent Brayard. *La vie des idées* 26.2.2009.
- "Internet Archive". Kirjoitus Internet Archive Blog -sivustolla 26.10.2012. <<http://blog.archive.org/2012/10/26/80-terabytes-of-archived-web-crawl-data-available-for-research/>>. 10.8.2016.
- "Katarina Gulliver". Verkkosivusto. <<http://katrinagulliver.com/>>. 10.8.2016
- "Lecture des sources historiques à l'ère numérique". Frédéric Clavertin blogi-kirjoitus 14.11.2012. <<http://histnum.hypotheses.org/1061>>. 10.8.2016.
- Mayer-Schönberger, Viktor; Cukier, Kenneth. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, Houghton Mifflin Harcourt 2013.
- Mitchell, Amy; Hitlin, Paul: *Twitter Reaction to Events Often at Odds with Overall Public Opinion*. Verkkootikkeli, Pew Research Center 4.3.2013. <<http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/>>.
- "Mise en données du monde, mise en données de l'histoire?" Frédéric Clavertin blogikirjoitus 12.7.2013. <<http://histnum.hypotheses.org/1718>>. 10.8.2016.
- "Mommsen 6-2006". H-NET-verkkosivusto 6/2006. <<http://h-net.msu.edu/>>

- cgi-bin/logbrowse.pl?trx=vx&list=H-German&month=0606&week=c&msg=0JJJDhrs/b%2b9SvUzb8eNSg&user=&pw>. 10.8.2016.
- Moretti, Franco: *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso, London 2007.
- "Ngrams". Google Books -verkkosivuston artikkeli. <<http://books.google.com/ngrams>>. 10.8.2016.
- "Open Edition". Verkkosivusto. <<http://www.openedition.org/>>. 10.8.2016.
- Pailler, Frédéric: #twittology: sociologie "just-in-time" et collecte de tweets. Verkkootikkeli, *Sociographie* 8.7.2013a. <http://blog.sociographie.net/post/2013/07/08/sociologie_just-in-time_et_collecte_de_donnees>.
- Pailler, Frédéric: *Préparer un corpus de tweets avec open/google Refine pour le visualiser dans Gephi*. Verkkootikkeli, *Sociographie* 9.4.2013b. <<http://www.clavert.net/lalgorithme-est-grand-je-suis-tout-petit/>>.
- "Peter Haber". Verkkosivusto. <<http://www.hist.net/peter-haber/>>. 10.8.2016.
- "PressForward". Verkkosivusto. <<http://www.pressforward.org/>>. 10.8.2016.
- "The New Republic". Verkkootikkeli. <<https://newrepublic.com/article/112734/what-big-data-will-never-explain>>. 10.8.2016.
- "Think again big data". *Foreign Policy* -lehden verkkojulkaisu. <http://www.foreignpolicy.com/articles/2013/05/09/think_again_big_data>. 10.8.2016.