Avoimet tutkimusmenetelmät yhteiskunta- ja humanistisissa tieteissä: R-kieli laskennallisen analyysin työkaluna

Markus Kainu

Syyskuu 2013

Sisältö

1 Avoimet tutkimusmenetelmät yleisesti

Uusi kappale Wikipedia (2013) sanoo tätä.

Keskustelu tutkimusmenetelmien avoimuudesta on ollut vähäistä yhteiskuntatieteiden ja humanististen tieteiden (YHT) piirissä. Keskeisin syy lienee se, että valtaosa YHT-tutkimuksesta on *laadullista*, jossa ei käytetä laskennallisia analyysimenetelmiä. Ajatus siitä että *diskurssianalyysin* käyttäminen velvoittaisi tiukkojen lisenssiehtojen hyväksymistä ja rahallisen maksun suorittamista on varmasti monelle vieras. Toisaalta määrällisessä tutkimuksessa analyysiohjelmistojen valintaa ovat ohjanneet muut seikat (perinteet, tuttuus) jolloin käytetyn ja opetetun analyysikielen valinnan eri ulottuvuuksia ja pitkäaikaisia vaikutuksia ei olla arvioitu. Joka tapauksessa 1) laskennallisen analyysin YHT-tieteellisten sovellusmahdollisuuksien lisääntymisen sekä 2) avointen analyysimenetelmien kehittymisen myötä on tutkimusmenetelmien avoimuus on nousemassa yhä tärkeämmäksi puheenaiheeksi.

1.1 Avoimen koodin kehittyminen

Avoimilla tutkimusmenetelmillä tarkoitetaan laskennallisia ohjelmointikieliä, joiden on lähdekoodi on avointa ja jotka on lisensointu vapaasti. Vapaan ja avoimen lähdekoodin ohjelmistokehityksen ideana on että kuka tahansa voi muokata lähdekoodia haluamallaan tavalla sekä kopioida ja levittää sekä alkuperäistä että muokattua versiota ja käyttää ohjelmaa mihin tahansa tarkoitukseen (Wikipedia 2013).

Avointen tutkimusmenetelmien kehityksen arvioiminen on vaikeaa. Menetelmien käytön yleisyys on yksi tapa, johon paneudutaan tekstin toisessa luvussa r-kielen osalta. Avoimen koodin ympärille rakentuneet projektit ovat tulleet yhä paremmin näkyville. Kenties tunnetuin avoimeen lähdekoodin pohjautuva projekti on mobiililaitteiden *Android*-käyttöjärjestelmä, joka käyttää linux-ydintä, on avoimesti lisensointu, ja josta on lyhyessä ajassa tullut mobiilikäyttöjärjestelmien markkinajohtaja. Avoimeen lähdekoodiin perustuvat myös internet-palvelimissa suositut linux-käyttöjärjestelmät, samoin kuin niille talletettua tietoa hallinnoivat tietokantaohjelmat (MySql,MariaDB). Nopeasti yleistyvät julkaisujärjestelmät, kuten wordpress, drupal, tai joomla ovat myös suosittuja ja tavalliselle käyttäjälle tuttuja avoimen lähdekoodin projekteja. Tieteellisessä tutkimuksessa avoin lähdekoodin on ollut jo pitkään valtavirtaa. Yli 95 % maailman supertietokoneista toimii linux-käyttöjärjestelmällä, *joku esimerkki geotieteistä* ja *esimerkki lääketieteestä*. Tilastollisen analyysin

Yhteiskuntatieteissä kuitenkin luotetaan edelleen siihen että kaupalliset ohjelmistoratkaisut (esim. Stata,SPSS,SAS,Matlab, jne.) ovat paras tapa vastata muuttuvan yhteiskunnan ja muuttuvan yhteiskuntatieteen analyysitarpeisiin.

1.2 Laskennallisen analyysin sovellusmahdollisuuksien lisääntyminen

Laskennallisen analyysin soveltamismahdollisuudet YHT-tutkimuksessa ovat lisääntyneet aineistojen digitalisoitumisen kautta. New York Times (Lohr 2013) käsittelin keväällä artikkelissaan *Dickens, Austen and Twain, Through a Digital Lens* Matthew L Jockersin tuoretta kirjaa *Macroanalysis Digital Methods and Literary History* (Jockers 2013). Kirjassaan Jockers analysoi vuosien 1780 - 1900 aikana 3592 englanniksi kirjoitettua novellia data-analyysiä ja tilastollisia menetelmiä yhdistelevällä metodilla paljastaen utta tietoa vaikutteiden leviämisestä sekä yksittäisten kirjailijoiden merkityksestä englantilaisen kaunokirjallisuuden historiassa.

Lohr (2013) jatkaa artikkelissaan tulevaisuuden visiointiaan olettaen että digitaaliset analyysimenetelmät tulevat leviämään internet-teollisuudesta ja luonnontieteistä yhä laajemmalle YHT-tutkimukseen.

> Uudet löytämisen työkalut tarjoavat tuoreen näkökulman kulttuuriin; pitkälti samaan tapaan kuin miten mikroskooppi auttoi näkemään elämän hienorakenteet ja miten teleskoopit avasivat näkymät kaukaisiin galakseihin.

Harvardin yliopiston *Institute for Quantitative Social Science*:n johtaja professori Gary King korostaa artikkelissaan *Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science* (s. 3, King 2013) uuden avoimen datan merkitystä perinteiselle määrälliselle yhteiskuntatutkimukselle:

> An important driver of the change sweeping the field is the enormous quantities of highly informative data inundating almost every area we study. In the last half-century, the information base of social science research has primarily come from three sources: survey research, end of period government statistics, and one-off studies of particular people, places, or events. In the next half-century, these sources will still be used and improved, but the number and diversity of other sources of information are increasing exponentially, and are already many orders of magnitude more informative than ever before.

Terveyden ja hyvinvoinnin laitoksen tutkimusprofessori Jussi Simpura (2012) kirjottaa Tiedepolitiikka lehdessä samasta aiheesta:

> Näyttää siltä, että tietotekniikan kehittymisen seurauksena tietovarannot ovat räjähtämässä globaaliksi pilveksi ja että ketjureaktio on jo käynnissä. Se koskettaa yksittäisiä tietovarantoja ja pakottaa niitten kanssa työskenteleviä miettimään uusia toimintatapoja. Avoimen datan aalto etenee globaalin tietopilven tuntumassa. Suomessakin aalto on jo liikkeellä: vuonna 2012 avoimen datan edistämisen ja hyödyntämisen ympärillä on ollut tapahtumia lähes viikottain ja etenemisvauhti on kova. Avoimen datan pelisääntöjä rakenetaan tilanteessa, jossa avoimen datan aalto uhkaa edetä sääntelijöiden tavoittamattomiin. Tällä kaikella alkaa olla kiire, ja tiedemaailmankin on pysyttävä aallon harjalla.

. . . Tilastollisen tutkimuksen käsitettä käytetään suomen kielessä usein melko laveasti. Tässä tekstissä puhutaan laskennallisesta analyysistä, joka nähdään koostuvan yhtäältä tietojenkäsittelytieteellisestä data-analyysistä että toisaalta tilastotieteellisestä data-analyysistä. Määrällisen tutkimusprosessin työmäärästä usein vain pieni osa on varsinaista tilastotieteellistä analyysiä ja suurin osa työstä kuluu datan keräämiseen sekä muokkaamiseen.

Matthew L. Jonckers

Manovich (2013, 22–23) sanoo että

Tässä esseessä käsittelen avointa sähköistä dataa yhteiskuntatieteellisen tutkimuksen näkökulmasta. Avoin data tai *big data* tarkoittaa tässä yhteydessä verrattain uutta sähköiseen muotoon kertyvää dataa, jota ei kerätä jotain tiettyä tutkimuskysymystä silmällä pitäen, vaan data ikäänkuin *kertyy* esimerkiksi dokumentaation tai teknisen sovelluksen keräämänä. Tyypillisinä esimerkkeinä aineistoista voidaan pitää mm. *sosiaalista mediaa* (esim. King, Pan, and Roberts 2012), *vaalidataa* (esim. Lahti, Lehtomäki, and Parkkinen 2013), *poliittisen*

päätöksenteon dataa (esim. Hetherington and Husser 2012) tai vaikkapa *rikollisuuteen ja oikeuslaitoksen toimintaan liittyvää dataa* (esim. Lipsky et al. 2012). Kaikille näille tutkimuksille yhteistä on avoiminen tietovirtojen tietotekninen hyödyntäminen tieteenalan perinteisiin tutkimuskysymyksiin vastaamiseksi.

Suomalainen määrällinen yhteiskuntatutkimus on edelleen vahvasti kiinni Kingin mainitsemissa *menneen puolivuosisadan* aineistoissa sekä menetelmissä. Menetelmällinen kehitys, sikäli kun sitä on ollut, on painottunut uusien *tilastollisten* menetelmien soveltamiseen aineistojen ja tutkimuskysymysten pysyttyä pitkälti ennallaan. Aineistojen osalta **viranomaisrekistereitä** on saatu tutkittavaksi yhä enemmän. Rekisteriaineistotutkimus on menetelmällisesti hyvin erityinen tutkimusala, jolla ei ole sovelluksia juuri pohjoismaita kauemmaksi.

Uudet sähköiset aineistot eivät ole ainoa syy päivittää teoreettista ja menetelmällistä osaamista. Etenkin survey-aineistojen käyttöä uhkaa yhtäältä resurssien väheneminen ja toisaalta ihmisten vähenevä osallistumisinto. Säästöpaineiden alla laajojen survey-aineistojen on uhattuna, etenkin kun uusien kyselytutkimusten vastausprosentit ovat tasolla, joka herättää kysymyksiä tulosten yleistämisestä ja siten koko tutkimusmenetelmän käyttökelpoisuudesta tulevaisuudessa.

Siirtyminen tulevan puolivuosisadan aineistojen hyödyntämiseen tulee olemaan kivinen tie. Se edellyttää merkittävää paradigma-tason muutosta läpi tieteenalan. Käytännössä muutos edellyttää tietotekniikan hyödyntämisen nostamista aivan uudelle tasolle. Menneen ajan aineistojen analysointiin on voinut hankkia yksityiseltä yritykseltä sovelluksen (esim. SPSS, Stata, SAS, jne.), jonka avulla kriteerit täyttävän mallin on voinut suorittaa. Nykyään tieteelliselle tutkimukselle relevantit uudet aineistot ja uudet analyysimenetelmät kehittyvät niin nopeasti, etteivät kaupalliset tuotteet edes yritä pysyä kehityksessä mukana, vaan keskittyvät palvelemaan yrityksiä joskin osin samoissa ison datan haasteissa.

Tieteellisten ohjelmistojen kehitys onkin siirtynyt yhä enemmän tiedeyhteisön varaan ja kehitystyöhön osallistumisesta tullut myös yksi keskeinen akateemisen meritoitumisen muoto. Viime vuosina yhdysvaltalaisten huippuyliopistojen professorirekrytoinneissa on toistuvasti painotettu aktiivisuutta avoimen koodin analyysiohjelmistojen kehitysyhteisöissä.

Meritoitumisen näkökulmasta uudet aineistot ja menetelmien kehittäminen ovat kiinnostavia myös ns. tietovirtojen *ohjaamisen* kannalta. Tietovirroilla voidaan tarkoittaa esimerkiksi Venäjän tilastoviranomaisen Rosstatin jatkuvasti julkaisevaa uutta tilastotietoa tai vaikkapa Twitteriä. Mikäli tutkija onnistuu kehittämään nokkelan tavan hyödyntää tätä tietovirtaa ja jakaa menetelmänsä avoimesti tiedeyhteisön kesken, siirtyvät tutkijat helposti käyttämään tätä menetelmää päästäkseen käsiksi ko. tietoihin. Tietovirtojen kanavoinnista on tulossa myös yksi yliopistojen vetovoimaisuutta lisäävä tekijä. Sama logiikka on R-kielen suosion nopean kasvun takana. R:n suosiosta kertoo paljon myös se, että parhaillaan Courserassa käynnissä olevalle *Computing for Data Analysis*-kursille oli ilmoittautunut yli 40 000 opiskelijaa.

Internetissä jaetun datan lisääntyminen ja kiinnostus sen hyödyntämiseen ovat yksi syy akateemisen julkaisutoiminnan ympärillä tapahtuneeseen liikehdintään. Avoimen tiedon helppo saatavuus on herättänyt kysymyksen siitä, miksi tutkijat eivät voisi yhtä lailla jakaa omia tutkimustuloksiaan avoimesti ilman tai minimaalisin välikäsin. *Elsevier*-kustannustalon liiketoimintamallin kritisoinnista lähtenyt Academic Spring-liike on kuluneen vuoden aikana yhdistänyt laajasti eri maiden tutkijayhteisöjä kyseenalaistamaan nykyistä akateemista julkaisumallia (ks. *Cost of knowledge*). Kritiikin kärki kohdistuu siihen, että isot kustantamot myyvät kovaan hintaan takaisin yliopistoille tutkijoiden veronmaksajien rahoilla tekemiä julkaisuja ja samalla pidättävät tekijänoikeudet itsellään rajoittaen tutkijoiden mahdollisuuksia jakaa omia tuloksiaan vapaasti netissä. Liikehdintä sai alkunsa matemaatikko Timothy Gowersin blogissaan julkaisemasta tekstistä. Blogin kommenteissa eräs lukija nasevasti tiivisti järjestelmän epäkohdat:

> It is utterly absurd that we still have publishers — we write for free (because we want our work read or known), we edit or referee for free and then pay large amounts of money to buy the work back. With the advent of the Web, authors should have eliminated publishers.

Perinteinen akateeminen julkaisumalli on monella tapaa jos ei tiensä päässä. Toiminnan liiketaloudellinen kannattavuus on monella tapaa uhattuna. Tutkimusjulkaisujen kysyntää siirtyy paperille painetusta sähköisiin dokumentteihin, joiden tuotannossa ei enää tarvita samanlaisia isoja pääomia kuin mitä esim. paperikirjojen. Uudet laitteet (tietokoneet, tabletit, älypuhelimet) yleistyvät nopeasti ja ovat osoittautuneet melko hyviksi artikkelikasojen korvaajiksi. Samalla myös kiinnostus open access -lehdet lehtiä kohtaan on lisääntynyt, olkookin että niiden arvo on vielä kaukana arvostettujen perinteisten lehtien tasosta. Kun taas ajatellaan kirjoja niin kustannustaloilla on vielä myös niiden *nimi* tai *maine*, jonka houkuttelemana ne säilyttävät asemansa vielä pitkään. Siitäkin huolimatta, että teknologinen kehitys on jo mitätöinyt kaikki muut kustantajien perinteiset vahvuudet.

Määrällisen tutkimuksen näkökulmasta sähköinen julkaiseminen on noussut yhä kiinnostavammaksi ns. vuorovaikutteisen grafiikan kehittymisen myötä. Grafiikan lisäksi on mahdollista luoda myös laskennallisia analyysiympäristöjä raporttien lomaan, joita voi käyttää suoraan selaimesta käsin. Varsinaisiin julkaisuihin vuorovaikutteisella grafiikalla tai virtuaalisilla laskentaympäristöillä on vielä matkaa, mutta erilaisissa *harmaissa julkaisuissa* ja itse tutkimusprosessissa ne tulevat olemaan hyödyksi jo lähitulevaisuudessa.

Uusien teknologisten sovellusten ohella sähköistä julkaisumuotoa puoltavat myös etenkin laskennallisten tieteiden kohtaamat vaatimukset analyysien *toistettavuudesta* (ks. esim. Peng 2011). Toistettavuudella tarkoitetaan datan ja analyysialgoritmien avointa julkaisemista siten, että lukijan on mahdollista arvioida analyysin pätevyyttä ja johtopäätöksiä. Toistettavuudesta on alettu puhua enemmän myös yhteiskuntatieteellisen tutkimuksen parissa kun samojen iso-

jen kansainvälisten tutkimusaineistojen parissa työskentelee tuhansia tutkijoita, joiden tutkimustulokset ja johtopäätökset ovat usein ristiriitaisia.

Palaan vielä johdannossa viittaamaani New York Times artikkeliin laskennallisista tutkimusmenetelmistä kirjallisuudentutkimuksessa (Lohr 2013). Artikkelissa pohdittiin paljon myös sitä, tarkoittaako *määrän* lisääntyminen laadullisen tutkimuksen aseman heikentymistä. Professori Jonckerin mukaan oikeiden tutkimuskysymysten löytäminen ja analyysin virheiden näkeminen edellyttävät syvää perehtyneisyyttä tieteenalaan yhä edelleen.

> Quantitative tools in the humanities and the social sciences, as in other fields, are most powerful when they are controlled by an intelligent human. Experts with deep knowledge of a subject are needed to ask the right questions and to recognize the shortcomings of statistical models.

> "You'll always need both," says Mr. Jockers, the literary quant. "But we're at a moment now when there is much greater acceptance of these methods than in the past. There will come a time when this kind of analysis is just part of the tool kit in the humanities, as in every other discipline."

Yksittäisen tutkijan on vaikea vastata tähän haasteeseen ja määrällisen analyysin yleistyminen tuleekin vaatimaan kaikilla tieteenaloilla uudenlaista yhteistyötä tutkijoiden ja tieteenalojen kesken. Gary Kingin (s 3., King 2013) mukaan myös Yhdysvalloissa yhteiskuntatieteiden tutkijat ovat jo siirtymässä tutkijankammioista yhteistyöhön kannustaviin, laboratorio-tyyppisiin monitieteisiin tutkimusryhmiin. Laajan sähköisen datan analysointi edellyttää tietoa ja osaamista, jota ei löydy miltään yhteiskuntatieteiden perinteiseltä tieteenalalta.

> Through collaboration across fields, however, we can begin to address the interdisciplinary substantive knowledge needed, along with the engineering, computational, ethical, and informatics challenges before us. (s 3., King 2013)

Uusi digitaalinen data myös hämärtää yhteiskuntatieteissä perinteistä jakoa laadullisiin ja määrällisiin tutkimusotteisiin. Gary King (s 4., King 2013) ennustaa molempien tutkimussuuntien yhdistyvän yhdeksi yhteiskuntatieteeksi, jossa samoja ongelmia ratkotaan yhteistyössä.

> Instead of quantitative researchers trying to build fully automated methods and qualitative researchers trying to make due with traditional human-only methods, both now are heading toward, using, or developing computer-assisted methods that empower both groups.

> This development has the potential to end the divide, to get us working together to solve common problems, and to greatly strengthen the research output of social science as a whole.

Tällainen lähetyminen on ainakin suomalaisessa yhteiskuntatieteessä vielä utopistinen visio. Sähköisen datan yhteiskunnallisen hyödyntämisen kenties kiinnostavin avaus on Aalto-yliopiston ja Helsingin yliopiston tietojenkäsittelytieteen jatko-opiskelijoiden perustama Louhos-projekti. Hyvin kuvaavaa on se, ettei tässä projektissa ole lainkaan yhteiskuntatieteilijöitä, ei määrällisesti tai laadullisesti orientoituneita.

Venäjän ja Itä-Euroopan yhteiskuntatieteellisen tutkimuksen näkökulmasta uudet avoimet aineistot ja menetelmät ovat erityisen ajankohtaisia. Venäjällä viralliset tilastot ovat olleet avoimia vasta muutaman vuoden ajan ja tekniset ratkaisut datan avaamiselle ovat edelleen hyvin takapajuisia. Voidaankin sanoa että jo *menneen puolivuosisadan* aineistojen tehokas käyttö edellyttää Venäjän kohdalla verrattain hyviä *tulevan puolivuosisadan* menetelmien osaamista.

Samaan aikaan internetin merkitys kaikessa yhteiskunnallisessa näyttää voimistuvan. Sosiaalisen median rooli yhteiskunnallisen keskustelun ja jopa mobilisaation kanavana näyttää voimistuvan kaikkialla entisen Neuvostoliiton alueella. Ja yhteiskunnan tutkijan tulee tietysti olla paikalla siellä, missä yhteiskuntaa tehdään.

2 Mikä on R? - Synty ja ominaispiirteet

> R has already won praise and plaudits from established media outlets such as the New York Times, Forbes, Intelligent Enterprise, InfoWorld and The Register. When you consider that R is a high-level computer programming language designed mostly for quants (the nickname for a subspecies of geeks who focus on quantitative analysis), the adoring media attention seems nothing short of astounding.

Joka tapauksessa Smith (2010 s.23) kirjoittaa että paska on aina paskaa.

2.1 Synty

- R-language was iniated by two
- open source

2.2 R-kieli ohjelmointikielenä

- object oriented, s-language bell laboratories (G)UI's, IDE Rstudio,
- contributed packages

2.3 R-kieli tilastollisen laskennan työkaluna

- structure
- visual
- open source, community, licensing, teaching

2.4 R-kielen suosio

- enterprise level services

3 Miten R on mahdollinen? - R-projekti

3.1 Projektin organisaatio

- development vs. user help

3.2 Kielen kehitys

3.3 Käyttäjätuki

- mailing lists - general vs. special interest groups blogs
- q & a sites

3.4 Kontribuoidut paketit

- CRAN - task views R-forge
- Github
- bioconductor

4 Mihin R-kieltä käytetään? - R käytännön työssä

Why R-language is popular

4.1 Tilastollisten menetelmien kehitys

- new methods implemented first in R

4.2 Soveltava tilastotiede

4.2.1 Bio/geo-tieteet

- Geograhical Information Systems

4.2.2 Yhteiskuntatieteet

4.2.3 Humanistiset tieteet - ihmiskielten analyysi

4.3 Yritysanalytiikka

- insurance, big data, banking, industry
- social media: facebook, google, twitter

4.4 Datajournalismi

- Guardian, New York Times, Chicago Herald Tribune

5 Miten R toimii käytännössä? - Yhteiskuntatieteellisten aineistojen analyysi R-kielellä.

5.1 Sanapilvet

5.2 Verkostokartat

5.3 Spatiaalinen visualisointi

5.4 Klusterointi

- insurance, big data, banking, industry
- social media: facebook, google, twitter

6 Kirjallisuus

Hetherington, Marc J., and Jason A. Husser. 2012. "How Trust Matters: The Changing Political Relevance of Political Trust." *American Journal of Political Science* 56 (2): 312–325. doi:10.1111/j.1540-5907.2011.00548.x. http://onlinelibrary.wiley.com/doi/10.1111/j.1540-5907.2011.00548.x/abstract.

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History*. 1st Edition. University of Illinois Press.

King, Gary. 2013. "Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science."

King, Gary, Jennifer Pan, and Molly Roberts. 2012. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review*.

Lahti, Leo, Joona Lehtomäki, and Juuso Parkkinen. 2013. "Louhos: Lähdekoodia Avointen Suomalaisten Datavirtojen Seulontaan." http://louhos.github.com/.

Lipsky, Sherry, Meg Cristofalo, Sarah Reed, Raul Caetano, and Peter Roy-Byrne. 2012. "Racial and Ethnic Disparities in Police-Reported Intimate Partner Violence Perpetration A Mixed Methods Approach." *Journal of Interpersonal Violence* 27 (11) (July): 2144–2162. doi:10.1177/0886260511432152. http://jiv.sagepub.com/content/27/11/2144.

Lohr, Steve. 2013. "Literary History, Seen Through Big Data's Lens." *The New York Times* (January). http://www.nytimes.com/2013/01/27/technology/literary-history-seen-through-big-datas-lens.html.

Manovich, Lev. 2013. *Software Takes Command*. New York: CONTINUUM PUBLISHING CORPORATION.

Peng, Roger D. 2011. "Reproducible Research in Computational Science." *Science* 334 (6060) (December): 1226–1227. doi:10.1126/science.1213847. http://www.sciencemag.org/content/334/6060/1226.

Simpura, Jussi. 2012. "Näkymättömien Sankarien Tiede: Tietovarannot Kansallisaarteena." *Tiedepolitiikka* (4). http://www.saunalahti.fi/~nl03449/tiedepolitiikka_lehti/tp4_12.htm.

Smith, David. 2010. "R Is Hot - How Did a Statistical Programming Language Invented in New Zealand Become a Global Sensation?" Revolution Analytics.

Wikipedia. 2013. "Avoin Lähdekoodi." *Wikipedia*. http://fi.wikipedia.org/w/index.php?title=Avoin_l%C3%A4hdekoodi&oldid=13074631.