Leidenin tutkimusmetriikkamanifesti

Diana Hicks, Paul Wouters, Ludo Waltman, Sarah de Rijcke, Ismael Rafols

22. huhtikuuta 2015

Suomennos Hicksin et al. artikkelista ”Leiden Manifesto for Research Metrics”, *Nature*, April 23, 2015, http://www.leidenmanifesto.org/, http://www.nature.com/news/bibliometrics-the-leiden-manifesto-for-research-metrics-1.17351. Kääntänyt SYN:in tutkimuksen tuen verkoston ohjausryhmän piirissä Johanna Lahikainen (ryhmän muut jäsenet puheenjohtaja Tua Hindersson-Söderholm, Mari Katvala, Anne Lehto ja Eva Tolonen). Ohjausryhmä kiittää Kimmo Tuomista, Pälvi Kaiposta ja Jussi Piipposta. Olemme kiitollisia Maria Forsmanille ja Susanna Nykyrille arvokkaista kommenteista. Erityiskiitos kuuluu manifestin yhteyshenkilölle Diana Hicksille.

Anna näiden kymmenen periaatteen ohjata tutkimuksen arviointia, kehottavat Diana Hicks, Paul Wouters ja kollegat.

Numeroita ja dataa käytetään yhä enenevässä määrin tieteen ohjailemiseen. Tutkimusarvioinnit, jotka aikaisemmin toteutettiin tilaustyönä tehtynä vertaisarviointina, ovat nykyisin rutiinia ja tehdään metriikan keinoin [1]. Ongelmana on, että arviointia johtaa nyt pikemminkin data kuin arvostelukyky. Metriikka on yleistynyt nopeasti: vaikka yleensä tausta-ajatus on hyvä, työtä ei aina osata tehdä asiantuntevasti, ja usein välineiden ja tulosten käyttö ontuu. On vaarana, että vahingoitamme järjestelmää juuri niillä välineillä, joiden tarkoituksena on parantaa sitä, sillä arviointia tekevät enenevässä määrin organisaatiot, joilla ei ole riittävää tietämystä tai jotka eivät ole saaneet opastusta hyvistä käytännöistä ja tulkintojen tekemisestä.

Ennen vuotta 2000 oli olemassa CD-ROM-levy, joka sisälsi Science Citation Index (ISI) –tietokannan, ja sitä käyttivät asiantuntijat erityisanalyyseihin. Vuonna 2002 Thomson Reuters julkaisi internet-portaalin, joka mahdollisti laajalle yleisölle pääsyn Web of Science –tietokantaan. Pian luotiin kilpailevia viittausindeksejä: Elsevierin Scopus (2004-) ja Google Scholar (beta-versio 2004). Tarjolle tuli verkkopohjaisia työkaluja, joilla on helppo vertailla instituutioiden tutkimusten määrää ja vaikuttavuutta. Näitä ovat esimerkiksi InCites (joka käyttää Web of Sciencen dataa) ja Scival (joka käyttää Scopuksen dataa) sekä työkaluja, joilla voidaan analysoida yksittäisten tutkijoiden viittausmääriä Google Scholarin datalla (Publish or Perish, 2007-).

Vuonna 2005 fyysikko Jorge Hirsch Kalifornian yliopistosta San Diegosta esitteli *h*-indeksin ja popularisoi viittausten laskemisen yksittäisille tutkijoille. Kiinnostus lehtien vaikuttavuuskertoimiin (impact factor) on kasvanut tasaisesti vuoden 1995 jälkeen (katso kaavio ”Impact-Factor Obsession”, ”Vaikuttavuuskerroinpakkomielle”).

Viime aikoina sosiaaliseen käyttöön ja verkkokommentointiin liittyvä metriikka on saanut jalansijaa: F1000Prime perustettiin vuonna 2002, Mendeley vuonna 2008 ja Altmetric.com (jota ylläpitää MacMillan Science and Education, Nature Publishing Groupin omistaja) vuonna 2011.

Tutkimusmetriikan asiantuntijoina, yhteiskuntatieteilijöinä ja tutkimushallinnon edustajina, olemme aina vain huolestuneempina todistaneet kaikkialle leviävää tutkimuksen arvioinnin mittareiden väärinkäyttöä. Seuraavat esimerkit ovat vain murto-osa suuresta joukosta. Ympäri maailman yliopistot ovat alkaneet suhtautua pakkomielteisesti asemaansa globaaleissa ranking-tilastoissa (esimerkiksi Shanghain ja Times Higher Educationin rankingit), vaikka meidän mielestämme sellaiset listat perustuvat virheelliseen aineistoon ja mielivaltaisiin mittareihin.

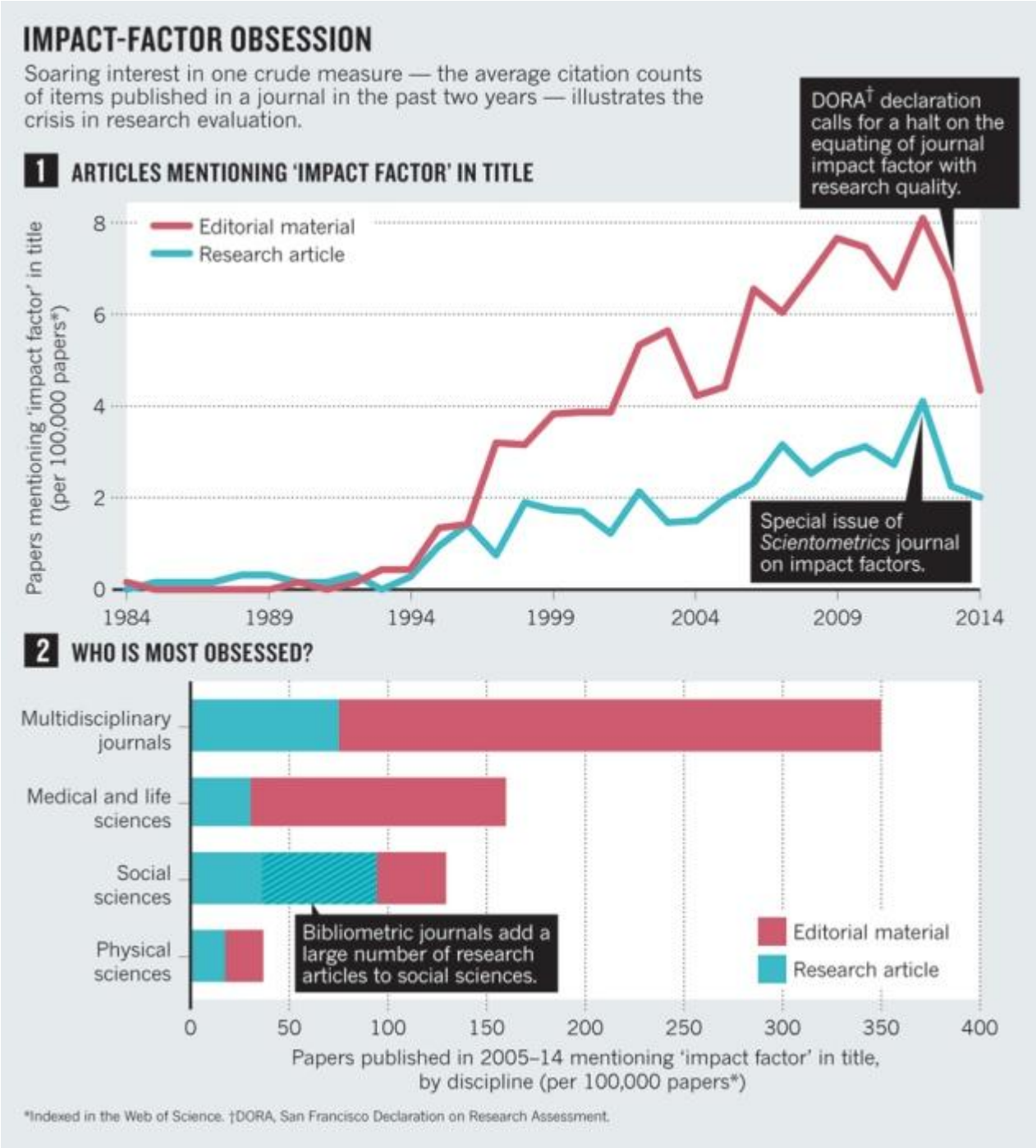
Jotkut rekrytoijat pyytävät työnhakijoiltaan *h*-indeksilukua. Useat yliopistot perustavat ylennyspäätöksensä *h*-indeksiin ja korkeiden vaikuttavuuskertoimien lehdissä julkaistujen artikkelien määrään. Erityisesti biolääketieteessä tutkijoiden ansioluettelot kerskuvat näillä luvuilla. Kaikkialla ohjaajat kehottavat väitöskirjan tekijöitä julkaisemaan korkeiden vaikuttavuuskertoimien lehdissä sekä hankkimaan ulkoista rahoitusta ennen kuin nämä ovat siihen valmiita.

Skandinaviassa ja Kiinassa jotkut yliopistot kohdentavat tutkimusrahoitusta tai bonuksia lukujen perusteella: esimerkiksi laskemalla yksilöllisiä vaikuttavuuslukuja kohdentaakseen ”työresursseja” tai antamalla tutkijoille bonuksia julkaisuista lehdissä, joiden vaikuttavuuskerroin on yli 15 (viite [2]).

Useissa tapauksissa tutkijat ja arvioijat käyttävät yhä monipuolista ja tasapainoista harkintaa. Tutkimusmetriikan väärinkäyttö on kuitenkin levinnyt niin laajalle, että on mahdotonta olla kiinnittämättä siihen huomiota.

Tämän vuoksi esittelemme Leidenin manifestin, joka on nimetty sen konferenssin mukaan, jossa se kiteytyi (katso http://sti2014.cwts.nl). Sen kymmenen periaatetta eivät ole uusia tutkimusmetriikan asiantuntijoille, vaikkei kukaan meistä olisi pystynyt luettelemaan niitä kaikkia, sillä niitä ei ollut aikaisemmin koottu yhteen. Alan merkkihenkilöt, kuten Eugene Garfield (ISI:n perustaja), ovat julkisesti tuoneet esiin joitain näistä periaatteista [3] [4]. Mutta he eivät ole läsnä silloin, kun arvioijat antavan selontekonsa yliopiston hallinnon edustajille, jotka eivät ole olennaisen metodologian tuntijoita. Kun tieteentekijät etsivät kirjallisuutta, jolla todistaisivat arvioinnin vääräksi, he huomaavat aineiston olevan hajallaan lehdissä, jotka ovat heille tuntemattomia ja jotka eivät ole heidän saatavillaan.

Me tarjoamme tämän metriikkapohjaisen tutkimuksen arvioinnin parhaiden käytäntöjen kiteytyksen, jotta tutkijat voivat tarkastella niitä tapoja, joilla heitä arvioidaan ja arvioijat voivat vastuullisesti perustella käyttämänsä mittarit eli indikaattorit.



KYMMENEN PERIAATETTA

1. Määrällisen arvioinnin pitäisi tukea laadullista asiantuntija-arviointia. Kvantitatiivinen metriikka voi kyseenalaistaa alttiutta puolueellisuuteen vertaisarvioinnissa ja helpottaa harkintaa. Tämän pitäisi vahvistaa vertaisarviointia, sillä kollegoiden arviointi ilman olennaista tietoa on vaikeaa. Arvioijien täytyy kuitenkin välttää kiusausta antaa numeroiden tehdä päätöksiä. Mittarit eivät saa syrjäyttää asiantuntevaa harkintaa. Jokainen kantaa vastuunsa omista arvioistaan.

2. Tarkastele suoritusta suhteessa instituution, ryhmän tai tutkijan omaan tutkimusmissioon. Tavoitteet tulisi ilmaista heti alussa ja mittareiden tulisi olla selvästi kytköksissä näihin tavoitteisiin. Mittarien valinnassa ja käyttötavoissa tulisi ottaa huomioon laajempi sosioekonominen ja kulttuurinen konteksti. Tieteentekijöillä on moninaisia tutkimusmissioita. Tutkimus, joka pyrkii akateemisen tiedon edistysaskeleisiin ja koettelee sen rajoja, on erilaista kuin tutkimus, joka pyrkii tuottamaan ratkaisuja yhteiskunnallisiin ongelmiin. Arviointi voidaan perustaa merkittävien akateemisten edistysaskelien sijaan olennaisiin meriitteihin lainsäädännössä, teollisuudessa tai yhteiskunnallisessa vuorovaikutuksessa. Yksi arviointimalli ei sovi kaikkiin tapauksiin.

3. Ota huomioon, että paikallisesti merkittävä tutkimus voi olla huippututkimusta. Monissa maailman kolkissa tutkimuksen laatu on yhtä kuin englanninkielisen julkaisemisen määrä. Esimerkiksi Espanjassa lainsäädäntö ohjaa espanjalaisia tutkijoita julkaisemaan korkean vaikuttavuuskertoimen lehdissä. Tämä vaikuttavuuskerroin lasketaan Web of Science –tietokantaan indeksoiduille lehdille, jotka ovat enimmäkseen yhdysvaltalaisia tai englanninkielisiä. Tämä puolueellisuus on erityisen ongelmallista humanistisissa ja yhteiskuntatieteissä, joiden tutkimus liittyy usein paikallisuuteen ja kansallisuuteen. Monilla muillakin tutkimusaloilla on kansallinen tai paikallinen ulottuvuus – esimerkiksi Saharan eteläpuoleisen Afrikan HIV-epidemiologian tutkimuksella.

   Tämä pluralismi ja yhteiskunnallinen tärkeys vaimentuvat, kun pyritään luomaan papereita, jotka kiinnostaisivat korkean vaikuttavuuskertoimen portinvartijoita: englanninkielisiä lehtiä. Web of Science -tietokannassa eniten viittauksia saaneet espanjalaiset sosiologit työskentelevät joko abstraktien mallien tai yhdysvaltalaisen tutkimusaineiston parissa. Tällöin katoaa se erityisosaaminen, jota esiintyy arvostetuissa espanjankielisissä tiedelehdissä julkaistuissa artikkeleissa: sellaiset aiheet kuin paikallinen työlainsäädäntö, perheissä tehtävä ikääntyneiden hoivatyö tai maahanmuuttajien työllistyminen [5]. Ei-englanninkieliselle korkealaatuiselle tutkimuskirjallisuudelle rakentuva metriikka mahdollistaisi paikallisesti merkittävän tutkimuksen tunnistamisen ja palkitsemisen.

4. Huolehdi, että aineiston kerääminen ja analyysiprosessit ovat avoimia, läpinäkyviä ja yksinkertaisia. Arviointia varten tarvittavien tietokantojen rakentamisessa tulisi seurata selkeästi ilmaistuja sääntöjä, jotka on määritelty ennen tutkimuksen valmistumista. Tämä oli yleinen toimintatapa niille akateemisille ja kaupallisille toimijoille, jotka rakensivat bibliometrisen arvioinnin metodologiaa useiden vuosikymmenten ajan. Nämä ryhmät toimivat vertaisarvioidussa

kirjallisuudessa esitetyn protokollan mukaisesti. Tämä läpinäkyvyys mahdollisti asian tarkastelun. Esimerkiksi vuonna 2010 käyty julkinen keskustelu yksikkömme (Leidenin yliopiston Centre for Science and Technology Studies -keskus Alankomaissa) erään ryhmän käyttämän tärkeän mittarin teknisistä ominaisuuksista johti tämän laskelman muuttamiseen [6]. Uusien kaupallisten tulokkaiden pitäisi yltää samoihin standardeihin; kenenkään ei pitäisi hyväksyä mustan laatikon kaltaista arviointikoneistoa.

Mittarin suhteen yksinkertaisuus on hyve, sillä se parantaa läpinäkyvyyttä. Mutta yksinkertaistava metriikka voi vääristää saavutuksia (katso periaate 7). Arvioijien täytyy tavoitella tasapainoa – yksinkertaisia indikaattoreita, jotka ovat uskollisia tutkimusprosessin monimutkaisuudelle.

5. Anna arvioitaville mahdollisuus tarkistaa data ja analyysit. Datan laadun varmistamiseksi kaikkien tutkijoiden, joihin arviointi kohdistuu, tulee voida tarkistaa, että heidän työnsä tulokset on tunnistettu oikein. Jokaisen, joka johtaa tai ohjaa arviointiprosesseja tulisi varmistaa datan virheettömyys joko varmistamalla se itse tai antamalla kolmannen osapuolen tarkistaa se. Yliopistot voisivat toteuttaa tätä tutkimustietojärjestelmissään ja sen pitäisi olla ohjaava periaate näiden järjestelmien valinnassa. Tarkan ja korkealaatuisen datan kerääminen ja käsittely vaatii aikaa ja rahaa. Sisällytä se budjettiin.

6. Ota huomioon tieteenalojen erot julkaisemisessa ja viittauskäytännöissä. Paras käytäntö on valita sarja mahdollisia mittareita ja antaa tieteenalojen itse valita niistä heille sopivin. Muutama vuosi sitten eurooppalainen historiantutkijaryhmä sai kansallisessa vertaisarvioinnissa verraten huonon luokituksen, koska he kirjoittavat kirjoja, eivätkä artikkeleita Web Of Science –tietokantaan indeksoituihin lehtiin. Epäonnekseen kyseiset historioitsijat kuuluivat psykologian laitokseen. Historian ja yhteiskuntatieteiden tutkijoille on tärkeää, että kirjat ja kansallisilla kielillä julkaistu kirjallisuus sisällytetään heidän julkaisumääriinsä, kun taas tietojenkäsittelytieteilijöille tulee laskea mukaan konferenssijulkaisut.

   Viittausmäärät vaihtelevat alakohtaisesti: matematiikan korkeimmalle rankattujen lehtien vaikuttavuuskertoimet ovat kolmen paikkeilla, kun taas solubiologian huippulehtien vaikuttavuuskertoimet ovat noin 30. Tarvitaan normalisoituja mittareita ja kaikkein järein normalisointi perustuu prosenttipisteisiin: jokainen julkaisu suhteutetaan siihen prosenttipisteytykseen, johon julkaisu kuuluu oman tieteenalansa viittausjakaumassa (esimerkiksi kärkijoukon ensimmäiseen prosenttiin kuuluvat, kymmenen prosentin joukkoon kuuluvat, 20 prosentin joukkoon kuuluvat). Yksittäinen paljon viitattu julkaisu parantaa hieman yliopiston

rankingia sellaisessa tilastossa, joka perustuu prosenttipisteindikaattoreihin, mutta viittauskeskiarvoihin perustuvassa rankingissa se saattaa työntää yliopiston keskitasolta kärkijoukkoon [7].

7. Perusta yksittäisten tutkijoiden arviointi heidän portfolionsa laadulliseen tarkasteluun. Mitä vanhempi ihminen on, sitä korkeampi *h*-indeksi hänellä on, vaikka uusia julkaisuja ei olisikaan. *H*-indeksissä on tieteenalakohtaisia eroja: biotieteissä saavutetaan luku 200, fysiikassa 100 ja yhteiskuntatieteissä 20–30 (viite [8]). Luku riippuu tietokannasta: tietojenkäsittelytieteiden tutkijoiden *h*-indeksi Web of Science-tietokannassa voi olla 10, mutta Google Scholarissa 20–30 [9]. Tutkijan töiden lukeminen ja arvostelu on paljon tarkoituksenmukaisempaa kuin yhteen lukuun luottaminen. Myös silloin, kun verrataan suuria määriä tutkijoita toisiinsa, yksilön osaamisen, kokemuksen, aktiviteettien ja vaikutuksen huomioonottava lähestymistapa on paras.

8. Vältä väärin kohdistettua konkretiaa ja epäluotettavaa tarkkuutta. Tieteen ja teknologian mittarit ovat alttiita käsitteelliselle monitulkintaisuudelle ja epävarmuudelle sekä vaativat olettamuksia, jotka eivät ole hyväksyttyjä kaikkialla. Esimerkiksi viittausmäärien merkityksestä on väitelty pitkään. Täten paras käytäntö on käyttää useita mittareita, jotta saavutetaan vankka ja moniarvoinen kokonaiskuva. Jos määrällistä epävarmuutta ja virheitä voidaan ilmaista esimerkiksi käyttämällä virhegraafeja, tämä tieto pitäisi sisällyttää julkaistuihin indikaattoriarvoihin. Jos tämä ei ole mahdollista, mittareiden valmistajien pitäisi ainakin välttää väärää tarkkuutta. Esimerkiksi lehden vaikuttavuuskerroin on ilmaistu kolmella desimaaliluvulla tasapelin eli monen samalle sijalle tulevan estämiseksi. Kuitenkaan – kun otetaan huomioon käsitteellinen monitulkintaisuus ja viittauslukujen satunnainen vaihtelevuus – ei ole järkevää erotella lehtiä, joiden erot vaikuttavuuskertoimissa ovat hyvin pieniä. Vältä epäluotettavaa tarkkuutta: vain yksi desimaali on perusteltavissa.

9. Myönnä järjestelmään kuuluvat arvioinnin ja mittarien seuraukset. Mittarit muuttavat järjestelmää luomillaan kannustimilla. Nämä vaikutukset pitäisi ennakoida. Tämä tarkoittaa, että sarja indikaattoreita on aina parempi – yksittäinen rohkaisee pelaamista ja tavoitteen syrjäyttämistä (mittaamisesta tulee tavoite). Esimerkiksi 1990-luvulla Australiassa tutkimusta rahoitettiin instituution julkaisemien julkaisujen määrään pohjautuvalla mallilla. Yliopistot pystyivät laskemaan julkaisun ”arvon” vertaisarvioidussa lehdessä: vuonna 2000 se oli 800 Australian dollaria (noin 480 Yhdysvaltain dollaria vuonna 2000) tutkimuksen rahoituksesta. Kuten arvata saattaa, australialaisten tutkijoiden julkaisujen määrä kasvoi, mutta ne julkaistiin vähemmän viitatuissa lehdissä. Tämä antaa viitteen siitä, että artikkelien laatu laski [10].

10. Tutki ja päivitä mittareita säännöllisesti. Tutkimuksen tehtävät ja arvioinnin tavoitteet liikkuvat, ja tutkimusjärjestelmä kehittyy yhdessä sen kanssa. Kerran hyödyllinen metriikka muuttuu riittämättömäksi, uusia ilmestyy. Mittarijärjestelmiä tulee tarkastella uudelleen ja ehkä uudistaa. Australia ymmärsi oman yksinkertaistetun mallinsa vaikutukset ja esitteli vuonna 2010 uuden, laatua painottavan ja monisyisemmän ”Excellence in Research for Australia” -hankkeen.

SEURAAVAT ASKELEET

Jos näitä kymmentä periaatetta noudatetaan, tutkimuksen arvioinnilla voi olla tärkeä osa tieteen ja sen yhteiskunnallisen vuorovaikutuksen kehittämisessä. Tutkimusmetriikka voi tarjota ratkaisevan tärkeää tietoa, jota yksilön asiantuntijuudella olisi vaikea kerätä tai ymmärtää. Mutta tämä määrällinen tieto ei saa muuttaa muotoaan työvälineestä itse tavoitteeksi.

Parhaat päätökset tehdään, kun yhdistetään vankkoja tilastoja sekä herkkyyttä tavoitteen ja arvioitavan tutkimuksen suhteen. Sekä määrällistä että laadullista näyttöä tarvitaan – molemmat ovat omalla tavallaan objektiivisia. Tieteen suhteen tehtävän päätöksenteon täytyy perustua laadukkaisiin prosesseihin, joita ohjaa mahdollisimman korkealaatuinen data.

Diana Hicks toimii yhteiskuntapolitiikan professorina Georgia Institute of Technologyssa Atlantassa, Yhdysvalloissa. Paul Wouters toimii skientometriikan professorina sekä johtajana, Ludo Waltman tutkijana ja Sarah de Rijcke apulaisprofessorina Centre for Science and Technology Studies-yksikössä Leidenin yliopistossa, Alankomaissa. Ismael Rafols toimii tiedepolitiikan tutkijana Espanjan kansallisessa tutkimusneuvostossa ja Valencian ammattikorkeakoulussa.
Sähköposti: diana.hicks@pubpolicy.gatech.edu

---

[1] Wouters, P. in *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (eds Cronin, B. & Sugimoto, C.) 47–66 (MIT Press, 2014).

[2] Shao, J. & Shen, H. *Learned Publ.* 24, 95–97 (2011).
[3] Seglen, P. O. *Br. Med. J.* 314, 498–502 (1997).
[4] Garfield, E. J. *Am. Med. Assoc.* 295, 90–93 (2006).
[5] López Piñeiro, C. & Hicks, D. *Res. Eval.* 24, 78–89 (2015).
[6] van Raan, A. F. J., van Leeuwen, T. N., Visser, M. S., van Eck, N. J. & Waltman, L. J. *Informetrics* 4, 431–435 (2010).
[7] Waltman, L. et al. *J. Am. Soc. Inf. Sci. Technol.* 63, 2419–2432 (2012).
[8] Hirsch, J. E. Proc. *Natl Acad. Sci. USA* 102, 16569–16572 (2005).
[9] Bar-Ilan, J. *Scientometrics* 74, 257–271 (2008).
[10] Butler, L. *Res. Policy* 32, 143–155 (2003).