Liikenne- ja viestintäministeriö

Lausunto kansallisesta big data-strategiasta

Dataan liittyen laadittu skenaario on monipuolinen ja sen pohjalta voidaan saada alkuun monenlaisia toimia. Ubiikin ongelmat on tekstissä kuitenkin käsitelty melko pintapuolisesti, koska massatiedonkeruuta ja tiedonlouhintaa voidaan käyttää moniin tarkoituksiin ja intresseihin (mm. kaupalliset, julkiset, rikolliset, tiedustelu, sodankäynti). Kiire hyödyntää datamassoja kaupallisiin ja julkisiin tarkoituksiin siten, että kartalla ei vielä ole oikein kokonaiskuva balanssissa esillä. Kuten EU:ssa, Suomessakin tarvittaisiin tutkimusta, jossa molemmat puolet tulevat realistisessa balanssissa esille. Muuten käydään kiireellä rakentamaan jotain, jonka massiiviset ongelmat tulevat esille vasta myöhemmin, kun niille ei ehkä voi tehdä enää mitään: ”Tarvitaan yhteiskunnallista ja juridista ymmärrystä siitä, miten aineistoihin liittyvää lainsäädäntöä voidaan kehittää tavalla, joka vie eteenpäin toivottuja tapahtumainkulkuja.”

Käsitteistössä kannattaa muistaa, että pelkkä data ei varsinkaan ihmisyksilölle riitä vaan tarvitaan datan jalostamista sen saattamiseksi käsitettävään muotoon, jotta siitä olisi aidosti hyötyä ihmisen oppimiselle ja tietämyksen kasvamiselle. Pelkän data-puheen rinnalla tulee siten nostaa datan hallinnan ja metadatan luomisen puhe, jossa kirjastolaitoksella on merkittävää osaamista. Oman, mittavan haasteensa muodostaa datamassojen pitkäaikaissäilytyksen hallinta: suurin osa tällä hetkellä tuotetusta datasta on laitteistoriippuvaa, joten esitetyt tavoitteet yhteisten käytänteiden ja standardien luomisesta tulisi aloittaa heti.

Useat maat pyrkivät big data-tutkimuksen ykköskaartiin, kuten strategialuonnoksessa todetaan. Mm. LERU ja LIBER ovat painottaneet, että eurooppalaisen tieteen kilpailukyky on vaakalaudalla esidigitaaliseen aikaan juuttuneen tekijänoikeussäätelyn vuoksi. Digitaalisessa maailmassa tieteelliset lehtiartikkelit ja tieteellinen kirjallisuus ovat arvokkaita tiedonlouhinnan materiaaleja: teksteissä voi piillä tietoa koneellisesti yhdistelemällä löydettäviä havaintoja ja hypoteeseja, joita yksittäisten tieteellisten artikkeleiden kirjoittajat eivät ole huomanneet tai edes voineet huomata. Näitä havaintoja on menestyksellisesti käytetty hyväksi mm. lääkeaineiden etsinnässä. Ihmistieteissä, joiden tutkimuskohteena ovat tekstit, big data avaa aivan uusia mahdollisuuksia tehdä tutkimusta ja löytää havaintoja: sen vuoksi tutkimusjulkaisujen avoimuus ja kaikkien julkaisujen avaaminen tutkimuskäyttöön datan louhinnalle on tieteen kehittymisen kannalta ensiarvoisen tärkeää.

Tällä hetkellä on arvioitu että eurooppalaiset yliopistot käyttävät yhteensä 2 miljardia euroja STM-alueella julkaistun tieteellisen kirjallisuuden hankintaan (suurimmaksi osaksi kyse on e-lehtien tilauslisensseistä).

Tiedeyhteisö tarvitsee oikeuden paitsi lukea tieteellisiä e-lehtiä ja -kirjoja, myös oikeuden käyttää tieteellistä kirjallisuutta systemaattisesti tiedonlouhinnan kohteena. Tällaisen oikeuden hankkiminen jälleen uusilla lisensseillä on hankalaa ja kallista. Tekijänoikeussäätelyyn tarvitaankin viipymättä ei-kaupallista tutkimusta koskeva poikkeus, joka oikeuttaa tieteellisen kirjallisuuden käyttämisen tiedonlouhintaan julkaisijalta tai välittäjältä lupaa kysymättä (e-lehden tilauslisenssi riittää; erityistä tiedonlouhintalisenssiä ei siis tarvittaisi). Jos tutkijat tai heitä palvelevat kirjastot joutuvat jatkuvasti miettimään tieteelliseen kirjallisuuteen liittyviä sopimusoikeudellisia ja muita velvoitteita, tehokkaan tutkimuskirjallisuutta koskevan tiedonlouhinnan mahdollisuudet minimoituvat. Strategiassa esimerkkinä käytetyn Watsonin kaltaisten sovellusten, jotka käyvät ”muutamassa sekunnissa tietokannastaan läpi hakusanojen avulla tuhansia oppikirjoja ja satoja tuhansia tieteellisiä artikkeleita” rakentamisen ja hyödyntämisen tulee olla myös juridisesti mahdollista.

Siirtyminen jälkidigitaaliseen yhteiskuntaan, jossa kaikki kansalaiset ja toimijat ovat riippuvaisia suurista datamassoista, vaatii sekä peruskoulutusta että laajaa täydennyskoulutusta. Nykyisen elinikäodotteen mukaan tänä vuonna keskimääräisessä iässä eläkkeelle jäävät ovat aktiivisia toimijoita vielä 2030 – 2040-luvuilla!

Termien suomentaminen on myös tarpeen: big data tarvitsee suomenkielisen nimen, jotta se hahmottuisi kaikille.

Kunnioittavasti,

Jarmo Saarti, FT, Dosentti

Suomen tieteellisen kirjastoseuran puheenjohtaja

jarmo.saarti@uef.fi