

Päiväys
17.6.2014

Liikenne- ja viestintäministeriö

Viite: LVM138/05/2014

Asia: LVM lausuntopyyntö kansallisesta Big data-strategiasta.

	CSC tieteen tietotekniikan keskus Oy
Lausunnon laatija:	Aleksi Kallio Pirjo-Leena Forsström

CSC kiittää mahdollisuudesta antaa lausunto LVM laatimasta ehdotuksesta kansallisesta Big data – strategiasta.

Strategian tarkoituksena on tunnistaa ja löytää Suomen kannalta oleellisia alueita, joilla big data voi tuoda arvoa niin yksityiselle kuin julkiselle sektorille. Lausunnolla oleva strategia on hyvä kooste keskeisistä kysymyksistä, mahdollisuuksista, edellytyksistä sekä erityisesti toimenpiteistä. Toimenpiteiden konkreettisuus ja selkeys ovat tämän strategian ansioita.

Strategia kaipaa kuitenkin käsiteiden selkeyttämistä. Jos ei lopullista määritelmää vaikkapa big datalle voi antaakaan, olisi hyvä, että kuvattaisiin selkeästi missä merkityksessä sitä on tässä raportissa käsitelty. Selvityksestä jää mielikuva, että kyse on kaikesta datasta – joka heikentää strategian arvoa ja mahdollisten toimenpiteiden fokusointia.

Toisaalta, moni ehdotettu toimenpide on osa yleisempää prosessien digitalisoitumisen mukanaan tuomaa tarvetta, eikä välttämättä tarvitse erityisiä big data –toimenpiteitä (vrt. sovellusosaajien koulutus, standardointi, infrastruktuurit). Liialla fokuisoimisella esim. osaajia koulutettaessa jäädään trendien muuttuessa tyhjän päälle.

Alun johdannossa kaipaisi hyvin teknologiavoittoisen näkemyksen rinnalle kuvausta siitä miten tiedon ja ymmärryksen tasolla (data-informaatio-tieto-ymmärrys) big data vaikuttaa. Tämän toimintatason ymmärryksen soisi vahvistuvan koko strategiassa.

Big dataan liittyvistä uhista nostetaan joitakin esiin. Kenties tärkein on toiminnan tasolla siirtyminen hypoteesien testaamisesta suoraan ilmiön mittaamiseen. Tämä tuo luottamuksen ongelman: big datan täytyy olla oikeaa ja luotettavaa ja edustavaa, jos suoraan siitä vedetään johtopäätökset. Yksi uhkista on ymmärtämättömyys big datan ennustekyvyyistä.

Kappaleessa 1.3 tulee tarkentaa että big data ja avoin data ovat osa Euroopan Komission Digital agenda for Europe -kärkihanketta. Siihen liittyvä puiteohjelma on nimeltään Horizon 2020, ja siitä rahoitetaan eurooppalaisia tutkimus- ja innovointihankkeita vuosina 2014–2020 yhteensä lähes 80 miljardilla eurolla. Komission tavoitteena on luoda tämän puiteohjelman avulla kasvua ja uusia työpaikkoja sekä parantaa yritysten asemaa globaalissa kilpailussa.

Kappaleessa 2.4 tai kappaleessa 2.8 olisi hyvä mainita, että tekijänoikeudelliset epäselvyydet toistaiseksi hankaloittavat Suomessa tekstin- ja datanlouhintaa (TDM), jolla on etenkin humanistisilla aloilla kuten kielitieteissä ja kirjallisuudessa vahvasti tutkimusta haittaava merkitys. USA:ssa sama asia on ratkaistu niin sanotulla Fair Use –säännöksellä, joka sallii TDM:n tutkimuskäytön.

Kappaleessa 2 kannattaa erikseen mainita koneluettavan metadatan tärkeys. Tieteenalakohtaisilla metadatatamalleilla kuvaillaan tietoaineistot ja mahdollistetaan tietoaineistojen oikea jatkokäyttö. Joillekin tieteenaloille kuten yhteiskuntatieteille on jo olemassa metadatastandardit, osalle niitä ollaan vasta kehittämässä.

Kappaleessa 2.9 puhutaan yhteistyöstä ja datan vaihdosta. Tieteessä ja tutkimuksessa datan jakaminen ei ole vielä ollen kovin yleistä. Sitä auttaisi jakamisesta annettava selkeä meriitti, samaan tapaan kuin tieteellisistä julkaisuista. Muutamat rahoittajat jo vaativat datanhallinnan suunnitelmaa (DMP), mutta datan avoimeksi saattaminen on vielä suosituksien varassa. Datan määrän kasvaessa saattaa tulla vastaan tilanne, jossa julkisella puolella joudutaan valitsemaan pitempään säilytettävät aineistot. Valintaan vaikuttavia kriteerejä voivat olla esimerkiksi aineistojen kypsyys ja arvo.

Kappaleessa 3.1 ehdotamme mainittavan eurooppalaisiin tutkimusinfrastruktuureihin kuuluvan ELIXIR –hankkeen, joka on keskeinen Suomen bio- ja lääketieteellisen tutkimuksen kannalta.

Kappaleessa 5 puhutaan tiedon etsinnästä. Tähän liittyy oleellisesti semanttinen verkko ja ontologiat teknologisenä ratkaisuna. Suomessa on lähdetty jo 2003 kehittämään semanttisen webin ideologian mukaisia teknologisia ratkaisuja, Aalto-yliopistolla Eero Hyvösen johtamassa tutkimusryhmässä. Parhaillaan Kansalliskirjastossa on menossa ONKI-projekti, jossa kiinnitetään huomiota ontologioiden käytettävyyteen ja sovellettavuuteen erilaisissa yhteyksissä.

Esitetyt toimenpiteet ovat kannatettavia. Seuraavassa joitakin tarkennuksia.

Koulutukseen ja osaamiseen liittyvät ehdotukset (kappale 6, d-g sekä h-k) ovat tärkeitä, mutta edellyttävät uudelleen muotoilua, kohdentamista tai poistamista. Termit, kuten ”big data opetustaidot” tulisi avata. Lisäksi tulisi varmistaa se, että ehdotettu toimenpide on ehdotetulla tavalla kartoitettavissa, koordinoitavissa tai osallistuttavissa.

Soveltavan osaamisen osalta tulisi tarkentaa ehdotukset keskeisiksi toimenpiteiksi alan osaajien ja osaamisen varmistamiseksi. Tässä kannattaa ottaa huomioon ICT2015 –toimenpideohjelman Big Dataa koskevat ehdotukset.

CSC:n näkemys on, että infrastruktuurin haasteiden ratkaiseminen on pakollinen ensiaskel, jotta suurten tietoa-aineistojen hyödyntämisessä päästään eteenpäin: big datassa vaatimus poikkeukselliselle infrastruktuurille on sisäänrakennettu. CSC kehittää aktiivisesti big data –infrastruktuuria oman alueensa tarpeisiin. Dataintensiivinen laskenta on tunnistettu merkittäväksi kasvualueeksi ja palvelun kehittämisestä vastaamaan on hiljattain perustettu oma ryhmänsä. Merkittäviä esimerkkejä jo tehdystä big data –palvelukehityksestä ovat mm. FIMM:in ja THL:n suurten biolääketieteiden tietomassojen käsittelyyn tarkoitettu pilvilaskentaympäristö, jolla käsiteltiin 7 000 teratavua dataa vuonna 2013, ja jo lähes 10 vuotta toiminnassa ollut dataintensiivisen laskennan pilvipalvelumuotoinen Chipster-järjestelmä. CSC:lle on rakennettu big data –osaamista määrätietoisesti mm. yhteistyössä Aalto-yliopiston kanssa. Näin ollen CSC näkisi mielellään, että sen keräämää kokemusta ja osaamista hyödynnetään myös tieteellisen laskennan ulkopuolella – yrityksen neutraali asema voittoa tavoittelemattomana valtio-omisteisena toimijana helpottaa tarvittavien yhteistyörakenteiden kehittämistä.

Tärkeitä big datan käytön kannalta ovat myös sääntelyyn liittyvät kohdat (t-v). Olisi tärkeää, että samassa yhteydessä pohdittaisiin myös tekijänoikeuksiin liittyviä kysymyksiä, sekä sopimusmalleista, joilla säädellään omistajuuden ja IPR:n muodostumisesta. Samassa yhteydessä olisi hyvä ottaa kantaa lisensseihin, jotka antavat mahdollisuuden big datan avaamiseen ja hyödyntämiseen muille. Koneluettavien lisenssien hyödyntäminen on yksi tärkeimmistä edellytyksistä big data –infrastruktuurin toimivuudelle. Toinen on riittävä hallinnollisten ja kuvailevien metatietojen liittäminen data-aineistoihin. Tällainen mahdollistaa laajan uudelleen käytön.

Kokonaisuudessaan strategia luo hyvän pohjan Big datan mahdollisuuksien hyödyntämiseen laajasti yhteiskunnan tarpeisiin. Onnittelemme liikenne- ja viestintäministeriötä hyvästä työstä.