Tiedonhallintasuunnitelma tehostaa tutkimusdatan käyttöä

Jussi Nuorteva

Kun Yhdysvaltain Department of Energy ja National Institute of Health vuonna 1990 yhdessä ilmoittivat pyrkimyksestään määritellä ihmisen geeniperimä, arvioitiin työn kestävän viisitoista vuotta. Jo kesäkuussa 2000 saattoivat presidentti George W. Bush ja Ison-Britannian pääministeri Tony Blair kuitenkin kertoa, että kansainvälinen suurhanke, Human Genome Project, oli saatu lähes päätökseen. Lopullisesti hanke valmistui vuonna 2003, kun ihmisen kaikki geenit oli saatu tunnistetuiksi ja DNA-sekvenssit kuvatuiksi.

Miten sitten oli mahdollista, että monilta osin ennakoitua laajemmaksi ja monimutkaisemmaksi osoittautunut hanke onnistuttiin viemään läpi jopa laskettua nopeammin? Suurin syy tähän oli tietoverkkojen ja tietotekniikan huikealla kehityksellä, jota 1990-luvun alussa ei ollut vielä osattu täysin ennakoida. Tieteellistä tietoa voitiin kuitenkin jo 1990-luvun lopulla välittää nopeasti koko tutkijayhteisölle käyttämällä yhtenäisiä datastandardeja ja hyödyntämällä tehokkaita tietoverkkoja ja tietokoneiden lähes eksponentiaalisesti kasvanutta laskentakapasiteettia.

> *" The emerging vision is to use cyberinfrastructure to build more ubiquitous, comprehensive digital environments that become interactive and functionally complete for research communities in terms of people, data, information, tools and instruments and that operate at unprecedented levels of computational, storage, and data transfer capacity."*
>
> *Revolutionizing Science and Engineering through Cyberinfrastructures*. National Science Foundation 2003.

Tutkimuksen tiedonhallinnan peruslähtökohdat muuttuivat radikaalisti 1990- ja 2000-luvulla kaikkialla maailmassa. Kyky hallita ja hyödyntää tehokkaasti tutkimuksen nopeasti kasvavaa tietomassaa julistettiin 2000-luvun alussa nopeasti yhdeksi modernin tieteellisen tutkimuksen ja teknologian suurimmista mahdollisuuksista Pohjois-Amerikassa.

Human Genome -projekti oli loistava näyttö siitä, mitä mahdollisuuksia sähköisen tutkimusdatan hallinta voi avata tieteelliselle tutkimukselle. Samana vuonna jolloin tuo suurhanke saatiin päätökseen, Yhdysvaltain National Science Foundation (NSF) julkaisi suurta huomiota saavuttaneen strategisen raporttinsa *Revolutionizing Science and Engineering through Cyberinfrastructures* (http://www.nsf.gov/od/oci/reports/atkins.pdf). Raportissa julistettiin, että tieteelle ja teknologialle on sarastamassa uusi aikakausi, jota kantavat laskennallisen tieteen sekä informaatio- ja kommunikaatioteknologian nopea kehitys. Raportissa todettiin, että monilla tieteenaloilla oli tapahtunut suoranainen vallankumous, kun tutkijat olivat ryhtyneet tehokkaasti hyödyntämään tieteellistä laskentaa ja digitaalisia tietoaineistoja sekä verkottuneita toimintamalleja.[1]

NSF vaati kyberinfrastruktuuristrategiassaan sekä Yhdysvaltain hallitusta ja tiedeyhteisöä että yksityisiä yrityksiä panostamaan uuteen osaamiseen, vahvistamaan sähköisen tiedonhallinnan vaatimaa infrastruktuuria sekä luomaan grid-pohjaisia ratkaisuja tietojenkäsittelyn voimavarojen paremmaksi hyödyntämiseksi.

1 Toinen digitaalisia tietoaineistoja koskeva keskeinen NSF:n raportti ilmestyi vuonna 2005 National Science Boardin julkaisemana: *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. National Science Foundation 2005. www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

Kanadan yhteiskuntatieteellis-humanistinen toimikunta (*Social Science and Humanities Research Council*) ja Kansallisarkisto julkaisivat puolestaan vuonna 2002 selvityksen, jossa voimakkaasti tuotiin esiin tarve säilyttää tehokkaammin tutkimusdataa ja perustaa sen käyttöä edistämään erillinen tietoarkisto. Raportissa kannettiin huolta tutkimuksen voimavarojen hyödyntämisestä sekä julkisen rahoituksen haaskaamisesta, johon puutteellisen tutkimusdatan hallinnan katsottiin olevan osaltaan syynä.

Yhdysvalloissa NSF perusti pian raporttinsa julkaisemisen jälkeen alaisuuteensa kyberinfrastruktuuritoimiston *(Office of Cyberinfrastructure*, OIC) edistämään tieteen yhdysvaltalaisten toimijoiden koordinaatiota ja kykyä hallita tieteellistä informaatiota. (www.nsf.gov/od/oci/about.jsp).

Pääpaino OIC:n toiminnassa on superlaskennassa, supertietokoneiden laskentakapasiteettia hyödyntävissä grid-tyyppisissä ratkaisuissa, digitaalisten aineistomassojen säilyttämiseen ja käyttöön liittyvissä kysymyksissä, tiedonhallinnan ohjelmien ja hakujärjestelmien kehittämisessä sekä uudenlaisen toimintaympäristön hallitsevien asiantuntijoiden koulutuksessa.

Digitaalisen tutkimustiedon hallinnan myötä tutkimukseen on syntynyt uusia ammattiryhmiä. Näitä ovat esimerkiksi *data scientist*, *data archivist*, *information manager* jne. Näiden ammattilaisten tehtävänä on kehittää tieteellistä mallintamista sekä luoda kansainvälisiä datastandardeja, jotka mahdollistavat verkottuneen tavan koota suuria sähköisten aineistojen tietopankkeja, joissa olevaa dataa voidaan käyttää yhtenäisten käytäntöjen avulla ilman ajan ja paikan asettamia rajoja. Metatietomallit, kehittyneet hakukoneet, semanttinen web ja ontologiat ovat aineistonhallinnan keskeisiä välineitä. Laskennallisten tieteiden osaamista tarvitaan muuttuneessa toimintaympäristössä yhtä lailla luonnontieteissä kuin humanistisillakin aloilla.

> *"…Although billions of dollars are spent each year collecting data, Canada lacks the necessary infrastructure to ensure these data are preserved and made publicly available. This limits the returns that can be made on our public investments in research and undermines good public stewardship… The core mission of a research data-archive is not to preserve the recorded memory of a group, organization or nation, but to provide a vital service to the research community."*
>
> *We build understanding. Final report. National Data Archive Consultation: Building Infrastructure for Access to and Preservation of Research Data*. Social Sciences and Humanities Research Council of Canada & National Archives of Canada. 2002. (www.sshrc.ca/web/about/publications/da_finalreport_e.pdf)

Kesällä 2006 tieteellisen laskennan kansainvälisesti tärkeimpiin keskuksiin lukeutuva San Diego Supercomputer Center (SDSC) ja Yhdysvaltain Kansallisarkisto (*The National Archives and Record Administration*, NARA) solmivat NSF:n kyberinfrastruktuuritoimiston myötävaikutuksella sopimuksen, jonka nojalla SDSC otti päävastuun sähköisen asiakirjahallinnon sekä tieteellisen ja teknologisen tutkimuksen tietodatan säilyttämisestä. NARA puolestaan vastasi aineistojen elinkaaren hallintaan liittyvistä kysymyksistä. SDSC toimii myös pohjoisamerikkalaisten grid-järjestelmien koordinaattorina. (www.nsf.gov/news/news_summ.jsp?cntn_id=107068).

Myös Euroopan unionissa huomattiin 2000-luvun puolivälissä, että tutkimusjärjestelmän kilpailukyvyn vahvistamiseksi tutkimuksen tiedonhallintaan tulee kiinnittää aivan erityistä huomiota. EU:n tietoyhteiskuntakehityksestä vastaava komissaari Viviane Reading korosti helmikuussa 2007 pitämässään puheessa, että tiedonhallinnan menetelmien kehittämisen ja uusien sovellusten etsimisen katsotaan yleisesti avaavan tieteelliselle tutkimukselle valtaisan uuden potentiaalin. Reading totesi myös, että jo nyt on havaittavissa kehitys tutkimustiedon elinkaaren hallintaan tutkimusdatan luomisesta tieteelliseen julkaisemiseen ("…a trend towards a continuum of the scientific information space, from data to publications"). Lisäksi komissaari muistutti, että monet tutkimusorganisaatiot ja tutkijayhteisöt ovat ryhtyneet kokoamaan erityisiä data-arkistoja ja rahoittajaorganisaatiot ovat puolestaan alkaneet edellyttää rahoittamil-

taan hankkeilta niiden tuottaman tutkimusdatan avointa säilyttämistä, jotta se olisi kaikkien saatavilla.

Open access -periaatetta tutkimusdatan hallinnassa edellytti myös OECD vuonna 2004 antamassaan suosituksessa. Sen pohjalta OECD julkaisi tarkennetut linjaukset vuonna 2007: *OECD Principles and Guidelines for Access to Research Data from Public Funding* (www.oecd.org/dataoecd/9/61/38500813.pdf). OECD:n linja-asiakirjassa korostetaan sähköisen tutkimustiedon vallankumouksellista merkitystä, mutta vaaditaan myös sen elinkaaren tehokkaampaa hallintaa. Asiakirja korostaa, että tutkimuksen rahoittajien ja tutkimusorganisaatioiden tulee vaatia rahoitettavilta hankkeilta tiedonhallintasuunnitelma, jossa tulee määritellä tutkimustiedon koko elinkaari. Hankkeiden tulee tehdä ehdotus pysyvästi säilytettävästä tutkimusdatasta ja säilytysratkaisusta, jotta julkisesti rahoitetut tietoaineistot ovat hankkeen päätyttyäkin tutkimusyhteisön käytettävissä. Mahdollisimman laajan saatavuuden katsotaan tukevan tutkimuksen innovatiivisuutta ja eri tieteenalojen tuottaman tutkimusdatan monipuolista käyttöä.

Euroopan yhteisöjen toimielinten tutkimuksen tiedonhallintaa koskevat linjaukset noudattavat pitkälti OECD:n asettamia tavoitteita. EU:n komissio korostaa 4.4.2007 julkaisemassaan Euroopan tutkimusalueen kehittämisen uusia näkökulmia koskevassa asiakirjassaan (*Green Paper – The European Research Area: New Perspectives* [SEC(2007)412]) yhteisten eurooppalaisten infrastruktuurien merkitystä maailman huipputasoa olevien tutkimusympäristöjen luomisessa. Näiden infrastruktuurien tulee olla integroituja ja verkottuneita, ja niiden tulee olla käytettävissä sekä kaikkialta Euroopasta että muualta maailmalta. Tämä käytettävyys voidaan toteuttaa tehostamalla verkottumista ja sähköisten viestintävälineiden käyttöä. Asiakirja korostaa myös sähköisessä toimintaympäristössä tapahtuvan tiedon luomisen, levittämisen ja hyödyntämisen muuttuvia mekanismeja ja metodeja tutkimusjärjestelmän ydintoimintoina.

Euroopan unionin neuvosto hyväksyi 2832. kokouksessaan marraskuussa 2007 päätöksen, joka käsittelee digitaalisessa muodossa olevan tieteellisen informaation saavutettavuutta, levittämistä ja säilyttämistä. Se perustuu OECD:n ohjeiden mukaisiin käytäntöihin, EU:n komission digitaalisen informaation käyttöä koskevaan tiedonantoon [SEC(2007)181] sekä EU:n neuvoston 7.12.2006 tekemään päätökseen kulttuuriperinnön digitoinnista, saatavuudesta ja säilytyksestä (2006/C/297/01).

Euroopan unionin neuvoston päätös velvoittaa jäsenvaltiot kehittämään kansallisia strategioitaan ja toimintarakenteitaan siten, että ne parantavat mahdollisuuksia käyttää, levittää ja säilyttää tieteellistä informaatiota. Päätös ei tee eroa sähköisessä muodossa olevien julkaisujen ja tutkimuksen sähköisen tietoaineiston hallinnan välillä. Tutkimuksen tiedonhallintaa ja sen rakenteita tuleekin tarkastella kokonaisuutena, jossa keskeistä on eri toimijoiden mahdollisimman hyvin koordinoitu yhteistyö. EU:n neuvosto kannustaa päätöksessään jäsenmaita lisäämään tiedonhallintaa koskevassa kehittämisessä yhteistyötä tutkimuksen rahoittajaorganisaatioiden, tutkimuslaitosten ja tiedonhallinnasta vastaavien organisaatioiden välillä. Erityistä huomiota kansallisissa ratkaisuissa pyydetään kiinnittämään tieteellisen informaation säilyttämiseen osana kansallisia sähköisen tiedon säilytysratkaisuja.

Digitaalisten tietoaineistojen hallinta alkoi löytää uusia muotoja jo ennen OECD:n ja EU:n suosituksia. Isossa-Britanniassa muodostettiin vuonna 2002 digitaalisten aineistojen käyttöä ja säilytystä ohjaamaan useiden toimijoiden muodostama kansallinen Digital Preservation Coalition (www.dpconline.org). Siihen kuuluvat jäseninä muun muassa Cambridgen ja Oxfordin yliopistot, UK Research Councils, superlaskennasta vastaava Lontoon yliopiston tietokonekeskus, Englannin ja Skotlannin kansallisarkistot, British Library ja vuonna 2004 perustettu digitaalisen tiedon käsittelyä edistävä Digital Curation Center. Liitännäisjäseniä ovat muun muassa BBC, National History Museum, yksityinen Wellcome Library sekä suuri joukko muita tieteen keskeisiä toimijaorganisaatioita. Yksi jäsenistä on opetuksen ja tutkimuksen sähköisten

aineistojen käyttöä, yhtenäistä sisällönhallintaa ja ICT-infrastruktuurien monipuolista käyttöä tukeva JISC (Joint Information Systems Committee). Muuttuva toimintaympäristö on näin synnyttänyt aivan uudenlaisia toiminnallisia rakenteita, vaikka monilla toimijoilla on takanaan jopa vuosisatojen mittainen historia.

Ison-Britannian kehitys ei ole aivan ainutlaatuista Euroopassakaan. Alankomaissa toimii vastaavanlainen useiden toimijoiden muodostama koalitio, Netherlands Coalition for Digital Preservation. Siihen kuuluvat muun muassa Alankomaiden kansallisarkisto ja kansalliskirjasto, perustutkimuksen rahoituksesta vastaava The Netherlands Organization for Scientific Research (NWO) sekä useita tutkimusdatan hallinnasta vastaavia laskentakeskuksia ja muita tieteen alan organisaatioita (www.ncdd.nl/en/).

Yleiseurooppalaisella tasolla tutkimuksen tiedonhallintaa edistämään perustettiin vuonna 2006 tiedejärjestöjen yhteenliittymä, Alliance for Permanent Access (www.alliancepermanentaccess.eu/). Siihen kuuluvat useat keskeiset tieteen kansainväliset toimijat, kuten The European Science Foundation (ESF), European Space Agency (ESA), CERN, saksalainen Max Planck-Gesellschaft sekä Ison-Britannian, Saksan ja Alankomaiden kansalliskirjastot. Ruotsin kansallisarkisto (Riksarkivet) oli yksi aloitteentekijöistä 2005.

Suomessa digitaalisen tiedonhallinnan alalla tapahtunutta nopeaa kansainvälistä kehitystä ei kovin laajalti tunneta. Tampereen yhteiskuntatieteellisen tietoarkiston johtaja Sami Borg ja tietoarkiston arkistonhoitaja Arja Kuula julkaisivat vuonna 2007 tärkeä selvityksen OECD:n datasuositusten toimeenpanomahdollisuuksista Suomessa. Selvityksessään *Julkisrahoitteisen tutkimusdatan avoin saatavuus ja elinkaari* (Yhteiskuntatieteellisen tietoarkiston julkaisuja 6/2007; www.fsd.uta.fi/julkaisut/julkaisusarja/FSDjs06_OECD.pdf) he korostavat julkisrahoitteista tutkimusdataa keräävien ja tuottavien tutkimusorganisaatioiden, tutkimusrahoittajien ja tiedeorganisaatioiden kiinteän yhteistyön tarvetta. Erityisesti tulisi selkiyttää aineistoihin liittyviä oikeuksia ja vastuita sekä kehittää tutkimusaineistojen elinkaaren hallintaa. Tutkimusaineistojen elinkaaren hallintaa on tarkoituksenmukaisinta kehittää osana sähköisten aineistojen hallinnan kansallisia kokonaisratkaisuja.

Borg ja Kuula ovat selvityksessään tarkastelleet myös eräiden merkittävien kansainvälisten tutkimusrahoittajien datapolitiikkaa. Kohteena ovat olleet yhdysvaltalaiset National Institute of Health (NIH) ja National Science Foundation sekä Iso-Britannian Medical Research Council ja Economic and Social Research Council. Vakiintuneen käytännön mukaisesti nämä edellyttävät tutkimusrahoituksen hakijoilta tiedonhallintasuunnitelmaa (*information/knowledge management plan*), jossa esitellään tutkimushankkeen tiedonhallinnan käytännöt sekä tutkimusdatan säilyttämiseen ja jatkokäyttöön liittyvät ratkaisut.

Tutkimuksen kansainväliselle tiedonhallinnalle on ominaista entistä voimakkaampi verkottuminen, yhteisten standardien kehittäminen aineistojen hallintaan sekä tutkimuksen tarvitseman ja tuottaman kirjallisuuden, tietoaineiston ja tutkimusdatan käyttö samassa käyttöympäristössä, usein tutkijan omalla tietokoneella. Siksi myös arkistolaitoksessa sekä yliopistojen ja tutkimuslaitosten kirjasto- ja informaatiopalveluiden suunnittelussa on kiinnitettävä huomiota tutkimuksen tiedonhallinnan käytäntöjen muuttumiseen, jotta tietoa voidaan tarjota ja hallita tavalla, joka on mahdollisimman tarkoituksenmukaista tutkimuksen ja opetuksen työprosesseissa. Yksi suurista haasteista on päättää siitä, mitä kannattaa tehdä kansallisesti ja miltä osin ryhdytään käyttämään kansainvälisiä ratkaisuja. Jälkimmäisiäkin on tarjolla runsaasti, ja yhä selvemmin kehitys kulkee kohti aikaisempaa yhtenäisempiä käytäntöjä ja datastandardeja.

Yliopistojen ja valtion tutkimuslaitosten tutkimuksessaan julkisella rahoituksella tuottamat aineistot kuuluvat arkistolain (831/1994) piiriin. Kansallisarkiston sen nojalla yliopistoissa ja tutkimuslaitoksissa tekemät julkisrahoitteisten tutkimusaineistojen pysyvää säilyttämistä koskevat päätökset perustuvat arvonmääritykseen, jonka kriteerejä ovat muun muassa aineistojen käytettävyys, tieteellinen arvo tai niiden yhteiskunnallinen ja historiallinen merkitys. Säilyttämistä ja

käyttöä rajoittavia tekijöitä ovat esimerkiksi henkilörekistereitä koskeva lainsäädäntö sekä eräissä tapauksissa tutkimuseettiset toimintaperiaatteet. Yleisohjeita tutkimusdatan säilyttämisestä ei ole, vaan vastuu aineistojen arvonmäärityksestä, dokumentoinnista ja mahdollisista säilytysratkaisuista on perustunut lähes poikkeuksetta tutkijoiden ja tutkimusryhmien omaan harkintaan.

Tutkimuksen rahoittajaorganisaatiot ovat Suomessa kiinnittäneet toistaiseksi hyvin vähän huomiota kansainvälisesti entistä tärkeämpään asemaan nousseeseen tutkimuksen tiedonhallintaan. Suomen Akatemia ei rahoittajaorganisaationa ole sekään tähän kenttään erityisesti puuttunut. Yleisperiaatteena on ollut, että Akatemian rahoituksella hankitut tarvikkeet, laitteet ja kirjallisuus jäävät suorituspaikan omistukseen ja hallintaan. Akatemian rahoituksen yleisissä ehdoissa suositellaan kuitenkin, että tutkimusta varten koottu yhteiskuntatieteellinen aineisto luovutetaan Tampereen yliopiston yhteiskuntatieteellisen tietoarkiston käyttöön. Opetusministeriön ja Suomen Akatemian väliseen tulossopimukseen vuonna 2008 sisällytetty vaatimus tiedonhallintasuunnitelman edellyttämisestä rahoitusta hakevilta tutkimushankkeilta parantaa sen vuoksi merkittävästi suomalaisen tutkimuksen kykyä tuottaa ja hallita sähköistä tutkimusdataa. Toivottavasti se lisää myös kykyä käyttää hyväksi tarjolla olevia kansainvälisiä data-aineistoja.

Tutkimuksen tiedonhallinnan keskeiset kansalliset kehittämistavoitteet on linjattu opetusministeriön työryhmämuistioissa *Suomen eScience-ohjelma* (OPM ts 2007:7), *Laskennallisen tieteen kehittäminen Suomessa* (OPM ts 2007:23) ja *Sähköisen asioinnin edistäminen korkeakouluissa* (OPM ts 2007:49). Niissä esitetyt päämäärät perustuvat opetusministeriön hallinnonalan vuosien 2006–15 tietohallintastrategiassa esitettyihin päämääriin. Vuoden 2008 alussa ilmestyneessä työryhmäraportissa *Sähköisen aineiston pitkäaikaissäilytys ja käyttö* (OPM ts 2008:2) on puolestaan luonnosteltu kansallista kokonaisratkaisua sähköisten aineistojen säilyttämisessä ja käytössä. Tieteen tietotekniikan keskus CSC Oy on puolestaan yksi eurooppalaisen grid-järjestelmän solmukohdista. Sen tarjoama laskentaosaaminen sekä tiedonhallinta- ja tiedonsiirtokapasiteetti ovat keskeinen osa suomalaisen tutkimuksen infrastruktuuria.

Primääriaineistojen tarkka dokumentointi ja tutkimuksen tuottaman aineiston systemaattinen hallinta ovat tutkimusprosessin keskeisiä osia. Muiden tutkijoiden mahdollisuus analysoida kriittisesti tutkimuksessa käytettyjä aineistoja on puolestaan tärkeää arvioitaessa johtopäätösten ja havaintojen luotettavuutta ja tieteellistä merkitystä. Tässä suhteessa sillä ei ole periaatteellista merkitystä, ovatko aineistot sähköisesti saatavilla tai muulla tavoin luotuja. Käytettävyydeltään materiaalit kuitenkin eroavat toisistaan.

Sähköisellä, tietoverkkojen välityksellä käytettävissä olevalla tutkimusdatalla on usein arvoa uuden tutkimuksen vertailuaineistona tai uudenlaisia käyttömahdollisuuksia tarjoavana raakadatana. Koko tutkimushankkeen elinkaaren kattavan tiedonhallintasuunnitelman laatiminen sekä tietoaineistojen omistamiseen, hallintaan ja tekijänoikeuksiin liittyvien näkökohtien selvittäminen ovat siksi entistä merkittävämpi osa tutkimuksen prosessia. Ne parantavat tutkimuksen päämäärätietoisuutta ja selkeyttävät aineistojen käyttöä verkottuneessa toimintaympäristössä.

Tutkimuksen tiedonhallinnan osaamista tulee kehittää suunnitelmallisesti erityisesti osana informaatiotieteiden koulutusta. Informaatiotutkimuksen, informaatiologistiikan, metatietomallien ja sähköisen aineistonhallinnan alueella onkin edelleen tarvetta tehostaa tohtorikoulutusta. Muilla tieteenaloilla on tarvetta tiedonhallintaan erikoistuneista tutkijoista (*data scientist*). Kansainvälisesti korkeatasoinen tiedonhallinta on keskeinen osa toimivaa tutkimusprosessia.

Kirjoittaja on arkistolaitoksen pääjohtaja ja tiedonjulkistamisen neuvottelukunnan puheenjohtaja. Kirjoitus on julkaistu lyhennettynä Suomen Akatemian *A propos* -verkkolehdessä 5.9.2008.