Tutkimuksen PAS-palvelu – ei vain säilöön vaan myös saataville

Julkaistu 21.04.2016 klo 12:42 Kirjoittaja Suvi Pousi

Tiede ja tutkimus perustuvat tutkimustulosten luotettavuuteen. Luotettavuuden arvioinnin mahdollistaa osaltaan aineistojen ja niihin liittyvien metatietojen huolellinen säilytys ja laaja saatavuus. Hyvin toteutettu pitkäaikaissäilytys takaa digitaalisten aineistojen käyttökelpoisuuden useiden vuosikymmenten tai satojen vuosien ajan. Pitkäaikaissaatavuudella taataan aineistojen saatavuus ja hyödynnettävyys eli varmistetaan, että tulevat käyttäjät pystyvät avaamaan aineistot, tulkitsemaan niiden sisällön ja hyödyntämään niitä tulevaisuuden työvälineillä.

Kehittämispäällikkö **Esa-Pekka Keskitalo** Kansalliskirjastosta ja tutkijatohtori **Enrico Glerean** Aalto-yliopiston neurotieteen ja lääketieteellisen tekniikan laitokselta kertoivat tutkimuksen pitkäaikaissaatavuuspalvelun pilotoinnista Avoimen tieteen osaajakoulutuksessa 8.4.2016 Aalto-yliopiston Design Factoryssa. Esa-Pekka Keskitalo johtaa Avoin tiede ja tutkimus -hankkeen Tutkimusdatan pitkäaikaissaatavuus(Tutkimus-PAS) -työryhmää. Enrico Glerean esitteli Brain & Mind -laboratorion kokemuksia ja avoimia kysymyksiä neurotieteen alan tutkimusaineistojen pitkäaikaissäilytyksen ja -saatavuuden suhteen.

Pitkäaikaissäilytyksen teknisiä perusratkaisuja on jo toteutettu Kansallinen digitaalinen kirjasto(KDK) -hankkeessa, jonka kohteena ovat arkistojen, kirjastojen ja museoiden aineistot. Hankkeessa tehtyä työtä voidaan pitkälti hyödyntää myös tutkimusdatan säilyttämisessä. KDK-hankkeen palvelua on kehitetty lakisääteistä työtä tekeville vakiintuneille laitoksille, kun taas Tutkimuksen PAS-palvelua kehitetään kiivastahtisille tutkimusryhmille. Tutkimuksen PAS-pilotoinnin tavoitteena onkin kehittää prosesseja, joilla tutkijoita voidaan palvella parhaalla mahdollisella tavalla. Pilotit koostuivat kolmesta vaiheesta: alkukyselystä, tallennettavan aineiston sopivaan muotoon käsittelystä ja paketoinnista tallennusta varten sekä loppukyselystä.

Esa-Pekka Keskitalo esitteli Tutkimuksen PAS-palvelun pilotointia Aalto Design Factoryssa 8.4.2016.

Tutkimuksen tulosten säilytys ja saatavuus liittyvät vahvasti yhteen. Keskitalon mukaan käytössä oleva data ei pilaannu tutkijan pöytälaatikkoon: sen käyttäjät pitävät sen elossa esimerkiksi havaitsemalla dataan mahdollisesti eksyneet virheet tai hankalat tiedostomuodot. Miten näin hyvään tilanteeseen päästään?

Kaikki pohjautuu hyvään aineistonhallintaan

Pitkäaikaissäilytys ei sinänsä vaadi tutkijalta erityisosaamista: ei ole olemassa mitään erityisiä pitkäaikaissäilytykseen liittyviä menetelmiä. Tutkijalta vaaditaan ainoastaan hyviä aineistonhallinnan käytäntöjä: datan on oltava teknisesti järjestyksessä ja tallessa, helposti käytettävässä muodossa ja sen sisällöt on säilytettävä kontekstitiedon, dokumentaation ja metatietojen avulla. Keskitalo korostaa, että datan on kestettävä teknologian muutokset, organisaatiomuutokset ja ennen kaikkea datan kanssa työskentelevien ihmisten vaihtuminen. Datan käytettävyys, säilyvyys ja saatavuus eivät saa olla yksittäisen ihmisen varassa.

Osa pilotoinnissa käytetyistä aineistonhallintaan liittyvistä käytännöistä on ollut tutkijoille uusia, mutta Keskitalon mukaan niiden opettelu ei ole ollut tutkijoille vaikeaa.

Suuret datamäärät ovat haaste monella tapaa

Myös Enrico Glerean korostaa hyvän aineistonhallinnan merkitystä. Hän antaa vastuullisille tutkijoille ohjeeksi edistää hyviä aineistonhallintatapoja jo datan keräämisvaiheesta alkaen. Näin säästyy sekä omaa että muiden tutkijoiden aikaa. Samalla helpotetaan vanhaan dataan palaamista, uuden datan jakamista ja yleisesti isojen tutkimusprojektien hallinnointia. Glerean muistuttaa myös menetelmien säilyttämisen tärkeydestä, sillä tutkimus ei ole toistettavissa, ellei sen menetelmiä tunneta tarkasti.

Enrico Glerean kertoi Tutkimuksen PAS-palvelun pilotointiin osallistuneen Brain & Mind -laboration kokemusta.

Glereanin mukaan säilytettävän datan ja metadatan valinta oli haastavaa, sillä *Big Data* ei ole vain hypetystä. Datamäärien kasvu on todellinen ilmiö niin neurotieteessä kuin muillakin tieteenaloilla. Jos aikaisemmin yhdessä tutkimuksessa kuvannettiin 20 koehenkilön aivot, voi nykypäivänä koehenkilöitä olla 200. Isojen datamäärien käsittely on työlästä ja vaatii automatisointia. Lisäksi

vastuullisen tutkijan on vaikeaa hahmottaa ja kontrolloida ryhmän tutkijoiden tuottamaa dataa ja sen metatietoja. Suurten datamäärien haasteet koskettavat myös yliopistojen IT-osastoja, hallintoa ja rahoittajatahoja, sillä myös heidän on vaikea kartoittaa, kuinka paljon dataa projekti on tuottanut, mikä osa datasta tulisi pitkäaikaissäilyttää ja minkä osan siitä voi tuhota.

Tarve standardeille

Glerean kertoo Nature-lehden haastatelleen Texasin yliopiston professori Russell Poldrackia neurotieteen alan datansäilytyskäytännöistä vuonna 2014. “Datan säilytyskäytäntöjä ei käytännössä ole. Dataa korkeintaan tallennetaan DVD-levylle tai nauhalle, jonka jälkeen tallenne jätetään johonkin lojumaan,” Poldrack kertoo. Glerean pitää totuttujen tapojen ja standardien puuttumista suurena haasteena väitöskirjantekijöille ja tutkijatohtoreille, sillä heillä ei ole valmiita työkaluja tai ohjeita suurten datamäärien käsittelyyn ja säilyttämiseen.

Glerean kertoo, ettei hänen tutkimusalallaan vielä ole standardoitu tiedostomuotojen käyttöä. Pilotoinnissa noudatettiin hyvää käytäntöä suosimalla avoimia formaatteja, joiden dokumentaatio on yleisesti saatavilla. Glerean kannustaa standardoitujen data- ja metadataformaattien käyttöön uskoen niiden helpottavan myös IT-, hallinto- ja rahoittajatahoja projektien ja niiden kustannusten tarkkailussa.

Aineistojen avaamiseen liittyy rajoituksia

Brain & Mind -laboratorion aivokuvantamisaineisto on tuotettu lääketieteellisessä kokeessa ja sen tuottamiseen on saatu lupa eettiseltä toimikunnalta. Eettiseltä toimikunnalta saatu lupa kattaa aineiston tuottamisen, mutta aivokuvantamisaineiston julkaisemiseen liittyvät juridiset vastuut ovat epäselviä. Brain & Mind -laboratorio päättää aineiston jakamisesta eteenpäin tapauskohtaisesti, koska yhtenäistä käytäntöä jakamisesta ei Suomessa ole. Tilanne on Glereanin mukaan ongelmallinen, sillä tällä hetkellä suomalaiset tutkijat joutuvat kieltäytymään datan jakamisesta myös tutkimusartikkelien ohessa tiedejulkaisuissa. Tutkimuksen PAS-palvelun pilotoinnin loppuraportin mukaan tilanteen odotetaan ajan myötä selkenevän lainsäädännön muutosten kautta.

Glerean jättää kuulijoille pohdittavaksi joukon pilotoinnin herättämiä kysymyksiä: Mitä tutkimusdataa tulisi pitkäaikaissäilyttää ja miten? Tulisiko datalla olla vanhentumispäivämäärä, jonka jälkeen sitä ei enää säilytetä, vai säilytetäänkö dataa ikuisesti? Pitäisikö dataan liittyvän lisenssin raueta joskus? Ja lopuksi: kuinka tarjota opetusta, jotta väitöskirjantekijät ja tutkijatohtorit voivat kehittyä aineistonhallinnassa?

Pilotoinnista käytäntöön

Tutkimuksen PAS-palvelun pilotoinnit jatkuvat yhä ja palvelua kehitetään edelleen. Tänä vuonna keskitytään mm. genomidatan erityiskysymyksiin. Menossa on myös eri tutkimusalojen tiedostoformaattien käyttöä koskeva kartoitus. Projektissa seurataan myös kansainvälistä kehitystä ja vahvistetaan kansainvälistä yhteistyötä. Tavoitteena on, että tutkimustulosten pitkäaikaissaatavuuspalvelu on vuonna 2017 täysimittaisessa toiminnassa.

Esa-Pekka Keskitalon ja Enrico Glereanin esityksen kalvot

Esa-Pekka Keskitalon ja Enrico Glereanin videoitu esitys