Ville Tenhunen

DATA YLITTÄÄ ORGANISAATIORAJAT

Data on ollut viimeiset vuosikymmenet tieteen, tutkimuksen, opetuksen ja niihin liittyvien tukipalveluiden haaste. Tällä saralla palveluita tuottavat muun muassa kirjastot sekä tietotekniikan ja tutkimushallinnon yksiköt.

Dataan liittyvässä haasteessa kyse on siitä, mitä data voi tuottaa tutkimukselle ja kuinka luotettavaa se on. Data saa arvonsa vasta, kun sitä tai sen osaa käytetään; hyöty syntyy siitä, että dataa prosessoidaan, tutkitaan tai käytetään hyväksi jossain sovelluksessa. Tämä on mahdollista kuitenkin vasta sitten, kun datalle on annettu konteksti.

Data on luonteeltaan dynaamista ja monimuotoista. Sitä versioidaan, pilkotaan, siivotaan, korjaillaan, anonymisoidaan ja niin edelleen. Yhteen julkaisuun liitettävä data on yksi ilmentymä datasetistä, joka on johdettu raakadatasta tai on otos jonkin laitteen datavirrasta. Datalla ei ole selvää ajan yli pysyvää tai stabiilisti toistuvaa muotoa.

Tietotekniikalle dynaamisuus ei varsinaisesti ole ongelma, sillä IT nuorena kulttuurina on luonut, käsitellyt, muokannut, integroinut, jakanut ja siirrellyt dataa aina – sekä hankkinut lisää kapasiteettia ja työvälineitä tarvittaessa. Datan määrän voimakas kasvu sen sijaan on IT:lle jatkuva ongelma.

Kirjaston traditioon ja palveluiden piiriin data on tullut oikeastaan vasta hetki sitten. Paljon ihan muuta on ehtinyt tapahtua ennen kuin kirjaston on edes tarvinnut ajatella jotain datan kuratoinnin kaltaisia asioita.

Kuinka sitten kehittää datan hallintaa? Yksi ratkaisu on liittää data julkaisuihin tai tehdä niin kutsuttuja datajulkaisuja. Kyse on kuitenkin staattisesta ratkaisusta dynaamiseen ongelmaan. Kehittyneempi askel on se, että kuvaillaan data, annetaan dataseteille PID eli pysyvä tunniste ja tallennetaan tiedostot repositorioon sekä tarjoillaan tutkijoille sopivasta rajapinnasta.

Monimutkaiseen haasteeseen on kehitetty myös vallitsevaa paradigmaa muuttavia ratkaisuja. Entäs jos ryhtyisimmekin puhumaan digitaalisista objekteista? Tähän ohjaa useampikin asia kuten se, että universaalit FAIR-periaatteet eivät koske pelkästään dataa, vaan myös muita objekteja kuten ohjelmistoja ja koneoppimisen malleja. Mitäs näiden kanssa sitten tehdään?

FAIR Digital Objects (FDO) -konsepti tarjoaa uudemman lähestymistavan, kun siirrytään pois pelkkien tiedostojen siirtelystä PID:llä varustettujen objektien käsittelyyn oman Digital Object Interfacing -protokollansa (DOIP) avulla. DOIP:in merkitystä on verrattu TCP/IP-protokollan vaikutukseen

Internetille. Ilman TCP/IP:tä ei nykyisessä Internetissä toimi mikään.

FDO:n kanssa tai ilman, tekoäly on tulossa mukaan automatisoimaan datanhallintaa. Esimerkiksi kuratointi ja metadatan tuottaminen ovat asioita, joissa tekoälystä voisi olla apua. Datan anonymisointia tehdään jo tekoälyn avulla.

Tulipa näitä uusia välineitä tai ei, pari asiaa on varmaa. Ensinnäkin datanhallintaa tekevien organisaatioiden on tehtävä yhteistyötä yli omien siilojensa. Lisäksi niiden on tarkasteltava omia perinteisiä toimintatapojaan kriittisesti uusien haasteiden ja mahdollisuuksien edessä.

Ville Tenhunen

Kirjoittaja toimii Helsingin yliopiston tietotekniikkakeskuksessa kehittämispäällikkönä. Yliopiston tutkimusdatan palveluiden kanssa hän on toiminut yli kymmenen vuotta ja osallistunut myös kansallisiin TTA- ja ATT-hankkeisiin. Tenhunen toimii Data Solutions Architect -roolissa eurooppalaisessa tutkimuksen infrastruktuuripalveluita tuottavassa EGI Foundationissa ja useissa Horizon 2020 -rahoitteisissa projekteissa WPleaderina.