# data science

Fakta Julkaistu 11.08.2015 klo 12:38 Kirjoittaja Eero Hyvönen

OKM:n Tieto käyttöön –haussa rahoitettavaksi valittu Linked Open Data Science (LODSci) -niminen hanke on käynnistynyt Aalto-yliopistossa. Hanketta vetää Semanttisen laskennan tutkimusryhmä ja hankkeen tavoitteena on edistää avointa tiedettä (Open Science) uusimman Linked Data -teknologian (LD) avulla [1]. Kehitystyön kohteena on "datatiede" (Data Science) ts. se, miten datasta voidaan muodostaa ihmisen ja koneen vuorovaikutuksella uutta tieteellistä tietoa.

Tutkimustyön perustana oleva data on nykyisin yhä useammin heterogeenista, peräisin eri lähteistä ja eri aloilta. Esimerkiksi ympäristön tutkimuksessa voidaan tarvita biologista tietoa eläimistä ja kasveista, ihmisten tuottamaa havaintotietoa lajeista, sekä fysikaalista mittaustietoa ilmakehästä, säästä, vesivaroista ja maaperästä. Kulttuurintutkimuksessa taas on tarpeen yhdistellä tietoa esineellisestä kulttuurista, kirjallisuudesta, historian tapahtumista, taiteista sekä ihmisistä ja näiden sosiaalisista verkostoista [2]. Tällöin nousee haasteeksi tutkimusdatan tuotannon, julkaisemisen ja hyödyntämisen kannalta monia teemoja, joihin LD-teknologia tarjoaa uusia lupaavia hyödyntämismahdollisuuksia:

1. **Datajoukkojen yhdistäminen**. Erillisistä datajoukoista koostuvan tiedon tutkiminen edellyttää, että eri aineistot saadaan yhdistettyä toisiinsa semanttisesti yhteismitalliseen muotoon.

2. **Datan laatu**. Heterogeenisen datan laatu on usein vaihtelevaa erityisesti kun sitä tuotetaan yhä enenevässä määrin automaattisin keinoin. Jotta datasta voidaan tehdä tieteellisesti luotettavia johtopäätöksiä, on tutkijan voitava validoida datan laatua.

3. **Datan analyysi ja visualisointi**. Monimuotoisen ja alaisen datan analyysi on vaikeampaa ja vaatii erilaisia menetelmiä kuin esimerkiksi yksittäisen taulukkodatan.

4. **Uuden tietämyksen muodostaminen**. Kun yhdistetään tietoa eri lähteistä, voidaan kokonaisuudesta usein löytää sellaista uutta tietoa, jota ei voida löytää datan osia erikseen tutkimalla. Uuden yllätyksellisen (serendipitous) tietämyksen muodostaminen, Knowledge Discovery, on LD-tutkimuksen keskeisiä uusia arvolupauksia.

LODSci-projekti kehittää näihin haasteisiin prototyypin avoimen linkitetyn tieteellisen tiedon julkaisupalvelusta tutkijoita varten, sekä interaktiivisen opetusaineiston linkitetyn datan julkaisemisesta. Järjestelmä perustuu W3C:n standardeihin ja parhaisiin käytäntöihin, ja hyödyntää Linked Data Finland -alustaa.

Palvelun hyödyntämistä pilotoidaan luonnontieteellisen havaintodatan julkaisuun liittyen sekä Digital Humanities -tutkimuksessa kahdessa laajassa yhteistyöverkostossa: Reassembling the Republic of Letters on Oxfordin yliopiston johtama laaja kansainvälinen EU COST -yhteistyöverkosto, jossa tutkitaan valistuksen ajan tieteellisen ajattelun kehittymistä datalähtöisesti kirjeenvaihto- ja elämäkerrallisen datan avulla. Sotasampo taas on kotimainen yhteistyöhanke, jossa julkaistaan Suomen toiseen maailmansotaan liittyvää dataa ja kehitetään sovelluksia tutkijoiden ja laajemman yleisön käyttöön.

**Lisälukemista**

[1] Tom Heath, Christian Bizer: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool, Palo Alto, 2011.

[2] Eero Hyvönen: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Morgan & Claypool, Palo Alto, 2013.