

Linkitetty data ja tiede - uusi mahdollisuus tutkimustiedon julkaisemiselle ja sovelluksille

Julkaistu 09.01.2015 klo 14:28 Kirjoittaja Eero Hyvönen

Ihmisen kannalta web on valtava toisiinsa linkitettyjen sivujen kokoelma, johon hakukoneet ja muut verkkopalvelut tarjoavat erilaisia näkymiä. Syvällisemmässä mielessä web on tiedon verkko, Web of Data, jossa erilaiset käsitteet assosioituvat toisiinsa, määrittyvät toistensa kautta tiedoksi ja rikastuvat isommiksi kokonaisuuksiksi.

Esimerkiksi Väinö Linna määrittyy henkilönä biografisten tietojen kautta, kuten syntymä vuonna 1920, jotka löytyvät vaikkapa Suomalaisen Kirjallisuuden Seuran Kansallisbiografiasta, ja yhdistyy Tuntematon sotilas -käsitteeseen romaanin kirjoittajana. Teoksen kuvaus löytyy mm. kirjastojemme Kirjasampo.fi-palvelusta ja yhdistyy edelleen aiheensa kautta samannimiseen elokuvaan, jonka Edwin Laine ohjasi 1955 ja jonka tiedot on kuvattu KAVAn Elonet.fi-palvelussa tai Wikipediassa. Tällaisesta webin ”sisällä” olevasta, käsitteitä ja tietoja yhdistävästä semanttisesta verkosta käytetään nimitystä linkitetty data, Linked Data, ja se on rakenteeltaan aivan erilaista kuin web-sivujen linkkien muodostava verkko.



Linkitetty data ja semanttinen web

Linkitetty data ja siihen liittyvä semanttisen webin teknologia ovat jo vuodesta 2001 olleet keskeinen kehityskohde [W3C-järjestössä](#), joka koordinoi WWW:n infrastruktuurin kansainvälistä kehitystä.

Semanttinen web on seuraavan sukupolven älykäs web, jonka sisällöt ovat myös koneluettavia ja linkitettyjä eli dokumentit sisältävät myös tietoa asioiden keskinäisistä suhteista. Semanttisessa webissä ohjelmistot pystyvät hyödyntämään sisältöjä käyttäjän tarpeiden mukaan, esimerkiksi etsimään ratkaisuja käyttäjän esittämään ongelmaan. RDF (*Resource Description Framework*) on W3C:n standardoima malli tiedon vaihtoon sovellusten välillä erityisesti web-ympäristössä. RDF-tietomalli ja sen tapa kuvata resursseja on keskeinen osa semanttista webiä.

Webin isot toimijat ovat ottaneet teknologiaa käyttöönsä hieman eri nimillä, kuten Googlen Knowledge Graph, Microsoftin Satori ja Facebookin Open Graph. Julkinen sektori on yhä

useammassa maassa ja tapauksessa valinnut linkitetystä datassa käytettävän RDF-tietomallin datansa avaamismuodoksi. Tähän on kaksi keskeistä syytä. Yhtäältä linkitetty data mahdollistaa nimensä mukaisesti erilaisten tietojen sisällöllisen yhdistämisen yhteentoimivasti (interoperability) isommiksi kokonaisuuksiksi, jolloin eri tiedontuottajien tietoa voidaan rikastaa automaattisesti toistensa avulla ”ilmaiseksi”. Toisaalta linkitettyä dataa voidaan sen hyvin määritellyn semantiikan ansiosta käsitellä helpommin tietokoneilla, esimerkiksi hakea tietoa ”älykkäästi” ja kieliriippumattomasti, päätellä uutta tietoa ja tuottaa tiedosta analyyseja ja visualisointeja ihmisten käytettäväksi erilaisissa sovelluksissa.

Nämä mahdollisuudet ovat hyvin lupaavia myös tieteellisen tiedon julkaisemisen ja tutkimuksen kannalta. Yhä useammin tutkimustyö perustuu suurten, eri lähteistä koottujen datamassojen analysointiin ja tämän perusteella kehitettyihin malleihin. Ilmiöstä käytetään eri aloilla eri nimityksiä, kuten Linked Science luonnontieteissä tai Digital Humanities humanistisilla aloilla.

Linked Data Finland

Suomessa ideaa on pilotoitu CSC:n vetämän Tutkimuksen tietoaaineistot -hankkeen ja Aalto-yliopiston ja Helsingin yliopiston [Linked Data Finland](#) -hankkeen pilotissa, jossa yhdistettiin Luonnontieteellisen keskusmuseon ja Tringa ry:n ornitologista havaintodataa, BirdLife-järjestön ja Luontoportti.fi-palvelun taksonomista tietoa linnuista sekä Ilmatieteen laitoksen säätietoja n. 30 vuoden ajalta. Säätietojen tiedetään olevan keskeinen lintujen muuttoon vaikuttava tekijä. Yhdistämällä ja julkaisemalla datat Linked Data Finland -palvelussa voidaan eri lajien muuton ja sääolosuhteiden vaikutuksia tutkia ja visualisoida uusilla tavoilla.

[Linked Data Finland -työ](#) jatkuu mm. [Your data -palvelun](#) kautta, jonka avulla halukkaat tiedon tuottajat voivat julkaista tutkimuksellisesti kiinnostavaa dataansa. Portaali luo RDF-muotoiseen dataan automaattisesti ohjelmalliset rajapinnat ja tarjoaa muita palveluita, joiden kautta datan tuottaminen, tutkiminen ja sovellusten kehittäminen voivat helpottua merkittävästi.

Kirjoittaja professori Eero Hyvönen johtaa Aalto-yliopiston [Semanttisen laskennan tutkimusryhmää](#).

[Avoin Suomi 2014 menestystarinavideo: Linked Data Finland](#)