Kirjastojen data avautuu ja linkittyy verkkoon – SWIB12 kokoontui Kölnissä

Posted on 18.12.2012 by helehilt

Semanttinen web ja linkitetty avoin data on ollut tapetilla viime vuosina ja herättänyt kiinnostusta ja kehityshankkeita monella sektorilla, esimerkkinä vaikkapa monimuotoiset FinnONTO-projektit. Kirjastomaailma on hidasliikkeisempi kuin web keskimäärin, mutta nyt näyttää siltä, että heräämistä on tapahtunut myös tällä sektorilla. Oireellista tässä suhteessa on se, että tänä vuonna on IFLA:aan on perustettu semanttisen webin special interest group -työryhmä.

Saksassa kirjastojen semanttisen webin kehityshankkeista on vaihdettu kokemuksia vuodesta 2009 alkaen, mutta kokoukset olivat alkuaikoina pääosin saksankielisiä. Tänä syksynä konferenssi vaihtoi kokouskielen englanniksi, mikä näkyi siinä, että osallistujia oli nyt 23 maasta. Jostain syystä konferenssin lyhenteenä on kuitenkin säilynyt SWIB, Semantic Web in Bibliotheken.

Lisää URIa

Avoimen linkitetyn datan erityispiirteenä, riippumatta siitä millä tekniikalla datan avaus toteutetaan, on datan linkittäminen muualle avoimeen verkkoon, ja erityisesti johonkin Linked Open Data -pilven tietoaineistoon. Tyypillinen linkityksen kohde on Wikipedia, jota tarjoaa hyödyllistä lisätietoa monesta asiasta, kirjastojen sovelluksissa esimerkiksi tekijöistä, aihealueista, teoksista tai paikannimistä. Sen sijaan, että linkitys tehtäisiin perinteisen Wikipedian sivuille, kohteena näissä sovelluksissa käytetään usein RDF-muotoon siirrettyä DBpediaa. Tästä on se etu, että yhdellä linkillä saadaan kootusti monikielinen kuvaus asiasta, esimerkkinä Alvar Aallon sivu tässä muodossa: http://dbpedia.org/page/Alvar_Aalto.

Niille, joita arveluttaa Wikipedian tietojen luotettavuus, on lohduksi se tosiseikka, että kenen tahansa on mahdollista parantaa ja korjata Wikipedian tietoja, toisin kuin kirjastojen auktoriteettitietojen puutteita ja virheitä. Wikipediaa ei ole tarkoitettu ensisijaiseksi julkaisufoorumiksi, vaan sinne kootaan tietoja muualla julkaistuista lähteistä.

Kirjastodatan rikastaminen linkityksillä tehdään mieluiten käyttämällä URI-tunnisteita, jotka ovat tyypillisesti pysyvämpiä kuin webin verkko-osoitteet keskimäärin. Käytännössä nämä ovat usein tavallisia URL-osoitteita, joilla asiasanoja täydennetään tai korvataan linkittämällä soveltuvan ontologian käsitteisiin. Linkitys voi myös olla tekijätietojen tarkentamista linkittämällä auktoriteetitietokantoihin kuten VIAF, tai paikannimien tarkentamista linkittämällä vaikka GeoNames-palveluun. Esimerkiksi Espanjan kansalliskirjaston projektissa bibliografista dataa on linkitetty RDF-muodossa aukoriteettitietokantoihin (VIAF, GND), yhteisluettelotietokantoihin (Sudoc, Libris), wikipediaan (DPpedia) sekä jatkossa myös Saksan, Ranskan ja Britannian kansallisbibliografioihin.

Tavoitteena tässä kaikessa datan linkityksessä on, että kirjastojen tarjoamat tietoaineistot eivät olisi enää erillisiä siiloja, joita voi käyttää vain kirjaston tarjoaman käyttöliittymän kautta, vaan aineisto olisi osa eri suuntiin rönsyävää verkkomaista tietorakennetta, joka tunnetaan myös nimellä GGG, Giant Global Graph.

API vai dumppi

Avoimen linkitetyn datan yhteydessä nousee aina keskusteluun erilaisia näkemyksiä siitä miten datan avaamista on tarkoituksenmukaisinta toteuttaa. Tässä kohden on kaksi pääkoulukuntaa, joista toinen korostaa sitä, kuinka keskeistä on, että koko aineisto on saatavilla yhtenä tai useampana tietokantadumppina jossain yleisesti käytetyssä formaatissa. Tilastodatan osalta tiedostomuoto voi olla esimerkiksi CSV ja kirjastodatan osalta esimerkiksi MARCXML.Tällöin datan avaaja ei tee etukäteen valintaa siitä mikä osa datasta voisi olla käyttäjän kannalta kiinnostavaa ja mikä ei. Käyttäjä tässä yhteydessä voi olla sovelluskehittäjä, joka yhdistää kirjastodataa omaan sovellukseensa.

Tätä filosofiaa noudattaa Oslon kaupunginkirjaston palvelu data.deichman.no, jossa koko kirjastoluettelo tarjotaan RDF-muodossa. Kirjasto on myös kehittänyt ja ja julkaissut työkalun marc-datan muuntamiseen RDF-muotoon (marc2rdf). Harmi vain, että norjalaiset ovat pysyneet omassa kansallisessa NORMARC-formaatissaan, minkä vuoksi työkalun käyttö Suomessa vaatii sen sovittamista MARC21-formaattiin.

Vaihtoehtona tietokantadumpeille tai niiden rinnalla tarjotaan usein API-ohjelmointirajapintaa, joka kautta data on saatavilla. Tämä on sinänsä toimiva ratkaisu, mutta usein rajapinta on suunniteltu jotain tiettyä käyttötapausta silmällä pitäen, eikä API:n toimintatavan muuttaminen ole datan soveltajille yleensä mahdollista . Myös ohjelmointirajapintojen toteutuksissa ja helppokäyttöisyydessä sovelluskehittäjän kannalta on suuria eroja. Saksalaisten kokemusten perusteella parhaaseen tulokseen päästään tarjoamalla yleisesti käytetty rest/json -rajapinta sen sijaan, että tarjottaisiin puhdasoppista sparql-rajapintaa, joka on monille kehittäjille vieras.

MARCille seuraaja

Kongressin kirjastosta esiteltiin Skypen välityksellä uutta Bibliographic Framework Initiative -hanketta, jossa tavoitteena on määritellä MARC-formaatin seuraaja ja strategia, jolla tähän voitaisiin siirtyä vaiheittain. Esityksen aikana kuitenkin ääniyhteys pätki niin pahasti, että tähän asiaan on helpointa tutustua vastikään aiheesta julkaistusta raportista.

Tutkimusdataa RDF-muodossa

Taloustieteellinen tutkimusdata on tyypillisesti tilastodataa, jota saksalaisessa hankkeessa ollaan myös muuntamassa RDF-muotoon. Tässä ei kovin pitkälle riitä se, että laajat tilastoaineistot kuvaillaan koko datajoukon tasolla, vaan ideaalitilanteessa taulukkomuotoisen aineiston jokainen solu saa oman URI:n, ja siihen voidaan viitata suoraan muualta, esimerkiksi julkaisuista. Tämäkään ei vielä riitä tutkimuksen toistettavuuden näkökulmasta, vaan lisäksi omasta URI-osoitteestaan pitäisi löytyä ne skriptit ja parametrit, joita datan analysointiin ja visualisointiin on käytetty. Haasteita tällä saralla riittää myös kirjastoille.

Kädet savessa RDF:n kimpussa

Konferenssiin liittyvässä työpajassa pääsivät kaikki osanottajat kirjoittamaan RDF:ää, ja tuotokset julkaistiin muidenkin nähtäville asiankuuluvalla avoimella lisenssillä. Kun luomukset vielä ajettiin validaattorin lävitse ja lisättiin puuttuvat pisteet ja välilyönnit, voitiin saada näkyviin kiinnostavia taustatietoja työpajaan osallistuvista, ja tähän joukkoon voitiin kohdistaa kyselyitä sparql-muodossa.

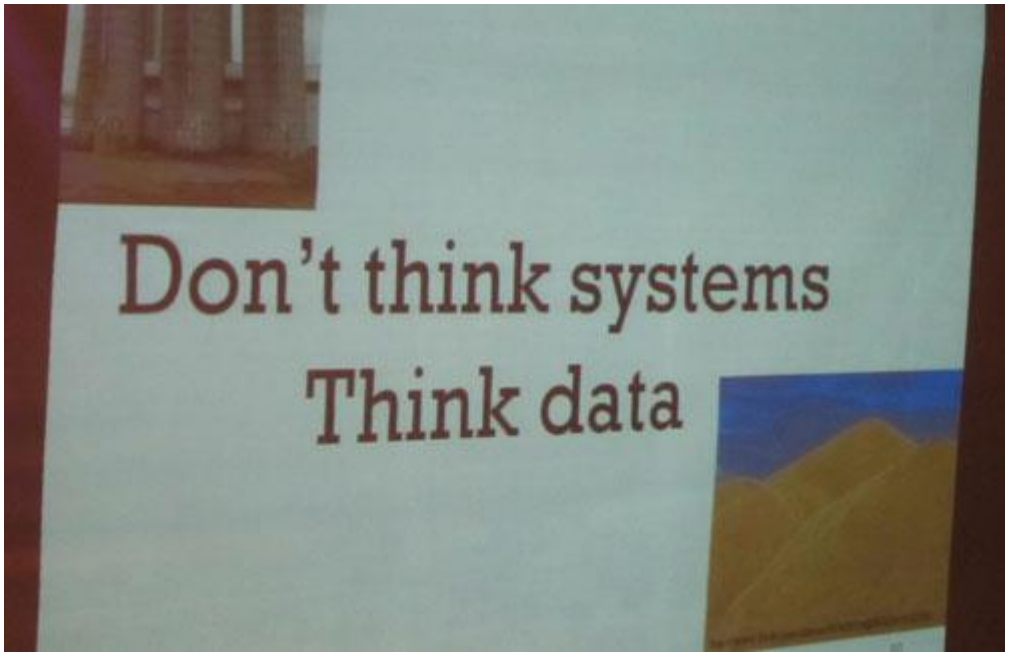
Työskentely oli hyvin antoisaa, kun käytännössä näki miten RDF-tiedot muodostuvat ja mahdollistavat tietojen yhdistämisen ja jakelun. Samalla myös tietojen laatuun liittyvät ongelmat konkretisoituivat. Pienessä mittakaavassa tehty harjoitus auttoi näkemään yhdistetyn tiedon mahdollisuuksia ja sudenkuoppia. Suosittelemme käytännön harjoittelua muillekin.

Mitä jäi mieleen?

Konferenssin pääviesti oli selvä. Yhdistetyn tiedon työkalut, RDF ja SPARQL, ovat käyttökelpoisia tiedon yhdistämiseen, jakamiseen ja hyödyntämiseen konekielisesti. Ne eivät kuitenkaan korvaa tiedon tuottamisessa ja tallennukseen käytettäviä järjestelmiä.

Tiedon muuttaminen RDF-muotoon onnistuu eri formaateissa olevasta tiedosta, mutta muunnos ei paranna tiedon laatua. Sama tietysti pätee muihinkin tiedon esitystavan muutoksiin. Virheellisten tai puutteellisten tietojen osalta lähtökohtana on, että korjaukset tehdään alkuperäisiin tiedon lähteisiin. Tosin ongelmana tietojen korjaamisessa voi olla useiden käsittelyjen, esimerkiksi tietojen yhdistäminen ja rikastaminen, ja linkitysten takana olevien tietojen alkuperän tunnistaminen ja viestiminen korjaustarpeesta. Toinen ongelma ovat samaa asiaa koskevat ristiriitaiset versiot tiedosta.

Esityksissä ja keskusteluissa tuli esille standardien ja auktoriteettitietojen kaipuu sekä niiden toteuttamisen ja käytön vaikeudet, esimerkkinä erilaisten sanastojen sisältämien samojen termien eri käyttötarkoitusten tunnistaminen myös koneellisesti. Tärkeäksi todettiin käsitteiden määrittely ja erityisesti se, miten edetään käsitemallista konkreettiseen järjestelmien sisältämien tietojen käsittelyyn. Esimerkiksi Library of Congressin BIBFRAME-projektissa on edetty RDF-muotoisen tiedon tuottamiseen ja törmätty käytännön ongelmiin (kalvo 38).



Jos konferenssin anti pitäisi kiteyttää yhteen lauseeseen, yksi hyvä ehdokas on lainaus **Lukas Kosterin** (Amsterdamin yliopisto) esityksestä: ”*Don’t think systems – Think data*”

Konferenssipaikka Kölnissä oli monitoimitalo Bürgerhaus Stollwerck.

SWIB12-konferenssin esitykset on taltioitu tänne.

Teksti ja kuvat

Pauli Assinen
tietojärjestelmäpäällikkö
Verkkopalvelut
Helsingin yliopiston kirjasto

Kimmo Koskinen
kehityspäällikkö
Verkkopalvelut
Helsingin yliopiston kirjasto