

ABSTRAKTI

Kaisa Hypén & Eetu Mäkelä

RDF ja FRBRoo: Kirjasammon skeemasta

Kaisa Hypén, Turun kaupunginkirjasto, kaisa.hypen@turku.fi.

Tyypillistä kaunokirjallisuuteen liittyville, kirjastoissa esitetyille kysymyksille on muun muassa, että niissä yhdistellään useita teoksen kirjailijaan, sisältöön ja julkaisuhistoriaan liittyviä näkökulmia, etsitään teosta jonkin siihen liittyvän yksityiskohdan perusteella tai kirjallisuutta jostakin aiheesta. Usein kysytään yleisesti vain ”hyviä kirjoja” tai ”samankaltaista kirjallisuutta”.

Laajassa ja monialaisessa Kirjasampo-projektissa on tutkittu sitä, miten kirjastojen tietojärjestelmissä voitaisiin ottaa huomioon näitä kaunokirjallisuuden haun erityispiirteitä. Projektissa kehitetään kaunokirjallisuuteen keskittyvää interaktiivista verkkopalvelua, joka tarjoaa kirjastojen tuottamalle datalle aivan uudenlaisen käyttöympäristön ja -sovelluksen. Kirjasammossa luodaan myös tiedontallentamisen ja -tuottamisen tapaa, johon voivat osallistua niin kirjastoalan ammattilaiset kuin lukijatkin. Tavoitteena on tallentaa, kumuloida ja jakaa kirjallisuuteen liittyvää tietämystä, kokemuksia ja hiljaista tietoa. (Hypén 2009, 7-11.)

RDF-tietomalli

Palvelu rakennetaan uudennlaisilla semanttisen webin tekniikoilla. (Berners-Lee, Hendler & Lassila 2001, 34-43). Ne on kehitetty helpottamaan tiedon uudelleenkäyttöä, yhdistelyä ja muokkausta usean toimijan heterogeenisten tietovarantojen kontekstissa. Kirjasammossa sovellettava RDF-pohjainen tiedontallennusjärjestelmä (RDF Semantics 2004) eroaa useimmista perinteisistä indeksointijärjestelmistä siinä, ettei siinä tietomallitasolla ole ensisijaista indeksointikohdetta.

Vaikka Kirjasammon kuvailussa kirjat ja kirjailijat ovatkin pääasiallisia sisältökohteita, pinnan alla myös kaikkia näihin liitettäviä kenttien arvoja käsitellään samanarvoisina objekteina, joilla taas voi olla omia kuvailujaan. Kun jokaiselle objektille vielä luodaan yksikäsitteinen URI-tunnuste, muodostuu semanttinen verkko, jonka mistä tahansa solmusta voi ottaa kiinni.

Käyttäjää ei siis sidota valmiiseen näkökulmaan, vaan kannasta saa samalla logiikalla esimerkiksi kirjal-

lisuuspalkintoihin keskittyviä listauksia, tai listauksen kustantajista jaoteltuna sen mukaan onko heidän listoillaan enemmän nais- vai miespuolisia kirjailijoita. Voidaan tarjota laajempia näkökulmia aineistoon, esimerkiksi kysyä mitkä teemat olivat keskeisimpiä vuonna 1974 ilmestyneissä yhteiskunnallisissa romaaneissa.

Koska yhteydet kohteiden välille kirjataan URI-tunnuksin tekstimuotoisten viittausten sijaan, saadaan lisähyötynä järjestelmän kieliriippumattomuus. Kullekin URI:lle voidaan myöhemmin lisätä nimi kaikilla halutuilla kielillä, joista ohjelma sitten valitsee näytettävän käyttäjän tarpeen mukaan. Samaten esimerkiksi henkilöiden nimien kaikki mahdolliset eri kirjoitusmuodot voidaan sitoa samaan, tiettyä henkilöä tarkoittavaan URI-tunnisteeseen.

Skeeman vaiheet ja FRBRoo

Tietoiseen linjaukseen pohjautuen Kirjasampo keskittyy vain painoksista riippumattomaan, teoksen sisältöön kohdistuvaan kuvailuun ja tietoon. Mukaan hiipi alusta lähtien kuitenkin myös tietoa kääntäjistä, julkaisuvuosista, kustantajista ja julkaisusarjoista, ja aluksi ohjeistettiin tallentamaan vain ensimmäisen suomenkielisen laitoksen tiedot. Tämä yhden kohteen malli toimi hyvin siihen asti, että projektin haluttiin laajentuvan myös ruotsinkieliseen kirjallisuuteen. Tällöin tuli välttämättömäksi miettiä uudestaan miten teoksen eri käsitteelliset tasot erotetaan toisistaan.

Oppia lähdettiin hakemaan FRBRoo-mallista (FRBRoo Model 2010), joka tunnistaa neljä käsitteellistä tasoa:

- Work: teoksen abstrakti sisältö = teoksen platoninen idea [pääasiallinen luoja, asiasanat]
- Expression: teoksen konkreettinen sisältö = alkuperäinen/käännösteksti, näytelmäkäsikirjoitus, elokuvakäsikirjoitus [kirjoittaja, kääntäjä, ohjaaja]
- Manifestation: konkreettinen teos/tuote = kirja, kokoelmakirja, näytelmäesityksen koko konsepti, elokuvan DVD [kustantaja, julkaisija, isbn]
- Item: fyysinen kappale = yksittäinen kirja/kokoelma/esitys/DVD

Yksinkertaista mallia ei kuitenkaan haluttu korvata koko neljän tason mallin monimutkaisuudella. Kirjasammon tiedot eivät koskaan mene yksittäisen kappaleen tasolle, joten Item-taso oli helppo pudottaa pois. Work-tason piti taas ehdottomasti olla erillään, jotta erikieliset laitokset pystyivät viittaamaan samoihin sisältöasiasanoihin. Expression- ja Manifestation-tasot päätettiin kuitenkin yhdistää, sillä keskimäärin yhdellä käännöstekstillä on yksi julkaisija ja fyysinen olomuoto, eikä tästä syystä toisteisia kuvauksia jouduttaisi tekemään paljoakaan.

Siirtyminen yhden käsitteellisen tason ratkaisusta kahteen oli pääasiassa yksinkertaista ja kivutonta.

Pieneksi ongelmaksi muodostuivat kuitenkin novellit ja niiden suhteet novellikokoelmiin. Tässä alun perin kahdesta kohteesta syntyi neljä, joiden väliset suhteet oli syytä määritellä tarkkaan. Samoin korjattiin skeemaa aikojen ja paikkojen tallennuksessa. Aluksi ne tallennettiin kulttuuritiedon tallennusmalli CIDOC-CRM:n (Doerr 2003) ja biografisen tiedon BIO-skeeman (Davis & Galbraith 2002-2010) hengen mukaisesti tapahtumiksi. Käyttäjätutkimuksessa kävi kuitenkin ilmi, etteivät tapahtumat ensisijaisina sisältökohteina ole helposti ymmärrettäviä indeksoijille kuin loppukäyttäjillekään. Tällöin kuvailussa siirryttiin takaisin perinteisempään malliin, jossa tiedot tapahtuma-ajoista ja paikoista tallennetaan suoraan esimerkiksi kirjailijan attribuuteiksi. (Mäkelä 2010, 65.)

Tähän skeeman pohjautuen räätälöitiin Kirjasammon tiedontallennusta varten annotointityökalu Kirjasampo-SAHA. Siinä on huomioitu kaunokirjallisuuden sisällönkuvailun erityispiirteet, ja siihen on integroitu kuvailussa käytettävät ontologiat. (Kurki & Hyvönen 2010.) SAHAN kautta ovat tiedontallentajan käytössä myös Kulttuurisampo-palvelun noin 550 000 kulttuurikohdetta.

Lähteet

- Berners-Lee, Tim, Hendler, Jim, Lassila, Ora (2001). The semantic web. *Scientific American* 284, no. 5, pages 34-43.
- Davis, Ian & Galbraith, David (2002-2010). BIO: a vocabulary for biographical information. Saatavana verkossa: <http://vocab.org/bio/0.1/.html>. (28.9.2010)
- Doerr, Martin (2003). The CIDOC CRM - an ontological approach to semantic interoperability of metadata. *AI Magazine* 3, vol. 24, s. 75-92.
- FRBRoo Model (2010). Chryssoula Bekiari, Doerr, Martin & Le Bœuf, Patrick (eds.) Saatavana verkossa: http://www.cidoc-crm.org/docs/frbr_oo/frbr_docs/FRBRoo_V1.0.1.pdf. (28.9.2010)
- Hypén, Kaisa (2009). Sisällöistä konteksteihin: Kaunokirjallisuuden verkkopalvelu semanttisen webin sovelluksena. *Informaatiotutkimus* 1, vol. 28, s. 7-18. Saatavana verkossa: <http://ojs.tsv.fi/index.php/inf/article/view/1834>. (30.9.2010)
- Kurki, Jussi & Hyvönen, Eero (2010). Collaborative metadata editor integrated with ontology services and faceted portals. Saatavana verkossa: <http://www.seco.tkk.fi/publications/2010/kurki-hyvonen-saha3-2010.pdf>. (28.9.2010)
- Mäkelä, Eetu (2010). View-based user interfaces for the semantic web. Dissertation, Aalto University, School of Science and Technology, Espoo. D.Sc. dissertation, accepted for publication. Saatavana verkossa: <http://www.seco.tkk.fi/publications/submitted/makela-drsc-thesis.pdf>. (29.9.2010)
- RDF Semantics (2004). Patrick Hayes, (ed). World Wide Web Consortium, W3C Recommendation. Saatavana verkossa <http://www.w3.org/TR/rdf-nt/>. (28.9.2010)