Informaatiotutkimuksen päivät 2014
6. - 7. marraskuuta, Oulun yliopisto, Oulu

ABSTRAKTI

Susanna Nykyri & Osma Suominen

YSAsta YSOon : merkkijonoista käsitteisiin

Yhteystiedot: Susanna Nykyri, Helsingin yliopisto, Kansalliskirjasto, susanna.nykyri@helsinki.fi
Osma Suominen, Helsingin yliopisto, Kansalliskirjasto, osma.suominen@helsinki.fi

Tausta

FinnONTO-projekti toi ontologiat kirjastomaailmaan (ks. Hyvönen 2005, Karhula 2005, Seppälä & Hyvönen 2014) jo lähes kymmenen vuotta sitten. Kehitystyö on pitkäjänteistä. Suuria sanastoja ja niiden välisiä yhteyksiä sekä käytäntöjä ei uudisteta hetkessä. Nyt ollaan kansallisen sanasto- ja ontologiapalvelu Finton (ks. Kansalliskirjasto 2014; Suominen ym. 2014) aktiivisen kehittämisen myötä kuitenkin jo tuotantovaiheen kynnyksellä.

Yleisen suomalaisen ontologian (jatkossa YSO) sisältö eli käsitteistö perustuu Yleiseen suomalaiseen asiasanastoon (YSA). YSOn alkuvaiheen (2003-2012) ontologisointityö tapahtui Aalto-yliopiston SeCo-tutkimusryhmässä (ks. Hyvönen ym. 2008; Hyvönen 2014). Vuodesta 2013 YSOa on kehitetty Kansalliskirjastossa sisällönkuvailun ja tiedonhaun apuvälineenä (ks. Lappalainen, Nykyri, Palonen 2013; Lappalainen, Frosterus, Nykyri 2014). Näkökulman muutos sanaston sisällöllisessä kehitystyössä perustuu Kansalliskirjaston ONKI-projektin tekemään arviointityöhön ja mm. käyttäjähaastatteluihin (ks. Nykyri & Palonen 2014).

YSAsta YSOon

YSA ja siten YSO perustuvat nykyisellään vahvasti kirjallisuustakuu-ajatteluun. - Niissä esitetyillä käsitteillä ja termimuodoilla pitää olla jo todennettu ja melko vakiintunutkin tarve päästäkseen kontrolloituun sanastoon. Sanastojen käsitteistö on päivittynyt melko hitaasti, ja se ei ole välttämättä heijastanut dynaamista ja kirjavaakin käyttäjädiskurssia. Koska YSO on perustunut käsitteistöltään jo todennettuun sisällönkuvailutarpeeseen (pääasiallisena kontekstinaan suomalaiset kirjastotietokannat), on se varsin kaukana ontologioista merkityksessä maailmaa mallintava käsitteistö (vm. tavasta ymmärtää ontologia ks. esim. Niemi 2008). Sen vuoksi esimerkiksi

“viikonpäiville” on annettuna vain kaksi suppeampaa käsitettä, “lauantai” ja “sunnuntai”, sillä muita eivät sisällönkuvailijat ole sanastoon ehdottaneet.

YSAn ja YSOn erot ilmenevät niin semanttisella kuin teknisellä puolella. Merkittävimmät erot YSOssa YSAan nähden ovat:

1. kaikenkattava ylähierarkia (ylimmällä tasolla ovat laaja-alaisimmat käsitteet: oliot, ominaisuudet sekä tapahtumat ja toiminta), jonka alle kaikki käsitteet on sijoitettu (vrt. YSA, jossa monilla termeillä ei ole laajempia termejä joten rakenne on hyvin litteä)
2. vahva käsitepohjaisuus: termit ovat vain nimityksiä käsitteille, jotka yksilöidään URI-tunnisteella (vrt. YSA, jossa pääsääntöisesti käytetään termejä)
3. monikielisyys: YSOn rakenne perustuu suomalaiseen kulttuuriin (ml. ruotsin kieli) ja käsitteet nimetään suomeksi, ruotsiksi ja englanniksi (vrt. YSAn yksikielisyys, jota Allärs osittain täydentää ruotsin kielen osalta)
4. semanttisen webin ja linkitetyn datan teknologia-alusta: tieto esitetään ensisijaisesti RDF-muodossa käyttäen mm. SKOS-mallinnustapaa (Baker ym. 2013) (vrt. YSAn termit MARC-tietueita), mikä mahdollistaa mm. automaattisen laadunvalvonnan (Suominen, Mader 2014) sekä linkitetyn datan julkaisun ja rajapintapalvelut (Hyvönen ym. 2008; Suominen ym. 2014)

YSOn hierarkia on pyritty muodostamaan “is a” -periaatteen mukaisesti. Tämä tarkoittaa sitä, että käsitteet nähdään luokkina, joiden merkitys supistuu liikuttaessa hierarkiassa ylätasolta kohti alempia tasoja. Tätä kautta hierarkia myös määrittää käsitteiden merkitystä ja YSA-käsitteiden sijoittelu YSOn hierarkiaan paljastaa usein monimerkityksisiä termejä. Voidaankin sanoa, että YSOssa käsitteiden merkitys on tarkemmin määritelty kuin YSAssa, vaikka varsinaisia määritelmiä käytetään harvoin. Tarkemmat merkitykset mahdollistavat niin tiedonhakutulosten tarkkuuden parantamisen kuin käsitteiden linkityksen muihin sanastoihin, tesauruksiin ja ontologioihin. Työn alla on YSO-käsitteiden linkittäminen LCSH-sanastoon. Lisäksi monet YSO-pohjaiset erikoisalojen ontologiat laajentavat YSOn sisältöä ja linkittävät omat käsitteensä YSOn käsitteisiin tuoden mukanaan tarkempia merkityksiä.

Sisällönkuvailu ja tiedonhaku

Perinteisessä termipohjaisessa sisällönkuvailussa, johon YSA ja VESA lukeutuvat, termien tarkka muoto on äärimmäisen tärkeää, koska kuvailtavan tietueen yhteys sanastoon ja muihin samaan asiaan liittyviin tietueisiin edellyttää auktorisoidun kirjoitusasun noudattamista. Tesaurus-standardeissa (ks. ISO 25964) määritellään tarkkoja sääntöjä sille, milloin on käytettävä yksikkö- tai monikkomuotoja, milloin isoa alkukirjainta jne, jotta yhteentoimivuus ei vaarantuisi. Korvattujen termien käyttö sisällönkuvailussa on kielletty, koska niitä käyttämällä merkkijonojen välistä yhteyttä ei synny.

Käsitepohjaisessa sisällönkuvailussa yhteys tietueiden ja sanaston välillä perustuu pysyviin tunnisteisiin (YSOn URI-tunnisteet). Vaikka käsite muuttuisi esimerkiksi suositeltavan termin vaihtuessa, sen tunniste pysyy samana. Termit ovat edelleen tärkeitä tarttumapintoja käsitteille, mutta ne eivät enää tietojärjestelmien tasolla toimi linkkeinä. Korvattujen termien asemesta voidaan puhua vaihtoehtoisista termeistä tai kielellisistä ilmaisuista. Ratkaisu mahdollistaa esim. sen, että yleisen sanaston puolella voidaan halutessa puhua “siirtoväestä” ja kaunokirjallisen sanaston puolella “evakoista”, ja silti voidaan viitata samaan käsitteeseen ja tavoittaa samassa haussa yhtä käsitettä käyttäen kummallakin ilmaisulla sisällönkuvaillut aineistot. Vastaavaan tapaan sisällönkuvailu voidaan tehdä suomeksi ja haku ruotsiksi, tai päinvastoin. Käsitteiden tunnisteet eivät riipu kielestä, joten erikielisiä ilmaisuja voidaan vapaasti käyttää sisällönkuvailussa ja

tiedonhaussa. Jos ontologiaan on lisätty sekä termien yksikkö- että monikkomuoto (YSOssa näin ei vielä ole, mutta esim. FinMeSH-sanastossa tähän on pyritty), molempia muotoja voidaan käyttää.

Semanttisessa linkittyvässä webissä sisällönkuvailua tuottavien ja sitä käyttävien toimijoiden kirjo on ennennäkemätön. Jo viime vuosituhannen vaihteessa ja silloin ns. perinteisessä suljetussa tietokantaympäristössä tiedonhakua ja -tallennusta pidettiin haastavina yksistään kielellisistä syistä johtuen: jokainen tiedonhaku tietokannassa käsittää ainakin viisi erilaista kieltä - kirjoittajan, sisällönkuvailijan, synteettisen rakenteen, tiedonhakijan ja tiedonstrategian (Buckland 1999). Nämä kaikki edustavat toisistaan poikkeavaa diskurssia, jotka ontologia voi yhdistää tai sillata. Perinteisten asiasanastojen tapaan ontologia ei ole vain silta tiedonhakijan ja tiedontallentajan osin eriävien puhetapojen välillä, vaan myös siis esimerkiksi eri sisällönkuvailijoiden kesken. Sisällönkuvailukaan kun ei ole termitasolla aina yhteneväistä, ei edes yhden ja saman sisällönkuvailua tekevän henkilön toiminnassa, mutta käsitetasolla yhteneväisyyden määrä on merkittävästi suurempi (Iivonen 1989 ja 1995, Iivonen & Kivimäki 1998).

Lopuksi

Ideatasolla ontologia antaa niin sisällönkuvailuun, tiedonhakuun kuin aineistojen linkittyvyyteen huikeat mahdollisuudet. Tekniset innovaatiotkin jo mahdollistaisivat paljon nykyistä rikkaammat ratkaisut, mutta käytännön kehitystyö ja toimintatavoista sopiminen sekä toimintakulttuurin muutos vievät aikaa.

Kontrolloidun sanaston kehittämisessä esimerkiksi varsin tiukkaa kirjallisuustakuu-ajattelua (merkityksessä asiasanaehdotus sisällönkuvailua tekevältä) on myös kyseenalaistettu, ja siihen onkin ryhdytty tekemään ontologisointi-yhteistyön myötä lievennyksiä ja laajennoksia. Oleellinen kehityskohde on semanttisen webin heterogeenisuuden ja linkittyvyyden vuoksi mm. käsitteiden ja termimuotojen ehdotusjärjestelmän kehittäminen.

Lähteet

Baker, Thomas & Bechhofer, Sean & Isaac, Antoine & Miles, Alistair & Schreiber, Guus & Summers, Ed 2013: Key choices in the design of Simple Knowledge Organization System (SKOS), *Web Semantics: Science, Services and Agents on the World Wide Web*, Volume 20, May 2013, Pages 35-49, ISSN 1570-8268, http://dx.doi.org/10.1016/j.websem.2013.05.001.

Buckland, Michael 1999: Vocabulary As A Central Concept In Library And Information Science. Preprint of paper published as "Vocabulary as a Central Concept in Library and Information Science" In: Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of the Third International Conference on Conceptions of Library and Information Science (CoLIS3, Dubrovnik, Croatia, 23-26 May 1999. Ed. by T. Arpanac et al. Zagreb: Lokve, pp. 3-12

Hyvönen, Eero 2005: Miksi asiasanastot eivät riitä vaan tarvitaan ontologioita. *Signum* 5:2005. URL: http://ojs.tsv.fi/index.php/signum/article/view/3358/3108

---------- 2014: FinnONTO-hanke loi ontologisen perustan kansalliselle webin tietoinfrastruktuurille. *Tieteessä tapahtuu*, no. 3, 2014. URL: http://www.seco.tkk.fi/publications/2014/hyvonen-finnonto-2014-02-19.pdf

Hyvönen, Eero & Viljanen, Kim & Tuominen, Jouni & Seppälä, Katri 2008: Building a National Semantic Web Ontology and Ontology Service Infrastructure--The FinnONTO Approach.

Proceedings of the European Semantic Web Conference ESWC 2008, Springer, Tenerife, Spain, June 1-5, 2008. URL: http://www.seco.tkk.fi/publications/2008/hyvonen-et-al-building-2008.pdf

Iivonen, Mirja 1989: *Indeksointituloksen riippuvuus indeksointiympäristöstä*. University of Tampere, Department of Library and Information Science, Studies No. 26/1989.

---------- 1995: Hakulausekkeiden muotoilun yhdenmukaisuus onlineviitehaussa. Tampereen yliopisto, Acta Universitatis Tamperensis. Ser. A, vol. 443, Tampere.

Iivonen, Mirja & Kivimäki, Katja 1998: "Common Entities and Missing Properties: Similarities and Differences in the Indexing of Concepts." Knowl. Org. 25(1998) No.3, pp.90-102

ISO 25964-1:2011 = Information and documentation - Thesauri and interoperability with other vocabularies -- Part 1: Thesauri for information retrieval. 1st ed. Geneva: ISO, 2011 (ISO 25964-1-2011-08-05).

ISO 25964-2:2013 = Information and documentation -- Thesauri and interoperability with other vocabularies -- Part 2: Interoperability with other vocabularies. 1st ed. Geneva: ISO, 2013 (ISO 25964-2- 2013-03-04).

Kansalliskirjasto 2014: *Finto : Suomalainen sanasto- ja ontologiapalvelu*. URL: https://wiki.helsinki.fi/display/ONKI/ONKI-projekti

Karhula, Päivikki 2005: Ontologiat kääntävät maailmankuvan. Signum 5:2005. URL: http://ojs.tsv.fi/index.php/signum/article/view/3357/3107

Lappalainen, Mikko & Frosterus, Matias & Nykyri, Susanna 2014: "Reuse of library thesaurus data as ontologies for the public sector". Paper presented at: IFLA WLIC 2014 – Lyon - Libraries, Citizens, Societies: Confluence for Knowledge in Session 86 - Cataloguing with Bibliography, Classification & Indexing and UNIMARC Strategic Programme. In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France. URL: http://library.ifla.org/819/1/086-lappalainen-en.pdf

Nykyri, Susanna & Palonen, Tuomas 2014: "Ontologioiden käytön moninaisuudesta ja tulevaisuusnäkymistä: Finto-palvelun sidosryhmien ontologiapalvelulle kohdistamat kehitystoiveet." Kansalliskirjasto. URL: http://www.doria.fi/handle/10024/94507

Seppälä, Katri & Hyvönen, Eero 2014: *Asiasanaston muuttaminen ontologiaksi : Yleinen suomalainen ontologia esimerkkinä FinnONTO-hankkeen mallista.* Kansalliskirjasto. URL: http://urn.fi/URN:ISBN:978-952-10-9883-3

Suominen, Osma & Mader, Christian 2014: Assessing and Improving the Quality of SKOS Vocabularies. *Journal on Data Semantics*, vol. 3, no. 1, pp. 47-73, 2014. URL: http://www.seco.tkk.fi/publications/2014/suominen-mader-skosquality.pdf

Suominen, Osma & Pessala, Sini & Tuominen, Jouni & Lappalainen, Mikko & Nykyri, Susanna & Ylikotila, Henri & Frosterus, Matias & Hyvönen, Eero 2014: Deploying National Ontology Services: From ONKI to Finto. *Proceedings of the ISWC 2014 industry track*, Riva del Garda, Italy, October, 2014. URL: http://www.seco.tkk.fi/publications/2014/suominen-et-al-deploying-onki-finto-2014.pdf