Yhteinen tiedon hallinta -hanke

Sanaston metatietomallin määrittely -työpaja

Aika	02.11.2016 klo 9.00–12.00
Paikka	VM, nh Loppupeli / etäyhteys* *etäyhteystiedot kalenterikutsussa*

Osallistujat

Miika Alonen, CSC
Kristiina Asp, UM
Riitta Autere, PRH
Matias Frosterus, Kansalliskirjasto
Katri Haverinen, Lingsoft Oy
Outi Hermans, Helsingin kaupunki
Mikael af Hällström, Vero, etänä (klo 10 >>)
Jari Kallela, VM (klo 10.30 asti)
Virpi Kalliokuusi, THL
Simo Kankkunen, VNK, etänä
Anne Kauhanen-Simanainen, VM, etänä
Elisa Kettunen, Kuntaliitto (klo 10.30 asti)
Kaisa Kuhmonen, VNK
Jussi Kurki, THL
Jaakko Laakso, Tilastokeskus
Marko Latvanen, Valtionkonttori
Tarja Myllymäki, MML, etänä
Vesa Nissinen, Kela
Tarja Pykälä, MML
Suvi Remes, VM (koordinointi)
Walter Rydman, CSC
Katri Seppälä, TSK
Saku Seppälä, TSK
Jari Ylikoski, Kuntaliitto

Työpajan konkreettisena tavoitteena on olemassa oleviin suosituksiin (JHS-175, ISO-standardit terminologisesta sanastotyöstä, SKOS ym.) ja keskustelussa esiintuotuihin muihin huomioihin pohjautuen sopia ”määrittelevän /terminologisen sanaston metatietomallista”, jotta erityyppisten sanastojen saatavuutta sekä ihmis- että koneluettavana ja laajaa hyödyntämistä voidaan jatkossa tukea. ”Metatietomalli” tässä tarkoittaa sanastolta vaadittavien ominaisuuksien ja tietojen määrittämistä.

Työpaja liittyy Yhteinen tiedon hallinta -kärkihankkeen kehittämistoimenpiteisiin (sanastotyön ja sanastojen käytön edistäminen ja tähän liittyvä työkalutuen kehittäminen, ns. Yhteentoimivuuden välineistö).

Muistio

Aloitettiin työpaja lyhyellä kertauksella aamupäivän työskentelyn tavoitteista ja taustarakenteena toimivasta Yhteinen tiedon hallinta -kärkihankkeesta sekä esittäytymiskierroksella. Työskentelyn tuloksena syntyvän määrityksen ja JHS-175:n liitoksesta todettiin, että mahdollinen muutostarve JHS-175-suositukseen arvioidaan erikseen.

Työpajassa syntyvän määrityksen käyttökohdetta konkretisoitiin lyhyellä, THL:n omiin tarpeisiinsa kehittämän ”Termieditorin” väline-esittelyllä. Editoriin liittyen esitettiin kysymyksiä, joita käsiteltiin nykyisen toteutuksen näkökulmasta. Editoriin liittyvistä rajoitteista todettiin, että

- nykyinen käyttöliittymäratkaisu ei pysty käsittelemään termiin liitettyä tietoa ja sitä pitäisi siis kehittää, mutta tietokanta taipuu laajennukseen kyllä.
- nykyisessä toteutuksessa ei ole graafista tukea, esimerkiksi käsitekaavioiden näyttämiseen.
- kenttien arvojen järjestyksen Editori säilyttää.
- Editori mahdollistaa luokitukset ja viittauksen aineistojen yli, tämä toteutus ei ole sidot tu SKOS-formaattiin.
- Editori ei nykyisellään tue muotoiludataa (esimerkiksi kursiivit, lihavoinnit ym.), mutta

tämän huomioiva kehitys olisi mahdollista.
- käsitehiearkkiapuu rakentuu hierarkkiasuhteista ja näyttää yläkäsitteet.
- aineiston saa nykyisellään Editorista ulos JSON-muotoisena (käyttöliittymä tukee), RDF-exporttina (suositus, mutta hukkaa mm. mainitun kenttien arvojen järjestyksen) ja CSV-tiedostona.

Keskusteltiin mainituista kysymyksistä myös Finto-palvelun näkökulmasta. Kenttien arvojen järjestyksen säilyttäminen on vaikeaa Fintonkin hyödyntämässä RDF-muodossa; toteutus on mahdollinen, mutta erittäin työläs se tulisi arviolta olemaan. Finto tukee jo nykyisellään hmtl-muotoiluja, mutta yksikään aineisto ei ole ominaisuutta hyödyntänyt; tarve ilmaantunut kansainvälisestä kehitysyhteistyöstä. Terminologiselle sanastolle ominaiset viittaukset literaalikentän sisällä resurssiin (ts. määritelmä ja huomautus sisältävät tiedon ko. käsitteen ymmärtämisen kannalta keskeisistä käsitesuhteista) ovat vaikeita. Staattinen viittaus onnistuu kyllä, mutta dynaaminen on haastavampi, ts. jos viitatun käsitteen termi vaihtuu, toinen resurssi ei automaattisesti tätä tiedä tai tietoa muutoksesta saa. Todettiin, että päivittämisprosessi on sanastotyön puolella olemassa jo nyt sanastotyön editoreissa ja voisi olla jatkossakin näin eli olisi syntyvän Termieditorin tehtävä, ei julkaisu- ja hakupalvelun.

Työskentelyn pohjaksi laadittua ehdotusta käytiin läpi IOW-välinettä hyödyntäen: http://iow.csc.fi/model/st/.

Alustuksena todettiin, että ko. malli sopii THL:n kehittämän Editorin jatkotyön pohjaksi (teknisessä mielessä), ja mahdollistaa myös rakenteiset termit. Käsite ja termi on nyt määritelty kahtena eri luokkana, jotta tarkempaa termikohtaista tietoa voidaan myös esittää. Tähän todettiin, että käsitteen ja termin tietoja täytyy pysyä katsomaan kokonaisuutena ja yhtä aikaa. Editorin käyttöliittymää tarvitsisi siis kehittää, jotta mahdollistettaisiin myös termin tietojen editointi. Termillä on nyt tietoja, jotka kaikki eivät sisälly SKOS-formaattiin (mm. termin tila).

Keskusteltiin mallin määrittelyn tarkkuudesta, että onko se tarkoitettu kaikille käyttöön sellaisenaan vai voiko laajentaa sanastokohtaisesti. Todettiin, että tavoitellaan kaikille yhteistä ns. minimimääritystä. Tämä tarkoittaa, että muutamiin potentiaalisesti haasteellisiin kohtiin on löydettävä toimintatapa; esimerkiksi sanaston tietojen järjestyksen näyttäminen (toteutusvaihtoehdot: indeksi, järjestysnumero, RDF-listat tai koodilla järjestäminen) ja sen ylläpito, joka pitäisi olla käyttäjän määritettävissä. Tähän liittyvät synonyymit ja kieliversioiden järjestys sekä huomautusten järjestys. Toiminnallisuudesta todettiin, että tekninen toteutus tulee piilottaa sisällönkehittäjältä, esim. RDF-listat. Toiseksi, tällä metatietomallilla tuotettuja sanastoja voidaan siirtää editorista toiseen tai käyttää metatietomallia pohjana sanaston kehitystyössä.

Lähdettiin käymään läpi luonnosta pääluokat kerrallaan. Pääluokat tulevat SKOS-formaatista pääkäsitteistä ja suomennokset TSK36 Terminologian sanastosta. Mikäli lisämäärittelyjä tarvitaan, ne annetaan YSR-ryhmälle tehtäväksi.

Käsitejärjestelmällä tarkoitetaan ”sanastoa”, mallinnusteknisesti SKOS-tietomallissa skos: ConceptScheme; kyse on tiedoista, jotka sanastosta halutaan antaa ja näyttää esimerkiksi julkaisuvaiheessa.
- Muutettu luokan nimi käsitejärjestelmä >> *Terminologinen sanasto*
  - vastaa esim. Finto-palvelussa sanastosta annettavia tietoa sivulla ”Tietoja sanastosta”
  - lisätään luokkaan vastaavat tiedot kuin Finto-palvelussa tulee sanastosta antaa
  - omistajuus, tekijä ja lisenssitiedot ovat merkittävä mm. YTI-hankkeen tavoitteiden näkökulmasta, kun rakennetaan tiedon hallinnan hallintamallia
  - kuvaus oltava vähintään yhdellä kielellä
  - terminologinen sanasto korostaa sanaston taustalla olevia sanastotyön menetelmiä sanaston laadinnassa
  - määrittelevä sanasto ymmärretään liian monella tavalla, ja ei siksi sovellu

Käsitevalikoima tarkoittaa yhteen sanastoon liittyvien käsitteiden ryhmittelyä jollakin perusteella jotakin tarkoitusta varten. SKOS-tietomalli (skos:Collection) ei rajoita käsitteiden ryhmittelyä yhteen sanastoon, mutta tässä tehdään rajaus sanaston sisäiseen korpukseen. Käsitevalikoima viittaa esimerkiksi aiheenmukaiseen luokitteluun, print-muodossa voi olla esimerkiksi yhden otsikon alle tulevat käsitteet. Esimerkiksi OKSA-sanasto ja käsitteiden ryhmittely (Opetus-, koulutus- ja varhaiskasvatussanasto) on rakennettu koulutuksen laatuympyrän pohjalle tai Palkkahallinnon sanastossa palkanerien käsittely palkkahallinnossa

muodostaa käsitevalikoiman. Fintossa käytetään asiasanastoille tyypillisesti teemoittaisina ryhminä; musiikin osalta voisi tarkoittaa esimerkiksi tyylisuunta-, esittämistilanne- tai instrumenttiperusteista jakoa. Tutkimuksen sanastotyöstä tuotiin esille, että sama termi voi esiintyä useammassa käsitevalikoimassa riippuen määrittelyn kontekstista ja valitusta määrittelyn näkökulmasta. Keskusteltiin onko tämä ongelma, että käsitevalikoimien muodostamisen taustalla vaikuttavat erilaiset käytänteet aineistotyyppikohtaisesti ja todettiin, että on hyvin potentiaalinen hämmennyksen aiheuttaja. Ratkaisuna voi toimia sanastoresursseille tehtävä ”tyyppiluokitus”, jolloin pitäisi olla selvää kunkin luokituksen arvon (=sanastotyypin) osa mihin menettelyyn käsitevalikoima perustuu. Lisäksi esimerkkiä voisi yrittää tarkentaa (nyt lyhyesti: alkuperän mukaan).

Käsite-luokan ominaisuuksista sovittiin, että

- suositeltava termi pakollinen antaa yhdellä kielellä
- määritelmä ei voi olla pakollinen tieto, sillä Editoria tullaan hyödyntämään myös tilanteissa, jossa korpusta vasta kasataan eikä määritelmää siis vielä ole. Määritelmän pakollisuuden voisi kytkeä myös käsitteen tila-tietoon, eli esimerkiksi keskeneräisiltä käsitteiltä sitä ei vaadittaisi, mutta julkaistavalta käsitteeltä kyllä; tila-tietoa voisi näin hyödyntää myös sanaston käsittelyssä hakutoiminnallisuuksissa, kun halutaan nähdä esim. kaikki käsitteet, joilta puuttuu määritelmä. Tilatiedon käyttö ei kuitenkaan saisi merkittävästi työllistää sanastotyön prosessia ja sanaston ylläpitäjää. Tähän mietittävä hallintamalli Editorin kehitystyössä.
    - Fintossa sanaston julkaisu- ja kehitysversio ovat erikseen, näiden välillä ”muunnospalikka”. Uudet käsitteet listataan kuukausittain erikseen tiedoksi, muokkauspäivämäärän perusteella.
    - Yhtenä sanaston kehitysversion käyttötapauksena on esimerkiksi lainsäädännön muutosvalmistelu, jossa käsitteitä vasta hahmotellaan vastaamaan tulevaa lakia/asetusta/säädöstä. Huomautettiin myös, että ns. säädösperusteiset termit eivät ole yksi yhteen tietomallissa käytettyjen käsitteiden ja luokkien kanssa. Myös tähän haasteeseen ns. yhteentoimivuusmenetelmä pyrkii vastaamaan, mutta lainsäädännön valmisteluun liittyvän yhteistyön kanssa ollaan vielä aivan alussa.
- Lähde
    - viittaa nyt normaalin käytänteen mukaisesti käsitemäärityksen lähteenä käytettyyn resurssiin, joka voi olla esimerkiksi kansainvälinen sanasto
    - nähdään tärkeäksi pystyä myös linkittämään omat sanastot kansainvälisiin vastaaviin, siis viittaamaan niihin koneluettavasti; tämä tieto ei kohdistu lähde-ominaisuuteen.
- Kommentti, käytetään yleensä luonnosvaiheessa käsitesisältöön liittyvässä hahmottelussa
    - toimii eräänlaisena ”muutoslokina” sanastoprosessissa (kuka sanoi, mitä sanoi jne.)
    - tärkeä huomata, että ei tarkoita ”huomautusta” >> huomautus täytyy lisätä käsitteen tietoihin
    - kommentti voi kohdistua mihin tahansa käsitteen tiedoista; ei välttämättä ole rakenteisessa muodossa
    - työkommentti >> ratkaistava miten liitetään tiettyyn kenttään (RDF) >> ovat ”muistilappuja”/perustelutietoa; käyttö helpompaa, jos kohdistettu
    - käyttöesimerkki >> kala = hauki -esimerkki viittaa metametametatieto ja metatieto -keskusteluun, mikä on eri asia kuin terminologinen sanasto >> poistetaan koko kenttä (Ylläpidon kommentti -tieto riittää)
        - terminologisessa sanastossa huomautukseen voi sisältyä esimerkki/esimerkkejä
- käyttöala >> ilmaistaan esimerkiksi yleiskielisyyttä tai sidosta lainsäädäntöön
- ei-suositeltava termi
    - on eri asia kuin piilotettu termi
    - on eri asia kuin vanhentunut termi
- piilotettu termi
    - käytetään käyttöliittymässä haun optimointiin >> tarvitaan siis sanaston tehokkaaseen hyödyntämiseen, mutta sisällöllisestä näkökulmasta ei merkittävä
    - printtijulkaisussa saattaa näkyä hakemistossa
      >> ei kuitenkaan ole sama käsite kuin mihin sanastossa on linkitetty
    - Fintossa myös käytössä hakutoiminnallisuuden taustalla, hyödynnetään mm. yksikkömuotojen tallennukseen
- joskus käytetään myös ”explan” (explanation, NTRF-tietomalli) -metatietokenttää ilmaisemaan pidempää, ei terminologista selitettä käsitteelle, on löyhempi selite määritelmän sijaan
    - jos käsitteellä on määritelmä, selite-tietoa ei voi olla
    - joskus hyödyllinen sisäisen työprosessin aikana

  - ei ole käyttöalasta riippumaton tieto
  - ei oteta mukaan tähän metatietomallin ainakaan tässä vaiheessa; voi olla monille käyttäjille vaikea erottaa määritelmästä
  - selkeytettävä skos:note ja skos:conceptNote -kenttien käyttö tämän metatietomallin yhteydessä >> skosnote ei välttämätön terminologisessa työssä
- vaihtoehtoinen termi, korvattu termi >> ”korvattu” perustuu aineiston käyttöön
  - sanastotyyppikohtaista variaatiota käytössä; asiasanastoissa ohjaustietona
  - tässä terminologisen sanaston kontekstissa voisi olla myös ”synonyymi”
    - skos:altlabel:alakäsite
- muutoshistoria
  - selkeytettävä metatiedon määrittelyyn skos:changeNote ja skos:historyNote -tietojen ero
    - toinen on käsitteen sisällön (käsitepiirteen/piirteiden muutos) ja toinen kuvaa muuta käsitteen ylläpidossa tapahtunutta muutosta
    - käsitteen merkityksessä tapahtuneella muutoksella yleensä kauaskantoiset seuraukset ja yhteentoimivuuden näkökulmasta muutokset ongelmallisia; yleensä tarvitaan/halutaan myös tieto miksi muutettu
    - Fintossa muutosta ilmaistu seuraavasuhteella; esimerkiksi ATK-ohjelmista tuli tietokoneohjelmia
- termin tila >> tarkoittaa itse asiassa käsitteen tilaa >> muutetaan, myös kieliversiot (termstatus)
  - liitetty luokitus käytössä tietomallieditorin (IOW) puolella
    - onko ero keskeneräisen ja luonnoksen välillä riittävän selvä; tarkoitettu, että ”keskeneräinen” viittaa aivan 1. versioon, ”luonnos” työryhmän jo työstämään versioon
      - JHSmetassa ollut käytössä ”ehdotus”, tässä luokituksessa vastaa ”luonnos”
    - tarvitaan myös deprecated poistetuille; käsitteitä ei voi linkitetyssä maailmassa vain poistaa; esim. Fintossa on oma käsitetyyppi, jolla on sanastossa omat suhteet (myös ne usein muuttuvat, kun sanastoa muokataan), ja johon pääsee käsiksi vain URI-linkin tietämällä.
- Keskusteltiin vielä määrittelyn näkökulmista ja tarpeista ilmaista samakin käsite eri sanoin eri kohderyhmille; erityisesti ”asiakaskeskeisyys” voi aiheuttaa tarvetta yksinkertaistaa/kansankielistää määrittelyjä
  - esimerkinomaisesti katselmoidussa Termieditorin luokituksessa oli kenttiä potilassegmenttikohtaisesti, voisi miettiä voisiko toteutustapaa hyödyntää jatkossa eri käyttäjäryhmien huomioimisessa sanastojen linkittämiseen rinnalla. Osallistuneilla ei ollut tietoa kuinka ko. luokituksen ko. tietoja käytännössä hyödynnetään.

AP (huomioitava, että IOW-väline näyttää aina uusimmin tiedon; muutoshistoria tallentaa aiemmat tiedot)

- lisätään sanastolle metatiedot Finton esimerkin mukaisesti (http://finto.fi/koko/fi/ >> Tietoa sanastosta)
- tarkennetaan käsitevalikoimalle esimerkki
- lisätään käsitteelle muokkauspäivämäärä
- lisätään käsitteelle huomautus; voi olla usealla kielellä
- ratkaistava kommenttien kohdistaminen käsitetietueen kulloinkin merkittävään osaan
- käyttöesimerkki-tieto poistettava
- selkeytettävä skos:note ja skos:conceptnote -kenttien käyttö terminologisen sanaston yhteydessä
- selkeytettävä metatiedon määrittelyyn skos:changenote ja skos:historynote -tietojen ero
- muutetaan termin tila >> käsitteen tila
- lisätään poistettu/deprecated -tilakoodistoon (käsitteen tila)
- jatketaan työ loppuun työpaja kakkosessa, mielellään ennen joulua; Remes kutsuu ja doodlella ajankohta
- viedään tarkentamista vaativat käsitteet määriteltäväksi YSR:lle

Liitteet Agenda ja YTI-hanke-info
Taustamuistio ja keskustelun pohjamateriaali