Yhteinen tiedon hallinta -hanke

Sanaston metatietomallin määrittely -työpaja, osa 2

Aika	08.12.2016 klo 9.00–11.30
Paikka	VM, nh Markka / etäyhteys* *etäyhteystiedot kalenterikutsussa*

Osallistujat

Miika Alonen, CSC
Dea Crichton-Turley, TEM
Matias Frosterus, Kansalliskirjasto
Katri Haverinen, Lingsoft Oy
Outi Hermans, Helsingin kaupunki, etänä
Virpi Kalliokuusi, THL
Simo Kankkunen, VNK, etänä
Kaisa Kuhmonen, VNK
Jaakko Laakso, Tilastokeskus
Marko Latvanen, Valtionkonttori, etänä
Sini Lehto, PRH
Tarja Myllymäki, MML, etänä
Vesa Nissinen, Kela
Kirsi Pispa, CSC, etänä
Tarja Pykälä, MML
Suvi Remes, VM (koordinointi)
Walter Rydman, CSC
Katri Seppälä, TSK
Saku Seppälä, TSK
Osma Suominen, Kansalliskirjasto
Jari Ylikoski, Kuntaliitto, etänä

Sisältönä työskentelyn jatkaminen 2.11.2016 pidetyn työpajan pohjalta.

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

Muistio

Remes muistutti edellisen työpajan muistiosta; siihen tulleet muokkauspyynnöt on huomioitu, eivät varsinaisesti muuta asiasisältöä. Nämä asiakirjat ja jatkossa hankkeen materiaalit löytyvät Julkict:n wikistä: https://wiki.julkict.fi/julkict/yti.

Keskusteltiin vielä sanastojen kieliversioinneista. Tuotiin esiin tarve ruotsinkielisille sanastoille ja huoli niiden puutteesta, mm. KaPA-palveluiden kontekstissa. Todettiin, että työpajan aiheena oleva metatietomalli ja siihen jatkossa pohjautuva väline mahdollistavat monikielisen sanastotyön. Kyse on enemmän sanastotyön prosesseista ja etenkin niihin liittyvästä hallintamallista. Remes vie tiedon monikielisistä sanastoista ja niiden valmisteluun liittyvästä tarpeesta YTI-hankkeessa valmisteltavaan ”hallintamallityöhön” ja/tai sen kautta muille asiaan liittyville toimijoille.

Jatkettiin työskentelyä osoitteesta http://iow.csc.fi/model/st/ löytyvän ehdotuksen parissa. Käytiin aluksi läpi edellisellä kerralla kesken jääneet *käsitteen* attribuuttitiedot:

- systemaattinen merkintätapa (skos:notation)
  - todettiin, että sallitaan vain yhdenlaisia notaatioita per sanasto
  - samalla vahvistetaan skos:notation -kentän systemaattista käyttöä ja vältytään jo tunnistetuilta SKOSXL-laajennosongelmilta tässä asiassa
  - ”merkintätapa” kentän nimessä liian laaja ilmaus >> kyseessä on eräänlainen luokitus, ”luokitustapa” parempi (yhdenlainen ”notaatiokoodi”, mutta terminä huonoa suomea)

- lähde
  - käytetään ilmaisemaan, mistä ”käsite on peräisin”, jos vastaava merkityssisältö löytyy jo toisesta, esimerkiksi kansainvälisestä, olemassa olevasta sanastosta
  - on tilanteita, joissa on useammalle lähdemerkinnälle tarve, esimerkiksi erikielisten vastineiden lähteiden merkitseminen; kielikoodistoissa lähde voi olla eri riippuen esimerkiksi mitkä koodit valitaan käytettäväksi; kemikaalitiedoissa voi olla useita lähteitä, ovat usein vahvalla mandaatilla kansainvälisesti standardoituja >> pystyttävä ilmaisemaan
  - voidaan ilmaista vapaana tekstinä, voi olla linkki lähteeseen; ei edellytä mitään tiettyä notaatiota (voi olla viittaus tiettyyn lainkohtaan tai kirjallisuuteen)
- lisätty käsitteelle attribuutti ”viimeksi muokattu (päivämäärä)”
- sovittiin lisättäväksi käsitteelle myös ”luotu (päivämäärä)”

- keskusteltiin vielä tietotyypeistä
  - nyt tietotyypit välineistössä XSD-standardin pohjalta; ei ota kantaa tiedon semantiikkaan
  - muutamia lisäyksiä tehty, muuan muassa hmtl
  - nykyisellään ei mahdollista laajennettua tietotyyppiä; voitaisiin mallintaa skos:notaatiolle alliominaisuuksiksi

- kaikille sanastomallin luokille on lisätty URI-kenttä pysyvän tunnisteen käytön mahdollistamiseksi

Käsiteltiin seuraavaksi *Termi*-luokka. Tavoitteena on siis pystyä ilmaisemaan myös termiin liittyvää, termikohtaista tietoa rakenteisena ja tietenkin termin suhde käsitteeseen. Mallissa on tässä kohtaa hyödynnetty SKOSXL-laajennoksia. Termiin liittyen keskusteltu, että:

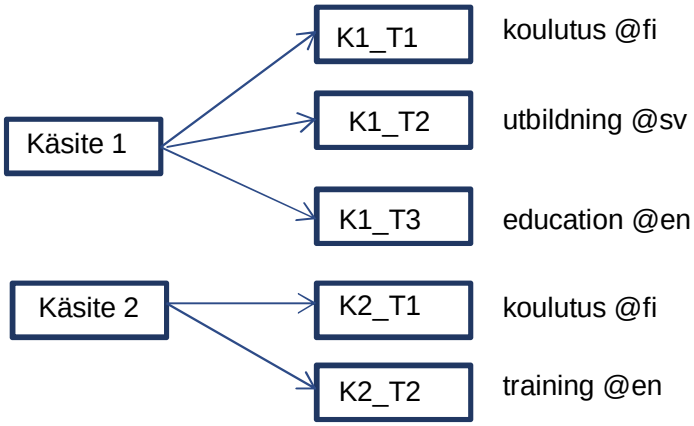
- termin URI, mahdollistaa pysyvän tunnisteen antamisen; voidaan välineistössä jatkossa generoida automaattisesti
- termi, kieliversioitu teksti (skosxl:literalForm) viittaa siis tekstimuotoiseen termiin
  - onko tekstimuotoisuutta tarpeen korostaa; joskus myös symboli voi olla termi >> ”literalform” pitää sisällään myös symbolit; unicode-merkki katsotaan myös symboliksi
  - useammin symboli termin synonyyminä, katsotaan termin osaksi
  - voisiko nimitys siis olla vain yksinkertaisesti ”termi” tai ”termin arvo”?
- lähde, ehdotus ok
- käyttöala
  - todettiin, että yhtenäisen kansallisen luokituksen laatiminen käyttöaloja kuvaamaan ei ole mahdollista; käyttöalojen merkitsemisen tarve arvioidaan yleensä sanastotyön edetessä kunkin sanaston osalta erikseen ja osin käyttöalat itsessäänkin muodostuvat osana sanastoprosessia ja sanaston käyttötarpeen mukaisesti
    - tulevaisuudessa ehkä mahdollista sanastoja koneluettavasti hyödyntämällä (viittaamalla) määritellä tarkemmin ja tehokkaammin myös käyttöalana käytetyt käsitteet; nyt on usein jokin samassa sanastossa määritelty käsite
    - käyttöalan määrittelyn vaatimus johtaa nykyisellään helposti ”tuplatyöhön”, jos käyttöalana tarvittava käsite ei kuulu ko. sanaston teeman piiriin (ts. tarvitaan vain käyttöala-tietoa kertomaan)
  - käyttöala merkitään yleensä sanana tai sanaliittona, ei ole juoksevaa tekstiä (huomautuksessa voidaan kertoa tarvittaessa käytöstä enemmän) >> tarkastettava tietotyyppi
    - kielivastineista erityisesti kerrotaan yleensä enemmän kuin pelkkä käyttöala, käyttökontekstikohtaista tietoa merkitään yleensä lisätietokenttään; voi sisältää myös kieliopillista tietoa, esim. käytöstä kielellisessä rakenteessa
  - tarvitaan muun muassa erikielisten vastineiden käyttöä ohjamaan, ts. termin kieliversiot voivat vaihdella käyttöalasta riippuen
  - tärkeää huomata ero asiasanastojen suhteen >> skos:scopeNote-kenttää käytetään niissä yhdenlaisena käyttöhuomautuksena (esimerkiksi tyyppiä ”käytetään eläimistä”) >> väärinymmärrysten välttämiseksi ei käytetä siten terminologisten sanastojen kontekstissa samaa standardikenttää eri tehtävään
- ääntäminen >> poistetaan tästä mallista tarpeettomana
  - otettu mukaan esimerkkinä siitä miten tätä mallia voidaan tarvittaessa laajentaa (käytettyjä standardeja laajemmaksi ja/tai niitä yhdistelemällä)

    - ääntämisen ohje otetaan yleensä muualta, ts. ei määritellä sanastotyössä; käyttötarve hyvin vähäistä terminologisessa sanastotyössä, viittaa enemmän esimerkiksi akateemisen kielentutkimuksen puolelle ja sen hyödyntämiin sanastoihin
- muistiinpano >> sovittu muotoon ”ylläpitäjän muistiinpano” (skos:editorialNote)
    - viittaa termistä tehtyjen huomioiden kirjaamiseen työprosessin aikana, voi olla hyvin vapaamuotoistakin
    - keskusteltiin attribuutin nimestä, huomio, huomautus vai muistiinpano >> valittu viimeisin, sillä kyseessä on sanaston tuottamisvaiheen kirjaus, eikä näy esim. valmiissa julkaisussa
- kommentti >> sovittu muotoon ”luonnosvaiheen kommentti”
    - tämän kentän tieto liittyy vain termiin, käsitteellä on oma kommentti-kenttänsä
        - käsitteeltä poistettu ”työkommentti”
    - hyödynnetään muun muassa sanaston kommentointikierroksilla
    - on sanastoprosessin työstämisvaiheen kenttä, ei näy esim. valmiissa julkaisussa
        - liittyy näin myös termin tila-tietoon
- lisätty ”termin käytön historiatieto” (skos:historyNote)
    - vastaava kenttä kuin käsitteellä; auttaa hahmottamaan käytön kontekstia, kun kieli elää
- termin tila
    - luokitus tulee kieliversioida
    - lisätään luokitukseen poistettu=depricated
    - tila pitää valita jokaiselle termille ja käsitteelle, jos halutaan esimerkiksi sanastosta julkaista vain ns. valmiit käsitteet >> prosessia tukemaan tarvitaan yhteinen luokitus
        - täytyy olla välinetuki prosessiin, ts. sisäänrakennettu editoriin, jotta käyttäjän työmäärää saadaan vähennettyä
- järjestysnumero >> käytetään termin synonyymien järjestämiseen tiettyyn, sanaston käyttöä sopivalla tavalla tukevaan järjestykseen
    - editorin käyttöliittymän tuettava toiminnallisuutta, että synonyymien järjestystä voidaan myöhemmin myös vaihtaa
- lisätään termille kentät:
    - tyyli (tyypillisesti esim. luokitus ”juhlallinen”, ”puhekieli” jne.)
    - kieliopillinen tieto (mm. ruotsin kieliversioissa suvut)
    - lisätieto-kenttä
        - voidaan antaa termiin sidottua lisätietoa vapaana tekstinä; julkaistavaa tietoa termistä
    - maatunnus (mm. kieliversioinnissa eri kielivarianttien osalta, ruotsi ja suomenruotsi) >> voidaan ilmaista kielikoodilla, ei vaadi omaa kenttää; ei pakollinen kertoa vaan tyyppiä voidaan käyttää tarvittaessa
        - huomattava, että tieto voi liittyä kieliversioissa vain esimerkiksi yhteen (kaikista annetuista) vastineista
        - voidaan hyödyntää esim. skos:literalForm -kenttää >> on käyttöliittymäasia mahdollistaa merkitseminen
- keskusteltu lisäksi:
    - lyhenteet: onko tarpeen eritellä synonyymilistassa (ts. tieto mistä on lyhenne)?
        - merkitään synonyymeinä, perässä tarpeen mukaan tyylimerkintä (lyhenne)
    - pysyvät tunnisteet: onko nyt tarpeen luoda sääntö miten URI muodostetaan, sillä käsitteiden yhteiskäyttö lisääntyy jatkuvasti ja siihen myös voimakkaasti kannustetaan?
        - tarvitaan automatisoitu tapa, ei käyttäjän luotavissa
        - varmistettava myös tunnistekäytänteiden yhteentoimivuus, jos käytetään jotain muuta editoria kuin tulevaa Yhteentoimivuuden välineistön versiota

Keskusteltu assosiaatioista:

- käsitteen ja termin välinen suhde
    - käsitteeseen voi liittyä useita termejä ja niiden semantiikka voidaan ilmaista
    - käsitteeseen liittyvä suositeltava termi >> onko tarpeen ilmaista vielä termien välisenä synonymiarakenteena? >> termi-inventaariovaiheessa tarpeellinen (suhteita hahmotellaan ennen varsinaista käsiteanalyysia); kun synonymiasuhteita tunnistetaan, ne liitetään käsitteeseen, irrallisia termejä ei jää

- homonymian ilmaiseminen >> on verraten yleinen ilmiö sanastotyössä, käytettiin esimerkkinä OKSA-sanaston koulutus-käsitettä, joita sanastossa neljä kappaletta
    - terminologisessa sanastotyössä homonyymiset termit esiintyvät käsitteidensä kanssa, ts. käsite ja sen nimitykset muodostavat termitietueen perustan ja termi esiintyy eri käsitteen yhteydessä niin monta kertaa kuin on tarve; termin valintaan vaikuttavat monet tekijät ja sekä valintaprosessi että itse termi on usein sosio-kulttuuris-poliittisesti latautunut
    - SKOSXL:ssä termit ovat alisteisia käsitteille, termit käsitetään merkkijonoina (ks. kuva alla):



    - keskusteltiin siitä vaihtoehdosta, että (homonyyminen) termi luodaan sanastoon vain kerran ja hyödynnetään tätä yhtä termiä viitatessa erisisältöisiin käsitteisiin
        - mallinnustapa tuntuu hieman ”akateemiselta”; huomioitava, että eri käsitteeseen viitatessaan termillä on usein erilaiset kielivastineet
        - editorin käyttöliittymässä olisi pystyttävä selkeästi ilmaisemaan, että esimerkiksi termiin tehtävä muutos vaikuttaa moneen käsitteeseen, miten tämä hallitaan?
        - todettiin, että mallinnustavalla saavutettava hyöty on todennäköisesti pienempi kuin aiheutettavan hämmennyksen määrä; oleellista, että
            - eri sanastoissa eri termit, tuodaan tämä näkyville
            - koneymmärteinen linkitys
            - ihmisymmärteinen merkitsemistapa
        - homonyymin numeroinnin mahdollistaminen (esimerkiksi: koulutus(1) ja koulutus (2))? >> tarvitaan tapa erottaa homonyymit, termin osana
            - Finto/Skosmos mahdollistaa nykyisin määritelmässä (tai huomautuksessa) sisäisen linkityksen, jos hyödynnetään sanastossa määriteltyä käsitettä toisen käsitteen määritelmän osana
- skosxl:altLabel (synonyymi rakenteisena) muutettava, sillä kyseessä voi standardimielessä olla muukin suhde kuin synonyymia
- skosxl:labelRelation >> määriteltävä mitä ovat
        - esimerkiksi tanssi-tanssit >> onko jälkimmäinen kieliopillinen muoto vai tapahtuma?
        - termien välisiä relaatioita? >> lyhenne ja sen aukikirjoitettu muoto?

- käsitteiden väliset suhteet (osin muokattu malliin jo työpajassa)
    - kaikki muut paitsi isPartOf-suhde tuotu malliin SKOS-standardista
    - hiearkkinen ylä/alakäsite – ok
    - vastaava käsite / täysi vastaavuus (exactMatch)
        - ei itse asiassa tarkoita, että ovat samat vaan että käytettävissä toisensa sijaan >> määritelmää tarkastettava
        - yllä oleva on tärkeä huomio, sillä esim. tiedonhakumielessä eivät tuota samaa tulosta
    - liittyvä käsite toisessa sanastossa >> määritelmää tarkennettava
    - liittyvä termi >> liittyvä käsite

        - yksikkö- ja monikkomuotojen hallintaan asiasanastoissa hyödynnetty
        - tarvitaan tarkentamaton vastaavuus >> kun käsitellään samaa ilmiötä, mutta eri näkökulmasta, esim. kasviohuoneilmiö ja ilmastonmuutos; voi olla saman sanaston sisälläkin, esim. tulovero ja tuloverotus
        - sama käsite termien välillä; ei termi käsitteiden välillä
- osittainen vastaavuus >> toisessa sanastossa; liittyy mm. käsitteen käyttöalaan
- termien nimet assosiaatiosuhteissa tarkastettava >> viittaus useimmissa pitäisi olla toiseen sanastoon ja toiseen käsitteeseen termin sijaan
    - SKOSin suhteen ei ole tiukasta sanottu, että olisi/mahdollistaisi viittauksia eri sanastojen välillä; tarkastettava linjaan myös Finton ohjeistuksen kanssa
    - käytännössä vaikuttaisi siltä, että samat suhteet käytössä, kun viitataan saman sanaston sisällä ja eri sanastojen välillä; käyttöliittymätasolla estettävä virheellinen käyttö
- isPartOf >> viittaa osista kokonaisuuteen
    - entä toisin päin? >> ISO25964:2011, kehitetty asiasanastoille, mutta voisi ehkä hyödyntää tässäkin; pakottaa aina valitsemaan hiearkkisen suhteen; pidettiin hyvänä systemaattista toimintatapaa suhteiden määrittelyssä

- edellisen työpajan perusteella lisätty malliin:
    - kieli-luokka >> sanaston kieli
        - Lexvo-urit Fintossa (ISO639-3), voisi hyödyntää myös tässä
        - voi olla monta kieltä per sanasto, ts. monikielinen sanasto
    - lisenssi-luokka (Finton käytänteen mukaisesti)
        - viittaus URI:lla
        - lisenssin nimi tekstinä
        - mistä otetaan >> voi olla ja olisi hyvä olla yhtenäinen lisenssiluokitus taustalla

Tulevasta keskusteltu, että tämä työpajoissa aikaansaatu tietomalli toimii siis Yhteentoimivuuden välineistöön liittyvän ”sanastoeditorin” kehitystyön pohjana. Kehitystyön toteutuksesta vastaa keväällä 2017 CSC YTI-hankkeen ohjauksessa. Työssä hyödynnetään THL:n kehittämää Termieditoria ja THL:n asiantuntija osallistuu kehitysprojektiin. Syntyvä tuotos on avoimen valmistelun periaatteella kiinnostuneiden nähtävillä ja testattavissa; Kansalliskirjaston Finto-projekti on kiinnostunut katselmoimaan tuotosta mahdollisimman pian FINTO-yhteentoimivuuden varmistamiseksi. Työn etenemisestä saa tietoa eri jakelulistojen kautta (erityisesti YSR/KMR) ja viime kädessä tiedustelemalla Remekseltä.

AP:t

- tehdään sovitut korjaukset/muokkaukset/lisäykset malliin työpajamuistioiden mukaisesti
- hyödynnetään malli välineistökehityksen osana; työpajamuistiot myös toteutusprojektin käyttöön
- kehitys avoimella valmistelulla, jotta tuotosta pääsee kommentoimaan jatkossakin
- useita ns. tietomääritysten hallintamalliin liittyviä seikkoja nostettu työpajoissa esiin, Remes vie nämä osaksi YTI-hankkeessa toteutettavaa organisoitumisen valmistelua; sieltä edelleen tarvittaessa muille tahoille