KANSALLISKIRJASTO
NATIONALBIBLIOTEKET

Dublin Core perehdytys

Käytännön soveltaminen Kansalliskirjaston ylläpitämissä julkaisuarkistoissa

Samu Viita

11.3.2022

Dublin Core ja Dspace (1/2)

- Dspace käyttää aineistojensa kuvailuun Dublin Core -formaattia
- Dspacen keskeisin käsite on tietue (item). Tietue koostuu DC-metadatasta ja 0-n:stä tiedostosta.
- Esimerkkejä:
  - https://julkaisut.valtioneuvosto.fi/handle/10024/163835
  - https://taju.uniarts.fi/handle/10024/44839
- Huom!
  - Tietue voi siis sisältää useita tiedostoja, kaikilla on kuitenkin yhteinen DC-kuvailusetti

Dublin Core ja Dspace (2/2)

- Dspace mahdollistaa myös kielimääreen lisäämisen metadatakentälle, esim. dc.description.abstract [**fi**], dc.description.abstract [**swe**]
  - Dspace-syöttölomake ei kuitenkaan tue kielimääreiden syöttöä, vain syöttörajapinnoilla ja tietuetta jälkikäteen editoimalla voi syöttää niitä
  - Tämän vuoksi KK:ssa käytetty “purkkaratkaisua”, eli samaan kenttään kirjoitetaan eri kieliset vastineet ja näitä sitten puretaan käyttöliittymässä ja rajapinnoissa tarpeen mukaan
    - Esim. *fi=suomeksi*|*sv=ruotsiksi*|*en=englanniksi*|

Dublin Coren vahvuudet / heikkoudet (1/2)

+ DC on intuitiivinen ja QDC (qualified dublin core) laajennettava. Tämä alentaa kynnystä kuvailla, mahdollistaa monipuoliset tarpeet ja helpottaa käyttöä mm. rajapintojen kautta
- “Flätti” metadataformaatti. Ei mahdollista helposti sellaisen tiedon kuvaamista, jossa kentillä on riippuvuuksia keskenään
  - Esim: Jos tietueella on kuvailtu sarja kenttään X ja sarjan numero kenttään Y ja nämä tiedot ovat toistuvia, (eli tietueella on kaksi sarjatietoa), sarjan nimet ja numerot menevät sekaisin keskenään:
    - dc.relation.ispartofseries=sarja1, dc.relation.ispartofseries=sarja2
    - dc.relation.numberinseries=1, dc.relation.numberinseries=2
    - Onko sarja1:n numero 1 vai 2?

Dublin Coren vahvuudet / heikkoudet (2/2)

- Flätin ongelmat jatkuu:
  - dc.subject.yso kenttään olisi kiva tallettaa myös ontologian uri
  - Tässä käy kuten sarjan kanssa, useampia asiasanoja sisältävät kentät menevät sekaisin.
  - Ontologia-url:ien tallentamiseen liittyy muitakin haasteita:
    - Miten loppukäyttäjä voi päivittää sanaa niin, että myös ontologian uri päivittyy (tai päinvastoin).
  - Tällä hetkellä siis osittain DC:n rajoituksista johtuen dc.subject.yso-kentässä ei kulje URI mukana.

Erityisroolissa olevat kentät DSpacessa

- Osa kentistä täydennetään automaattisesti järjestelmän toimesta
  - dc.date.accessioned: ISO 8601 sekunnin tarkkuudella, tietueen syötön hetkellä annettu aikaleima, esim. 2022-03-08T15:32:52Z
  - dc.date.available: Ei toimi nimensä mukaisesti. Käytännössä kertoo vain sen, koska embargo-aika on lauennut, eli milloin tiedostot ovat avoimesti saatavilla. Myös sekunnin tarkkuudella kirjattu tieto, kuten accessioned.
  - dc.description.provenance: Toistuva kenttä, johon kertyy tietueeseen kohdistuvia tapahtumia: Luonti, hyväksyminen workflowsta, muokkaus… (kenttä ei ole julkisesti näkyvissä sen sisältämien henkilötietojen ja sähköpostiosoitteiden vuoksi)
  - dc.identifier.uri: Tietueen verkko-osoite, joka on muotoa: *julkaisuarkiston verkko-osoite/handle/handleosuus*
    - Huom! Tämä on mahdollista poistaa manuaalisesti, joten ei täysin luotettava, vaikka kattavuus onkin yli 99%

Erityisroolissa olevat kentät KK:n Dspace-instansseissa (1/2)

- Osan kentistä oletetaan olevan tietyssä formaatissa.
- dc.contributor.author / editor yms. nimikentät
  - Sukunimi, Etunimet
- dc.rights
  - Vahva suositus käyttää 4.0 version CC-lisenssejä, Rights statements yms. sanastoja. Käyttöliittymässä tietyllä tavalla merkityt lisenssit näytetään kuvalinkkeinä, mutta pohjimmiltaan ovat vain merkkijonoja. Esim CC BY 4.0: https://urn.fi/URN:NBN:fi-fe2021092847307
- dc.identifier.urn: URN:-alkuinen merkkijono. Tästä muodostetaan pysyvä tunniste. Käytetään vain jos tietueen ensisijainen säilytyspaikka (omistajuus) ko. julkaisuarkistolla. Muutoin käytetään dc.relation.urn -kenttää

Erityisroolissa olevat kentät KK:n Dspace-instansseissa (2/2)

- dc.relation.url: Linkki ulkoiseen resurssiin, joka näytetään tietueessa klikattavana linkkinä
- dc.subject.yso: Vain kontrolloituja asiasanoja, syöttölomake huomaa tarkenteen ja pakottaa käyttämään ysoa

QDC vs. DCTERMS

- Julkaisuarkistoissa ei käytännössä käytetä Qualified Dublin Coren DCTERMS.nimiavaruuksia tietorakennetasolla, ainoastaan dc.element.qualifier-tyyppistä tallentamista
- DCTERMS-nimiavaruus QDC:n yhteydessä uudempi suositus, mutta vaikeasti laajennettava ja käytettävä. Käytännössä monet toiminnot Dspacessa eivät edes tue DCTERMS-nimiavaruutta.
  - Ovat käytännössä mäppäyksiä dc.element.qualifier -> DCTERMS
  - Tiedot kuitenkin tallennetaan vain muodossa dc.element.qualifier
  - Kansallinen metadatasuositus samoilla linjoilla: Käytetään peruselementtejä ja niiden qualifiereitä ja tarvittaessa laajennokset tehdään qualifiereinä sopiviin elementteihin
- DCTERMS vaatii tietämystä sen säännöistä, mikä vaikeuttaa esim. ei kirjastoalalla olevien tahojen integraatioita ja on käytännössä rajoittava ja epäintuitiivinen sekä lisää kompleksisuutta
- DCTERMS-nimiavaruutta voidaan kuitenkin käyttää mäppäyksillä / crosswalkeilla esim. rajapinnassa, esim. QDC_Finnassa se on käytössä osassa metadataa. Tämä kuitenkin aiheuttaa ongelmia suhteessa metadatasuosituksen tarkkuuteen, kun mäppäykset eivät mene aina välttämättä yksi yhteen tai esim. laventaa tietoa

OAI-PMH ja muut rajapinnat (1/2)

- OAI-PMH pohjautuu Dublin Coren käyttöön. Kansalliskirjastossa on tehty kolme formaattia perus-DSpacen tarjoamien OAI-PMH formaattien lisäksi:
- OAI-formaatit Kansalliskirjaston Dspacessa:
  - DC: Ei hyvä, latistaa qualifierit pois hukaten dataa:
    - https://www.doria.fi/oai/request?verb=ListRecords&metadataPrefix=oai_dc
    - Minimivaatimus OAI-PMH:lta tarjota tämä perus DC-formaatti
  - QDC: Vähän parempi. Ei näytä kaikkia kenttiä, mutta osan qualifiereistä konvertoiden ne dcterms-muotoon (huomaa hämmentävä sekaannus qdc / dcterms välillä, monesti qdc:lla tarkoitetaan DC:tä, joka on laajennettu qualifiereilla, eli tarkentimilla, esim. dc.date.**issued**)
  - https://www.doria.fi/oai/request?verb=ListRecords&metadataPrefix=qdc

OAI-PMH ja muut rajapinnat (1/2)

- OAI-formaatit jatkuu….
  - **qdc_finna**: Finnan tarpeisiin qdc-formaatista räätälöity formaatti. Finna ja jo Nelli käyttivät qdc-formaattia alun alkaen, jonka Dspace tarjoaa oletusasennuksessaan. Kuten qdc, qdc_finnakaan ei näytä kaikki kenttiä, mutta enemmän kuin edeltäjänsä, eikä latista qualifiereitä pois ja tarjoaa myös tiedostolinkkejä sekä valitsee parhaan pysyvän tunnisteen tietueelle
  - **qdc_kk**: qdc_finnasta jalostettu formaatti, joka pilkkoo kieliversiot fi=| -tyyppisesti koodatusta metadatan arvosta erillisissä elementeissään
    - Lisäksi asettaa myös puretun kokotekstin saataville esim. Finnan kokoteksti-indeksointia varten:
  - **kk**-formaatti: Antaa kaikki julkiset metadatat siten kuin ne on tallennettu (dc.element.qualifier, ei dcterms-mäppäyksiä) ja tietueen tiedostot

qdc_finnan mäppäykset tarkemmin syynissä (11.3.2022)

- Esimerkkitietue Tajusta: https://taju.uniarts.fi/handle/10024/7512
- qdc_finna: https://taju.uniarts.fi/oai/request?verb=GetRecord&metadataPrefix=qdc_finna&identifier=oai:taju.uniarts.fi:10024/7512
- qdc_finna näyttää vain osan metadatasta poimien sieltä Finnalle pureksittuja tietoja Finnan pyytämissä kentissä:
  - Title.* -> <dc:title type="*">,
  - dc.contributor.author-><dc:creator>,
  - contributor.* -> <dc:contributor type="*">
  - dc.subject.* -> <dc:subject type="*">
  - dc.description -> <dc:description>
  - dc.type.* -> <dc:type type="*">
  - dc.language.* -> <dc:language type="*">
  - dc.relation.* -> <dc:relation type="*">
  - dc.rights.* -> <dc:rights>
  - dc.bitstream.format ja dc.format.extent -> <dc:format>
  - dc.coverage -> <dc:coverage>
  - dc.publisher.* -> <dc:publisher type="*">
  - dc.source.* -> <dc:source type"*">
  - dc.description.abstract -> <dcterms:abstract>
  - dc.description.tableofcontents -> <dcterms:tableOfContents>
  - dc.date.available -> <dcterms:available>
  - dc.date.issued -> <dcterms:issued>
  - dc.programme -> <qdc_extended:programme>
  - dc.keyword -> <qdc_extended:keyword>
  - dc.identifier.uri/urn-> jos arvo alkaa merkkijonolla urn tai URN -> <kk:permaddress>http://www.urn.fi/urn</kk:permaddress>
  - dc.identifier.doi -> <kk:permaddress>http://dx.doi.org/doi</kk:permaddress>
  - Tiedostot: <kk:file> Attribuutit: bundle, href, name, type, length, sequence, accessrights (jos dc.embargo.lift-kentässä arvoja)

qdc_kk ja kk-formaattien esimerkit

- qdc_kk: Kuten qdc_finna, mutta pilkkoo fi=arvo|sv=arvo|en=arvo| -tyyppisesti koodatut arvot omiin tageihin, esim: https://taju.uniarts.fi/oai/request?verb=GetRecord&metadataPrefix=qdc_kk&identifier=oai:taju.uniarts.fi:10024/7512
- Lisäksi antaa myös kokotekstiaineiston puretun raakatekstin kokoteksti-indeksoinnin tarpeisiin, sekä tiedostojen tarkistussummat
- kk-formaatti: Antaa kaikki julkiset Qualified Dublin Core -metadatat siinä muodossa, kun ne on tallennettu (ilman dc_terms yms. mäppäyksiä: https://taju.uniarts.fi/oai/request?verb=GetRecord&metadataPrefix=kk&identifier=oai:taju.uniarts.fi:10024/7512

Muita syöttötapoja kuin Dspace-syöttölomakkeet

- Rajapintojen kautta
  - Tutkimustietojärjestelmät
  - Opinnäytteiden hallintajärjestelmät
  - Web of science
  - Julkaisujen ja artikkelien toimitusjärjestelmät
- Eräajona
  - SAF- CSV ja REST-ajot, joilla ajetaan esim. vanhaa konvertoitua aineistoa.
- Massa-editoinnit
  - Käytetään erityisesti metadata-uudistusten aikana tai muussa metadatan korjailussa
  - https://www.kiwi.fi/pages/viewpage.action?pageId=45782290
  - CSV-muotoinen metadata on helposti muokattavissa LibreOfficen työkaluilla
  - Rajoituksia: Vain pääkäyttäjät voivat tehdä täysin itsenäisesti. Max 10 000 tietuetta per ajo, käyttöliittymän kautta itsenäisesti ajettuna raja vain 2000

DC:n historiaa ja nykytila KK:n julkaisuarkistoissa (1/2)

- Alunperin julkaisuarkistojen metadata tuotettiin vapaammin sillä periaatteella, että asiakaskirjasto on paras asiantuntija päättämään käytetyt DC-kentät
- Kansalliskirjaston ainoa ohjeistus oli “Kansallinen metadataformaatti opinnäytteille”, jota monet hyödynsivät suunnittelussa
  - https://urn.fi/URN:NBN:fi-fe201303132561
- Asiakkaat suunnittelivat myös konversiomäppäyksiä omille aineistoilleen

DC:n historiaa ja nykytila KK:n julkaisuarkistoissa (2/2)

- Aineistojen kasvun ja metadatan eroavaisuuksien myötä ylläpito ja uusien ominaisuuksien ja rajapintojen hyötykäyttö hankaloitui
- Pitkään tiedostettiin tarve laajemmalle suositukselle (linkki vie asiakaswikiin), ja sellainen onneksi saatiin vihdoin vuonna 2017
  - Helpotti huomattavasti sekä asiakkaita että ylläpitoa ja uusien ominaisuuksien tekoa
  - Metadatauudistuksia on tehty ahkerasti myös takautuvasti, mm. Julkari, Theseus, Valto… Uudet arkistot suunnitellaan alun perinkin noudattamaan suositusta, kuten TAJU, Siirtolaisuusinstituutti yms…
- Tilanne siis nykyään jo paljon parempi, mutta joitain kompromisseja on jouduttu tekemään varsinkin vanhemmalle aineistolle
  - Näissäkin tapauksissa oleellisimmat puutteet on pyritty korjaamaan ja konversion pohjana käytetään syöttölomakkeiden määrityksiä mahdollisimman pitkälle, jotta vanha ja uusi data olisi yhteensopivaa

Tulevaisuus

- Nykyisin käytössä Dspacen version 5, työ alkamassa Dspace 7:n osalta.
- Seiskaversio mahdollistaa mm. entiteettien määrittämiset:
  - Suhteet entiteettien välillä mahdollista kuvata, samoin entiteeteillä voi olla omat DC-metadatat
  - Ratkaiseeko tämä esim. YSO-asiasanoituksen URL:ine koodaamiset, emojulkaisun yms. suhteiden kuvaamiset?
  - Miten OAI-PMH ja muut muiden järjestelmien käyttämät rajapinnat taipuvat näihin muutoksiin?
  - Onko tarvetta avata uusi Rest-rajapinta esim. Finnalle?
  - Paljon kysymyksiä, vielä vähän vastauksia, työ alussa!
  - Iso päivitys, lähes uusi ohjelmisto

Kiitos!