Julkaisuarkistojen metadatasuosituksen hahmottelu

Tanja Vienonen, tietoasiantuntija Tajua-hanke

Tajua-asiantuntijaverkoston tapaaminen
16.11.2016

Suosituksen lähtökohtia (1)

- Tavoitteena luoda suositus suomalaisille julkaisuarkistoille soveltuvaksi sisäiseksi metadataformaatiksi
  - Soveltuu etenkin DSpace-arkistoille (mutta auttaa myös muita)
- Formaatti ei sinällään pyri olemaan suoraan yhteensopiva kaikkien kansainvälisten tai kansallisten suositusten kanssa
  - Suosituksen mukaisen metadatan pitäisi kuitenkin olla niin rikasta ja tarkasti eroteltua, että sen pohjalta voidaan tuottaa keskeiset rajapintojen kautta ulospäin välitettävät formaatit
  - Huomioon otetaan ainakin unqualified DC, OpenAIRE, Finna ja OKM:n tiedonkeruu
  - Sisäisen formaatin kenttien ei tarvitse välttämättä olla tiukasti standardien mukaisia, kunhan ne ovat loogisia ja niitä käytetään yhdenmukaisesti
  - Yhtenäistäminen koskee kaikkea aineistoa, myös jo tallennettua

Suosituksen lähtökohtia (2)

- Dublin Coressa rajoituksensa, MARCin kaltaiset bibliografiset tiedot eivät mahdollisia
  - Pyrkimyksenä yksi asia/kenttä, esim. numerot ja nimet erikseen
  - Sanastoista tulevat tiedot omiin alakenttiinsä
- dc. vs. dcterms-nimiavaruudet
  - dctermsin käyttö ollut pitkään suositeltua
  - Miten laajasti metadatan aggregoijat valmiita dctermsiin?
  - Sisäisessä formaatissa ei kuitenkaan välttämätöntä
- dc.*-nimiavaruus vs. paikalliset nimiavaruudet (esim. local.*)
  - dc.*:n venyttäminen kaikkiin paikallisiin tarpeisiin ei mielekästä

Suosituksen tilanne tällä hetkellä

- Tässä esityksessä suosituksen tämänhetkiset kentät
  - Kenttiä ja niiden sisältöjä käyty läpi noin puolessa tusinassa kokouksessa
  - Eivät vielä lopullisessa muodossaan
  - Laajempi GoogleDocs-dokumentti kommentteineen ja pohdintoineen saatavilla verkossa: http://bit.ly/2fP6e21
- Kenttiä on tällä hetkellä 53, joista murto-osa ns. helppoja
  - Ensimmäiset kentät käydään kursorisesti läpi, koska tietojen esittäminen jo melko vakiintunutta
  - Eniten päänvaivaa aiheuttavat käyttö- ja pääsyoikeuskentät, koska eivät ole käytössä kovin laajalti
  - Haasteita liittyy myös mm. OKM:n julkaisutyyppeihin

Vakiintuneet kentät

- Otsikko: dc.title
- Vaihtoehtoinen otsikko: dc.title.alternative
- Kenttien sisällön esittäminen vakiintunut muotoon Otsikko : alaotsikko

- Tekijä: dc.contributor.author / dc.creator (DSpace varannut kentän järjestelmän sisäiseen käyttöön)
- Toimittaja: dc.contributor.editor
- Nimi muodossa Sukunimi, Etunimi

- Julkaisija (kustantaja): dc.publisher
- Julkaisijan nimen vakiintuneen muodon voi tarkistaa esim. Julkaisufoorumista, Sherpa/RoMEOsta tai Ulrichswebistä

Vakiintuneet kentät jatkuu

- Julkaisu/ilmestymisvuosi: dc.date.issued
  - ISO 8601-standardin mukaisesti VVVV-KK-PP, pelkkä vuosi riittää.

- URN-tunnus: dc.identifier.urn
- DOI-tunnus: dc.identifier.doi
- Pelkkä tunnus vai koko osoite (http:// tai https://)?

- Tiivistelmä: dc.description.abstract
- Huomautukset: dc.description.notification
- Molemmat vapaatekstikenttiä, muualta kopiointi UTF-8 mukaisesti

Julkaisuun liittyvät verkko-osoitteet

- Monessa julkaisuarkistossa dc.identifier.uri käytössä
  - Kentän ei pitäisi olla dc.identifier-alkuinen, koska kyseessä on vain linkki, ei varsinainen identifier
  - Osoitteet eivät ole pysyviä vaan voivat mennä ajan kuluessa rikki / johtaa jonnekin aivan muualle, siksi jokin muu kuin identifier-kenttä parempi
  - DSpace tallentaa .uri-kenttään automaattisesti tietueen verkko-osoitteen; joissain organisaatioissa URN-tunnus tässä kentässä
- Tarvitaan ns. kaatokenttä verkko-osoitteille, esim. dc.relation.url
  - Linkit rinnakkaistallenteiden julkaistuihin versioihin ja muihin tietueeseen liittyviin verkkodokumentteihin

Kielikoodit ja asiasanat

- Kielikoodi: dc.language.iso
  - Suositellaan ISO 639-x standardin mukaista muotoa, joka on julkaisuarkistotasolla yhtenäinen
  - Toivotaan siirryttävän 3-merkkiseen koodiin, 2-merkkinen ei pian enää suositeltava rajallisen kielivalikoiman vuoksi
  - ISO 639-2 Kongressin kirjaston ylläpitämä suppeampi 3-merkkinen kielikoodisto, koodeja ~500
  - ISO 639-3 vapaammin laajeneva 3-merkkinen standardi, kielikoodeja ~7700
- Asiasanat: dc.subject
  - Vapaamuotoiset avainsanat ilman tarkennetta dc.subject-kenttään?
  - DCMI suosittelee kontrolloitujen asiasanastojen käyttöä
  - Asiasanat dc.subject.[asiasanaston lyhenne], esim. dc.subject.yso

Affiliaatiot / teokseen liittyvät organisaatiot

- Muut tekijät: dc.contributor.? Mikä olisi kentän tarkenne?
  - Halutaanko määritellä tarkat kentät muillekin kuin tekijöille (author) ja toimittajille (editor), mitä ne olisivat?
    - Väitöskirjoissa ohjaajat, vastaväittäjät, kustokset?
  - Muita, esim. kääntäjät, kuvittajat, säveltäjät (Sibelius-akatemia)?
- Artikkelin affiliaatiotieto: dc.contributor.? dc.relation.affiliation?, dc.relation.organization?
  - DC:ssa ei pysty yhdistämään affiliaatiota tekijään, mutta DCMI ehdottaa, että affiliaatio koskisi tallennettavaa artikkelia
    - Artikkelin affiliaatio ei muutu, vaikka tutkija vaihtaisi tutkimuslaitosta
    - http://dublincore.org/documents/dc-citation-guidelines/ (kohta 5.1.)
- Opinnäytteen tekopaikka samaan vai omaan kenttäänsä?
  - Voidaanko pitää affiliaationa?
  - Opinnäytteen vai tutkinnon hyväksynyt organisaatio?

Sivut ja tiedoston koko

- Format = fyysinen olomuoto/laajuus, description = aineiston esittely (abstraktit, sisällysluettelot)
- Sivumäärä: dc.format.extent
  - Numerokenttä, sivujen kokonaislukumäärä
- Sivunumerot: dc.format.pagerange
  - Numerokenttä, artikkelin sivunumerot emojulkaisussa
  - Sivumäärä ja sivunumerot hyvä erottaa omiin kenttiinsä
- Tiedoston koko: dc.format.size
  - Tiedoston koko esim. megatavuissa
- OpenAIREssa dc.format suositeltu, ei pakollinen kenttä

Julkaisupaikka ja -maa

- Julkaisupaikka: dc.publisher.place
  - Julkaisun kustantajan kotipaikka
- Julkaisumaa: dc.publisher.country
  - Julkaisun kustantajan kotimaa
  - Suositellaanko ISO-3166:n mukaisten maakoodien käyttöä?
  - Vai olisiko maan nimi datan käytön kannalta hyödyllisempi?
  - OKM:n tiedonkeruuohjeet: Tilastokeskuksen valtiot ja maat 2012 -luokituksen mukainen 3-numeroinen arvo, esim. 246. https://www.tilastokeskus.fi/meta/luokitukset/valtio/001-2012/index.html

Aineistotyypit (1)

- Suositellaan tarkenteiden käyttöä, sanastot eroteltava toisistaan tarpeiden mukaan
- DCMI aineistotyyppi: dc.type.dcmi
  - Lista aineistotyypeistä: http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=dcmitype#H7
- OKM aineistotyyppi: dc.type.okm
  - OKM:n tiedonkeruuohjeiden mukaan; suositellaan julkaisulajien mukaisten kirjain-numerolyhenteiden käyttöä
  - Ihanteellisinta, jos tieto tulisi julkaisuarkistoon suoraan tutkimustietojärjestelmästä (jos käytössä)

Aineistotyypit (2)

- OpenAIRE aineistotyyppi: dc.type.publication
  - OpenAIREn mukaiset aineistotyypit, pakollinen tieto
  - https://guidelines.openaire.eu/en/latest/literature/field_publicationtype.html
- OpenAIRE aineistoversio: dc.type.version
  - Suositeltu, ei pakollinen tieto
  - https://guidelines.openaire.eu/en/latest/literature/field_publicationversion.html
  - Horizon 2020 -suositus: https://www.kiwi.fi/x/vgIoB

Monografiat ja kokoomateokset

- Monografian verkkoversion ISBN: dc.identifier.isbn
  - Vakiintunut käyttöön, monografian URN muodostetaan tästä
  - Aiheuttaisiko sekaannusta, jos samaa kenttää käytettäisiin kokoomateoksen ISBN:n merkintään?
- Painetun monografian ISBN: dc.relation.isversionof
  - Ehdotus käytettäväksi kentäksi, käytössä esim. Valtossa
- Emojulkaisu: dc.relation.ispartof
  - Artikkelin emojulkaisun eli kokoomateoksen nimi
- Emojulkaisun ISBN: dc.relation.isbn
  - Ehdotus käytettäväksi kentäksi, jos ei voi käyttää samoja kenttiä monografioiden kanssa

Julkaisujen bibliografiset tiedot

- Viittausohje / alkuperäinen ilmestymiskanava: dc.identifier.citation
  - Lähdeviite, tiedot alkuperäisestä kirjasta tai lehdestä, jossa julkaisu on ilmestynyt
  - Sisällön formaattia ei ole määritelty DC-ohjeistuksessa
- Yleisesti käytössä, muotoilu vaihtelee
  - Vapaasanakenttä tai automaattisesti tietueen muista kentistä generoituva
- Miten muotoillaan suosituksessa?
  - Tieteenaloilla omat tapansa merkitä lähdeviitteet
- Käsin syöttäminen ei suotavaa, sillä virheet todennäköisiä, esim. viitteiden siirtäminen RefWorksiin vaikeaa

Sarjatiedot

- Sarjatieto, nimi: dc.relation.ispartofseries
  - Käytössä joissain julkaisuarkistoissa (esim. Valto, Lauda, Jyx)
  - Vapaatekstikenttä sarjan ja lehden nimelle? Järjestysnumero eri kenttään
- Sarjatieto, järjestysnumero: dc.relation.numberinseries
  - Tässä muuttuisi käytännössä vain prepositio of -> in
  - Sarjan ja lehden numero samassa?
  - Nimi ja numero hyvä erottaa omiin kenttiinsä yhdenmukaisuuden parantamiseksi
- Sarjan/lehden ISSN: dc.relation.issn
  - Numerokenttä, pelkkä ISSN
  - Relation-kenttä, koska koko sarjan/lehden tunniste, ei yksittäisen artikkelin
  - Identifier-kenttä ei suositeltava, vaikka paljon käytetty

Lehden tiedot

- Onko tarpeen eritellä sarjat ja lehdet toisistaan?
- Lehden nimi: dc.relation.ispartofjournal?
  - Vai mieluummin lehtien ja sarjojen yhteinen kenttä: dc.relation.ispartofseries?
  - Tarkista lehden vakiintunut nimi
- Lehden numero: dc.relation.issue
  - Yhteinen kenttä: dc.relation.numberinseries
  - Numerokenttä, vuosikerta erikseen (konversiot)
- Lehden vuosikerta: dc.relation.volume
  - Numerokenttä
  - Roomalainen numerointi?

Tekijän/käyttöoikeudet

- Tekijän/käyttöoikeustiedot: dc.rights / dc.rights.rightscode (ehdotus)
  - Tallennetun julkaisun käyttöoikeudet/lisenssi
  - Kenttään kirjataan lyhenne, jonka muoto on sovittava, esim. CC BY-NC-ND (lisenssin ehdot erotetaan toisistaan tavuviivoilla)
  - Onko pudotusvalikko (CC-lisenssit ja All rights reserved) mahdollinen toteuttaa?
  - Europeanan ja DPLAn Rights Statementsit eivät välttämättä haravoitavissa
    - http://rightsstatements.org/page/1.0/?language=en
  - Europeanalla ratkaisu ”Tekijänoikeudet voimassa - vapaa pääsy” -dilemmaan (RR-F): http://www.europeana.eu/portal/en/rights/rr-f.html
- Nyt yleisimmin käytössä oleva kenttä: dc.rights.accessrights, mikä on hiukan harhaanjohtava tarkenne
  - Kyse julkaisun käytöstä, ei vain pääsystä
  - Voisiko käyttää muiden pääsyoikeustietojen esittämiseen? Ehdotus myöhemmässä diassa

Verkko-osoite tekijän/käyttöoikeussivulle

- Kenttä: dc.rights.uri
  - Tarvitaanko? DSpacen käyttöliittymässä tietueen esittelysivulle voi lisätä automaattisesti linkin CC-lisenssisivulle
- Kenttä: dc.rights.license
  - KUMEAn suositus (jossa tosin virheitä): https://www.kiwi.fi/x/i4VDAg
- Linkki-kenttiä pohditaan edelleen, riittävän hyvän ratkaisun löytäminen haastavaa
- Miten otetaan huomioon linkkien rikkoutuminen tai vanheneminen?

Pääsyoikeudet

- Pääsyoikeustiedot: dc.rights.accessrights
  - Muu selite (kuin lisenssi/käyttöoikeus) julkaisun pääsyoikeudesta
  - Esim. ”Käyttö kirjaston tiloissa.” tai Elektra-aineiston pääsyoikeudet
- OpenAIRE-pääsyoikeustiedot: dc.rights.accesslevel
  - Kentässä käytettävät termit: https://guidelines.openaire.eu/en/latest/literature/field_accesslevel.html
  - Ei-pakollinen license condition: https://guidelines.openaire.eu/en/latest/literature/field_licensecondition.html

Vertaisarviointi ja tietueen sisältö

- Vertaisarviointi (ehdotukset): dc.description.version / dc.type.version
  - Vertaisarvioitu julkaisu vs. vertaisarvioitu versio julkaisusta?
  - Kenttä dc.description.version käytössä: pre-print, post-print / final draft, publisher’s version?
- Tietueen sisältö (ehdotus): dc.format.hasdescribedcontent
  - Miten erotetaan pelkän metadatan/tiivistelmän sisältävät tietueet kokotekstitietueista?
  - Kokotekstien tilastointia vaikeuttavat pelkän tiivistelmän sisältävät dc.format.mimetype: application/pdf -tietueet
  - Valikkovaihtoehdot:
    - Fulltext
    - MetadataOnly
    - AbstractOnly
  - Toinen vaihtoehto (jos pelkkä kokoteksti):
    - True/false / kyllä/ei

ORCID, ISNI ja EU-projektitunnus

- Julkaisuun liittyvien henkilöiden pysyvät tunnisteet: dc.contributor.tunnisteen-lyhenne?
  - Tekijöiden ORCID- tai muu tunniste, kuten ISNI
  - Esim. dc.contributor.orcid
  - DSpacen oletustoteutuksessa ORCID-tunnukset authority cachessa, eivät tietueessa
  - ORCID-tunnisteen oikeaoppinen lisääminen vaatii autentikoinnin ORCID-rajapinnan kautta, mikä aiheuttanee lisätyötä ylläpitäjille
  - Vaikea kohta, koska käytöstä ei ole kokemusta
- EU-projektit: dc.relation.projectid
  - OpenAIREn vaatima tunnus
  - syntax: info:eu-repo/grantAgreement/Funder/FundingProgram/ProjectID/[Jurisdiction]/[ProjectName]/[ProjectAcronym]

Kiitos! Onko kysymyksiä?

Lisätietoja:
Tanja Vienonen
tanja.vienonen@helsinki.fi