Ajankohtaisia standardiasioita

Juha Hakala
Kansalliskirjasto
2018-01-16

Sisältö

- Yleistä standardointivastuista yms.
- Luettelointiin vaikuttavien ISO:n standardisointihankkeiden nykytilanne
  - ISO 8601: Päivämäärän ja ajan esittäminen
  - Tunnisteet: ISBN- ja ISSN-standardien uudistaminen
  - Kielikoodit: ISO 639:n modernisointi
    - IETF:n Best Current Practice 47:n soveltaminen kuvailussa

Kansalliskirjaston standardointivastuista

- ISO & SFS Tietohuolto
  - Kansalliskirjasto koordinoi Tietohuolto-ryhmän toimintaa.
    - https://www.kiwi.fi/display/tietohuolto/Tietohuoltokomitea
  - Komitean toimiala on ISO:n Technical Committee 46 (Information and documentation), joka kehittää standardeja esim. kustannusalalle, kirjastoille ja museoille
- Muut kv. standardointijärjestöt
  - Muistiorganisaatioiden kannalta tärkeitä ovat esim. IFLA, ANSI/NISO, IETF, W3C
  - Näiden toimintaan osallistutaan resurssien asettamissa rajoissa ja tarpeen mukaan
- Standardien soveltamista koskeva ohjeistus
  - KDK-hanke, Kuvailustandardiryhmä, KUMEA,…
  - Keskeistä hallinnollisen metadatan (pitkäaikaissäilytys) kannalta

ISO 8601:n päivitys ja MARC 21

- Standardin nykyinen versio on vuodelta 2004; päivitys edennyt DIS-vaiheeseen (Part 1: Basic rules, Part 2: Extensions
  - https://www.iso.org/iso-8601-date-and-time-format.html
- Standardista on puuttunut kirjastojen kannalta tärkeitä ominaisuuksia, jotka lisättiin Kongressin kirjaston EDTF-profiiliin
  - http://www.loc.gov/standards/datetime/pre-submission.html
  - Profiilin tarjoamat lisäpiirteet löytyvät uudesta ISO-standardista
- Uusi ISO 8601 ja MARC-käytänteet eroavat toisistaan
  - Periodien merkintätapa: 1980/1984, ei 1980-1984 kuten 260 $c:ssä
    - Kuukaudet + vuodenajat aina koodeina: 1993-05, ei June 1993-
  - Extensions-ominaisuudet ja kenttä 263 (Arvioitu julkaisuaika)
    - 2018?, 2018~ ja 2018%: uncertain, approximate, uncertain & approximate
    - 201X unspecified

ISO 8601:n päivitys ja MARC 21

- Jotkin MARC-formaatin kentät ovat vanhentuneita
  - Kentän 045 (Ajankohta tai ajanjakso) viimeisin päivitys on vuodelta 1987, ajalta ennen ISO 8601 –standardia
  - -> ei ihme, ettei 045 ole yhteensopiva standardin kanssa
  - Sama ongelma on kuitenkin myös paljon tuoreemmassa 033-kentässä vuodelta 2010
- Eri MARC-kentissä on erilaisia tapoja aikamääreiden tallentamiseen
  - 045:ssä ajanjakson alku ja loppu tallennetaan eri osakenttiin, ja se sallii myös formaatin oman ajanjaksokoodin käytön
    - 045 $ay-y- (21. vuosisata)
  - Vastaavia piirteitä voisi toteuttaa yhteismitallisesti ISO 8601:n avulla

Pohdittavaksi

- MARC 21 -ohjeistus ajanjaksojen ja arvioidun julkaisuajan etc. tallennuksesta poikkeaa ISO 8601:n uuden version antamista linjauksista
- RDA on pääosin MARCin linjoilla
- Formaatin suomennoksesta puuttuu joidenkin kenttien käyttöä koskevia ohjeita (esim. 263), joiden lisäämistä voisi harkita
- Formaatin soveltamistapaa ei ole syytä muuttaa standardin mukaiseksi ennen kuin kansainvälinen ohjeistus päivittyy
  - … mutta jos Webissä sovelletaan ISO 8601:stä, meillä on ongelmia avoimen linkitetyn datan tuottamisessa, ja Web-sovelluksilla on haasteita metatietojemme käyttämisessä

ISO 639:n päivitys

- Kielikoodistandardista vastaavat yhteistyössä ISO TC 46/SC 4 (Technical Interoperability) ja ISO TC 37/SC 2 (Terminology workflow and language coding)
  - ISO 639-1: 2 merkin mittaiset kielikoodit (jäädytetty)
  - ISO 639-2: 3 merkin mittaiset koodit, suppea lista (noin 500)
    - https://www.loc.gov/standards/iso639-2/php/code_list.php
  - ISO 639-3 3 merkin mittaiset koodit, kattava lista (noin 7500)
  - ISO 639-4 General principles of coding; päivitettävänä
- ISO 639-2:n ylläpitäjä on Kongressin kirjasto
  - Uusista koodeista päättää ISO 639/Joint Advisory Committee
    - Montenegrolle hyväksyttiin koodi loppuvuodesta 2017
    - Periaatteessa koodit annetaan kielitieteellisin perustein, käytännössä politiikka sotkeutuu asiaan

Best Current Practice (BCP) 47

- Internet Engineering Task Forcen suositus kielikoodin tallentamiseksi; laajalti käytössä Webissä
- Kieli ilmaistaan ISO 639-1:n mukaisella kahden kirjaimen koodilla jos sellainen on olemassa; lisäksi voidaan kuvata kirjaimisto ISO 15924 –koodilla sekä alue jolla kieltä käytetään ISO 3166:n maakoodilla tai muulla tavoin
  - Amerikanenglanti: en-US
  - Jiddis latinalaisin merkein: yi-Latn
  - Skotlannin murre foneettisin aakkosin: en-scotland-fonipa
- BCP 47:n mukaista koodausta voisi jo nyt käyttää luetteloinnissa
- Kentässä 041 yksi sallituista koodilähteistä $2:ssa on RFC 5646. Se on BCP 47:n se osa, joka määrittelee koodit; toinen osa (RFC 4647) kuvaa vain koodien tulkintaa

BCP 47:n soveltamisesta

- BCP 47:n soveltaminen helpottaisi avoimen, Web-yhteensopivan datan tuottamista
- Erityinen lisäarvo kirjaimiston ilmaiseminen ISO 15924 –standardin mukaisesti
  - MARC-formaatin nykyinen käytäntö kirjaimiston ilmaisemiseen 066:ssa (koko tietue) ja 880 $6:ssa (ko. kenttä) on epästandardi ja kattaa kuusi kirjaimistoa; Unicodeen niitä sisältyy kymmeniä
    - http://unicode.org/iso15924/iso15924-codes.html
  - OCLC sallii WorldCatissa kirjaimiston ISO 15924-koodin tallentamisen sekä 066- että 880-kenttään, mutta ei ole selvää osaavatko muut kirjastojärjestelmät käsitellä tätä dataa oikein
- MARC-formaattiin tulisi saada yleinen, standardeihin perustuva ratkaisu kirjaimiston esittämiseen

BCP 47:n soveltamisesta (2)

- BCP 47:n haasteena on sen laajuus sekä se, että koodi perustuu kahden merkin mittaiseen koodiin
- Olemassa olevasta 008-koodista voisi useimmiten generoida koneellisesti BCP 47 –koodin kenttään 041, mutta:
  - Pitäisikö suomenruotsi koodata se-FI, erotukseksi riikinruotsista?
  - Entä muut kansalliset versiot, kuten Amerikan englanti ja Englannin englanti?
  - Milloin on tarpeen ilmoittaa kirjaimisto?
  - Miten toimitaan silloin, jos kirjaimisto ei sisälly MARC-formaattiin?

Sanastot ja tunnisteet

- ISO 5127:2017 Information and documentation – Foundation and vocabulary
  - Uudistettu ja laajennettu sanasto, sisältää esim. tutkimusdataan liittyvää terminologiaa, joka omaksuttiin osin Research Data Alliancelta
  - Suhde RDA-kuvailusääntöihin?
  - Standardin suomentamisesta neuvotellaan
- ISNI (International Standard Name Identifier) sekä ORCID (Open Researcher and Contributor ID)
  - Periaatteessa ISNIä voidaan käyttää kuvailussa, mutta harkinta tarpeen koska tunnuksessa voi olla ongelmia (kahden tai useamman toimijan tiedot yhdistyneet)
  - ORCID-tunnusta on vaikea yhdistää oikeaan toimijaan, jos metatiedot ovat puutteelliset (esim. pelkkä etunimi)

Lopuksi

- Perinteisiin standardeihin perustuvan datan muuntaminen avoimeksi linkitetyksi dataksi on ylipäätään työlästä
  - Osma Suomisen tuotantoketju avoimelle linkitetylle datalle: MARC 21 -> BIBFRAME -> Schema.org
- MARC-formaatin vaihtelevat ja osin epästandardit käytänteet kieliä ja aikamääreitä koskevan metatiedon tallennuksessa luovat lisähaasteita ja rajoittavat kirjastojen tuottaman avoimen linkitetyn datan käyttökelpoisuutta verkossa
- MARC-formaatin ja kuvailun periaatteiden muuttaminen Web-kelpoiseksi kertarysäyksellä on mahdotonta, mutta vähittäinen muutos kohden BCP 47:n soveltamista ja parempaa ISO 8601-tukea pitäisi olla toteutettavissa