

Nimitietopalvelun tietomalli

KAM-kuvailuryhmän ja Toimijakuvailuverkoston
yhteiskokous 20.11.2020

Jarmo Saarikko <https://orcid.org/0000-0002-6801-6151>

Okko Vainonen <https://orcid.org/0000-0002-2662-7333>

Taustaa

- **Nimitietopalvelu** on aineistoihin liittyvien toimijoiden kuvailutietoja sisältävä metatietopalvelu
 - Luonnolliset ja juridiset henkilöt, suvut, yhteisöt, (tapahtumat)
- Valmisteltu yhteistyössä jo vuosia
 - KAM-kuvailuryhmä: Tietomalli 2015-2019 ([Word dokumentti](#))
 - KDK:n tietoarkkitehtuuriryhmä
 - Selvitys ja jatkoselvitys ja tietomalli julkaistiin 2016
 - <https://www.digime.fi/yhteentoimivuus/nimitieto/>
- Projektin 2020
 - <https://www.kiwi.fi/display/TD/Nimitietopalvelu>

Jatkoselvityksen ehdotuksia

- KDK:n tietoarkkitehtuuriryhmän jatkoselvitys nimitietopalvelun toteuttamismahdollisuuksista 4.12.2017
- Tietomallin ja sektoreiden nykyisten kuvailukäytäntöjen yksityiskohtaisempi analyysi
 - Keskeisin ero on se, että kirjastot kuvailevat kaikki toimijan julkiset identiteetit erikseen, kun taas arkistot ja museot kuvailevat ne yhtenä entiteettinä (tietueena)
 - Myös nimenmuutosten käsittelyssä on eroja
- Nimitietopalvelun tietomalliluonnoksen ja sektoreiden nykyisin soveltamien metadataformaattien yhteismitallisuuden analyysi
 - Vaihtoehdot ovat MARC 21 Auktoriteettitiedot ja EAC-CPF
 - Pysyvät tunnisteet ja tietueiden linkitys
 - Semanttinen yhteentoimivuus

Lähteenä: Juha Hakala 19.5.2016



Nimitietopalvelu - projekti 2020

- Tietomallin edelleen käsittely
 - Tuotettu Excel-taulukko sekä tietomalli yhteentoimivuusalustalla
 - Linkitettyyn dataan siirtyminen edellyttäisi tunnisteiden käyttöä kaikenlaisille entiteeteille
- KAM-sektorin järjestelmien kuvailutietojen mappaus
 - Tietomalliin mapattu arkistojen Ahaa, museoiden MuseumPlus sekä kirjastojen Asterin toimijakuvailukentät.
- Sopivan tuotantoalustan testaus
 - Kansainvälisten esimerkkien mukaisesti lähdettiin testaamaan Wikibase-sovellusta
 - Avoimien tietojen jakelu linkitettyinä datana Finton kautta

Tietomalli

- Tietoelementit pohjautuvat pääosin [RDA Registryyn](#)
 - Tietoelementit linkitetään eri sektoreiden toimijakuvailujen tietomalleihin (RDA, EAC-CPF, CIDOD-CRM, ISAAR-CPF, EU Core, GNDO)
- Tietoelementeillä neljä tasoa
 - Ydintaso – kaikille pakolliset tietoelementit, (RDA ydinelementit)
 - Yhteinen taso – kaikkien käyttämät elementit
 - Jaettu taso – sektorikohtaiset elementit tarjolla kaikille
 - Sektoritaso – käytössä vain sektorin sisällä

RDA-luokat ja propertyt

- RDA Classes
 - Toimija, Henkilö, Suku, Yhteisö, Nomen, Ajanjakso, Paikka
 - Entiteetti, Teos, Ekspressio, Manifestaatio
 - NTP: Tunniste (Nomen), Tapahtuma (Yhteisö)
- RDA Properties
 - Kaikki tietomallin ominaisuudet on pyritty mappamaan vastaaviin RDA-ominaisuuksiin

Muut ominaisuudet

- Mikäli vastaavaa RDA-ominaisuutta ei löytynyt, pyrittiin etsimään sellainen jostakin muusta julkaistusta tietomallista
- Osa tietomallin ominaisuuksista on sekä
 - attribuuttina (merkkijono, owl:dataTypeProperty) että
 - assosiaationa (viittaus kohteeseen, owl:ObjectTypeProperty)
- Kaavio: kts. projektin sivuilla tietomallin liitteenä

Kenttien mappaukset

- Eri formaateissa käytettyjä elementtejä mapattiin nimitietopalvelun yhteisen tietomallin ominaisuuksiin yhteistyössä museoiden ja arkistojen asiantuntijoiden kanssa.
 - AHAA - RDA, eac-cpf (arkistot)
 - MuseumPlus (museot)
 - Asteri - Marc21 (kirjastot)
- Testiaineisto koottiin satunnaisesti valittujen yhteisöjen ja ennen 1900-lukua syntyneiden henkilöiden kuvailutiedoista. Tämän jälkeen laadittiin tiedontuontiprosessi ja testattiin aineiston latausta Wikibase-testiympäristössä.
 - Tietomallin propertyt luotu Wikibaseen
 - Taulukkomuotoinen aineisto ladattiin vastaaviin propertyihin

Tietomallin haasteita: ajan esitysmuoto

- Eri sektoreilla on vaihtelevat merkintätavat ajan ja ajanjaksojen kuvailussa
- Eroja epämääräisten ja epätarkkojen ajankohtien merkitsemisessä ajan tai ajanjakson tietotyyppiin.
 - Yhteinen toimintaympäristö edellyttäisi eri sektoreiden ajan merkitsemistapojen ja käytäntöjen yhdenmukaistamista
 - On sovittu, että ajanjaksojen merkitsemisessä käytetään ISO 8601 –standardia
 - Konvertoimalla kuvailutiedot jossakin vaiheessa standardoituun ajan esittämistapaan saavutetaan nykyistä huomattavasti tehokkaampi yhteentoimivuus ajan esittämisessä ja aikaan perustuvissa hauissa.

Tietomallin haasteita: nimimuodot

- Eri formaateissa nimet merkitään eri tavalla:
 - yhdessä formaatissa etu- ja sukunimet tallennetaan erillisiin kenttiin
 - toisessa ne merkitään samaan kenttään välimerkeillä erottaen
 - Yksiosaisten nimien kuvailutietoja pitää ensin analysoida, jotta voidaan määritellä, onko kyseessä etunimi vai pelkkä sukunimi
- Mikä on henkilön ensisijainen nimenmuoto?
- Miten varianttiniimen tyyppi merkitään?
- Luodaanko eri nimimuodot kohteina, joilla on oma tunnisteensa vai pelkästään merkkijonoina?

Tietomallin haasteita: identiteetit

- Todellinen identiteetti vs. toinen identiteetti
- Arkistoissa ja museoissa kaikki samaan henkilöön liittyvät nimet ja nimenmuodot on koottu yhteen kuvailuun
- Kirjastoissa on tapana erottaa todelliset ja toiset identiteetit (esimerkiksi pseudonyymit) omiksi kuvailuikseen, joilla on omat tunnisteensa
- Yhteisen tuotantoalustan on hallittava nämä erilaisuudet ja pystyttävä kokoamaan ja palauttamaan tietoja takaisin sektorien omiin järjestelmiin luotettavasti kunkin sektorin vakiintuneiden käytäntöjen mukaisesti.

Yhteentoimivuusalusta

- Tietomalli-työkalulla voidaan hyödyntää kansallisia yhteiseen käyttöön laadittuja elementtejä tai myös muiden sektoreiden tietomalleja.
- Kulttuurialan tietokomponenttimalli
 - https://tietomallit.suomi.fi/model/ntp_v2/
 - Elementit muiden sovellusprofiilien käytettävissä

Y-alustan tietomallit-työkalu

- Työkalun tallennuslomakkeella on rajallinen määrä elementtejä käytettävissä. Joitakin tarpeellisia elementtejä puuttui ja osaa ei tarvita tässä vaiheessa.
- Linkitys vain alustalla julkaistuihin sanastoihin ja koodistoihin
- Viittaukset ulkopuolisiin ominaisuuksiin rajoittuu vain owl:equivalentProperty sekä rdfs:subPropertyOf
- Käyttää luokan kuvauksessa ominaisuuksille shacl:ia arvojen validointiin
 - mm. sh:disjoint, sh>equals, sh:minCount, sh:maxCount, jne.
 - <https://www.w3.org/TR/shacl/> Shapes Constraint Language
- Kaaviokuvan piirrostyökalu ei näytä luokkien välisiä suhteita

Wikibase

- Kansainvälinen yhteistyö KAM-sektoreilla
 - Saksan ja Ranskan kansalliskirjastot sekä OCLC Yhdysvalloissa pyrkivät kaikki luomaan GLAM-sektoreille yhteistä toimijakuvailupalvelua ja ovat testanneet Wikibasea
 - Kokemukset ovat olleet positiivisia
 - Wikimedia Library Manifesto
- Tietomalli luotavissa vapaasti – vaikkakin aluksi työlästä
 - Tietojen tuonti ja vienti Wikibaseen vastaaviin propertyihin
- Kuvailunäkymä karu – vaatii kehittämistä
 - Kullekin ominaisuudelle kenttensä
 - Kohteisiin viittaavat ominaisuudet – tallennus pudotusvalikosta
 - Termien poiminta ulkopuolisesta lähteestä vaatii kehittelyä

Tietomallin julkaisutilanne

- Tietomallin tarkistuspiste
 - Palveleeko tietomalli nimitietopalvelun alkuperäisiä tavoitteita?
- Yhteentoimivuusalusta – Luonnos, v. 2.1
 - Kuvaukset: https://tietomallit.suomi.fi/model/ntp_v2/
 - Namespace: http://uri.suomi.fi/datamodel/ns/ntp_v2#
 - Ongelmana versionhallinta
- Github
 - <https://github.com/NatLibFi/ntp-model>
- Muita jakeluvaihtoehtoja, ml. dokumentointi
 - digime.fi, kiwi.fi

Jatkotoimia

- Rahoitusta jatkoon on haettu
- Jatkotoimia
 - Ominaisuuksien mappaukset muihin tietomalleihin
 - Toimijakuvailujen vienti Wikibase alustalle
 - Aineistojen yhdistelyn testaus
 - Päällekkäisyyden kartoitus, tuplien yhdistelytarve
 - Tuotantoversion suunnittelu ja toteutus kv. yhteistyönä
 - Editointiympäristö
 - Tietojen synkronointi sektorikohtaisten järjestelmien kanssa



**Tiedustelut:
toimijakuvalu-posti@helsinki.fi**