Datan avaaminen linkitettynä datana

Osma Suominen

Tiedonhallinnan ohjausryhmä 5.9.2017

Datan avaamisen tavoitteet / hyödyt

1. Avoin data parantaa tutkimuksen edellytyksiä ja tukee kansalaisyhteiskuntaa.
2. Kirjaston tietovarannot voivat saada uusia käyttäjiä ja uusia käyttötapoja. Kansalliskirjasto pyrkii datan käyttöluvissa sallimaan myös kaupallisen hyödyntämisen.
    a. Tutkijat pääsevät käsiksi dataan mahdollisimman matalalla kynnyksellä
3. Kirjaston palveluiden näkyvyys ja merkitys tulevat esiin.
    a. Aineistojemme relevanssi säilyy, kun ne voidaan liittää osaksi Internetin avoimia aineistoja

Toimenpiteet

1. **Datakatalogi**, joka kartoittaa ja dokumentoi olemassaolevat tietoaineistot
   a. sekä kokotekstiaineistot (mm. digitoidut) että metatiedot

2. **Aineistojen avaaminen sellaisenaan** koneluettavassa muodossa
   a. Finna API (avattu helmikuussa 2016)
   b. Z39.50- ja SRU-rajapinnat (Fennica, Melinda, Arto ja Viola avattu syyskuussa 2016)
   c. OAI-PMH-haravointirajapinnat?
   d. Ladattavat MARC-tiedostodumpit? (toistaiseksi annettu tutkijoiden käyttöön pyynnöstä)

3. **Aineistojen avaaminen linkitettynä datana**
   a. Finton sanastot, ontologiat, luokitukset ja auktoriteetit
   b. Fennica ja muut bibliografiset aineistot linkitettynä datana

tämä esitys

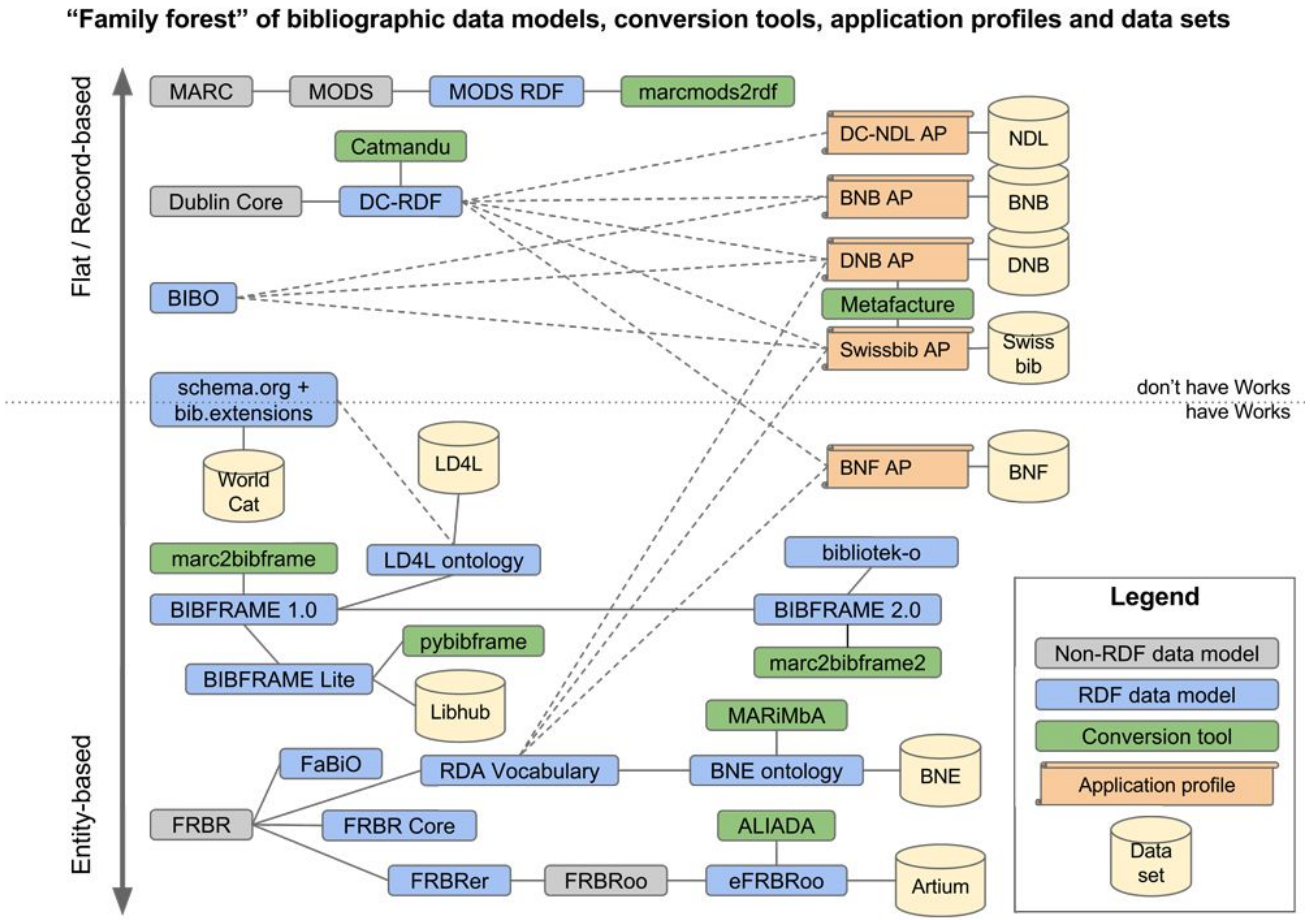
with apologies to Scott Adams

Kim Tallerås:
[Quality of Linked Bibliographic Data: The Models, Vocabularies, and Links of Data Sets Published by Four National Libraries]
Journal of Library Metadata, 2017

Fennica: ei erityisen linkitettyä

- Vain osa tietueista on WorldCatissa...
  - ...emmekä tiedä niiden OCLC-numeroita
- Bibliografisia tietueita ei ole ID-tasolla linkitetty auktoriteettitietueisiin
  - ...mutta tämä on pian tehty Melindassa!
- Vain osa nimiauktoriteeteistamme (henkilöt ja yhteisöt) ovat VIAFissa
  - ...emmekä tiedä niiden VIAF-tunnisteita
- Nimiauktoriteettimme eivät ole vielä ISNIssä
- Asiasanat (YSA) on YSO:n kautta kytketty LCSH:oon
  ja paikannimet YSO-paikkojen kautta Wikidataan

Tietomallin valinta



Osma Suominen & Nina Hyvönen: [From MARC silos to Linked Data silos?]
o-bib journal, 2017

Schema.org-malli pakottaa katsomaan dataa web-käyttäjän näkökulmasta

“Meillä on miljoona bibliografista tietuetta”

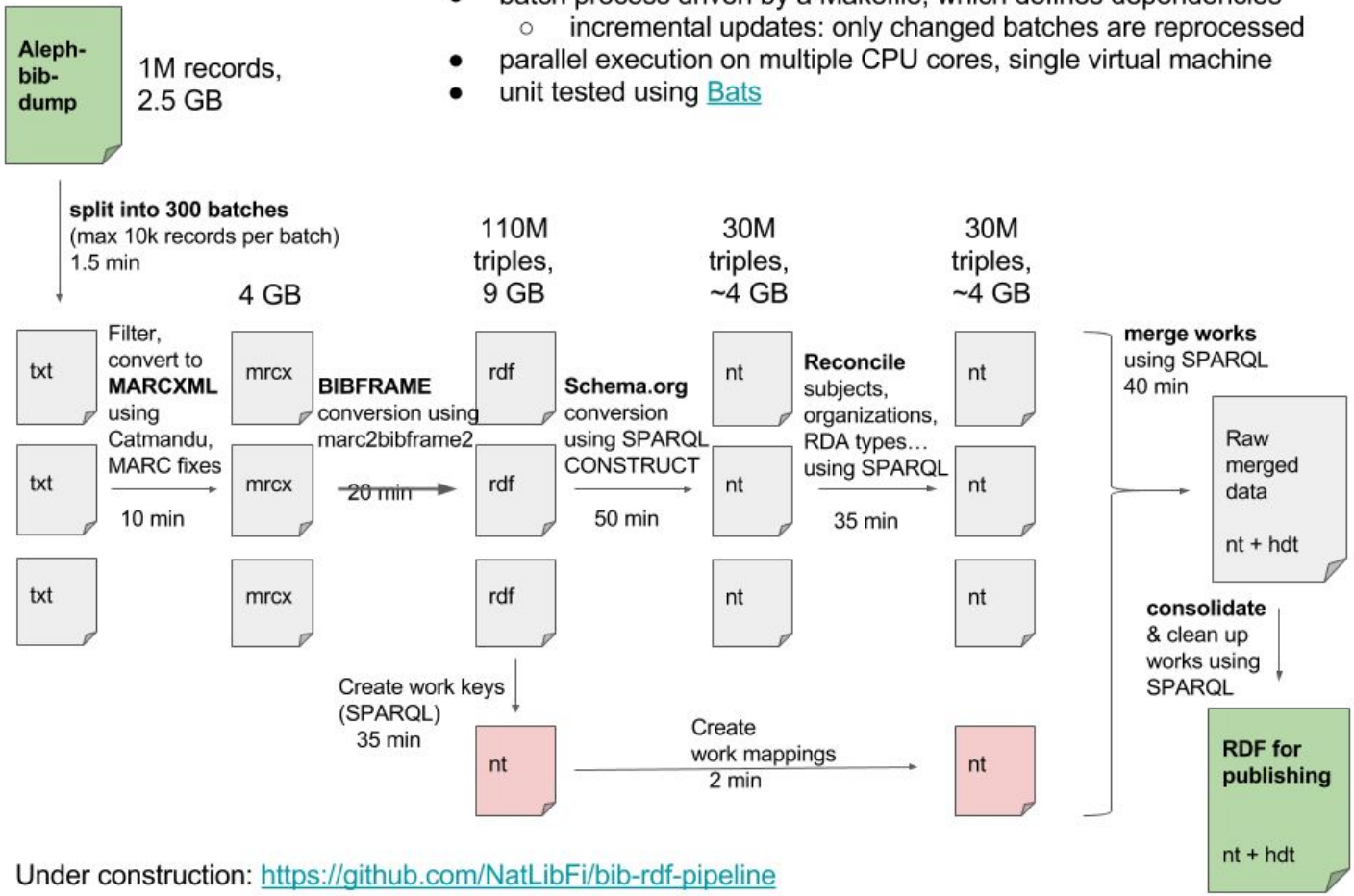
Schema.org-malli pakottaa katsomaan dataa web-käyttäjän näkökulmasta

~~“Meillä on miljoona bibliografista tietuetta”~~

“Kansalliskirjastolla on suuri kokoelmaa kirjallisia teoksia!
Kokoelmastamme löytyy teoksista nämä painokset ja kappaleet.
Niitä saa lukea (joissain tapauksissa myös lainata) ilmaiseksi
pääkirjastossamme (Unioninkatu 36, 00170 Helsinki, Suomi)
joka on avoinna ma-pe 9-18, paitsi keskiviikkoisin 9-20.
Sähköiset versiot teoksistamme ovat saatavilla näissä URL-osoitteissa.”

Fennica RDF conversion pipeline (draft)

- batch process driven by a Makefile, which defines dependencies
  - incremental updates: only changed batches are reprocessed
- parallel execution on multiple CPU cores, single virtual machine
- unit tested using Bats



Under construction: https://github.com/NatLibFi/bib-rdf-pipeline

COMHIS-yhteistyö: vuosilukujen putsaaja

COMHIS / **aleph-processor**

Watch 4 | Star 0 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Pulse | Graphs

apply metadata cleaning functions to aleph sequence

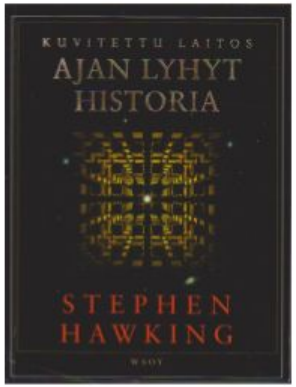
5 commits | 1 branch | 0 releases | Fetching contributors

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

Fetching latest commit…

- input
- lib
- output
- .gitignore
- README.md
- enrich_aleph.R
- enrich_aleph_functions.R
- enrich_aleph_options.R

Fennica using Schema.org

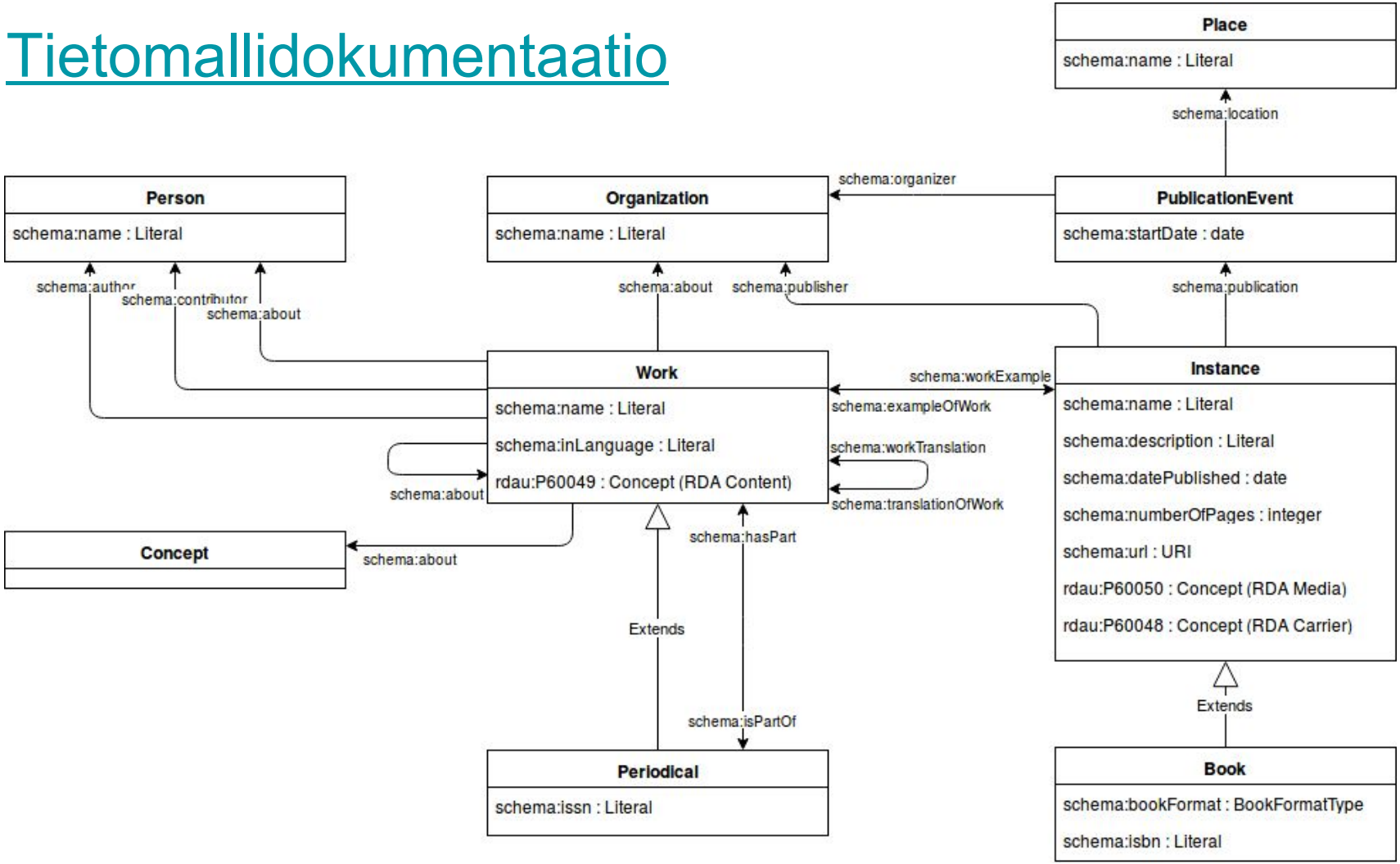



```
# The original English language work
fennica:000215259work9 a schema:CreativeWork ;
  schema:about ysa:Y94527, ysa:Y96623, ysa:Y97136,
               ysa:Y97137, ysa:Y97575, ysa:Y99040,
               yso:p18360, yso:p19627, yso:p21034,
               yso:p2872, yso:p4403, yso:p9145 ;
  schema:author fennica:000215259person10 ;
  schema:inLanguage "en" ;
  schema:name "The illustrated A brief history of time" ;
  schema:workTranslation fennica:000215259 .

# The Finnish translation (~expression in FRBR/RDA)
fennica:000215259 a schema:CreativeWork ;
  schema:about ysa:Y94527, ysa:Y96623, ysa:Y97136,
               ysa:Y97137, ysa:Y97575, ysa:Y99040,
               yso:p18360, yso:p19627, yso:p21034,
               yso:p2872, yso:p4403, yso:p9145 ;
  schema:author fennica:000215259person10 ;
  schema:contributor fennica:000215259person11 ;
  schema:inLanguage "fi" ;
  schema:name "Ajan lyhyt historia" ;
  schema:translationOfWork fennica:000215259work9 ;
  schema:workExample fennica:000215259instance26 ;
  rdau:P60049 rdacontent:1020 .
```

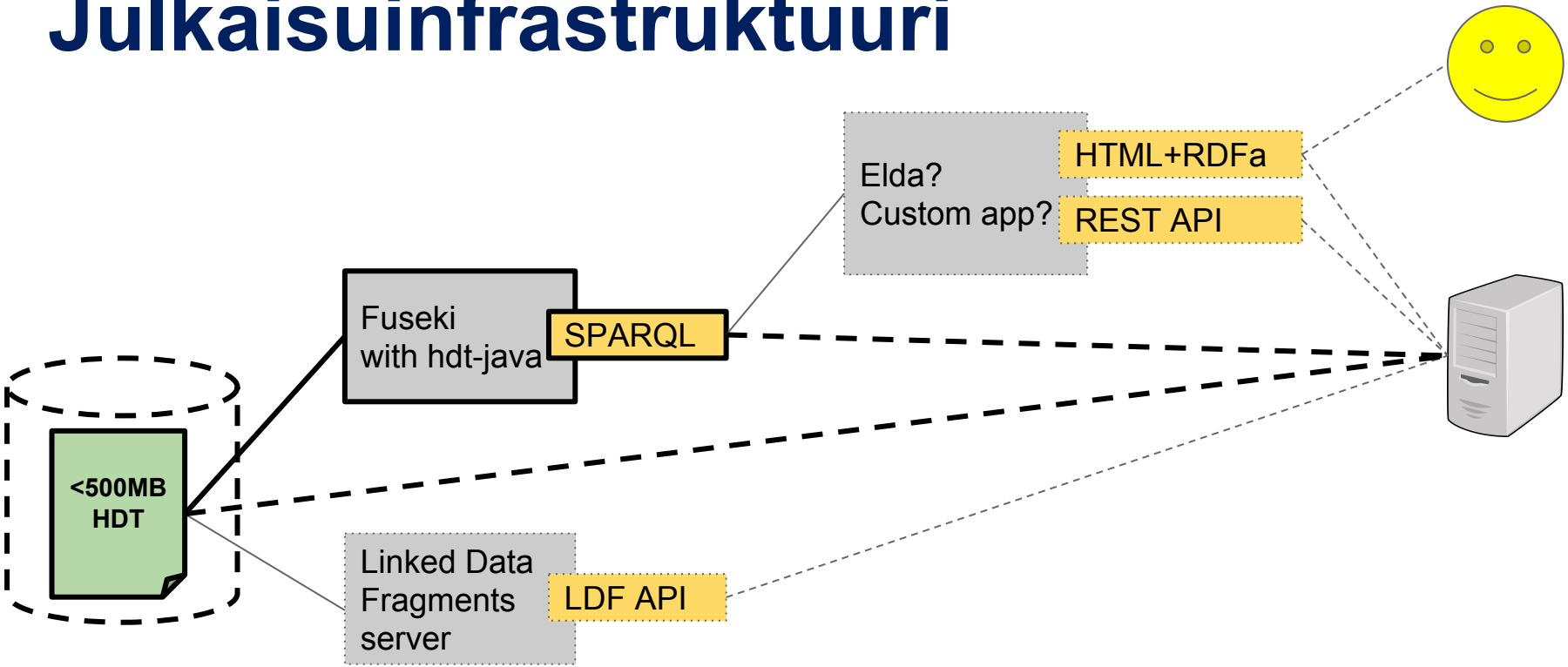
```
# The manifestation (FRBR/RDA) / instance (BIBFRAME)
fennica:000215259instance26 a schema:Book, schema:CreativeWork ;
  schema:author fennica:000215259person10 ;
  schema:contributor fennica:000215259person11 ;
  schema:datePublished "2000" ;
  schema:description "Lisäpainokset: 4. p. 2002. - 5. p. 2005." ;
  schema:exampleOfWork fennica:000215259 ;
  schema:isbn "9510248215", "9789510248218" ;
  schema:name "Ajan lyhyt historia" ;
  schema:numberOfPages "248, 6 s. :" ;
  rdau:P60048 rdacarrier:1007 ;
  schema:publisher [
        schema:name "WSOY" ;
        a schema:Organization
  ] .

# The original author
fennica:000215259person10 a schema:Person ;
  schema:name "Hawking, Stephen" .

# The translator
fennica:000215259person11 a schema:Person ;
  schema:name "Varteva, Risto" .
```

Special thanks to Richard Wallis
for help with applying schema.org!

Tietomallidokumentaatio



Julkaisuinfrastruktuuri



Lihavoidut osat on jo toteutettu!

http://linkeddata-kk.lib.helsinki.fi (väliaikainen osoite)

Esitys SWIB-konferenssissa (Hampuri 6.12.2017)

09:45 - 10:10 **Finnish National Bibliography Fennica as Linked Data**

Osma Suominen

National Library of Finland, Finland

Abstract

The National Library of Finland is making our national bibliography Fennica available as Linked Open Data. We are converting the data from 1 million MARC bibliographic records first to BIBFRAME 2.0 and then further to the Schema.org data model. In the process, we are clustering works extracted from the bibliographic records, reconciling entities against internal and external authorities, cleaning up many aspects of the data and linking it to further resources. The Linked Data set is CC0 licensed and served using HDT technology.

The publishing of Linked Data supports other aspects of metadata development at the National Library. For some aspects of the Linked Data, we are relying on the RDA conversion of MARC records that was completed in early 2016. The work clustering methods, and their limitations, inform the discussions about potentially establishing a work authority, which is a prerequisite for real RDA cataloguing.

This presentation will discuss lessons learned during the publishing process, including the selection and design of the data model, the construction of the conversion pipeline using pre-existing tools, the methods used for work clustering, reconciliation and linking as well as the infrastructure for publishing the data and keeping it up to date.

Suomi Finland 100

Kiitos

http://tinyurl.com/thohry-ld