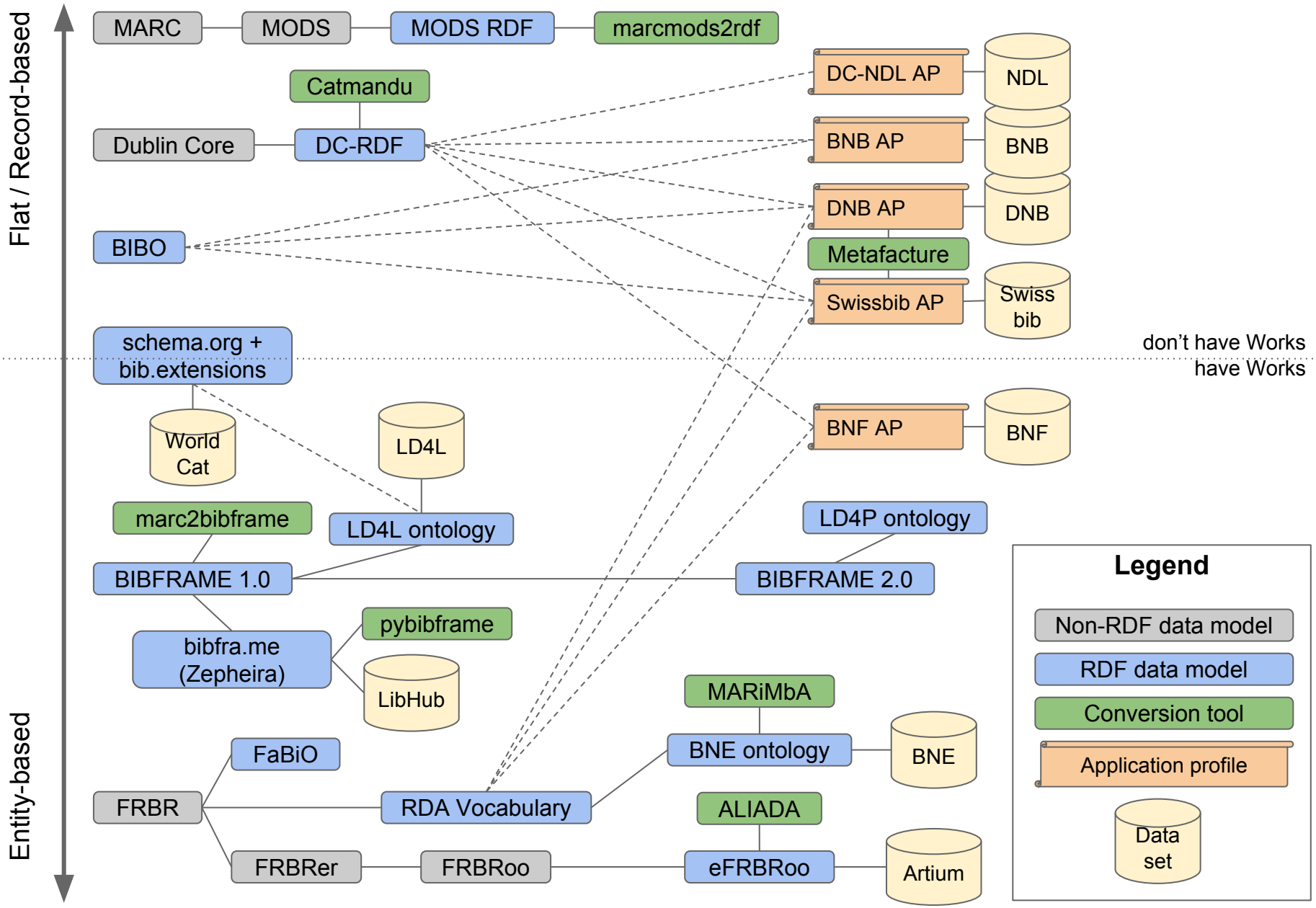
Fennican RDF-konversio

Osma Suominen
31.10.2016

Linkitetyn kirjastodatan tietomalleja

“Family forest” of bibliographic data models, conversion tools and data sets

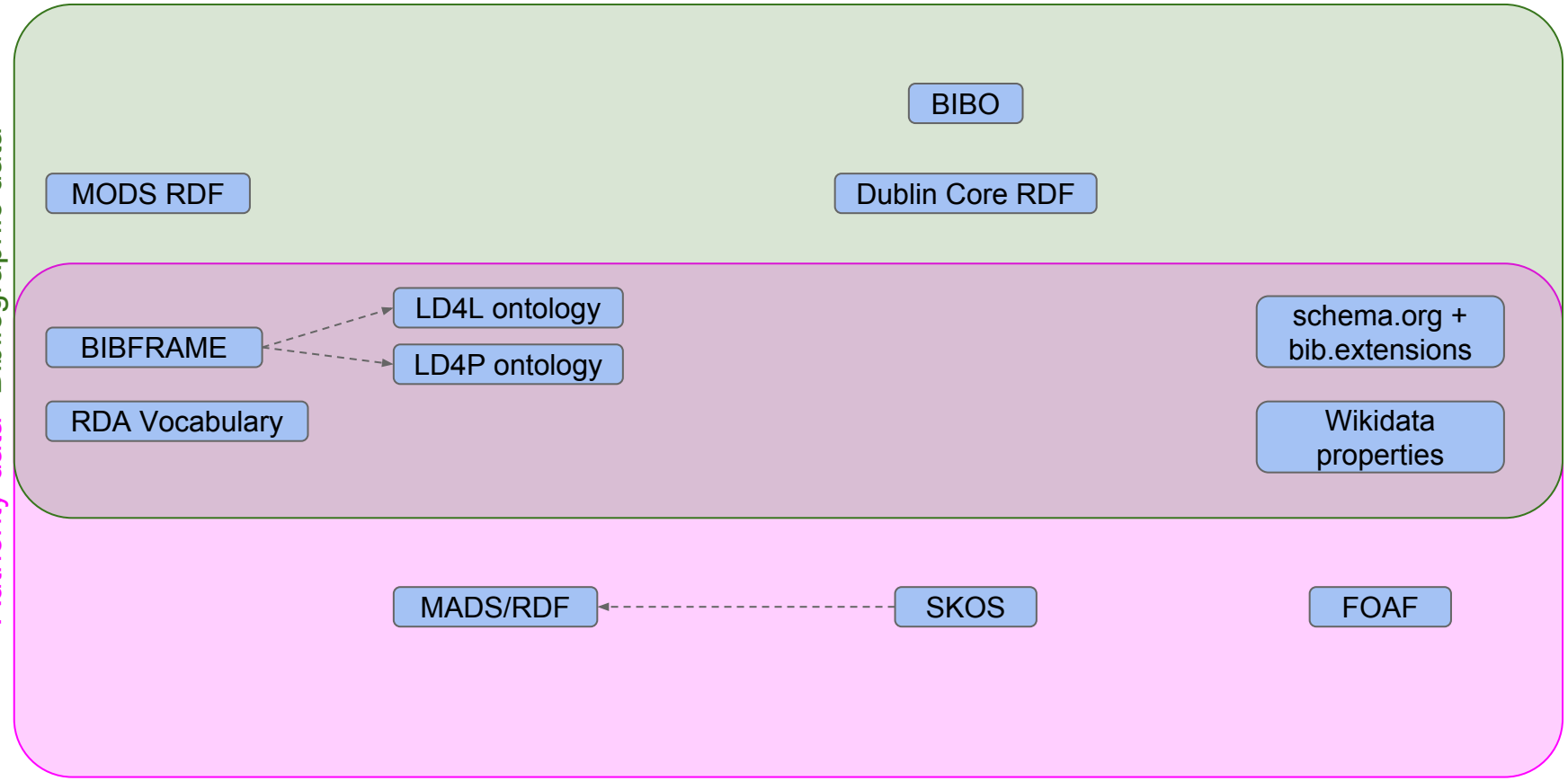


Libraryish

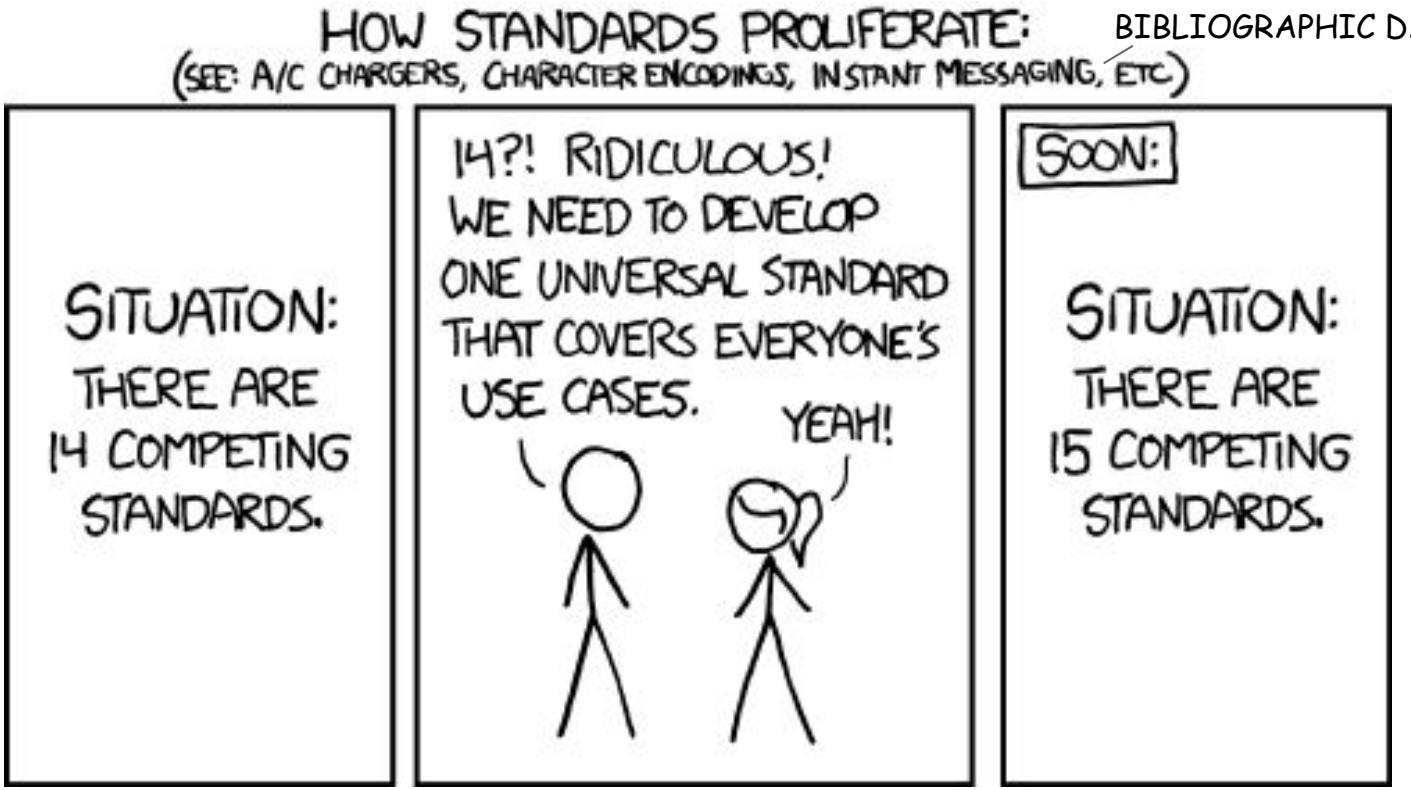
- used for **producing** and **maintaining** (meta)data
- **lossless conversion** to/from legacy formats (MARC)
- modelling of **abstractions** (records, authorities)
- **housekeeping metadata** (status, timestamps)
- favor self-contained modelling over reuse of other data models

Webbish

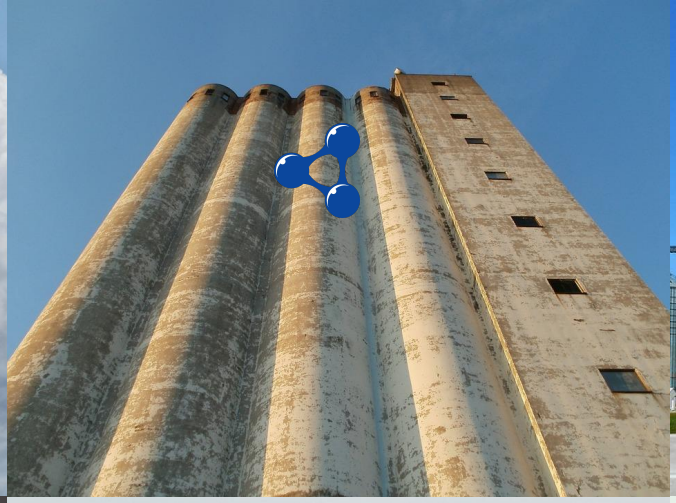
- used for **publishing** data for others to reuse
- **interoperability** with other (non-library) data models
- modelling of **Real World Objects** (books, people, places, organizations...)
- favour **simplicity** over exhaustive detail

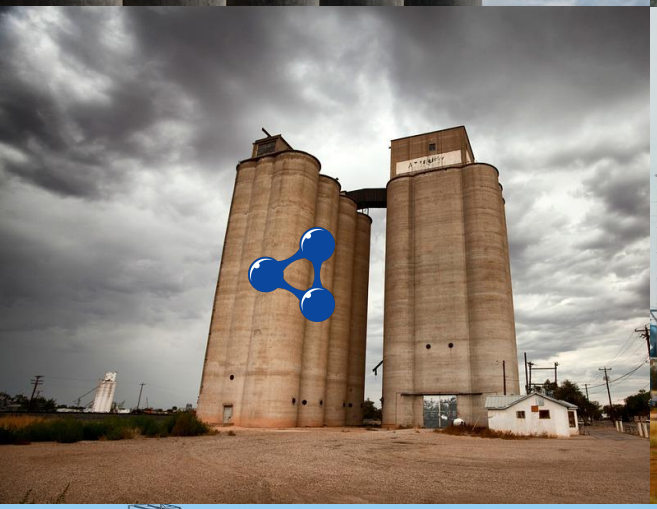


BIBLIOGRAPHIC DATA MODELS



https://xkcd.com/927/



SHARMAN
SHARMAN
SHARMAN



Choosing a data model

1. Want to have Works, or just records?

2. Libraryish or Webbish?

Fennica RDF:nä

My assignment

with apologies to Scott Adams

Not very Linked to start with

- Only some of our bibliographic records are in WorldCat
  - ...and we don't know their OCLC numbers
- Our bibliographic records don't have explicit links to authority records
  - ...but we're working on it!
- Our person authority records are not in VIAF or ISNI
- Our corporate name authority isn't linked anywhere either
- Our main subject headings (YSA) are linked via YSO to LCSH

Targeting schema.org

- schema.org + bibliographic extensions allows surprisingly rich descriptions
- modelling of Works is possible, similar to BIBFRAME [1]
- forces to think about the data from a web user's point of view:

  ~~"We have these 1M bibliographic records"~~

  "The National Library maintains this amazing collection of literary works!
  We have these editions of those works in our collection.
  They are available free of charge for reading/borrowing
  from this library building (Unioninkatu 36, 00170 Helsinki, Finland)
  which is open Mon-Fri 10-17, except Wed 10-20.
  The electronic versions are available online from these URLs."

[1] Godby, Carol Jean, and Ray Denenberg. 2015. ***Common Ground: Exploring Compatibilities Between the Linked Data Models of the Library of Congress and OCLC.*** Dublin, Ohio: Library of Congress and OCLC Research.
http://www.oclc.org/content/dam/research/publications/2015/oclcresearch-loc-linked-data-2015.pdf

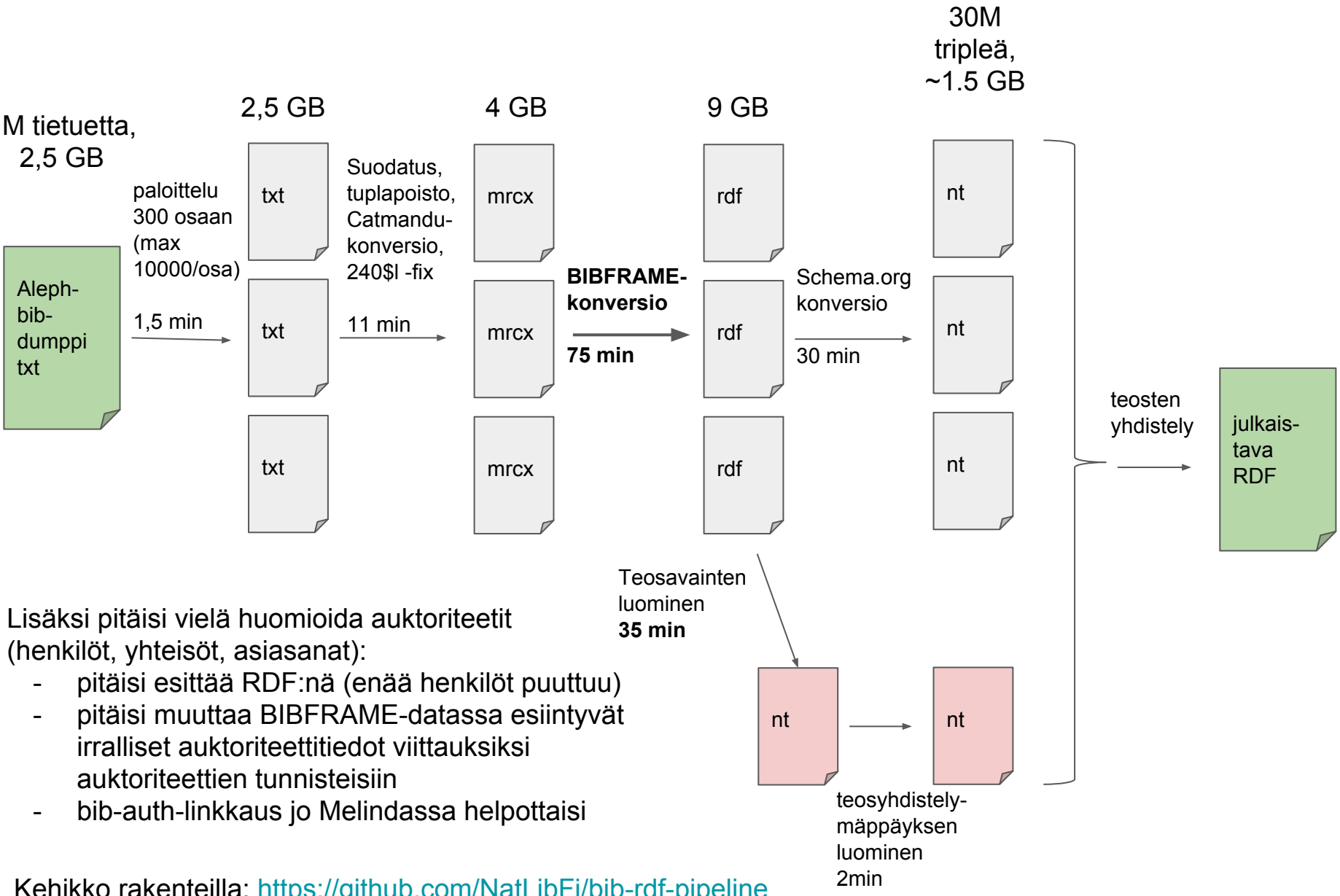
Fennica-dataa schema.org:illa

```
# This represents the original English language work
fennica:000215259work9  a schema:CreativeWork ;
  schema:author   fennica:000215259person15 ;
  schema:inLanguage  "eng" ;
  schema:name    "The illustrated A brief history of time" .

# This is the Finnish translation work (expression in FRBR/RDA)
fennica:000215259 a schema:CreativeWork ;
  schema:about              "maailmankaikkeuden synty" , "kvarkit",
                            "mustat aukot" , "maailmankaikkeus" ,
                            "aika" , "suhteellisuusteoria" ;
  schema:contributor        fennica:000215259person11 ;
  schema:creator            fennica:000215259person10 ;
  schema:inLanguage         "fin" ;
  schema:name               "The illustrated A brief history of time" ;
  schema:translationOfWork  fennica:000215259work9 ;
  schema:workExample        fennica:000215259instance29 .

# This is the manifestation (FRBR/RDA) / instance (BIBFRAME)
fennica:000215259instance29 a schema:Book , schema:CreativeWork ;
  schema:creator            fennica:000215259person10 ;
  schema:datePublished      "2000" ;
  schema:exampleOfWork      fennica:000215259 ;
  schema:isbn               "9510248215" , "9789510248218" ;
  schema:name               "Ajan lyhyt historia" ;
  schema:numberOfPages      "248, 6 s. :" ;
  schema:publisher          [ a       schema:Organization ;
                              schema:name  "WSOY"
                            ] .

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

# The original author again - should be merged with above
fennica:000215259person15 a schema:Person ;
  schema:name  "Hawking, Stephen." .

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

Konversiotyökaluista

- Catmandu: liian rajoittunut RDF-konversioon, sopii MARC-esikäsittelyyn
- ALIADA: paperilla hieno, käytännössä vaikea asentaa ja buginen
- pybibframe: sidoksissa Zepheiran omaan bibfra.me-versioon, hidas
- marc2bibframe: kömpelö mutta toimiva, paljon tiivistettyä MARC-tietoutta
  - tehty LoC:n esimerkkikoodin pohjalta marc2bibframe-wrapper
  - wrapperin avulla voi tehokkaasti käsitellä 10000 tietueen eriä kerrallaan, konversionopeus >200 tietuetta/sekunti (4 CPU)
- tulossa myös uusia BIBFRAME-muuntimia:
  - LoC on ilmeisesti tehnyt Index Datan kanssa sopimuksen BIBFRAME 2.0 -muuntimen kehityksestä; ei vielä mustaa valkoisella
  - LD4P-projekti on tekemässä oman muuntimen, jonka kohdemalli on LD4P-ontologia: *“a new, robust, efficient, well-documented, well-tested, open-source MARC to BIBFRAME converter to support the revised BIBFRAME ontology”*

Fennican BIBFRAME-muunnosketju (alustava)



Lisäksi pitäisi vielä huomioida auktoriteetit (henkilöt, yhteisöt, asiasanat):

- pitäisi esittää RDF:nä (enää henkilöt puuttuu)
- pitäisi muuttaa BIBFRAME-datassa esiintyvät irralliset auktoriteettitiedot viittauksiksi auktoriteettien tunnisteisiin
- bib-auth-linkkaus jo Melindassa helpottaisi

Kehikko rakenteilla: https://github.com/NatLibFi/bib-rdf-pipeline

Tämän hetken haasteita

- rikkinäiset URLit MARC-tietueissa
  - [Tähän kirjaston linkki]
  - http://urn.fi/URN:ISBN: 978-951-53-3352-0
  - http://urn.fi/URN:ISBN:978-951-784-608-0 (PDF)
  - http:// www.maailmalle.net
  - http://ethesis.helsinki.fi/julkaisut/maa/skemi/vk/mentula/ base target=_blank
  - http://www.etk.fi/Binary.aspx?Section=44857&Item=64774 ‡z Linkki verkkoaineistoon (PDF)
  - http://helda.helsinki.fi/bitstream/handle/10138/15810/Tutkimuksia108.pdf?sequence=1 |y Linkki |q PDF
  - http://formin.finland.fi/public/download.aspx?ID=96845&GUID={E3C53F54-3FA3-4A33-BA1E-C55F5CA16703}
  - jne., yhteensä reilut 100 kpl virheellisiä URLeja jotka aiheuttavat syntaksivirheitä RDF-konversion jälkeisessä käsittelyssä
- teosten eristäminen: alustava toteutus olemassa
  - pitää vielä yhdistää samaa teosta koskevat tiedot älykkäästi
- replikointidirektiivien FENNI<KEEP> ja FENNI<DROP> huomiointi
  - haluamme varmaankin julkaista Fennica-tietueet Fennican mukaisina, ei sellaisina kuin ne ovat Melindassa?
- linkitys YSAan/YSOon, henkilö- ja yhteisöauktoriteetteihin

Julkaisu RDF:nä

- Testattu Apache Marmottaa
  - RDF-tietokanta on varsin hidas
  - kehityksen tilanne vaikuttaa epäilyttävältä
  - **ei jatkoon**

- HDT-tiedostomuoto ja Linked Data Fragments vaikuttaa lupaavalta
  - koko Fennica RDF:n voisi julkaista noin 1GB HDT-tiedostona
  - HDT:n pystyy helposti tarjoilemaan linkitettynä datana ja SPARQL-rajapintana

Kiitos!

osma.suominen@helsinki.fi