From MARC silos to Linked Data silos?

Data models for bibliographic Linked Data

Osma Suominen

DCMI webinar

February 28, 2017

About the National Library of Finland

- The National Library of Finland is the oldest and largest scholarly library in Finland. Our origins date back to 1640, when the Academy of Turku was founded.

- We are responsible for the collection, description, preservation and accessibility of Finnish printed national heritage and the unique collections under its care.

About me

Twitter:
@OsmaSuominen

LinkedIn:
osmasuominen
(I accept invites only
from people I've met)

GitHub:
@osma

Osma Suominen
Information Systems Specialist at the National Library of Finland

PhD thesis "*Methods for Building Semantic Portals*"
from the Semantic Computing Research Group
at Aalto University, 2013

Joined the National Library of Finland in 2013
to set up the Finto.fi thesaurus and ontology service

Currently working on opening up Finnish bibliographic
metadata as Linked Data

Open source software projects e.g.

- Skosify - Validation and QA tool for SKOS vocabularies
- Skosmos - SKOS vocabulary publishing tool

DCMI Governing Board member

SWIB conference Programme Committee member

Apache Jena project committer & PMC member

Contents

1. Overview of current data models for bibliographic data

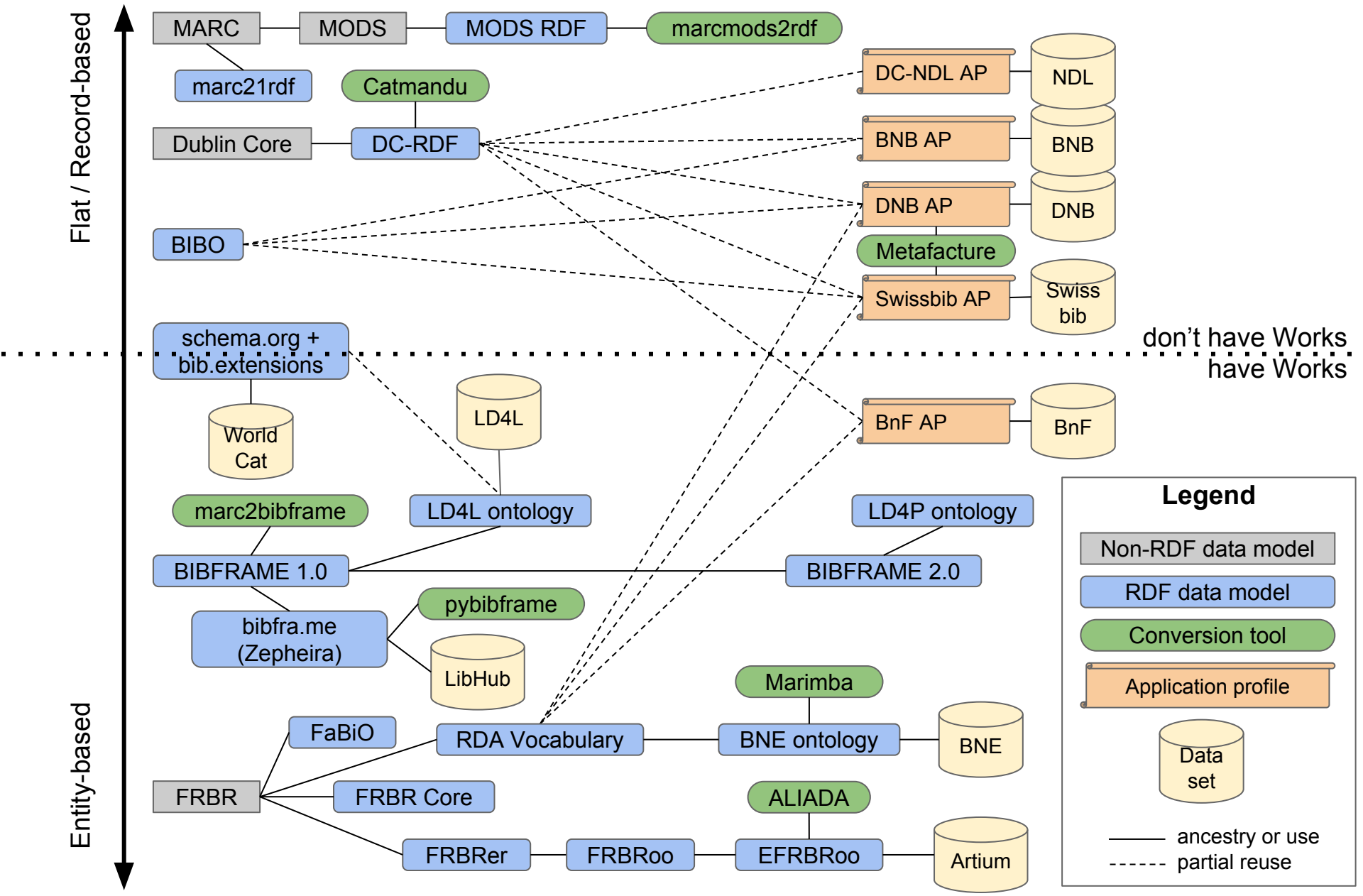
2. Publishing Finnish bibliographic data as Linked Open Data

Part I:
Overview of current data models for bibliographic data

data models

- **schemas**
- **vocabularies**
  (of classes and properties)
- **ontologies**
- **application profiles**

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

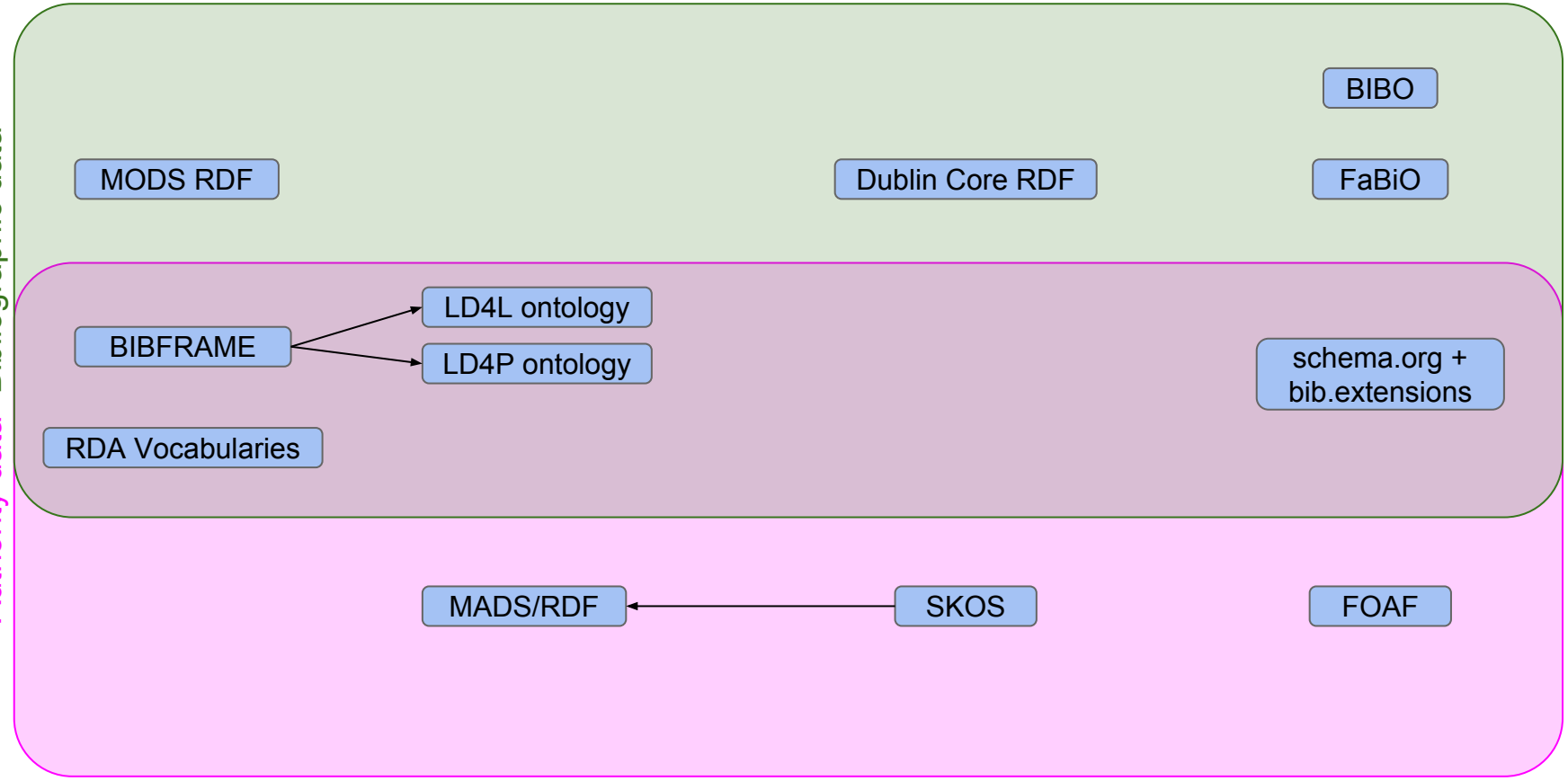


Libraryish

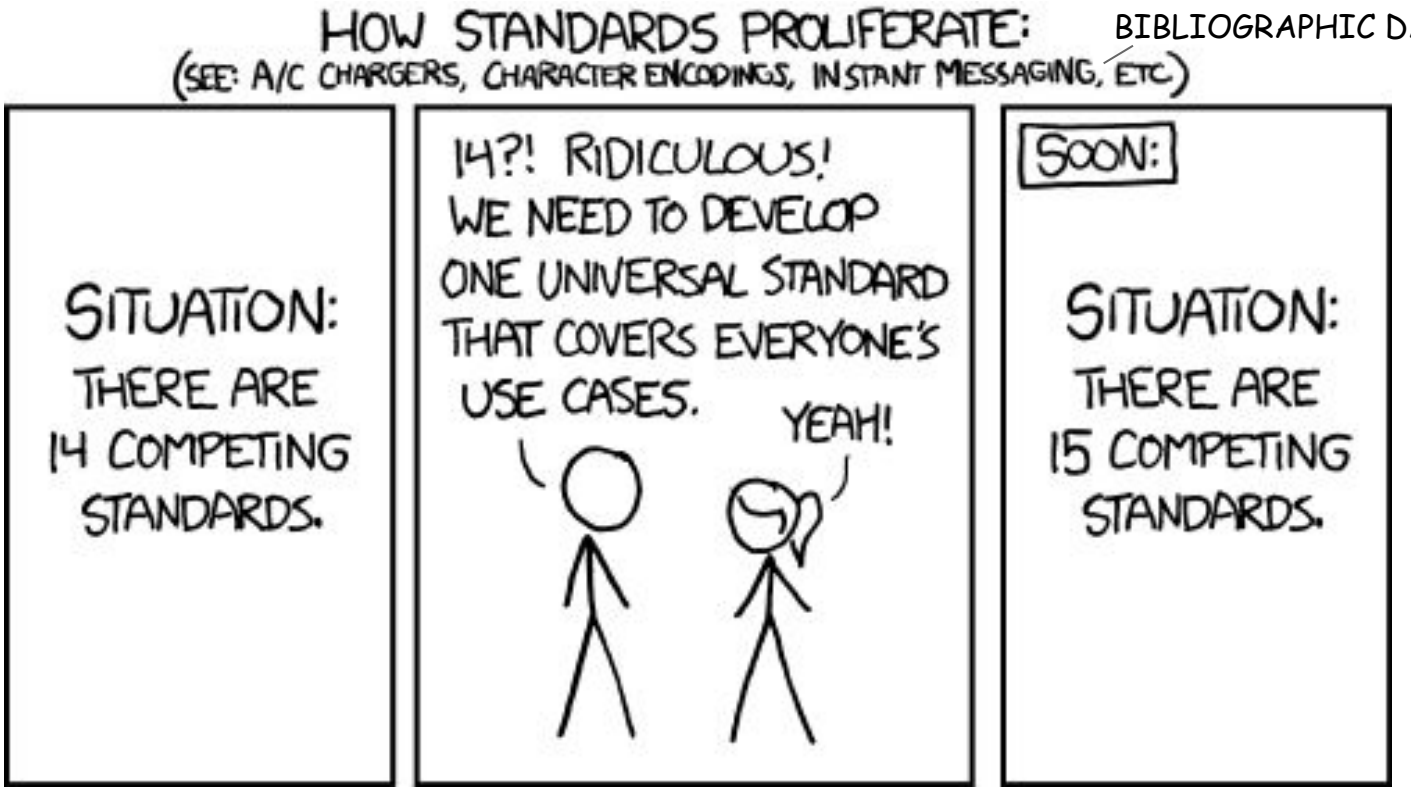
- used for **producing** and **maintaining** (meta)data
- **lossless conversion** to/from legacy formats (MARC)
- modelling of **abstractions** (records, authorities)
- **housekeeping metadata** (status, timestamps)
- favour **self-contained** modelling over reuse of other data models

Webbish

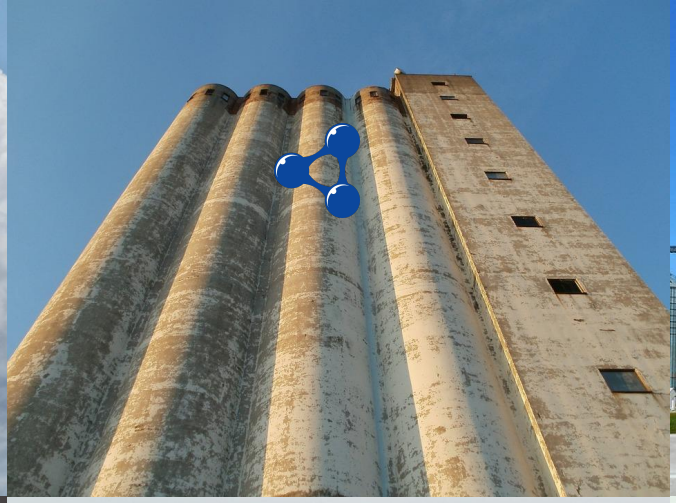
- used for **publishing** data for others to reuse
- **interoperability** with other (non-library) data models
- modelling of **Real World Objects** (books, people, places, organizations...)
- favour **simplicity** over exhaustive detail

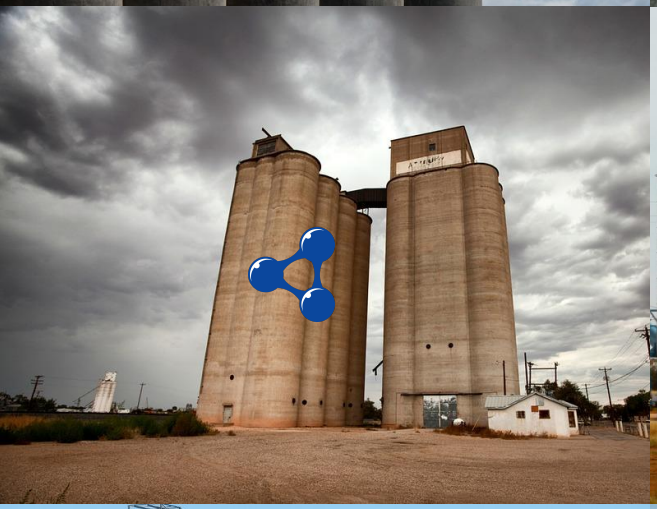


BIBLIOGRAPHIC DATA MODELS



https://xkcd.com/927/


SHARMAN
SHARMAN
SHARMAN
SHARMAN




Why does it have to be like this?

Reason 1

Reason 2

Reason 3

Reason 4

Different use cases require different kinds of data models. None of the existing models fits them all.

But surely, for basic MARC records (e.g. a “regular” national library collection) a single model would be enough?

Reason 1

Reason 2

Reason 3

Reason 4

Converting existing data (i.e. MARC) into a modern entity-based model is difficult and prevents adoption of such data models in practice for real data.

All FRBR-based models require “FRBRization”, which is difficult to get right. BIBFRAME is somewhat easier because of its more relaxed view about Works.

Reason 1

Reason 2

Reason 3

Reason 4

Libraries want to control their data - including data models.

Defining your own ontology, or a custom application profile, allows maximum control. Issues like localization and language- or culture-specific requirements (e.g. Japanese dual representation of titles as *hiragana* and *katakana*) are not always adequately addressed in the general models.

Reason 1

Reason 2

Reason 3

Reason 4

Once you've chosen a data model, you're likely to stick to it.

Choosing an RDF data model for a bibliographic data set

1. Want to have Works, or just records?

2. Libraryish (maintaining) or Webbish (publishing) use case?

For maintaining metadata as RDF, suitable data models (BIBFRAME, RDA Vocabulary etc.) are not yet mature.

For publishing, we already have too many data models.

What can we do about this?

Don’t create another data model, especially if it’s only for publishing.

Help improve the existing ones!

We need more efforts like LD4P that consider the production and maintenance of library data as modern, entity-based RDF instead of records.

How could we share and reuse each other’s Works and other entities instead of having to all maintain our own?

Will Google, or some other big player, sort this out for us?

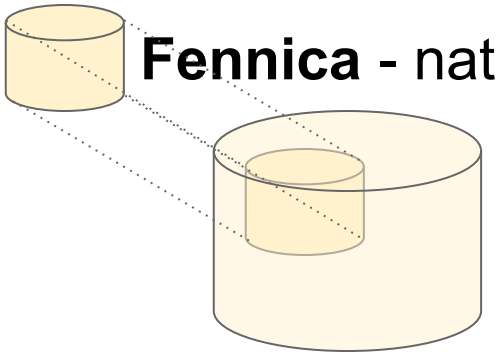
A big actor offering a compelling use case for publishing bibliographic LOD would make a big difference.

- a global bibliographic knowledgebase?
- pushing all bibliographic data into Wikidata?
- Search Engine Optimization (SEO) using schema.org?

This is happening for scientific datasets - Google recently defined a schema for them within schema.org.

Part II:
Publishing Finnish bibliographic data as Linked Open Data

Our bibliographic databases



Fennica - national bibliography (1M records)

Melinda union catalog (9M records)

Arto - national article database (1.7M records)

Viola - national discography (1M records)

All are MARC record based Voyager or Aleph systems.

The Z39.50/SRU APIs have been opened in September 2016

My assignment

with apologies to Scott Adams

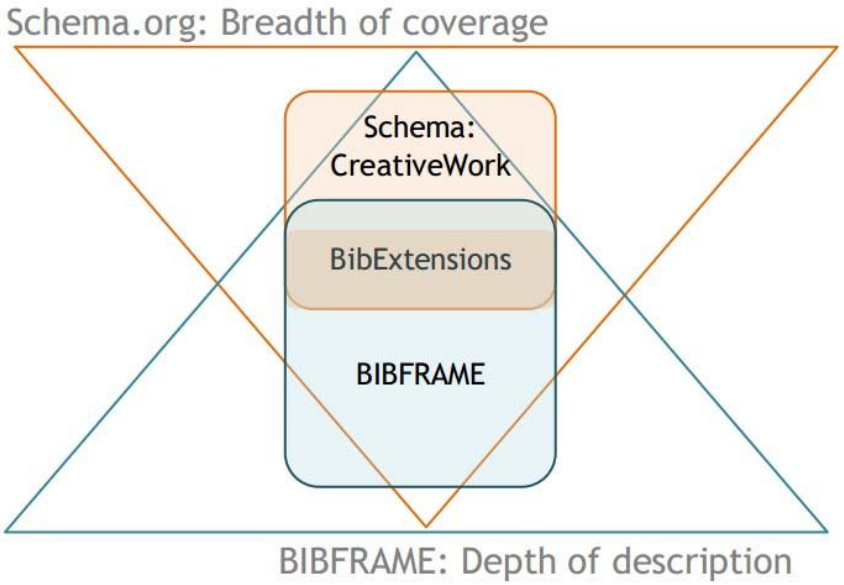
Not very Linked to start with

- Only some of our bibliographic records are in WorldCat
  - ...and we don't know their OCLC numbers
- Our bibliographic records don't have explicit (ID) links to authority records
  - ...but we're working on it!
- Only some of our person and corporate name authority records are in VIAF
  - ...and we don't know their VIAF IDs
- Our name authorities are not in ISNI 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]



[1] Godby, Carol Jean, and Denenberg, Ray. 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

Schema.org forces to think about data from a Web user's point of view

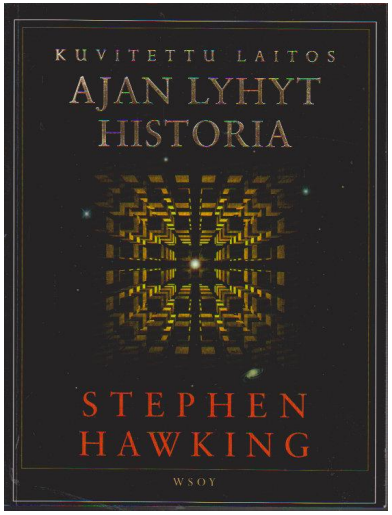
"We have these 1M bibliographic records"

Schema.org forces to think about 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 our 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."

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!

From MARC to Schema.org - via BIBFRAME!

To convert to Schema.org, we first need to break down the MARC records into some (any!) kind of RDF data, without losing any important information.

BIBFRAME converters do a fairly good job of this!

1. Zepheira's **pybibframe** was tested briefly. It was rather slow and seems to lose more information than I'd like. Does some internal reconciliation.

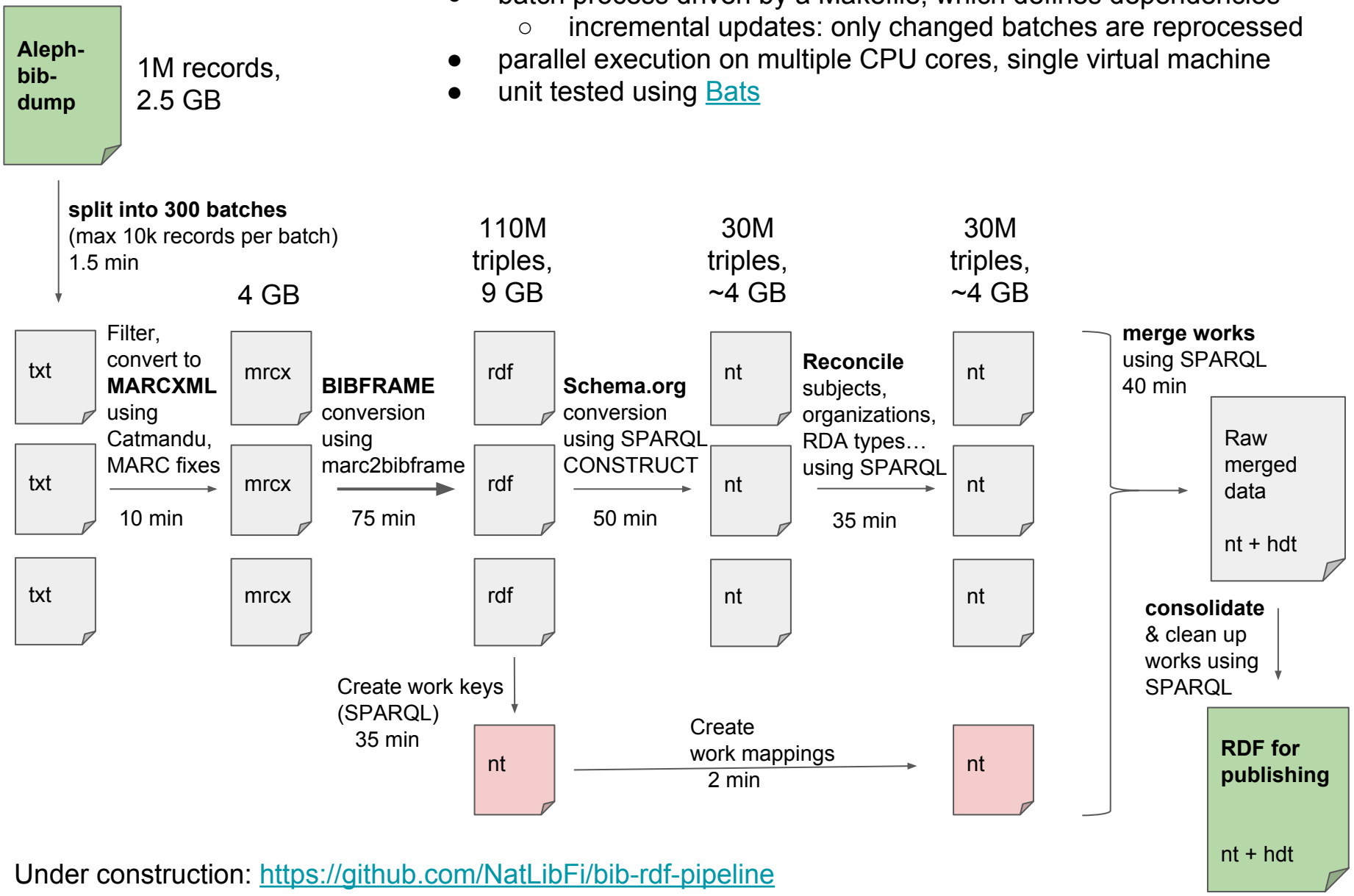
2. LoC's **marc2bibframe** is our current choice. Together with a wrapper, it has relatively good performance and consistent, but quite verbose RDF output. Not maintained anymore!

3. LoC is working on a BIBFRAME 2.0 converter together with a consultant. Not yet released, but I want to try it!

4. LD4L-Labs is working on the bib2lod converter, from MARC to their flavor of BIBFRAME 2.0. Following closely!

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



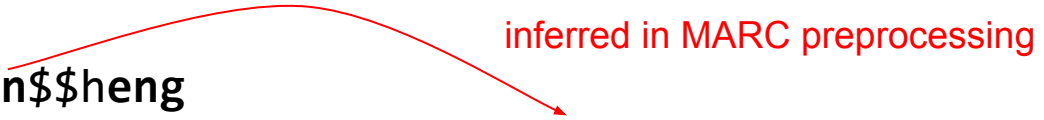
Under construction: https://github.com/NatLibFi/bib-rdf-pipeline

Modelling translated works

15% of Fennica records are translations (041 $h = original language)

Ideally, they should have

1. the name of the original work in the 240 field (uniform title) - only ⅔ do!
2. the name of the translator in a 700 field (contributor) - about 90% do!

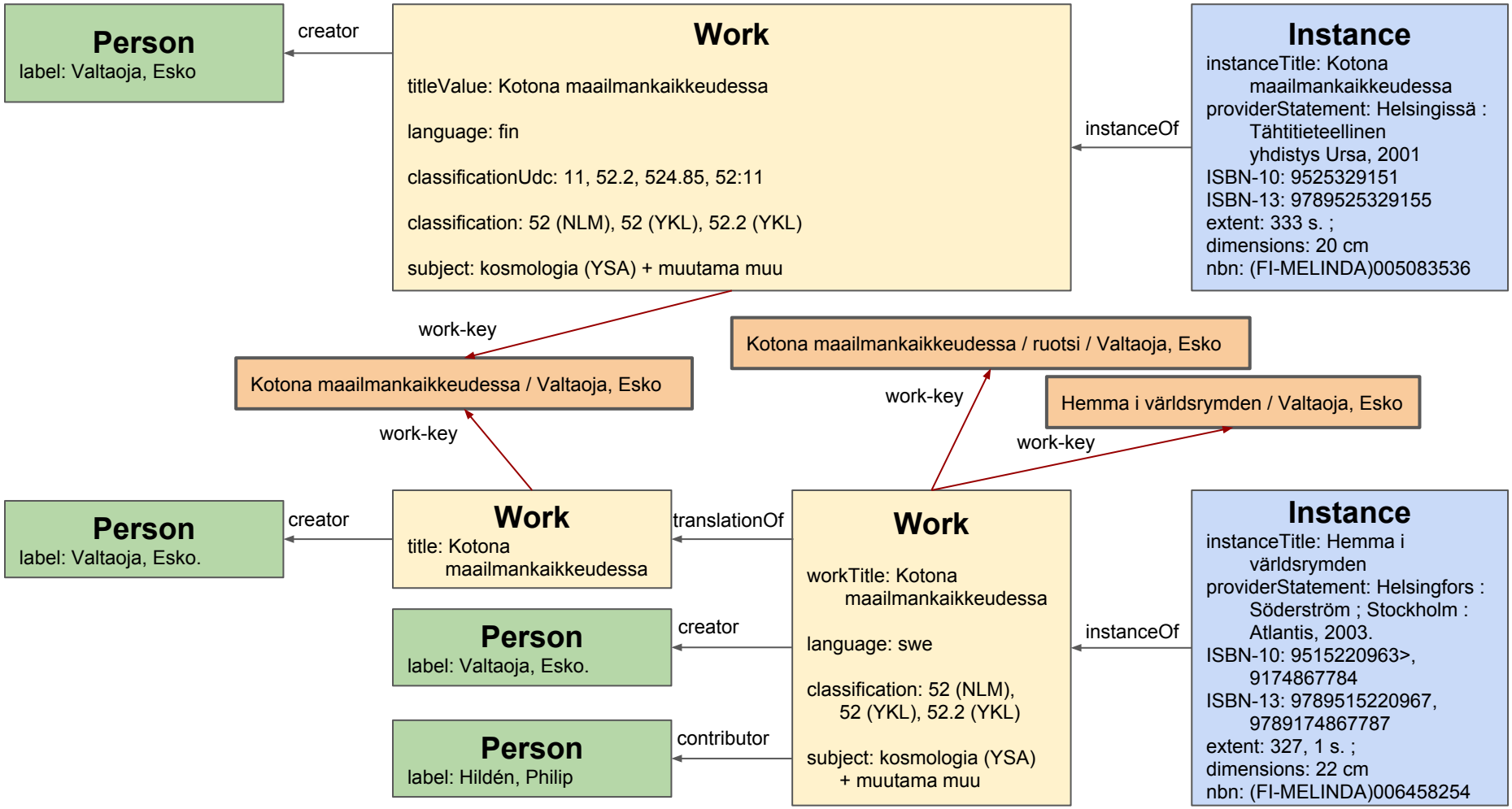



```
000095841 0411  L $$afin$$heng
000095841 24012 L $$aA brief history of time$$lsuomi
000095841 24510 L $$aAjan lyhyt historia :$$balkuräjähdyksestä
mustiin aukkoihin /$$cStephen W. Hawking ; alkusanat: Carl Sagan
; piirrokset: Ron Miller ; suomentanut Risto Varteva.
000095841 7001  L $$aSagan, Carl.
000095841 7001  L $$aVarteva, Risto.
```

(700 $e subfield could contain the specific role, e.g. “kääntäjä” = translator, but it only exists for <25% of translated records, making 700 ambiguous)

Calculating work keys

similar to OCLC's FRBR Work-Set algorithm [1]
but extracting the keys from BIBFRAME RDF using SPARQL queries



[1] Hickey, Thomas and Toves, Jenny. **FRBR Work-Set Algorithm**. OCLC Research, 2009.
http://www.oclc.org/content/dam/research/activities/frbralgorithm/2009-08.pdf

“Original works” by Aleksis Kivi, after merging works

Translations without link to original work look like independent original works

- Eriika (2)
- Halavan himmeän alla (1)
- Kanervala: runoelmia (3)
- Kanerwala: runoelmia (1)
- Karkurit : näytelmä viidessä näytöksessä (4)
- Kihlaus : komedia (5)
- Kihlaus : komedia 1:ssä näytöksessä (3)
- Kihlaus : komedia yhdessä näytöksessä (2)
- Kihlus : ühejärguline naljaäitlus (1)
- Kullervo : näytelmä (3)
- Kullervo : näytelmä viidessä näytöksessä (3)
- Lea : näytelmä (3)
- Lea : näytelmä yhdessä näytöksessä (2)
- Lea : oshinyandwa molweetho lumwe (1)
- Lea : skådespel i två akter (2)
- Lintukoto (2)
- Margareta : näytelmä yhdessä näytöksessä (3)
- Nummi-suutarit : komedia 5:ssä näytöksessä (2)
- Nummisuutarit (1)
- Nummisuutarit : komedia (5)
- Nummisuutarit : komedia viidessä näytöksessä (16)
- Nummisuutarit : näytelmä viidessä näytöksessä (1)
- Die Heideschuster : Bauernkomödie in fünf Akten (1)
- Olviretki Schleusingenissä : näytelmällinen… (1)
- Aleksis Kiven Seitsemän veljestä (3)
- Seitsemän veljestä (89)
- De zeven broeders : een Fins volksepos (1)
- Les sept frères (1)
- Sedm bratři (1)
- Vuoripeikot (1)
- Yö ja päivä : näytelmä viidessä näytöksessä (1)
- Yö ja päivä : näytelmä yhdessä näytöksessä (4)
- Yö ja päiwä : näytelmä yhdessä näytöksessä (2)
- Natt och dag : skådespel i en akt (1)

Actually 14 works, but looks like 34 in the data!

Translations of “Seven brothers”: 110 records

15 of these records (in 7 different languages) lack original work name in 240 field. **What happens?**

Dutch	245: De zeven broeders : een Fins volksepos	1 rec (1956)	not merged (spelling differs)
	245: De zeven broeders : een finsch volks-epos 240: Seitsemän veljestä, hollanti	2 recs (1941-1943)	
German	245: Die sieben Brüder	4 recs (1959-1962)	**merged**
	245: Die sieben Brüder 240: Seitsemän veljestä, saksa	3 recs (1954-1965)	
	245: Die sieben Brüder : Erzählung	1 recs (1935)	**merged**
	245: Die sieben Brüder : Erzählung 240: Seitsemän veljestä, saksa	3 recs (1935-1944)	
	245: Die sieben Brüder : Roman	3 recs (1950-1958)	**merged**
	245: Die sieben Brüder : Roman 240: Seitsemän veljestä, saksa	8 recs (1921-2006)	
French	245: Les sept frères	1 rec (1963)	not merged (subfield $b differs)
	245: Les sept frères : roman 240: Seitsemän veljestä, ranska	3 recs (1926-1985)	
Spanish	245: Los siete hermanos	1 rec (1951)	**merged**
	245: Los siete hermanos 240: Seitsemän veljestä, espanja	8 recs (1941-1988)	
Czech	245: Sedm bratří : román	1 rec (1941)	nothing to merge with
Estonian	245: Seitse venda	1 rec (1956)	**merged**
	245: Seitse venda 240: Seitsemän veljestä, viro	2 recs (1971-1983)	
	245: Seitse wenda : romaan	1 rec (1924)	**merged**
	245: Seitse wenda : romaan 240: Seitsemän veljestä, viro	1 rec (1935)	
Russian	245: Semero brat'ev : povest'	1 rec (1935)	**merged**
	245: Semero brat'ev : povest' 240: Seitsemän veljestä, venäjä	1 rec (1951)	

Out of 15 records lacking 240 field, 12 were **still merged** using the “Friend of a Friend” rule via other records!

(helps with 10% of records lacking 240)

Current challenges

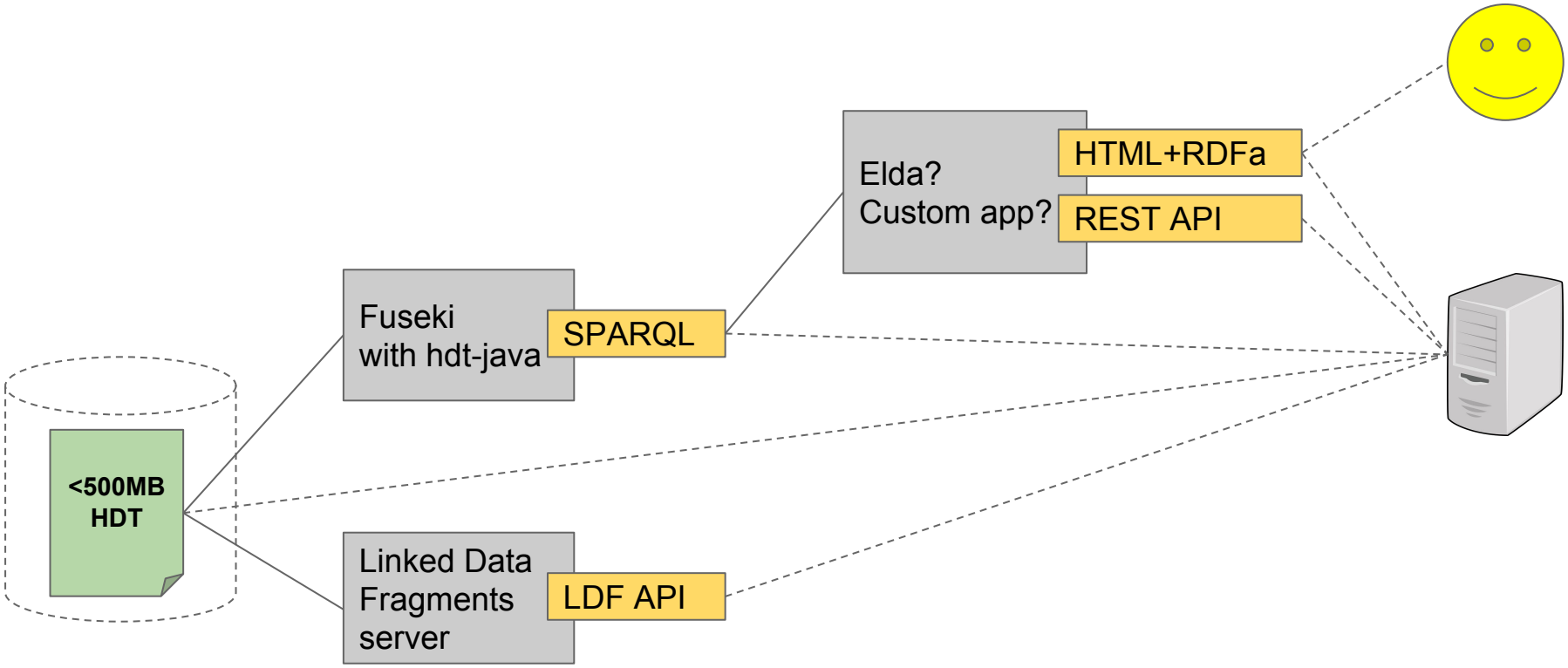
1. problems caused by errors & omissions in MARC records

2. dumbing down MARC to match Schema.org expectations
   - e.g. structured page counts: “`vii, 89, 31 p.`”
     -- schema.org only defines numeric `numberOfPages` property

3. linking internally - from strings to things
   - subjects from YSA and YSO - already working
   - organizations from corporate name authority - already working
   - person name authority

4. linking externally
   - RDA Vocabularies: content, carrier and media types - already working
   - linking name authorities to VIAF, ISNI, Wikidata...
   - linking works to WorldCat Works?

Publishing as LOD (draft plan)



Thank you!

osma.suominen@helsinki.fi

code: https://github.com/NatLibFi/bib-rdf-pipeline
these slides: http://tinyurl.com/linked-silos-webinar