From MARC silos to Linked Data silos?

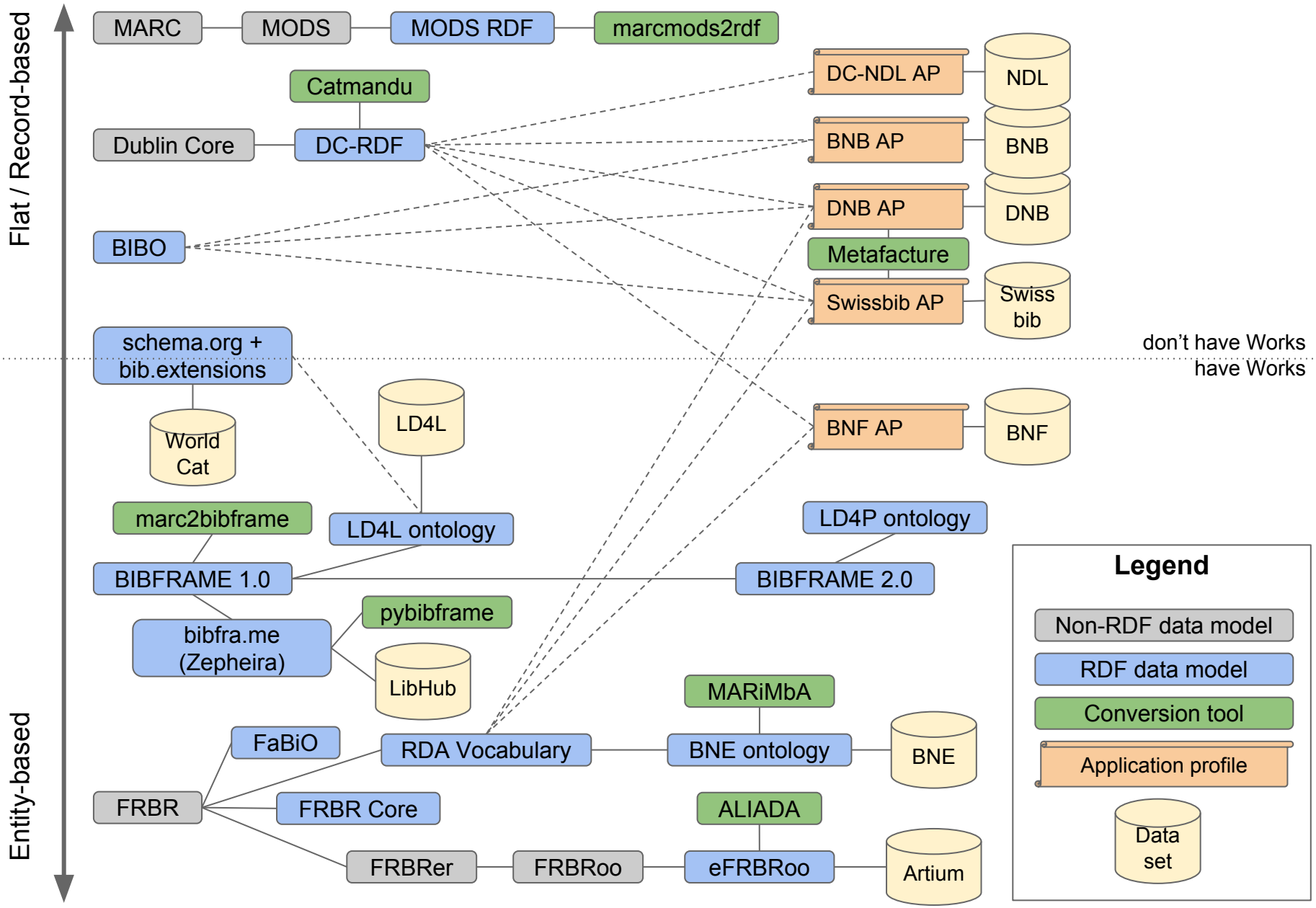
Osma Suominen and Nina Hyvönen

SWIB16, Bonn

November 30, 2016

Overview of current data models for bibliographic data

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

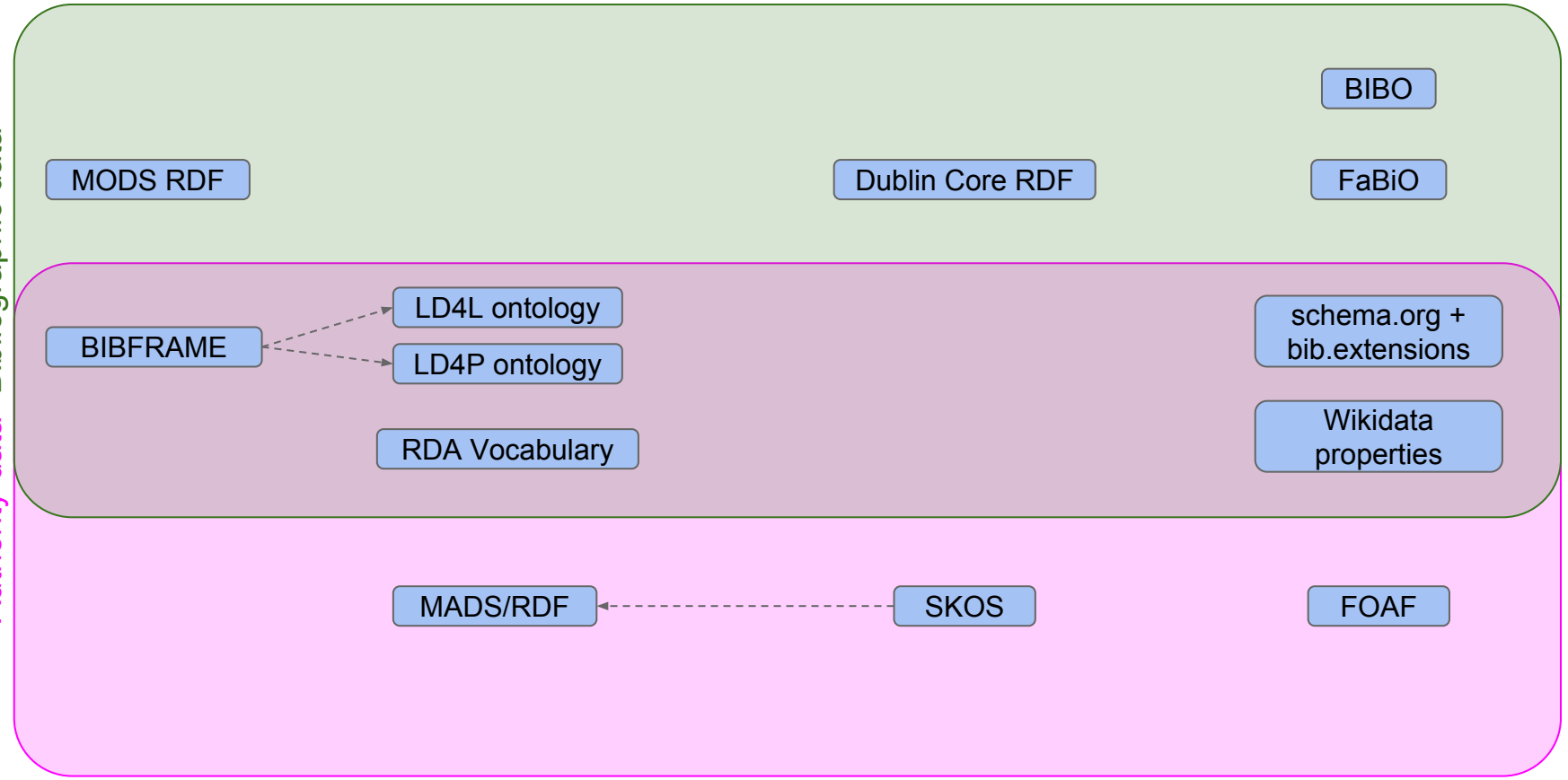


Libraryish

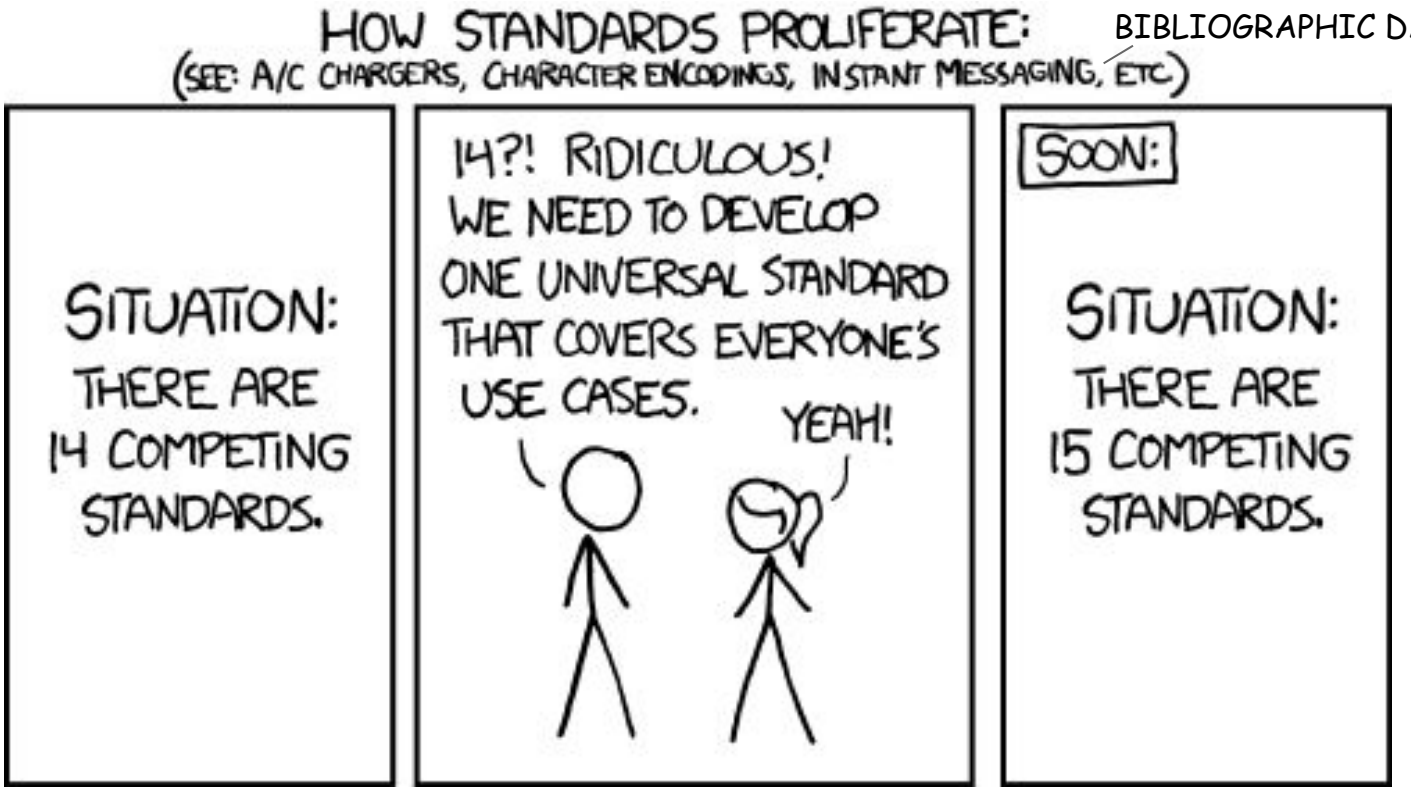
- used for **producing** and **maintaining** (meta)data
- **lossless conversion** to/from legacy formats (MARC)
- modelling of **abstractions** (records, authorities)
- **housekeeping metadata** (status, timestamps)
- favour **self-contained** modelling over reuse of other data models

Webbish

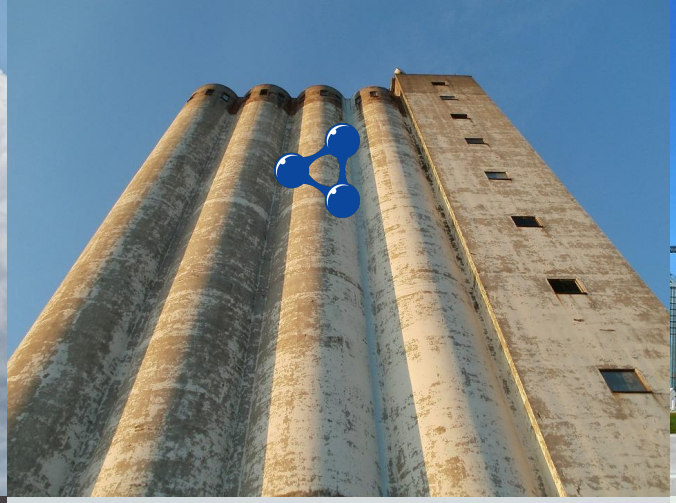
- used for **publishing** data for others to reuse
- **interoperability** with other (non-library) data models
- modelling of **Real World Objects** (books, people, places, organizations...)
- favour **simplicity** over exhaustive detail

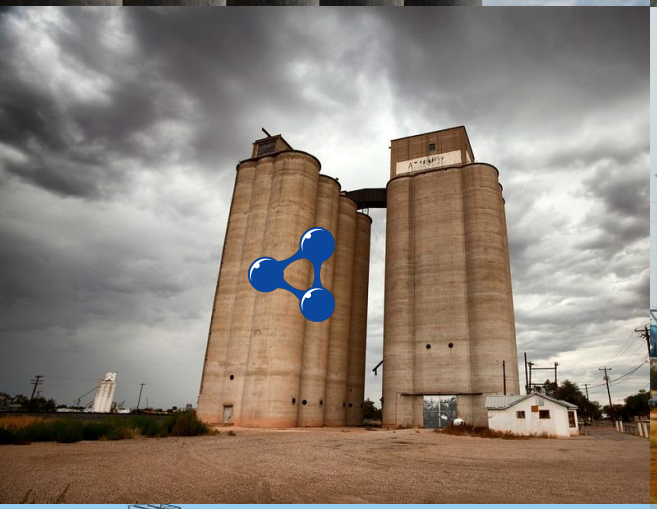


BIBLIOGRAPHIC DATA MODELS



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?

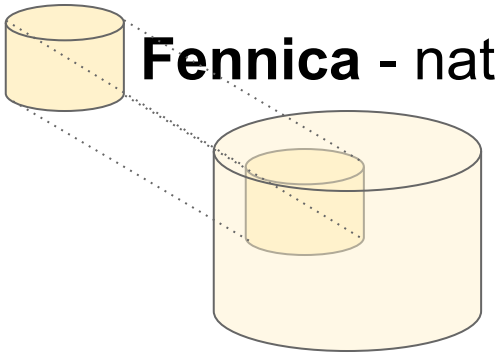
A big actor offering a compelling use case for publishing bibliographic LOD would make a big difference.

- 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.

Bibliographic data as LOD at the National Library of Finland

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

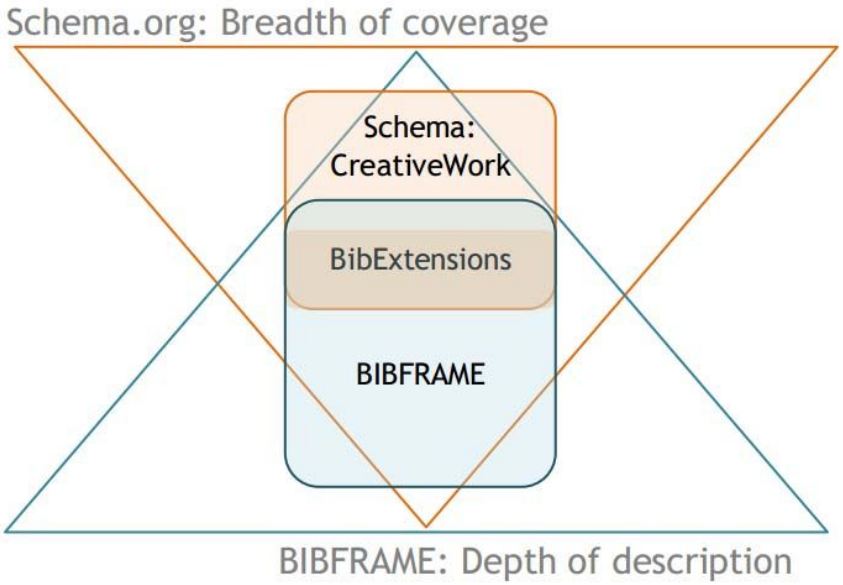
Not very Linked to start with

- 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

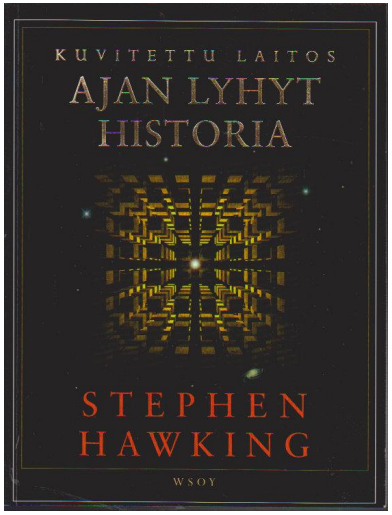
"We have these 1M bibliographic records"

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."

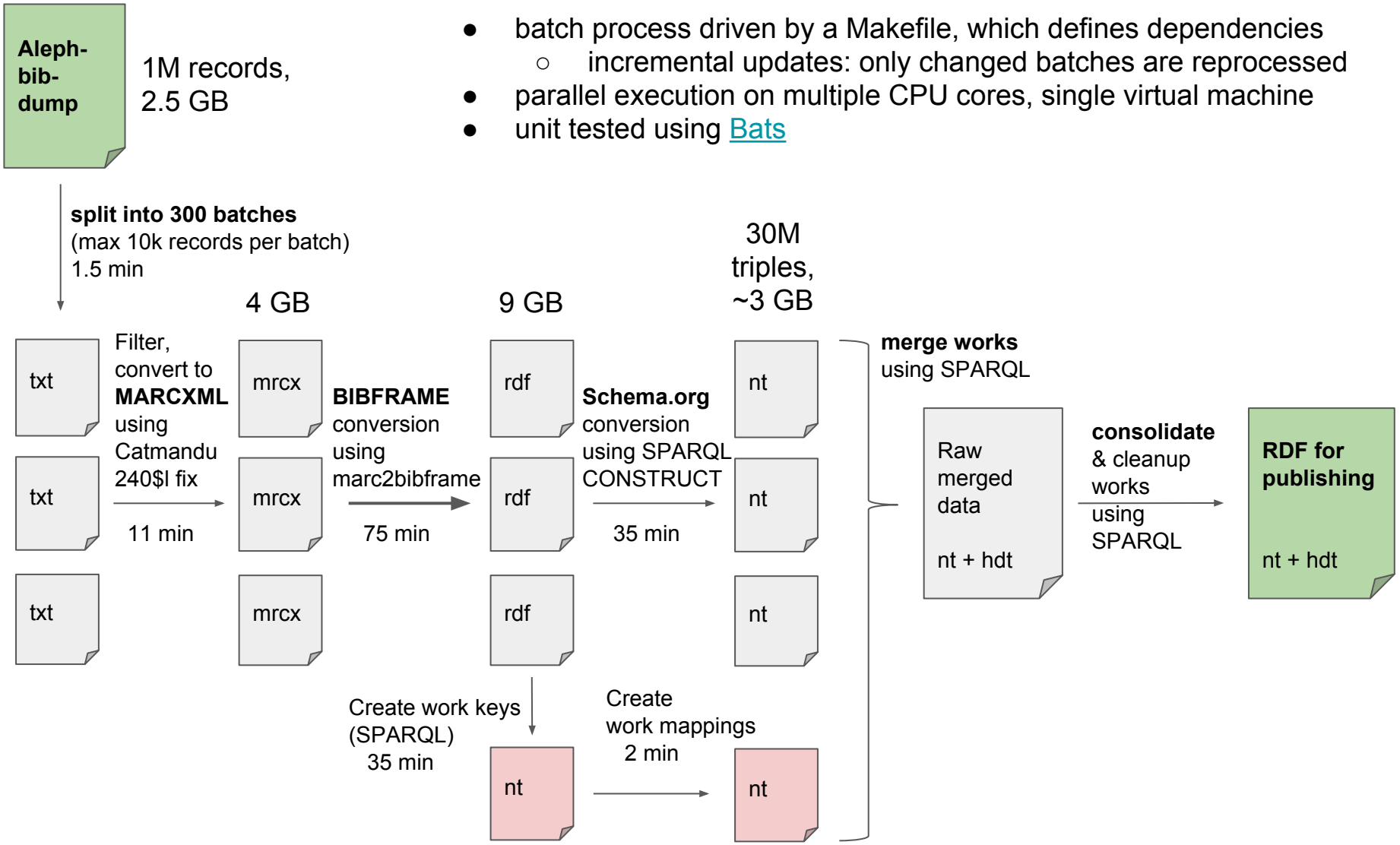
Fennica using schema.org




```
# 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 .

# 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. :" ;
  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!

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

Current challenges

1. problems caused by errors & omissions in MARC records

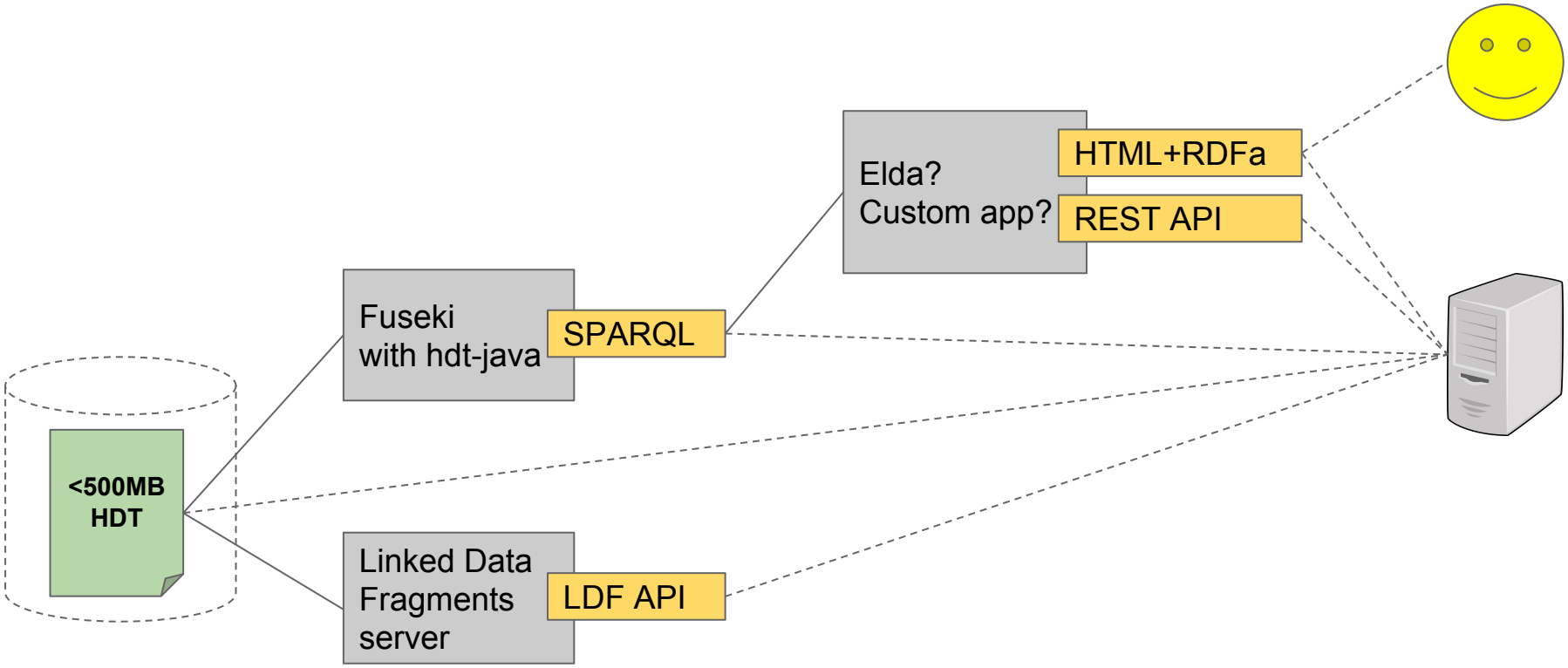
2. extracting works: initial implementation needs fine tuning
   - the result will not be perfect; establishing a work registry would help

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

4. linking internally - from strings to things
   - subjects from YSA and YSO - already working
   - using person and corporate name authorities

5. linking externally
   - 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