20 million URIs and the overhaul of the Finnish library sector subject indexing

Matias Frosterus, Jarmo Saarikko & Okko Vainonen
The National Library of Finland
SWIB19, Hamburg, Germany, 26-Nov-2019

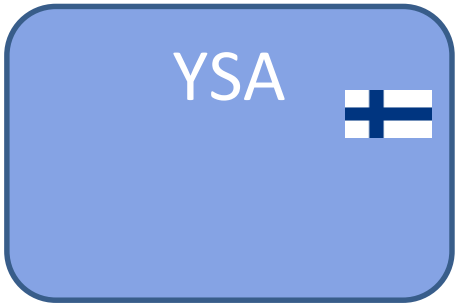
The Overhaul of subject indexing in Finnish libraries: 2019

- The goal:
- moving from monolingual thesauri to
  - multilingual,
  - machine-readable,
  - interlinked
  - SKOS vocabularies

The Overhaul of subject indexing in Finnish libraries: motivation

- The motivation:
  - Indexing in one language allows for searching in another
  - Links to other vocabularies allows for interoperability
  - Moving from terms to concepts with URIs makes updating easier

The vocabularies 1/5



- General Finnish Thesaurus YSA was the most used thesaurus in Finland
  - Developed since the 1980s
  - Used to describe all of the non-fictional literature published in Finland
  - Monolingual

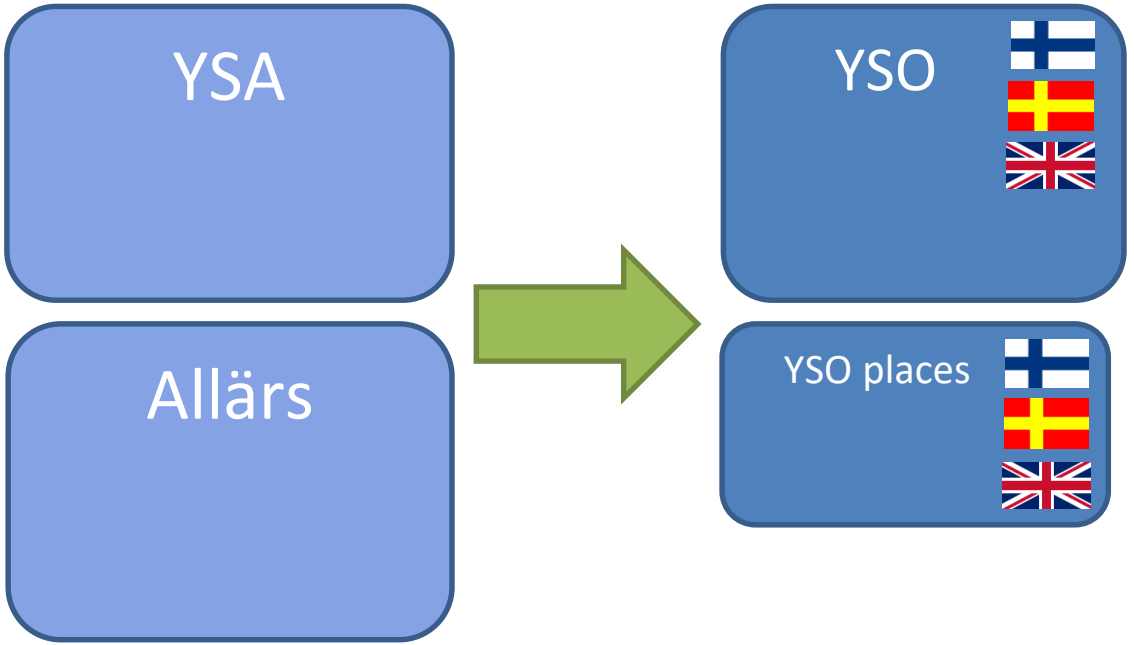
The vocabularies 2/5

- Swedish language counterpart called Allärs
  - Finnish-Swedish, to be precise
  - Very slightly different structure due to linguistic differences

The vocabularies 3/5

- In 2018 MUSA, a thesaurus of music terms was absorbed into YSA
  - Cilla, the Swedish language counterpart of MUSA, absorbed respectively into Allärs

The vocabularies 4/5

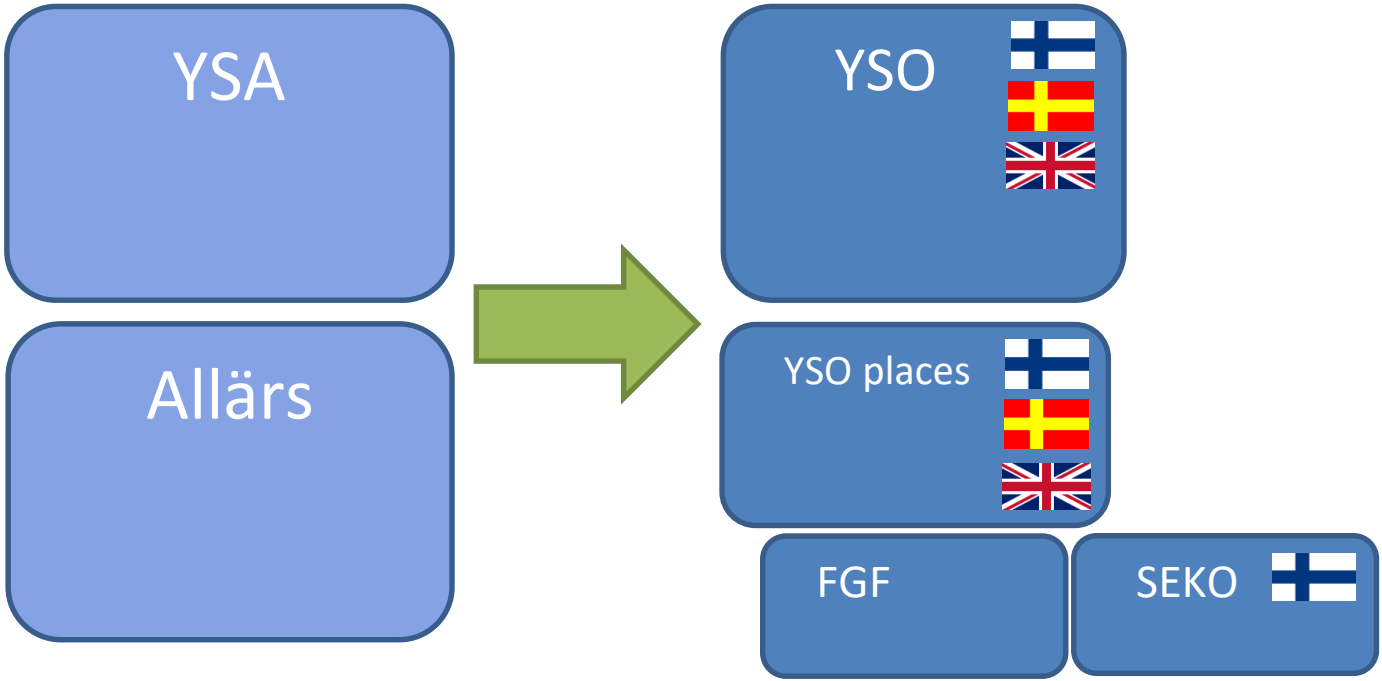


In 2003 FinnONTO research project began work on the General Finnish Ontology YSO

General Finnish Ontology YSO

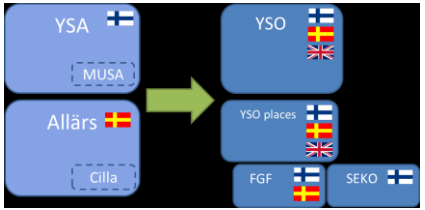
- Based on YSA and Allärs
  - Places as a separate vocabulary YSO Places
- From terms to concepts identified by URIs
- Concepts based on Finnish and Swedish
  - Translated into English
- Complete hierarchy and clearly defined semantics
- Linked
  - to Finnish ontologies of other domains
  - Library of Congress Subject Headings, Wikidata

The vocabularies 5/5



Two more vocabularies for the conversion

Scope expanded

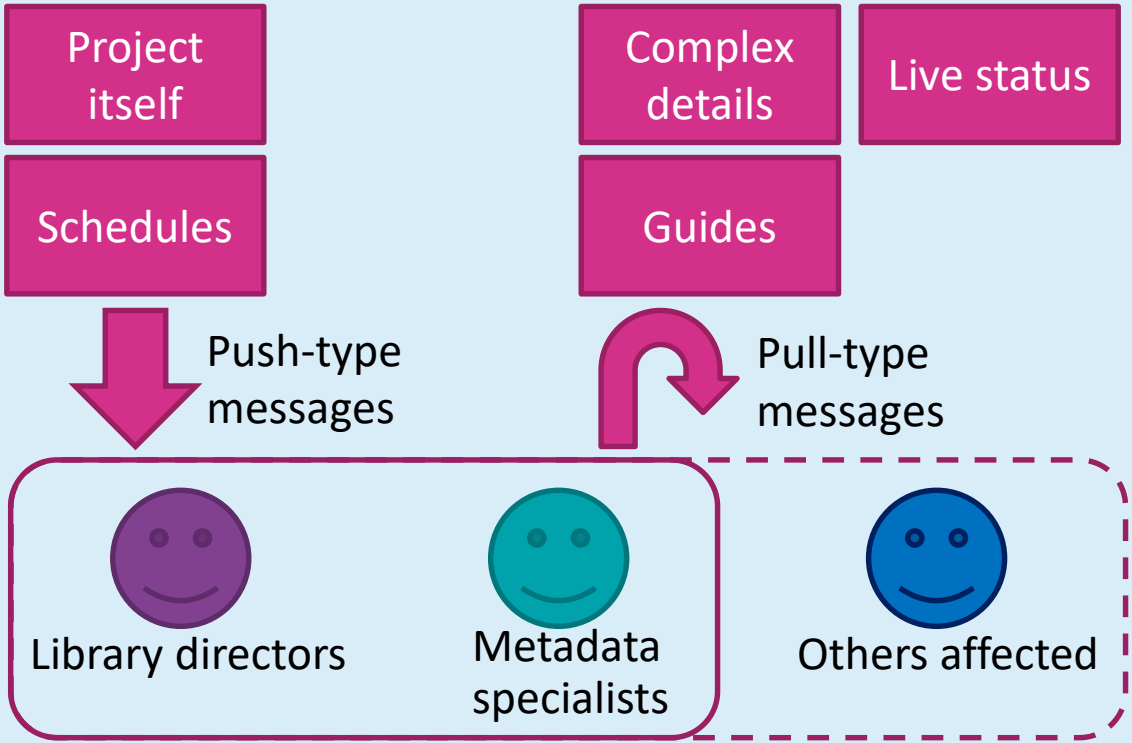


R | D | A
Resource Description & Access

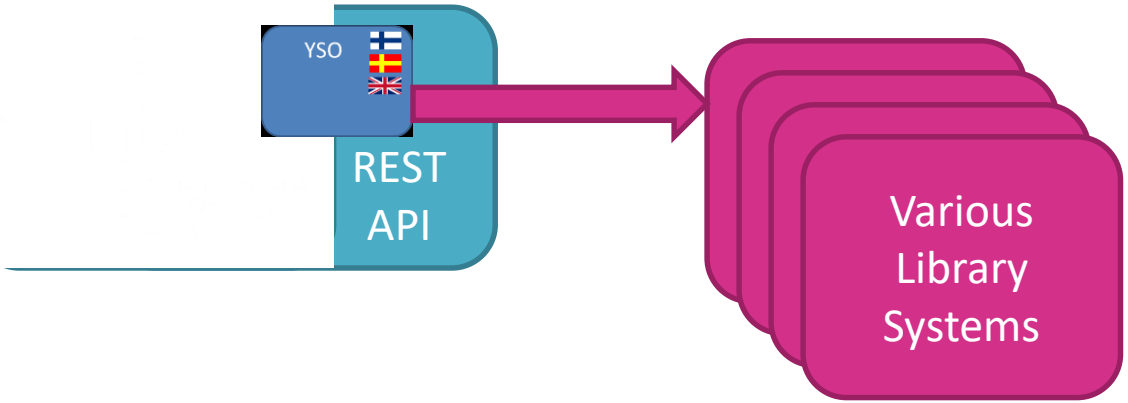
MARC 21

- Many vocabularies
- Dismantling subfields used in subject indexing ”chains”
- New MARC fields

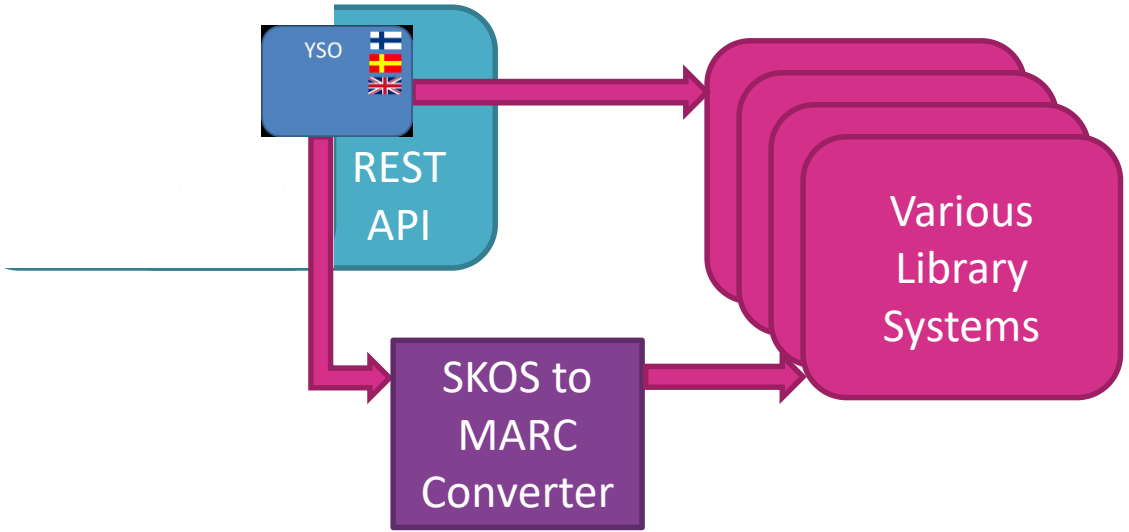
Lessons learned: Communication



Conversion: Authority Records 1/2



Conversion: Authority Records 2/2



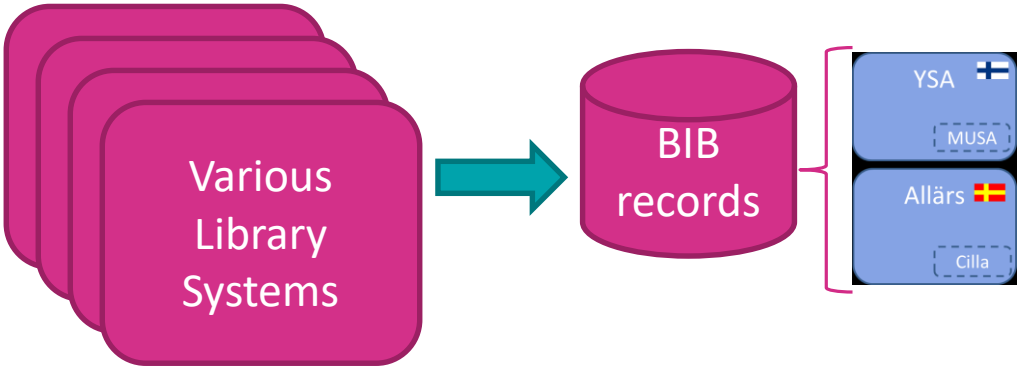
SKOS Record for yso:p16239

```
yso:p16239
 a skos:Concept, <http://www.yso.fi/onto/yso-meta/Concept> ;
 skos:prefLabel "morgon"@sv, "aamu"@fi, "morning"@en ;
 skos:broader yso:p5264 ;
 skos:exactMatch koko:p17356, ysa:Y109535, allars:Y23054 ;
 skos:closeMatch
<http://id.loc.gov/authorities/subjects/sh2004006540> ;
 dc:modified "2017-05-10"^^xsd:date ;
 skos:inScheme yso: .
```

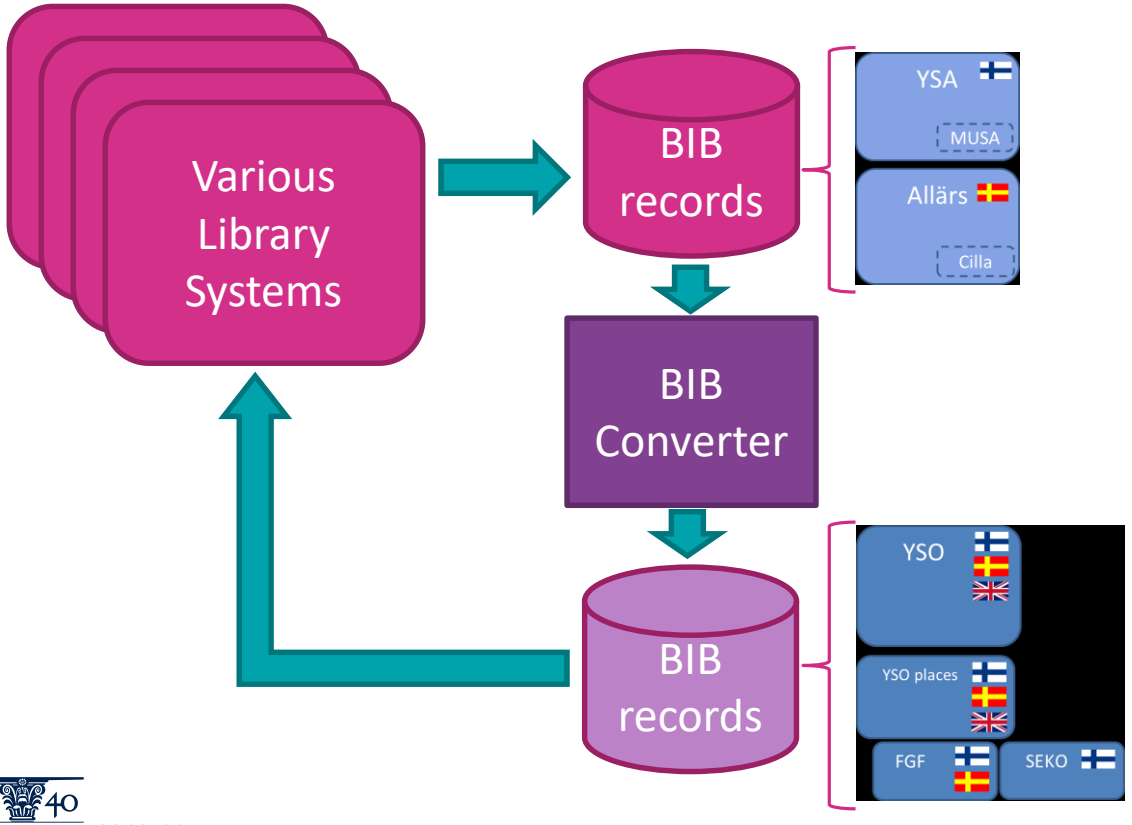
MARC Authority File for yso:p16239

FMT	AU
LDR	00000cz a2200000n 4500
001	000226463
003	FI-NL
005	20190522204644.0
008	800101|n|anznnbabn | ana
0247	|a http://www.yso.fi/onto/yso/p16239 |2 uri
040	|a FI-NL |b fin |f yso/fin
065	|a 06 |c Tähtitiede. Astronomia. Avaruustutkimus |2 yso |0 http://www.yso.fi/onto/yso/p26588
150	|a aamu |2 yso/fin |0 http://www.yso.fi/onto/yso/p16239
550	|w g |a vuorokaudenajat |2 yso/fin |0 http://www.yso.fi/onto/yso/p5264
688	|a Luotu: 1980-01-01
688	|a Viimeksi muokattu: 2017-05-10
750 7	|a morgon |4 EQ |2 yso/swe |0 http://www.yso.fi/onto/yso/p16239
750 7	|a morning |4 EQ |2 yso/eng |0 http://www.yso.fi/onto/yso/p16239
750 0	|a Morning |4 ~EQ |0 http://id.loc.gov/authorities/subjects/sh2004006540
CAT	|a LOAD-YSO |b 00 |c 20190522 |l FIN10 |h 2046
SYS	000226463

Conversion: Bibliographic Records 1



Conversion: Bibliographic Records 2



Two sets of rules

- An expert group made up of indexing specialists from various national groups and libraries
- Two sets of rules
  - SKOS to MARC for authority records
  - BIB conversion rules
    - Separate rules for fiction and non-fiction and music/film due to different indexing rules

SKOS to MARC Converter

BIB Converter

Dismantling subfields in subject indexing

- New subject indexing rules use only one subfield for each term
  - Existing records had not been converted
- All in all proved to be a very complex task
  - Same MARC fields and subfields but different conventions for different types of content
  - Specific “labels” that changed the meaning of subfields
  - The conventions had changed over time and older ones were difficult to re-engineer

Example of Conversion

650#7 **|a** hard rock **|z** Finland **|y** 2000-2009 **|2** allars

The publication **is about** Finnish rock music

648 #7 **|a** 2000-2009
650 #7 **|a** hard rock **|2** yso/swe **|0** http://www.yso.fi/onto/yso/p29778
651 #7 **|a** Finland **|2** yso/swe **|0** http://www.yso.fi/onto/yso/p94426

The publication **is a** music score, recording or video

370 #7 **|g** Finland **|2** yso/swe **|0** http://www.yso.fi/onto/yso/p94426
388 1# **|a** 2000-2009
655 #7 **|a** hard rock **|2** slm/swe **|0** http://urn.fi/URN:NBN:fi:au:slm:s828

Coverage of the conversion

- National union catalog **Melinda**
- Local library databases employing various library systems (Voyager, Koha, Axiell Aurora, etc.)
  - Both universities and public libraries
- Other systems that were using YSA/Allärs
  - E.g., government institutions

Lessons learned: Unwritten conventions

- History has a tendency to accumulate
- Including experts widely is key

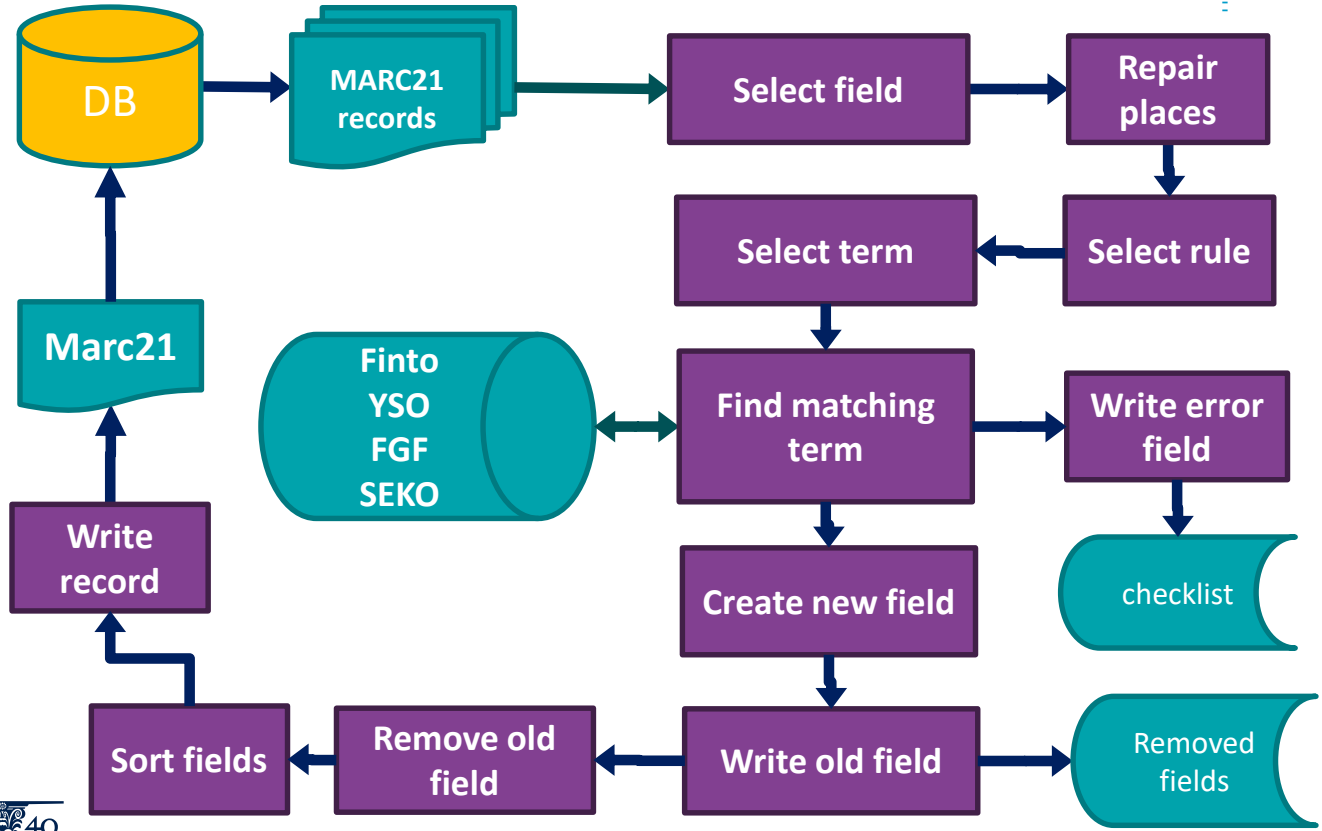
Coding the conversions

- 2 programs
  - SKOS to MARC authorities
  - Changing terms in MARC BIB-records
- Open source Python3 code
- Available to libraries and library system providers
  - https://github.com/NatLibFi/Finto-data/tree/master/tools/finto-skos-to-marc
  - https://github.com/NatLibFi/yso-marcbib

Lessons learned: Complexity of programming

- **Original plan**
  - Take each term and switch it to the label of the same concept in the other vocabulary
- **Reality**
  - Metadata in data
    - Meanings of terms were interdependent
    - Content type affected the use of MARC fields
  - Many analyses had to be done before selecting the ”correct” term

Process of BIB-record conversion



Conversion of MARC BIB records

- Conversion analyzed fields:
  - 648, 650, 651, 655
  - Field was analyzed only if subfield **|2** value was **ysa, allars, musa** or **cilla**
- Conversion created fields:
  - **257**, **370**, **382**, **388**, 648, 650, 651, **653**, 655
- For YSO and FGF terms we also added language independent concept URIs to the **|0**-subfield

Finding place subfields first

- Identify and concatenate place subfields that are concatenated in the vocabulary (e.g. city districts)

- 650#7 |aJAZZ **|zHelsinki |zEira**
  Search for ”**Helsinki - - Eira**” label in the SKOS-vocabulary

Repair places

Coding the conversion: matching the concepts

Repair places

YSA: Helsinki -- Eira (fi)

yso-paikat: Eira (Helsinki), **Allärs:** Helsingfors -- Eira (sv)

http://www.yso.fi/onto/ysa/Y116934

ysa:Y116934 skos:exactMatch yso:p116934 .

Eira (Helsinki)

Eira (en), **Eira (Helsingfors)** (sv)

http://www.yso.fi/onto/yso/p116934

370## |g Eira (Helsinki) |2 yso/fin |0 http://...
370## |g Eira (Helsingfors) |2 yso/swe |0 http://...

Using the subfield 8 to identify connected terms

- Example of a symphony composed in 1900 and performed in 2019

Create new field

650#7 |a sinfoniat |y 1900 |z Helsinki |2 ysa
650#7 |a sinfoniat |y 2019 |z Wien |2 ysa
650#7 |a sinfoniaorkesterit |2 ysa

370#7 |81\u |g Helsinki |2 yso/fin |0 http://www.yso.fi/onto/yso/p94137
370#7 |82\u |g Wien |2 yso/fin |0 http://www.yso.fi/onto/yso/p106956
382#1 |a sinfoniaorkesteri |2 seko |0 http://urn.fi/urn:nbn:fi:au:seko:00936
388#7 |81\u |a 1900 ‡ 2yso/fin
388#7 |82\u |a 2019 ‡ 2yso/fin
655#7 |81\u |82\u |a sinfoniat |2 slm/fin |0 http://urn.fi/URN:NBN:fi:au:slm:s917

- MARC21 subfield 8 links all related fields
- Years are not (yet) authorized in Finnish thesauri

Sorting the fields

- We tried to keep the original order of first occurence of terms
- New fields were sorted according to field number, 2nd indicator, vocabulary identifier
- We checked and removed any duplicate fields

Lesson learned: MARC

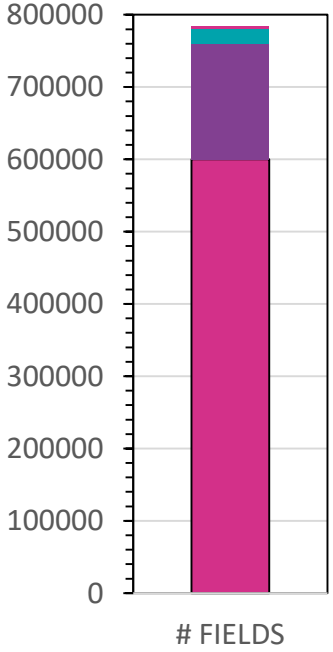
- Library systems did not automatically index the new fields
- Multiple language support (yso/fin, yso/swe)
  - Vocabulary identifier with a language qualifier
  - Confirm that systems support this
- Reserve enough time for testing
  - New conversion rules (SEKO-terms) were added at a very late stage of the process

Results of BIB-conversion of Melinda union catalogue

- **15** million records (about half are siblings)
  - **10** million records without terms from the four vocabularies → **no action**
- **4,9** million records were converted
- **23** million fields removed
- **45** million fields added
  - **22** million YSO and FGF terms were added in two languages
  - **<1** million SEKO terms to field 382

Non-converted terms

800,000 terms out of the 23 million removed fields were not converted to new terms



- Miscellaneous cases 3.400
- Removed helping terms 21.000
- Terms with multiple targets 160.000
  Manual editing needed
- Terms not found in vocabulary 600.000

Terms with multiple targets

finto

Find matching term

- Example of multiple matches for ”**ohjaus**”
  - **ohjaus** (hallinta) – **control** (steering)
  - **ohjaus** (neuvonta) – **direction** (instruction and guidance)
  - **ohjaus** (taiteet ja media) – **direction** (arts and media)
- Same entry term in multiple concepts
- Matching done with normalized terms
  - **CHAMPAGNE : Champagne** (place) vs. **champagne** (wine)
- ➔ manual corrections

Non-converted terms: Create new field without identifiers

- If the term was **not found** in the thesauri
  - Move the term to field 653
  - Set the 2nd indicator according to field/subfield

Create new field

- If term was found but **not exact string** OR
  If **multiple matches** in the target thesauri
  - Keep the term in the same field
  - Remove subfield **|2** identifier
  - Set the 2nd indicator to ”**4**”

Lessons learned 1/2

- Term normalization and use of multiple languages
  - **wrong matches** was considered a low risk
- Logfiles: removed fields, written fields
  - A thirde, more **complex logfile** was needed for conversion error tracking, e.g. when terms disappeard
- Multiple matches
  - **Manual editing** before and after conversion
- **Subfield 8** used for connected concepts was unnecessary in most cases
- **Deduplication** of fields did not always go through

Lessons learned 2/2

- Document the ”unwritten” subject indexing conventions
- Remove the old authority files so that they are not used any more

Thank you!

https://www.kiwi.fi/display/ysall2yso

finto-posti@helsinki.fi

Image by Bessi from Pixabay