

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



KANSALLISKIRJASTO

Matias Frosterus, information systems manager,

<https://www.helsinki.fi/en/people/people-finder/matias-frosterus-9131705>

Jarmo Saarikko, information specialist DOI: <https://orcid.org/0000-0002-6801-6151>

Okko Vainonen information systems specialist,

<https://www.helsinki.fi/en/people/people-finder/okko-vainonen-9378633>

The National Library of Finland

Finto service <https://finto.fi/>

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: 2019


- 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




- 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



YSA 

Allärs 

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

 16340
KANSALLIS KIRJASTO

2019-11-26 SWIB19 Frosterus, Saarikko & Vainonen

5


Finland has two official languages:


Finnish (fi-FI) and Swedish (sv-FI).

These vocabularies contained around 36,000 main terms and 20,000 entry terms. In the autonomous administrative region Åland, Swedish is the only official language.


In the Sami-region in Northern Finland the Sami languages may be used. "Act on the use of the Sami language before the authorities" (516/1991 English)

The vocabularies




YSA 

MUSA

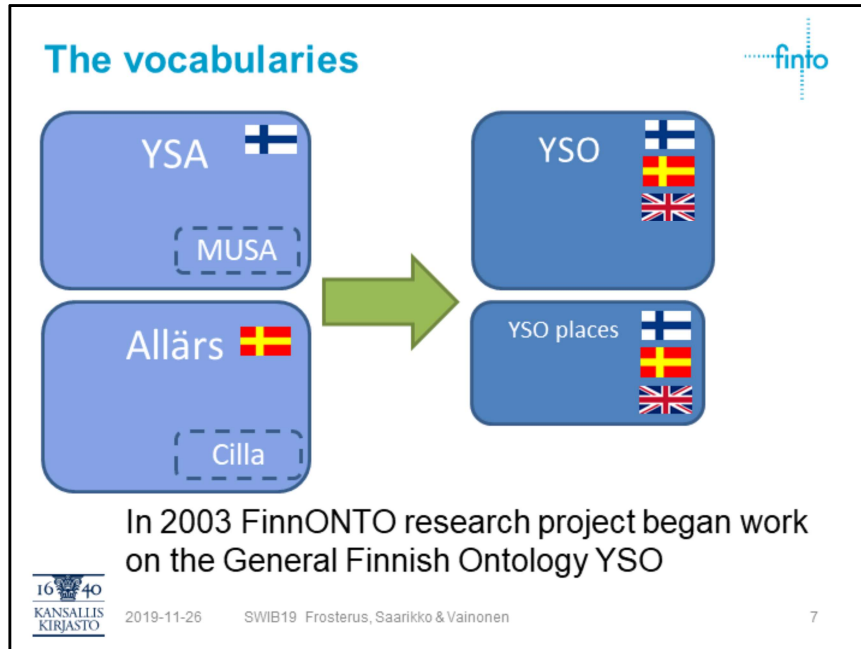
Allärs 

Cilla

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

 2019-11-26 SWIB19 Frosterus, Saarikko & Vainonen 6

Musa and Cilla covered about 950 concepts



National Semantic Web Ontology Project in Finland (FinnONTO), 2003-2012
<https://seco.cs.aalto.fi/projects/finnonto/>

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



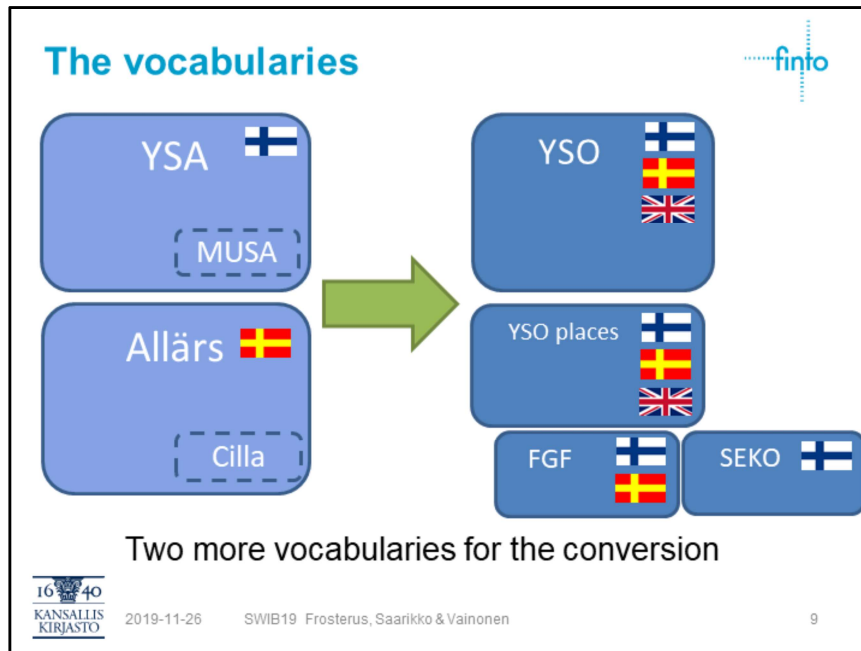
2019-11-26

SWIB19 Frosterus, Saarikko & Vainonen



8

Library of Congress Subject Headings (13.203 concepts)
Wikidata (5336 YSO-places concepts)
YSO-links in 16 Finnish domain ontologies
(numbers as of November 2019)



FGF – Finnish Genre and Form vocabulary SLM (1282 concepts)
SEKO – musical instruments, voices and ensembles (1241)

To make the conversion we had to guarantee that there was at least one skos:exactMatch from the old vocabularies to the new. Musa terms had exactMatch with YSA terms and then from that concept an exactMatch to YSO.

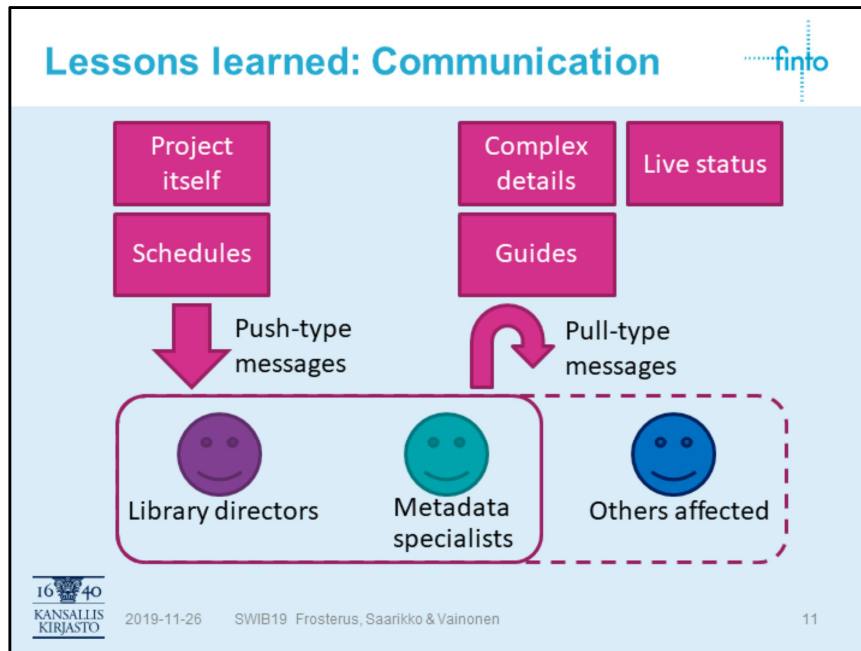
Scope expanded



R | D | A
Resource Description & Access



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



Push type messages:

- Project itself and its motivation; Schedules

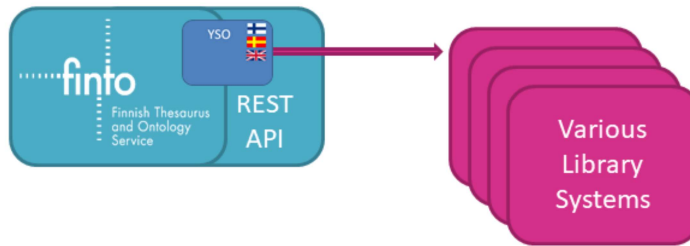
Pull type messages:

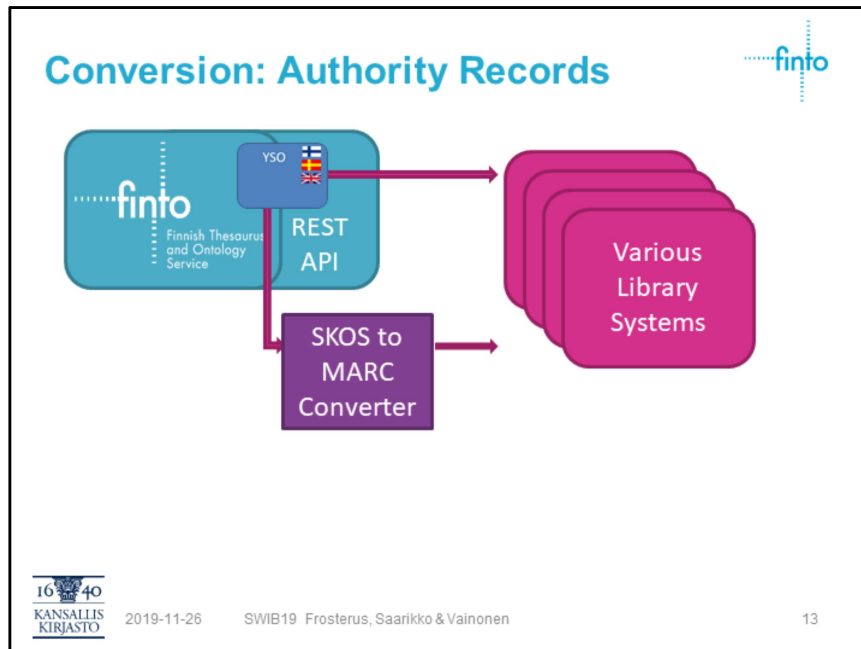
- Complex things and Details; Guides; status updates

Various audiences:

- Library directors
- Metadata specialists
- Others affected (who? We're not sure: difficult to reach – basically just a banner in Finto which is the service housing the vocabularies)

Conversion: Authority Records





From SKOS into MARC

- To support the subject indexing processes in the union catalogue
- One record for each preferred term in each language.
- Concept URI and the vocabulary id in the 024 field

Challenges

- Multilingual terms and language identification in MARC
 - Implemented as a qualifier: yso/fin, yso/swe
 - One record for each language
 - Linking to the other language terms
- Finding the correct terms for the record
- Updating all related records when labels change
 - BT, NT, RT records which include the changed term

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: .
```



KANSALLIS
KIRJASTO

2019-11-26

SWIB19 Frosterus, Saarikko & Vainonen

14

Triples for one concept

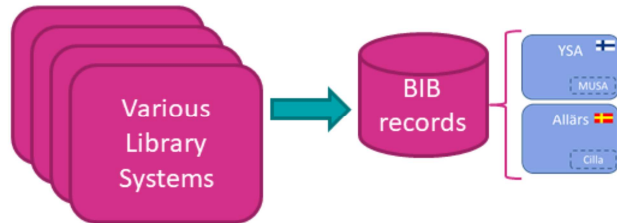
MARC Authority File for yso:p16239

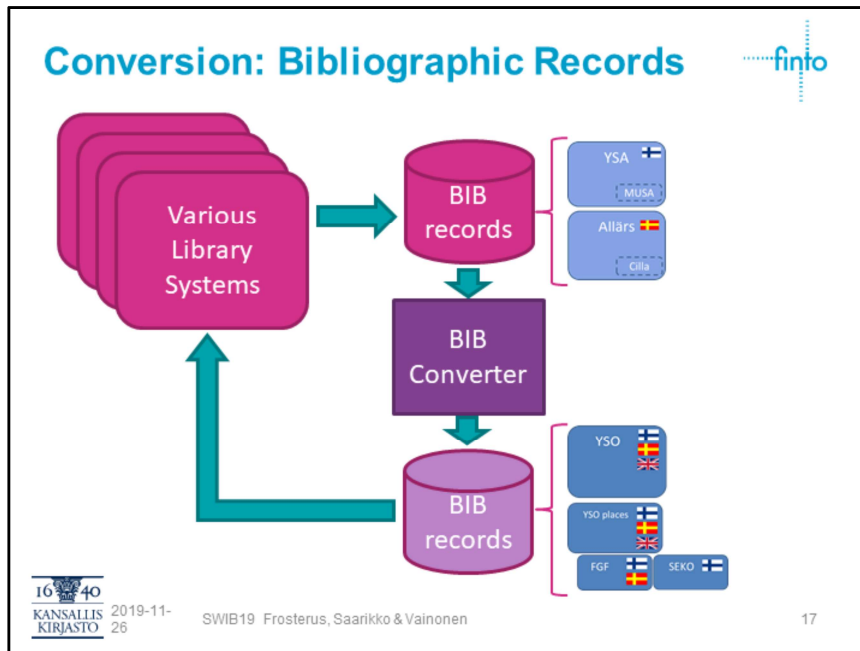


FMT	AU
LDR	00000cz a2200000n 4500
001	000226463
003	FI-NL
005	20190522204644.0
008	800101 n anznbnabn ana
0247	a http://www.yso.fi/onto/yso/p16239 2 uri
040	a FI-NL b fin f yso/fin
065	a 06 c Tähtitiede. Astronomia. Avaruustutkimus 2 yso 0 http://www.yso.fi/onto/yso/p26588
150	a aamu 2 yso/fin 0 http://www.yso.fi/onto/yso/p16239
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

MARC authority file created from the same concept
<http://www.yso.fi/onto/yso/p16239>

Conversion: Bibliographic Records





From YSA and Allärs annotations to YSO annotations with URIs
 From subfields to individual fields → removing term chains

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>

For YSO and SLM terms we also added language independent concept URIs to the |0-subfield
The fields were repeated for each language:
implemented as a qualifier: yso/fin, yso/swe, notice the same URI

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

<http://melinda.kansalliskirjasto.fi/>

Lessons learned: Unwritten conventions



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

- We wanted to preserve all the information and that proved to be quite challenging
- Fitting the modern linked data paradigm to library systems using MARC21 is also a challenge
- Sitting with the group of experts in meetings trying to memorize the unwritten conventions was not as efficient as in a workshop looking at existing content descriptions.
- New undocumented “indexing rules” kept popping up from the memories over several casual discussions with experts.

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>



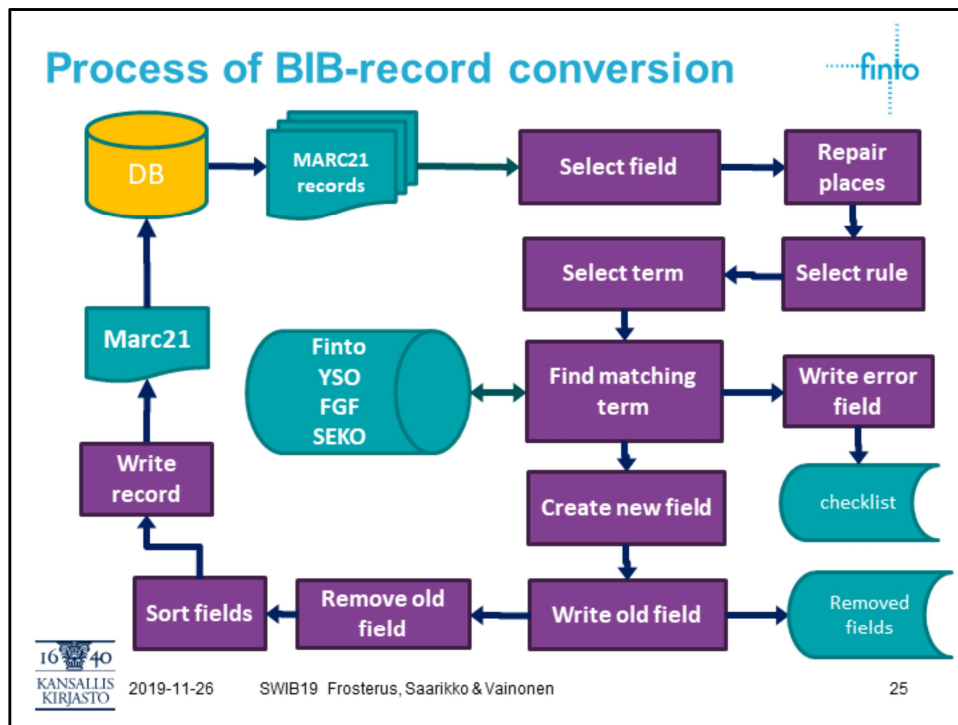
Lessons learned: Complexity of programming



- **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

E.g. the type of data was submerged within the data. Music genre was in the first subfield of 650.

- For fiction, movies and games, the genre terms were matched only with SLM (FGF).
- In non-fiction literature it was impossible to decide whether the term was expressing genre as the type or genre as the subject.



- Select field: Analyse only fields 385, 567, 648, 650, 651, 655
- Select field: Read field only if |2 subfield in [ysa|allars|musa|cilla]
- Repare places: Find place term combinations
- Select term: Find "special terms"
 - Special stop terms that affected other subfields: aiheet (subjects), musiikki (music), fiktio (fiction)
- Find matching term: Analyze all terms without diacritics & normalize
 - if a direct match was not found, both the vocabulary and the description term were normalized
- Create new field: Move all terms to separate fields (RDA)
 - Terms were moved to various fields according to content type, field and subfield
 - Time terms in the 648 field and \$y subfields were moved to 388 and 648 with only the |2 vocabulary identifier without links in the |0 subfield as the time concepts are not yet authorised in the vocabulary.

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

Select field

From multiple term subfields

- to one term subfield per field

New MARC fiels are being taken into active use

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

```
Helsingin seutukunta
Helsinki
-Aleksanterinkatu -- Helsinki
-Bulevardi -- Helsinki
-Eteläsatama
-Helsinki -- Ala-Malmi
-Helsinki -- Alppiharju
-Helsinki -- Alppila
-Helsinki -- Arabianranta
-Helsinki -- Aurinkolahti
-Helsinki -- Eira
-Helsinki -- Eläintarha
-Helsinki -- Faltipakka
-Helsinki -- Haaga
-Helsinki -- Hakaniemi
-Helsinki -- Hakuninmaa
```


Two consecutive place subfields in data could be one concept in the current YSA vocabulary (geographical concepts)

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://...


 KANSALLIS
KIRJASTO

2019-11-26 SWIB19 Frosterus, Saarikko & Vainonen

28

Because here the record type was music, the term jazz would be put to field 655 and the location to 370.

Before conversion, we had to make sure that there was at least one exactMatch for each concept in the old vocabularies.

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/ysop94137>

370#7 |82|u |g Wien |2 yso/fin |0 <http://www.yso.fi/onto/ysop106956>

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



2019-11-26

SWIB19 Frosterus, Saarikko & Vainonen

29

Hypothetical example

Terms are moved from field 650 to 370, 388 and 655.

The subfield 8 links the terms which were together in the same original field

Sorting the fields



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

Sort fields

The order of terms in the record had been used to indicate the importance of the subjects

Lesson learned: MARC



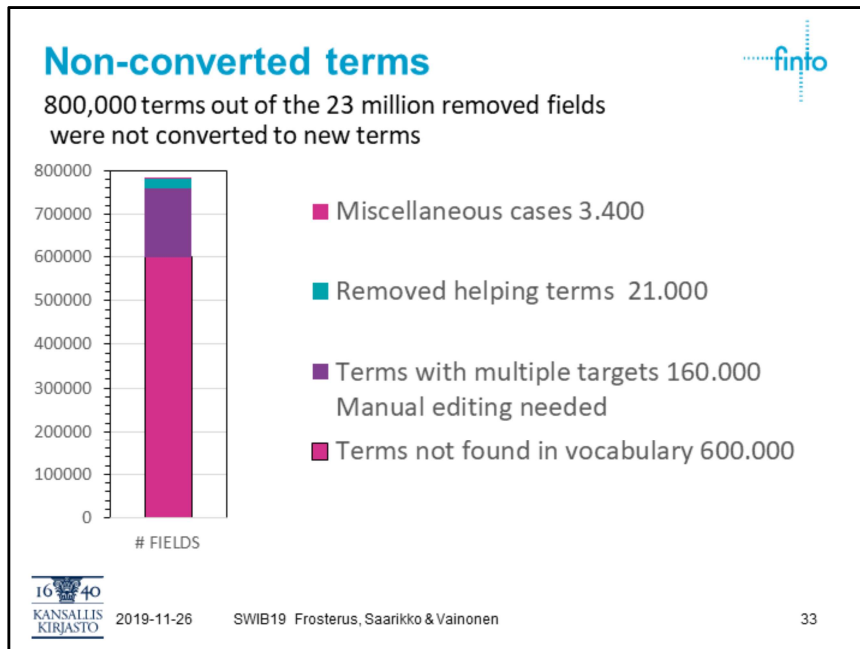
- Library systems did not automatically index the new fields
- Multiple language support (yso/fin, yso/swe)
 - Vocabulary identifier with a language qualifier
 - Confirm that systems support this
- Reserve enough time for testing
 - New conversion rules (SEKO-terms) were added at a very late stage of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

- Code of the conversion programme was corrected and updated according to any errors found in the logfiles.
- Vocabulary identifier qualifier was not supported by some library systems at first.
- Also the SKOSMOS API needed a new variable to provide the vocabulary identifier for MARC systems
- First conversion with the largest dataset, however, provided valuable feedback for error and term handling before converting the local databases.
- The music records are indexed with quite different conventions so that, originally, a separate conversion was planned at a later date. However, for administrative reasons, we had to push and write the conversion rules for music records and include their conversion in the same conversion schedule. So the time after finishing coding of the conversion programme to the actual conversion time window was thus dropped from three months to few weeks.

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

Conversion of bibliographic records in the national union catalogue Melinda



Terms not found in the vocabularies :

- 600,000 terms in 435,000 records
- manual correction of all these is not possible
- YSO or SLM – matching was depended on the field and subfield.
- If the subfield was |v or the field was 655, terms were converted only to SLM terms – irrespective of the term being a legal YSA-term.
- 170,000 different terms
 - 135,000 very rare terms with only 1-2 fields;
 - 300 very common terms with more than 100 fields
 - The most common cases were genre and form terms not included in the SLM vocabulary


Terms with multiple matches in the ontology :

- 160,000 terms (1,400 different terms)
- e.g. term had been changed by adding a qualifier in parentheses, etc.

Minor issues : 40,000 terms

- e.g. "helper" terms - like 'music' for music data - were removed
- Terms with an erroneous sufield label, etc.


Terms with multiple targets



- Example of multiple matches for ”ohjaus”

Find matching term

 - **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



16340
KANSALLIS
KIRJASTO

2019-11-26 SWIB19 Frosterus, Saarikko & Vainonen

34

Multiple targets in the mappings between thesauri and ontologies

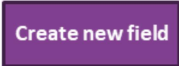

- Relations of 1 – many, (same skos:altLabel in several concepts)
- This was due to differences in the division of concepts and terms between languages

These terms have to be checked and updated by humans


Text normalization

- some terms would become equal if normalized, although meaning was different. Especially if the term was written in upper case.
- Capitalization, e.g. Kemi (a city) vs. kemi (chemistry in Swedish)

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

 2019-11-26 SWIB19 Frosterus, Saarikko & Vainonen 35

Conversion programme had to handle many types of errors in the original data

- Misspelling
- Terms in a wrong field or subfield
- Terms with a wrong vocabulary identifier
- Wrong language etc.
- Some subfield codes had changed during former MARC conversions

Lessons learned



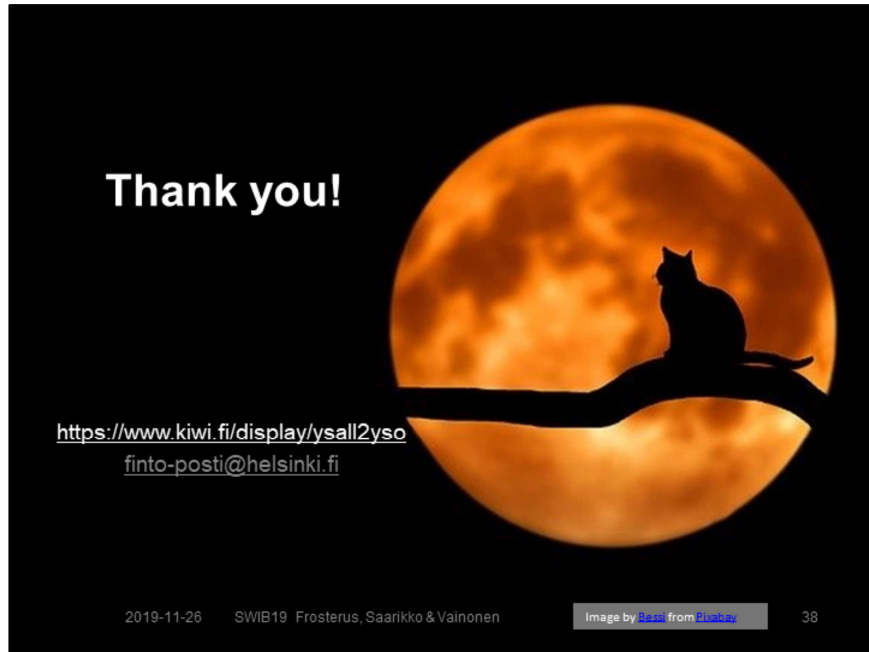
- Term normalization and use of multiple languages
 - **wrong matches** was considered a low risk
- Logfiles: removed fields, written fields
 - A third, more **complex logfile** was needed for conversion error tracking, e.g. when terms disappeared
- Multiple matches
 - **Manual editing** before and after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

- We created two logfiles: removed fields and new fields.
- If some terms disappeared from records it was still difficult to locate what had happened.
- A third logfile for non-converted terms and other errors was created
 - Listed the record-id, record type, the term, the old field and new field all on the same line.
- Listing of removed fields with the corresponding new field was not always possible to match if the same concept term was in the record in two languages and the code was both creating new fields and removing duplicates at the same time.

Lessons learned

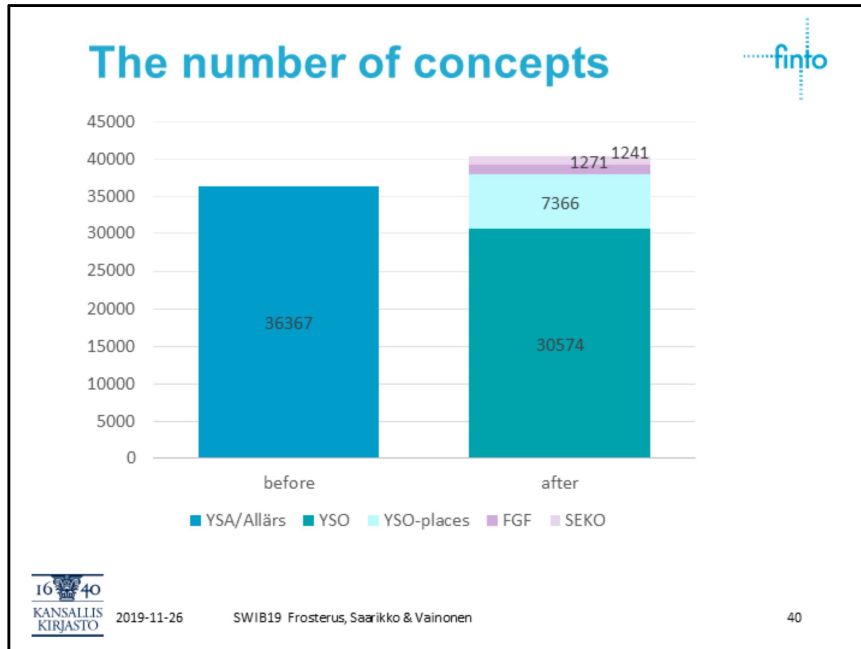


- Document the "unwritten" subject indexing conventions
- Remove the old authority files so that they are not used any more



More information by email: finto-posti at helsinki.fi
Project website: <https://www.kiwi.fi/display/ysall2yso>

Additional slides



The number of concepts in YSA in spring 2019.
 The number of concepts in YSO, YSO-paikat, SLM and SECO in autumn 2019

Coding the conversion: specific rules



- In music type (recordings, sheet music, scores)
 - Check terms with SEKO vocabulary of instruments and ensembles
 - If found, move term to field 382

Analysing the terms: examples of music records



- 650#7 |a jazz |y 1930 |z Helsinki |2 ysa
 - 1st term is in the genre vocabulary
 - Time and place are of creation or performance
- 650#7 |a konsertit |y 1930 |z Helsinki |2 ysa
 - 1st term is not a genre term
 - All fields are subjects
 - Exception: instruments in SEKO were put to 382

In music type of records:

If the first concept in field 650 is in the genre vocabulary SLM

- interpret rest of the terms as linked to the creation of work
- the musical work was performed in the year 1930 in Helsinki

If the first concept is not in the genre vocabulary SLM

- Terms are the subjects of the work about concerts in the year 1930 in Helsinki

In cinema type of records, the location was the producer's country and the term was put to field 257

**Most common types:
in terms moved to 653**



293 most common terms (>100)

▪ subjects	81
▪ forms	75
▪ places	75
▪ organisations	21
▪ works	11
▪ persons	8
▪ SEKO-term	7
▪ products	3
▪ errors	3
▪ term in English	6
▪ term in German	3

Reasons:

- misspelling
- wrong language
- wrong field
- term missing from vocabulary

 2019-11-26 SWIB19 Frosterus, Saarikko & Vainonen 43

From the Melinda conversion:

- The 293 most common terms that were not found in the corresponding vocabularies divided according to the type of concept.
- These terms were moved to field 653.
- Some of these may be added to the new vocabularies and new concepts or as entry terms.