Thesauri Quality Assessment:
Analyzing the Rijksmuseum Library Thesaurus

By: Daan de Ruijter
Supervised by: Jacco van Ossenbruggen &
Chris Dijkshoorn

Context

The Rijksmuseum Library thesaurus is currently maintained in the MARC format

- → This is a somewhat **dated format**
- → The thesaurus has **manual data entry**
- → Entries have a **lack of quality assurance**

How can we assess the quality of such a thesaurus?

Research Questions

1. What **changes** are caused **when converting** from MARC to SKOS?
2. What are the **quality issues** of the thesaurus expressed in SKOS?
3. How well can the Rijksmuseum collection and library thesauri be **aligned with** each other?

Research Layout

1. Convert MARC to SKOS

SKOS supports tools and standards for quality analysis.

Done by directly mapping XML tags with an XSL Transformations.

2. Analyze SKOS Quality Issues

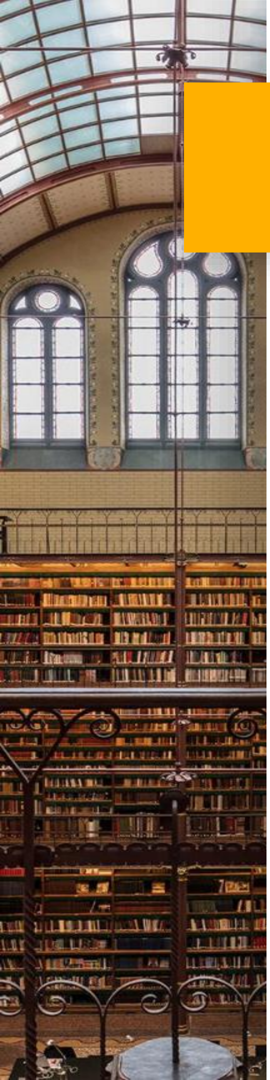
With formalized methods defined by previous studies.

Done with standard SKOS tools and custom python scripts.

3. Align Library and Collection Thesauri

To see how well the two can be integrated.

Mainly done through string matching.



MARC 550 Tag Errors (N = 14828)

Each error represent a hierarchical relation that cannot be converted to SKOS

	code "w"	code "a"	code "0"
Error count	9	37	**875**
Correct entry example	h	boekwetenschap	(NL-AmRIJ)126543
Entry error examples	w	NULL	NULL
	9		mariaverering
	(NL-AmRIJ)131820		(NL-AmRIJ)#129341
	hippodromen		(NL-AmRIJ)

Quality Analysis Results (for all: N = 7826)

Quality Issue	Count in MARC	Count in SKOS	After Skosify
Omitted or Invalid Language Tags	0	0	0
Incomplete Language Coverage	7766	7766	7766
No Common Language	0	0	0
Overlapping Labels	29	29	29
Empty Labels	0	0	0
Orphan Concepts	**976**	**1391**	1364
Cyclic Hierarchical Relations	unknown	23	0
Valueless Associative Relations	unknown	183	0
Omitted Top Concepts	unknown	2043	0

Library Thesaurus Alignment onto the Collection Thesaurus Using Exact String Matching (N = 7826)

Selected Label Type	Selected Languages	Aligned Concepts	Aligned Concepts after Stemming	Percentage of Aligned Concepts after Stemming
skos:prefLabel, skos:altLabel	nl, en	844	1030	13.16%
	nl	**840**	**1024**	**13.08%**
	en	3	4	0.05%
skos:prefLabel	nl, en	729	894	11.42%
	nl	**726**	**890**	**11.37%**
	en	3	4	0.05%
skos:altLabel	nl, en	13	16	0.20%
	nl	**13**	**16**	**0.20%**
	en	0	0	0.00%

Conclusion

- The mapping from MARC to SKOS tags proved to be **a viable conversion method**.

- SKOS provided both better **insight into quality issues**, and was supported by tools to fix them.

- The **amount of alignments** between the Rijksmuseum thesauri was **low**, and this could be improved.

THANK YOU FOR YOUR ATTENTION

Are there any questions?

Follow this project on Github:

Scan me

Special thanks to:

Jacco van Ossenbruggen (VU - supervision)

Chris Dijkshoorn (Rijksmuseum - supervision)

Contact me:

d.a.c.de.ruijter@student.vu.nl