Thesauri Quality Assessment: Analyzing the Rijksmuseum Library Thesaurus

Daan de Ruijter

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

Abstract. Describing the structure and relation between objects by means of a controlled vocabulary is commonplace in the cultural heritage domain. The vocabularies created to describe historical objects are often large and sparse in nature. This hinders the functionality of traditional tools used to either assess the quality of such vocabulary, or the amount of overlap vocabulary have between each other. This report assesses the quality, and identifies overlap with another vocabulary, of a vocabulary from the cultural heritage domain. This is done by utilizing tools and methods designed to overcome the shortcomings of these traditional tools. By doing so, further applications and limitations of current practices used for quality assessment and alignment of vocabulary from the cultural heritage domain are identified

1 Introduction

Controlled vocabularies have been used in the cultural heritage domain in recent decades to describe collections of artistic and historical objects. These controlled vocabularies aim to remove the ambiguities that may arise when using natural language to describe either the content of such an object, or the relations between different objects [1]. Thesauri are a widely adopted type of controlled vocabulary, which distinguishes itself from other controlled vocabulary types by having a hierarchical structure between the different entries.

Individual entries in a thesaurus typically represent either a term or a concept. The difference between the two is subtle and often confusing [2]. However, either approach does have relevant implications on how the thesaurus should be designed and interpreted [3]. One way to think about concepts is as them being units of thought, things that exist as ideas inside the heads of people. If these ideas need to be communicated, ordered or linked to each other they are commonly expressed as terms. Terms, in the context of this research, are thus lexical labels that aim to describe a concept. The idea behind thesauri is that they try to map each concept to a specific term, so that users may reliably find a concept by searching for this term [2].

Sometimes the need arises for two thesauri to be merged. A common practice to facilitate this merging is to look for corresponding data points between the thesauri. These correspondences can be labeled as alignments between the thesauri. In some cases, finding these alignments can be as straightforward as matching the labels or descriptions of the concepts from different thesauri [4]. However, the quality of such straightforward alignments is not guaranteed. Exact matching of labels or descriptions might wrongly match concepts when terms are used ambiguously. For example, a term might have a homonym that describes another concept entirely. Furthermore, finding a complete set of alignments has proven to be a difficult task to automate in domains such as cultural heritage [5].

One such institute in the cultural heritage domain is the Rijksmuseum in Amsterdam. The Rijksmuseum mainly deals with Dutch artistic and historical objects from the Middle Ages onward, as well as object from the most significant aspects of European and Asian art.[1] The Rijksmuseum has also collected a large number of artistic and historical literature, which is presented in the Rijksmuseum Research Library.[2] The Rijksmuseum currently maintains, among others, a thesaurus which describes their collection of artworks and another thesaurus which describes the books contained in their library. This paper will focus on a subset of the library thesaurus, namely the scheme which describes the topic terms of the library, such as ”archaeology” or ”music”.

This paper aims to further identify and address problems regarding thesaurus conversion from a term-based format to a concept-based format, as well as problems regarding thesaurus quality assessment and alignment in the cultural heritage domain. Additionally, analyzing the library thesaurus might result in actionable insight on how to improve the thesaurus by reducing quality issues. A practical implication is that this research helps Rijksmuseum make progress towards a single unified thesaurus, by identifying quality issues which might be preventing this thesaurus from being properly linked to other thesauri. A unified thesaurus would allow for a more integrated and uniform data representation of their objects.

2 Problem Statement

The main objective of this paper is to assess the current state of the library thesaurus maintained by Rijksmuseum, which will be done using two methods. One method is replicating a previous analysis done on the Rijksmuseum collection thesaurus [6]. This analysis consisted of identifying and measuring SKOS quality metrics in both the collection thesaurus and an external thesaurus, and comparing the results. The other method is analyzing how well alignments between these two thesauri can be found using current practices. This section provides a summary of the problems that arise with the implementation of these two methods.

1. The analysis done on the collection thesaurus assumes the thesaurus to be expressed in SKOS. The library thesaurus is currently expressed in a different format, MARC.[3] If the analysis done on the collection thesaurus is to be

[1] https://www.rijksmuseum.nl/en/organisation/vision-and-mission
[2] https://library.rijksmuseum.nl/
[3] https://www.loc.gov/marc/

replicated on the library thesaurus, the latter must first be converted from MARC to SKOS. However, this conversion is non trivial:
   (a) MARC and SKOS maintain different identifiers for information on individual entries. MARC stores information as a combination of tags and codes, SKOS does so by utilizing triples.
   (b) Whereas MARC is a term based format, SKOS is concept based. Automatically mapping MARC terms to SKOS concepts can be problematic, given that one must somehow infer to which specific concept the term belongs.
2. Given that the library thesaurus can be converted to SKOS:
   (a) It would still be naive to exactly replicate the analysis done by Van Langen [6]. It is unclear if the code used for the collection thesaurus analysis would function as expected on the library thesaurus.
   (b) Automatically finding alignments with the collection thesaurus will still prove difficult given their origin in the cultural heritage domain [5]. Van Ossenbruggen et al. found that many alignment tools are not able to process the large number of concepts typically found in this domain. Additionally, even if a tool is able to produce a result, the quality of this result is often hard to assess.

3 Related Work

Converting a thesaurus from MARC to SKOS has proven to be possible by linking certain MARC tags to corresponding SKOS labels [7]. This method of creating a mapping between a term-based thesaurus and SKOS has been proven to be viable, though more difficult than creating a mapping between another concept based thesaurus and SKOS [3]. Additionally, MARC being term-based could result in some information loss when converting to the concept-based SKOS [3].

Assessing the quality of thesauri is important for maintaining, reusing and integrating different thesauri [8]. Quality issues in SKOS vocabularies can be classified as either labeling and documentation issues, structural issues or linked data specific issues [8, 9]. Labeling and documentation issues include problems such as invalid or missing language tags, or two concepts that have the same preferred label for a given language. Structural issues concerns topics addressed in graph theory, for example disconnected concepts or cyclic relationships between concepts. Errors or inconsistencies regarding the use of internal or external URIs fall under the linked data specific issues. For almost half of the identified quality issues, a correction heuristic is developed [8].

A previous study [6] analyzed the Rijksmuseum collection thesaurus on a combination of quality measurements suggested by [8] and [10]. In addition to this analysis, the possible overlap with regards to an external thesaurus (the Art & Architecture Thesaurus (AAT) created by the Getty Research Institute[4]) is found to be almost 11%, which is a significant increase compared to the current links to external sources made by the collection thesaurus.

[4] http://www.getty.edu/research/tools/vocabularies/aat/

Tordai et al. found that aligning large SKOS-like vocabulary could consist of a trade-off between alignment precision and the total coverage of the found alignments [4]. Furthermore, the difference between vocabulary with respect to characteristics such as the amount of alternative labels or vocabulary domain influence the performance of different alignment techniques.

In order to provide a solution to the commonly encountered problems when aligning structured vocabularies from the cultural heritage domain, Van Ossenbruggen et al. propose an interactive approach to vocabulary alignment [5]. This approach provides a framework called Amalgame[5] on which alignments can be produced in small and discernible steps. This framework allows users to build a workflow consisting of building blocks such as selections, matches and merges between thesauri. For each individual step, the framework produces a result which can be analyzed by the user. This intermediate analysis helps the user to assess the quality of the alignments found thus far, and additionally to identify the next available steps in the alignment process.

4 Research Questions

In order to address the main objective of this paper presented in section 2, as well as the problems regarding the chosen methods, the following research questions are defined:

1. What changes in data structure does converting the library thesaurus from the term based MARC format to the concept based SKOS format have?
   (a) Is there any information presented in MARC that SKOS is unable to capture?
   (b) What errors in data representation are being introduced by the conversion?
2. What kind of quality issues does the library thesaurus expressed in SKOS have?
   (a) To what degree did the library thesaurus expressed in MARC have these quality issues?
   (b) How do the correction heuristics identified by [8] affect these quality issues?
3. Given the current state of the art, how well can the collection and library thesauri be aligned with each other?

The first research question relates to the problem of converting thesauri between different formats, and how this might be influenced by factors such as converting from a term based thesaurus to a concept based one. The second research question covers the problem of how the analysis by Van Langen might be applied to the library thesaurus, as well as how a conversion from MARC to SKOS might affect the overall quality of a thesaurus. The third research question relates to the problem of merging thesauri in the cultural heritage domain.

[5] https://semanticweb.cs.vu.nl/amalgame/

5 Methodology

This section will describe the methods used to answer the research questions. It will do so by linking practical solutions to underlying research discussed in previous sections.

5.1 Converting from MARC to SKOS

As stated in the related work section, the current methodology to convert a term-based thesaurus to SKOS is to find a mapping between different labels [3]. A mapping between the specific MARC file used for this paper and SKOS has already been identified [7]. This paper will adopt this mapping with some slight modifications, namely the addition of the 680 and 942 MARC tags, which can be found in Table 1. Given that both formats are expressed as XML files, the conversion can be done using an XSL transformation.

Table 1. Overview of the Modified MARC to SKOS Mapping

MARC tag	**Meaning**	**SKOS label**
001	Identifier of the record	rdf:about
005	Date and time of last transaction	dcterms:modified
008	Date of creation	dcterms:created
150	Topical Term	skos:prefLabel with xml:lang=”nl”
450	Alternative name	skos:altLabel
550 code = w ¿ g	Term is broader than the topical term	skos:broader
550 code = w ¿ h	Term is narrower than the topical term	skos:narrower
550 no code = w	Term is related to the topical term	skos:related
680 with ’Vertaling: ’	English translation of the topical term	skos:prefLabel with xml:lang=”en”
942	Thesaurus data scheme	skos:inScheme

5.2 Analyzing quality issues

The MARC to SKOS mapping produces a SKOS vocabulary that is of similar structure to the Rijksmuseum collection thesaurus. This allows the code used to analyze the collection thesaurus by [6] to be run on the library thesaurus, which produces an output from which the quality issues described in [10] can be identified. The quality issues identified by [8,9], and explained in further detail here[6], will be analyzed using the external SKOS analysis tool Skosify.[7]

The severity of the found quality issues found in the SKOS library thesaurus will be compared to the SKOS conversion after applying relevant correction heuristics described in [8]. The different quality issues found in this analysis can be found in Table 2. This is a subset of the quality issues presented by Suominen and Mader.

Table 2. Overview of Relevant Thesauri Quality Issues

Quality Issue	Description
Omitted or Invalid Language Tags	Labels without a valid language tag
Incomplete Language Coverage	Concepts without labels of as much languages used in the vocabulary as possible
No Common Language	Absence of a single language tag shared across all concepts
Overlapping Labels	An identical label shared by different concepts in the same concept scheme
Empty Labels	A label without any textual information
Orphan Concepts	Concepts without any associative or hierarchical relations
Cyclic Hierarchical Relations	Concepts connected via a hierarchical cycle
Valueless Associative Relations	Concepts that are both directly related and connected by an associative relation
Omitted Top Concepts	Concepts without a broader term but not labeled as a top concept

5.3 Finding alignments

Initial alignments between the Rijksmuseum collection and library thesauri will be produced by performing lexical matching between concept labels. The possibility to improve the quality of the found alignments will be explored by adopting the step wise procedure suggested by [5]. The quantity and quality of the found alignments will be discussed. Additionally, an assessment of the relevance for Rijksmuseum of the final alignment result will be made.

6 Results

This section presents the results from the analysis described in the methodology, after giving a general description of the used dataset. The research questions will be addressed using these results.

[6] https://github.com/cmader/qSKOS/wiki/Quality-Issues

[7] https://skosify.readthedocs.io/en/latest/index.html

As stated in the introduction, this research was performed on a thesaurus maintained by Rijksmuseum which describes different topics of literature. The thesaurus is stored on an XML file following the MARC format. This file contains 7826 records, each record contains information on a topic as described in Table 1.

6.1 Identified conversion challenges

Manual inspection of the MARC file revealed a number of noteworthy aspects that are to be taken into account when converting from MARC to SKOS.

Firstly, some data tags do not necessarily share a common label. That is to say that sometimes different records can have different spellings for the same label, presumably due to the fact that most input was done manually and that MARC lacks quality assurance. An example of this is a subfield in the 550 tag, from which the spelling alternates between ’(NL-AmRIJ)” and ’(NLAmRIJ)’. When performing an XSL transformation from MARC to SKOS one must thus take extra care to ensure that all relevant tags are selected when performing string based selection.

Secondly, the manual data entry of the unusual mapping of the 550 MARC tag, as seen in Table 1, has introduced many types of errors into the vocabulary from which the total amounts and a number of examples are presented in Table 3. Each of these errors prevents that particular hierarchical relation to be converted to SKOS. The most numerous error is the code ’0’ containing no value, which made up 800 out of the 875 identified errors in the code ’0’. In total, 6.2% of the hierarchical relations can not be converted to SKOS due to an input error in the 550 tag.

Table 3. MARC 550 Tag Errors (N = 14828)

	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)

Thirdly, MARC presents the creation date of a record with only the final 2 digits of a year (YYMMDD), while SKOS includes the other digits of a year (YYYY-MM-DD). Converting from the MARC date format to SKOS requires the missing digits to be assumed, given that the records were created in both the 20th and the 21st century. The assumption is made that MARC record creation dates starting with either a 0 or a 1 are from the 21st century, and that the remaining creation dates are from the 20th century. This assumption is fair for this specific case given that no concepts have a creation date before 1980, but this may not be the case for other MARC thesauri or at a later date.

Finally, Table 1 shows that English translations of a preferred label are to be found in the 680 MARC tag. However, this is only the case when the data inside the tag starts with 'Vertaling: ' (dutch for 'translation'). Other 680 tags, for example those starting with 'Volledige term: ' (full term) or 'Omschrijving: ' (description), should not be included as English translation.

6.2 Quality analysis

Analysis of the thesaurus converted to SKOS revealed a number of insight with regards to the Quality issues from Table 2. Table 4 shows that while all concepts are labeled with a dutch preferred label, only 60 (0.8%) also have a preferred label for the English language. Of the 1149 alternative labels, 407 labels are spread across 182 concepts.Which makes the number of unique concepts with at least 1 alternative label 924 (11.8%). The results from the analysis on the collection thesaurus done by [6] show a similar relation between the amount of preferred labels and alternative labels, with the exception of a larger portion of the concepts also having an English preferred label (5.1%).

Table 4. Language Coverage

library thesaurus		(N = 7826)	collection thesaurus [6]		(N = 38150)
	nl	**en**		**nl**	**en**
prefLabel	7826	60	**prefLabel**	38150	1964
altLabel	1149	0	**altLabel**	2225	0

Table 5 shows the identified quantities of the quality issues described in Table 2. The amount of overlapping labels and incomplete language coverage issues remains unchanged throughout this entire process. Due to the lack of tool support for MARC, the structural issues for the original data could not be quantified. Manual inspection using Python XML libraries was able to reveal the original number of overlapping labels and orphan concepts in the MARC data. The only observable change in thesaurus quality from the conversion of MARC to SKOS was an increase in orphan concepts of 42.5%, increasing the total percentage of orphan concepts from 12.5% to 17.8%.

Applying the quality improvement heuristics from Skosify reduced the number of structural issues to 0. The reduction of orphan concepts in the new scenario is due to the fact that 27 concepts in the original SKOS had no outgoing

Table 5. 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

hierarchical relations, but did have incoming hierarchical references from other concepts. These missing outgoing relations are added by Skosify. Furthermore, all identified top concepts were labeled as such.

6.3 Thesaurus alignment

[present a general alignment strategy as suggested by the Amalgame Paper]

[present the results of this strategy in terms of number and percentage of concepts aligned]

Table 6. Library Thesaurus Alignment onto the Collection Thesaurus Using Exact String Matching (N = 7826)

Selected Label Type	**Selected Languages**	**Count of Aligned Concepts**	**Count of 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%

7 Discussion

The results presented in the previous section show that quality issues are present in the MARC representation of the different Rijksmuseum library topics thesaurus, and that conversion to SKOS has an impact on the severity of these quality issues. This section will try to rationalize the implications of the results, present research choices and limitations that may have impacted the results. Finally, based on the results of this research some practical suggestions will be made to the Rijksmuseum in the hopes that these may help improve the quality of the current thesaurus and help Rijksmuseum in their goal to link individuals with art and history.

7.1 Implications of the results

In line with the research presented in [3], the mapping between different tags from MARC to SKOS proved to be a viable method to convert between the term based MARC to the concept based SKOS. The mapping allowed for all relevant data held by the thesaurus to be represented in SKOS. This mapping made evident that the conversion from MARC to SKOS allowed for a wide range of quality assessment and improvement techniques to be applied. There are numerous SKOS specific tools like Skosify and qSKOS. Additionally, the more human readable format of SKOS also ensures a better user experience when dealing with generic data analytic and XML manipulation tools.

The increased accessibility to analytical tools led to an increase in insight into the structural quality of the vocabulary, as shown in Table 5. Furthermore, these tools support the implementation of quality improvement heuristics. As suggested by [8] insight into these quality issues is important for maintaining, reusing and integrating different thesauri. For example, the results show that there is a relatively high number of orphan concepts present in the thesaurus. Adding new concepts or more relations between the concepts reduces the amount of orphan concepts which in turn improves query results and thesaurus navigation.

The observed amount of top concepts (2043) could be the result of a number of factors. By definition, orphan concepts have no broader relation and as such are labeled as a top concept. An overall decrease in the number of hierarchical relations can not decrease the number of orphan concepts. Thus, it would stand to reason that the errors in the 550 MARC tag which decreased the amount of hierarchical relations led to an overall increase in the amount of orphan concepts, which is in line with the results of this research. In turn, this increase in orphan concepts has most likely led to an increase in the amount top concepts in the SKOS version compared to MARC. However, the current results of this research can not substantiate this claim.

Comparing the results of this research with those of [6], which analyzed another thesaurus from the Rijksmuseum, shows a number similarities. The distribution of preferred and alternative labels of both the library and collection thesaurus from the Rijksmuseum is similar. Van Langen compared this distribution to that of an external thesaurus and concluded that the Rijksmuseum thesaurus had less language coverage, as well as less usage of alternative labels which could be used to expand upon a label by adding synonyms or plural forms. The results from Van Langen also showed that the collection thesaurus did not make use of the notion of a top concept, which is in line with the results from this research.

7.2 Research choices and limitations

While the mapping method proposed by [3] has proven to be viable, a number of explicit design choices in creating that mapping affect the eventual results. For example, does one try to accommodate and adjust for human error in the creation of these thesauri (E.g. the '(NL-AmRIJ)" and '(NLAmRIJ) misspelling) or does one map only the correct input. The first option results in a more accurate representation of the intended quality of the thesaurus, while the second option results in a more accurate representation of the actual thesaurus. Additionally, this research chose not to include the hierarchical relations which would have resulted from the errors in the 550 tag as presented in Table 3. It would be fair to assume that this has had an impact on the amount of identified structural issues.

Given the amount of errors identified in the 550 MARC tag, the question arises as to the actual quality of the hierarchy in the thesaurus. If the input of the actual data has been done incorrectly to this degree, how does one identify if the input of the intended hierarchical structure between the different concepts is without error? This research lacks the time and resources to answer such a question. Nevertheless, it is an important aspect that should be taken into account in the interpretations of the results and the overall evaluation of the quality of the Rijksmuseum library topics thesaurus.

7.3 Insight on matching and alignment

[Discuss how using the alignment methods from Amalgame affected the ability to align the thesauri from the cultural heritage domain.]

[Exact string matching does not work as intended when there is limited quality assurance.]

[Lack of alternative label usage in general prevents aligning certain concepts to other thesauri. For example when one thesaurus uses a different synonym or verb tense.]

[Lack of English preferred or alternative labels prevents matching and alignments with international vocabularies.]

7.4 Suggestions to Rijksmuseum based on the results

In general, this research found that a conversion from MARC to SKOS is not without implications. However, the gained insight in quality and the ability to improve this quality by use of supported tools makes such a conversion worthy of consideration. Regardless of the decision to apply such a conversion, a number of quality improvements could be made by Rijksmuseum such as the addition of English labels or the reassessment of the inputs in the 550 MARC tag.

Should the Rijksmuseum choose to convert this thesaurus in a similar fashion as discussed in this research there are a number of relevant factors to take into account. Firstly, a suggestion would be to address the problem of the 550 MARC tag before the conversion in order to preserve as much of the original thesaurus as possible. Secondly, there are a number of ways in which one can structure a SKOS vocabulary while still keeping it in line with the SKOS semantics presented in the primer.[8] Rijksmuseum should evaluate their options carefully to ensure that the final thesaurus fits their needs as much as possible.

8 Conclusion

The mapping between different tags from MARC to SKOS proved to be a viable method to convert from a term based to a concept based vocabulary. The SKOS representation was able to capture all MARC data relevant to the usage of such a thesaurus. However, some data could not be converted due to the initial quality of the MARC thesaurus. This led to an increase orphan concepts present in the thesaurus in SKOS format.

The increased accessibility to analytical tools of SKOS as compared to MARC led to an increase in insight into the structural quality of the vocabulary. The data shows that of the observable quality issues, the initial conversion to SKOS most likely introduced a number of structural issues. However, the correction heuristics available to SKOS were able to reduce the amount of quality issues in the thesaurus.

[General conclusion on alignment, answer to the research question]

8.1 Future work

While this research presents favorable results for a mapping between MARC and SKOS in the sense that all MARC records could be converted to SKOS concepts, further research on converting different kind of thesauri is needed to identify possible drawbacks of this method.

More examples are needed on the practical implications of the SKOS quality issues discussed in this research. While it certainly is a positive thing to be able to identify different kinds of labeling, documentation and structural issues. The effects on a thesaurus having for example more or less orphan concepts is not explored in this research.

[Further research needed on vocabulary alignment]

8.2 Acknowledgments

[Jacco en Chris, Data Rijksmuseum]

[8] https://www.w3.org/TR/skos-primer/

References

1. S. van Hooland and R. Verborgh, *Linked Data for Libraries, Archives and Museums.* 06 2014.
2. S. Dextre Clarke and M. Zeng, “From iso 2788 to iso 25964: the evolution of thesaurus standards towards interoperability and data modeling,” *Information Standards Quarterly*, vol. 24, p. 20, 12 2013.
3. M. van Assem, V. Malaisé, A. Miles, and G. Schreiber, “A method to convert thesauri to skos,” pp. 95–109, 06 2006.
4. A. Tordai, J. Van Ossenbruggen, G. Schreiber, and B. Wielinga, “Aligning large skos-like vocabularies,” *Journal of The American Society for Mass Spectrometry - J AMER SOC MASS SPECTROM*, 01 2010.
5. J. Van Ossenbruggen, M. Hildebrand, and V. de Boer, “Interactive vocabulary alignment,” vol. 6966, pp. 296–307, 09 2011.
6. Q. Van Langen, “Thesauri usage in musea a case study of the rijksmuseum thesaurus. master thesis,” 2018.
7. M. Brink and J. Hoeksema, “Structured vocabularies at the rijksmuseum: Transforming marc into skos,”
8. O. Suominen and C. Mader, “Assessing and improving the quality of skos vocabularies,” *Journal on Data Semantics*, vol. 3, 03 2014.
9. C. Mader, B. Haslhofer, and A. Isaac, “Finding quality issues in skos vocabularies,” 06 2012.
10. D. Adams and S. Milton, “Towards quality measures for evaluating thesauri,” vol. 108, pp. 312–319, 10 2010.