Quality Criteria for Controlled Web Vocabularies

Christian Mader
University of Vienna, Faculty of Computer Science
christian.mader@univie.ac.at

Bernhard Haslhofer
Cornell University, Information Science
bernhard.haslhofer@cornell.edu

1 Introduction

In recent years, the Simple Knowledge Organization System (SKOS) [5] has emerged as the de-facto standard for expressing controlled vocabularies on the Web. Representative examples include AGROVOC[1], EuroVoc[2], GEMET[3], and the STW[4]. With the adoption of the Web of Data idea and the Linked Data principles, the number of controlled vocabularies on the Web is growing and it becomes increasingly complex to manually assess their quality. Vocabulary quality assessment is needed when institutions have to decide whether or not to adopt or align with an existing vocabulary or when they are creating new vocabularies. We believe that this can, to a certain extent, be supported by automated mechanisms.

2 Motivation and Proposed Approach

The notion of *quality* regarding controlled vocabularies is a highly subjective one. Most vocabularies are designed for a specific use case and target a certain application domain. A vocabulary that fulfills the needs of one institution might be totally inappropriate for others. Designers of Web-based information systems, who are typically encouraged to reuse existing controlled vocabularies as much as possible, are burdened with the tedious task of evaluating and selecting potential vocabularies that meet their requirement(s).

In order to facilitate this decision process, we propose an approach that is capable of automatically assessing the quality of controlled Web vocabularies with respect to a defined set of formal quality criteria. These criteria indicate

[1] http://aims.fao.org/website/AGROVOC-Thesaurus/sub
[2] http://eurovoc.europa.eu/drupal/
[3] http://www.eionet.europa.eu/gemet
[4] http://zbw.eu/stw/versions/latest/about

- the suitability of a controlled vocabulary for a specific use case (e.g., indexing, classification, query expansion)
- how well a controlled vocabulary is maintained and reused by others
- if the vocabulary is consistent in itself
- readability for both human users and machines

We believe that quality assessment should also be integrated into collaborative controlled vocabulary development processes. This can help vocabulary designers to reach a state of the vocabulary that conforms to their understanding of quality.

3 Related Work and Methodology

In order to assess the quality of a SKOS vocabulary, we propose a set of criteria for computing quality metrics on a given input vocabulary. Thesaurus quality has been studied before (e.g., [4], [7]) and standards (e.g., [6], [1], [10]) have evolved. However, existing thesaurus quality notions either rely on a solely manual quality assessment process or are bound to specific use-cases and domain-specific development processes (e.g., [2]).

Related work in the area of ontology engineering exists (e.g., [9], [8], [3]), but hardly focuses on instance-level quality criteria, as it would be interesting for assessing thesauri or controlled vocabularies. Furthermore, a vocabulary is domain dependent and metrics that are based solely on structural properties, such as the number of distinct SKOS classes and properties, are insufficient for determining vocabulary quality. To the best of our knowledge, the applicability of ontology quality metrics on thesauri and controlled vocabularies in general has not been studied so far.

The ability to automatically evaluate quality criteria is a key point for integrating quality assessment into existing vocabulary development tools and processes. Therefore, we focus on specifying criteria that are inexpensive to compute but still give valuable feedback to vocabulary designers. They should serve two purposes: (i) helping users to quickly assess existing vocabularies on the Web and (ii) guiding vocabulary developers to successively improve their vocabularies.

In the following, we briefly outline five groups of quality assessment criteria we identified so far. More detailed descriptions and illustrative examples are available on our Github Wiki page[5].

- **Graph-based criteria**: SKOS is based on RDF, which is a graph-based data model. Therefore we can consider the vocabulary's structure for assessing the quality of SKOS vocabularies and apply graph- and network-analysis techniques. The criteria we identified to be relevant for vocab-

[5] `https://github.com/cmader/qSKOS/wiki/Quality-Criteria-for-SKOS-Vocabularies`

ularies involve the degree of nodes, weakly connected components and cycles.

- **Linked Data specific criteria**: Since SKOS vocabularies play an important role in many Linked Data sources, we developed a set of criteria that assess Linked Data specific aspects of SKOS vocabularies. We measure the degree of external links, i.e., to resources in other vocabularies. We also check the percentage of "working" links to external resources and estimate a concept's usage by examining the number of external resources referencing these concepts.
- **SKOS-specific criteria**: The SKOS reference document[6] lists some consistency issues and bad practices. These issues can be detected automatically and reported as quality problems since non-conformance to the SKOS "standard" will render parts of the vocabulary's information unusable for many applications.
- **Labeling issues**: An important property that helps humans in interpreting vocabulary concepts, is the presence of consistent and unambiguous labels, possibly in multiple languages. Carefully maintained labels and synonyms are important when vocabularies are used for indexing documents.
- **Domain-specific / other criteria**: Especially when vocabularies are created collaboratively, redundant or semantically related concepts might be introduced, but not connected with each other. If these concepts are not detected, they can lead to even more inconsistent or redundant entries when the vocabulary is growing.

4 Preliminary results

In a preliminary study, we computed some of the above-mentioned quality criteria for the following SKOS vocabularies:

- STW Thesaurus for Economics[7]
- Ontology for Representing Network Entities (ORNE)[8]
- Press Contacts Information (PCI), University of Southampton[9]
- The New York Times People directory[10]
- LVAk Thesaurus developed by the Austrian Armed Forces (not publicly available)

[6] http://www.w3.org/TR/skos-reference/
[7] http://zbw.eu/stw/versions/8.06/download/
[8] http://river.styx.org/ww/2010/10/network
[9] http://data.southampton.ac.uk/dataset/pressinfo.html
[10] http://data.nytimes.com/people.rdf

We found that the results vary between the different vocabularies. In the PCI thesaurus, for instance, all of the 1125 concepts are loose concepts, which means that none of them has attached SKOS properties. The New York Times People vocabulary contains 4979 concepts, but, as the STW thesaurus, shows no loose concepts, i.e., every concept has at least one attached SKOS property. The ORNE vocabulary, despite having only 11 concepts, features one loose concept.

It is evident that the PCI thesaurus also shows 1125 weakly connected components (WCC) since every loose concept also constitutes a WCC. The STW and NYT People vocabularies show exactly one WCC which indicates that they consist of one "giant component" where every concept is connected to at least one other concept. The ORNE ontology shows 4 and the LVAk thesaurus 32 weakly connected components. Consulting the creators of the LVAk thesaurus, the identified WCC were identified as "forgotten" test data.

With qSKOS[11] we provide a quality analysis tool that implements our SKOS quality criteria. Development is currently in progress and the tool will be published as an open source library targeted for integration into existing vocabulary development environments.

5 Future Work

At the moment, we are still implementing the quality criteria in our qSKOS library. Once this is done, we will set up a detailed survey on existing vocabularies and evaluate their quality. We expect the criteria need to be further adjusted and optimized according to the community feedback we receive during that work.

Furthermore, we will investigate how continuous quality monitoring can improve a collaborative vocabulary building process. In that context, the qSKOS library is expected to serve as an essential part of a system that, at regular intervals, identifies quality issues and reports them to the contributors.

Defining formal criteria that give feedback on the best usage scenario(s) for a controlled vocabulary is another area of research. Such criteria could complement the vocabulary quality assessment process and be beneficial to users developing a specific vocabulary or trying to find one that best serves their requirements. Vocabulary analysis results obtained using qSKOS will therefore include suggestions for the best use-cases of the vocabulary (e.g., for query expansion, indexing or classification).

We also plan to provide a public Web service for assessing the quality of SKOS vocabularies. We believe that vocabulary creators as well as vocabulary adopters could benefit from such a tool.

[11] `https://github.com/cmader/qSKOS`

References

[1] ISO TC 46. *Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval.* Geneva, Switzerland, 2011.

[2] Sherri de Coronado, Lawrence W. Wright, Gilberto Fragoso, Margaret W. Haber, Elizabeth A. Hahn-Dantona, Francis W. Hartel, Sharon L. Quan, Tracy Safran, Nicole Thomas, and Lori Whiteman. The nci thesaurus quality assurance life cycle. *Journal of Biomedical Informatics*, 42(3):530–539, 2009.

[3] Rim Djedidi and Marie-Aude Aufaure. Onto-evoal an ontology evolution approach guided by pattern modelling and quality evaluation. In Sebastian Link and Henri Prade, editors, *Foundations of Information and Knowledge Systems*, volume 5956 of *Lecture Notes in Computer Science*, pages 286–305. Springer Berlin / Heidelberg, 2010.

[4] Daniel Kless and Simon Milton. Towards quality measures for evaluating thesauri. In Salvador Sánchez-Alonso and Ioannis N. Athanasiadis, editors, *Metadata and Semantic Research*, volume 108 of *Communications in Computer and Information Science*, pages 312–319. Springer Berlin Heidelberg, 2010.

[5] Alistair Miles and Sean Bechhofer. *SKOS Simple Knowledge Organization System Reference*, 2008.

[6] NISO. *ANSI/NISO Z39.19 - Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*, 2005.

[7] Dagobert Soergel. Thesauri and ontologies in digital libraries. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '05, pages 421–421, New York, NY, USA, 2005. ACM.

[8] K. Supekar, C. Patel, and Y. Lee. Characterizing Quality of Knowledge on Semantic Web. In *Proceedings of AAAI Florida AI Research Symposium (FLAIRS-2004)*, Miami Beach, Florida, May 17-19 2004.

[9] Samir Tartir, I. Budak Arpinar, Michael Moore, Amit P. Sheth, and Boanerges Aleman-Meza. OntoQA: Metric-based ontology quality analysis. In *Proceedings of IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources*, 2005.

[10] UNESCO. *Documentation – Guidelines for the establishment and development of multilingual thesauri Documentation – Guidelines for the establishment and development of multilingual thesauri Documentation – Guidelines for the establishment and development of multilingual thesauri.* Geneva, Switzerland, 1985.