Quality Assurance in Collaboratively Created Web Vocabularies

Christian Mader

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

Abstract. In recent years, controlled vocabularies have become available on the Web using SKOS[1], i.e. they are linked to each other in order to be used in an interoperable way. Well-crafted controlled vocabularies are beneficial for, e.g., search and retrieval systems that provide functionalities like search term completion, query expansion or the ability for inter-domain queries. Some of these vocabularies are created collaboratively by experts, holding expertise in different domains. In order to support vocabulary contributors to create high quality vocabularies, we propose a method that semi-automatically ensures vocabulary quality in collaborative authoring processes. The proposed approach tackles this issue by (i) defining a set of criteria that serve as a metrics to measure vocabulary quality and (ii) introducing a method to continually assess and improve this quality. As a result of our approach, the developed vocabularies are expected to better fit the intentions of the contributors and are more useful for reuse and adoption on the Web of Data.

1 Motivation

Most institutions that build and publish controlled vocabularies create them for search and retrieval purposes, with specific functionalities in mind like, e.g., query expansion or faceted search [9]. However, shortcomings during the vocabulary creation process impinge upon these functionalities, affecting the effectiveness of the systems backed by these vocabularies, e.g., in terms of recall and precision. Among the problems arising in that context are missing relations between concepts, ambiguous labeling or lack of documentation. Furthermore, duplicate or abandoned entries, or logical contradictions might also be introduced in the vocabulary creation process.

However, when publishing a vocabulary on the Web, additional requirements have to be taken into consideration. Contributing to the Linked Data cloud involves providing references to other data sources in order "to connect disparate data into a single global data space" [6]. With the increasing availability of vocabularies expressed in SKOS, finding and utilzing a well-accepted and well-maintained vocabulary becomes even more challenging. Furthermore, as a consequence of the ever-changing nature of the Web, resources might also become unavailable, introducing the problem of "broken links".

[1] Simple Knowledge Organization System, http://www.w3.org/2004/02/skos/

E. Simperl et al. (Eds.): ESWC 2012, LNCS 7295, pp. 870–874, 2012.

All of these issues could be subsumed under the term "controlled vocabulary quality". It is important for various reasons: as mentioned above, quality assurance measures primarily aim to improve search and retrieval use-cases, since this traditionally has been a very common motivation for creation of controlled vocabularies. However, they can also serve to enhance the usage experience for human users who directly interact with the vocabulary itself, e.g., for getting an overview about the covered domain or incorporating changes.

Especially in open linked environments, vocabulary quality is crucial for acceptance of a vocabulary by others, which in turn is a key concept of the Linked Data principles. Once published as Linked Data, controlled vocabularies can and should be referenced, enhanced, and reused, and with the "building blocks" being of high quality, this is expected to happen to a much greater extent.

Research questions addressed in the proposed approach encompass: (i) what does "vocabulary quality" mean in open, collaboratively maintained environments and how can it be measured? (ii) how can quality assessment be integrated with collaborative vocabulary development environments? (iii) how does vocabulary quality assessment affect the quality of collaboratively created vocabularies?

2 Related Work

Existing standards for thesaurus construction [2,10] and manuals [3,7] propose guidelines and best practices for testing and evaluating controlled vocabularies. Many of them are hardly suitable for automatic assessment because additional knowledge about the creation process, target user group or intended usage would be required. [1] mentions vaguely formulated guidelines like, e.g., inclusion of "all needed facets" or adherence of the term form to "common usage", whereas others, like "both BT and RT relationships occur between the same pair of terms" [3] are more precise and better suited for algorithmic evaluation. However, these guidelines are not specific for a concrete representation (e.g., SKOS) or form of publication (e.g., Linked Data, relational database, hardcopy).

In [8], Kless & Milton provide a list of measurements constructs for the intrinsic quality of thesauri, examining a thesaurus as an artifact itself, i.e. isolated from an application scenario. As stated by the authors themselves, the constructs (e.g., "Conceptual clarity" or "Syntactical correctness") are "solely based on theoretical analysis" and application to existing thesauri is subject to their future work. Although undeniably useful for assessment by humans, algorithmic methods to measure the defined constructs are not covered. Furthermore, multilinguality and, since the paper focuses on intrinsic quality metrics, collaborative aspects of the creation process were not taken into consideration.

In the field of ontology engineering, metrics have been developed to evaluate and validate ontologies [5,11]. Common to these metrics is the fact that they are designed to be applied to general ontologies and instance data. As a consequence, they either do not deal with specific requirements in development of controlled vocabularies and applicability of the metrics for measuring vocabulary quality is still unclear.

Various initiatives that create controlled vocabularies publish details of their construction and validation process on the Web, such as the National Cancer Institute thesaurus (NCIt) or Food and Agriculture Organization (FAO). Despite employing different guidelines and (proprietary) tools [4], certain methods (e.g., duplicate checks) can be abstracted that prove useful in other domains.

3 Proposed Approach

In recent years, SKOS has been adopted by many organizations[2] as a technology for expressing vocabularies on the Web in a machine-readable format. As a consequence, our approach focuses on processing vocabularies represented in the SKOS language.

The overall goal of the approach is to ensure iterative improvement of a controlled vocabulary's quality in a collaborative development process. The "View" in Figure 1 constitutes contributors taking part in the collaborative process. At the core of the work is the "Quality Controller" component, which is based on a catalog of quality criteria (cf. Table 1) and acts as a proxy between view and model, managing quality assessment, user notification, and concurrency issues.

Upon instantiation the quality controller is parameterized with a vocabulary (the model). Every contributor has to register at the quality controller by providing contact information and gets her own "working copy" of the vocabulary. The quality of this working copy is analyzed on every relevant modification. Based on this analysis, notification messages are created, containing information about quality issues. These messages are then disseminated to the contributor who can now decide to fix the issues or keep the current state of the vocabulary. Eventually the changes of the contributor are synchronized with the model.

Fig. 1. Conceptual overview of the proposed approach

Quality assessment of the vocabulary is not only triggered on user interaction, but also after a certain period of time. This is due to the fact that (i) the model might also be changed by contributors bypassing the quality controller and (ii) because the quality of the vocabulary may be affected by changes and independent evolution of other vocabularies on the Web.

Based on existing research and evaluation of SKOS vocabularies available on the Web, a (preliminary) catalog of quality criteria (cf. Table 1) has been identified that can be automatically evaluated. The fact that at least one violating

[2] e.g., AGROVOC, GEMET, Standard Thesaurus Wirtschaft.

Table 1. Identified quality criteria

Qual. Criterion	Description	Example of Impact
Loose Concepts	Concept with no hierarchical or associative relations to other resources.	No hierarchical query expansion
Weakly Conn. Components	Subgraphs within the vocabulary that are not connected to each other.	Obstructive for understanding and querying
Cyclic Relations	Cycles break the hierarchy, might reveal logical problems.	No drill-down search possible
Lack of External Links	No linkage to resources in foreign namespaces (*external* resources).	Gathering knowledge from other domains
Unavailable Resources	Resources must be dereferencable via their HTTP URIs (no broken links).	Information content
Low Concept Indegree	Concepts that serve as link targets in other vocabularies.	Impression on the vocabulary acceptance
SKOS Inconsistency	Conflicts with consistency criteria of SKOS reference or invention of new terms within SKOS namespace.	Standard conformance
Deprecated Property Usage	Some properties have been removed from the current SKOS version.	Interoperability
Poor internationalization	Inclusion of language tags and concepts labeled with same set of languages.	Translation use-cases
Ambiguous Labeling	SKOS labels are pairwise disjoint; avoid identical labels for different concepts.	Retrieval precision
Unconnected Related Concepts	Concepts labeled (slightly) different but mean the same and are not hierarchically or associatively connected.	Expose structural misconceptions
Lack of Documentation	Usage of properties documenting vocabulary concepts in human-readable form.	Disambiguation

vocabulary for each criterion could be found on the Web indicates practical relevance of this catalog.

4 Research Methodology

In a first step, as a **problem definition and state-of-the-art survey**, existing publications targeting data quality, vocabulary and thesaurus development as well as ontology building principles will be reviewed. It is important to find out to what extent quality criteria in these areas exist and to elaborate on their importance to controlled vocabularies. Based on the findings in the first step, we **propose a set of general quality criteria** for controlled vocabularies. The result of this step is a list of criteria together with algorithms that allow for programmatic evaluation. After that, **implementation of the tools**, i.e. a library (API) that creates a metrics based on the quality criteria, will be started. In the course of an **analysis** step, existing vocabularies available on the Web will be evaluated against the quality criteria. Community feedback collected in this step might lead to adjusting and reformulating the quality criteria which target research question (i). To

evaluate the approach, instantiating (or integrating into) a collaborative vocabulary development process is essential, addressing research question (ii). A valid setting would be to assign two comparably skilled groups of users with the task of concurrently creating a vocabulary. The continuous quality assessment of the approach will be activated for both groups, with only one group getting support by the automatic feedback mechanism. That way it is possible to track the evolution of vocabulary quality in both groups, obtaining information how the quality assessment influences development and contribute to research question (iii). If those groups that are supported by automatic quality feedback, develop higher-quality vocabularies, the proposed approach is said to be successful.

5 Preliminary Results and Conclusion

Based on the identified criteria, qSKOS[3], an open source library for vocabulary quality assessment, has been created. First results[4] in utilizing the library on various vocabularies were promising and it will be continually updated based on the community's feedback and as research progresses. qSKOS also provides the basis for further research regarding quality implications in collaborative settings.

References

1. Proceedings of ACM/IEEE 2003 Joint Conf. on Digital Libraries (JCDL 2003), Houston, Texas, USA, May 27-31. IEEE Computer Society (2003)
2. Information and documentation – thesauri and interoperability with other vocabularies – part 1: Thesauri for information retrieval. Norm (Draft) ISO 25964-1, Int. Org. for Standardization, Geneva, Switzerland (2011)
3. Aitchison, J., Gilchrist, A., Bawden, D.: Thesaurus construction and use: a practical manual. Aslib IMI (2000)
4. de Coronado, S., et al.: The nci thesaurus quality assurance life cycle. Jour. of Bio. Inf. 42(3), 530–539 (2009)
5. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Ontology evaluation and validation: an integrated formal model for the quality diagnostic task. Tech. rep., Lab. of Applied Ontologies – CNR, Rome, Italy (2005)
6. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space, 1st edn. Morgan & Claypool (2011), http://linkeddatabook.com/
7. Hedden, H.: The accidental taxonomist. Information Today (2010)
8. Kless, D., Milton, S.: Towards Quality Measures for Evaluating Thesauri. In: Sánchez-Alonso, S., Athanasiadis, I.N. (eds.) MTSR 2010. CCIS, vol. 108, pp. 312–319. Springer, Heidelberg (2010)
9. Nagy, H., Pellegrini, T., Mader, C.: Exploring structural differences in thesauri for skos-based applications. In: I-Semantics 2011, pp. 187–190. ACM (2011)
10. NISO: ANSI/NISO Z39.19 - Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies (2005)
11. Tartir, S., Arpinar, I.B.: Ontology evaluation and ranking using ontoqa. In: ICSC 2007, pp. 185–192 (2007)

[3] The source code can be downloaded from https://github.com/cmader/qSKOS/
[4] https://github.com/cmader/qSKOS4rb/raw/master/results/qskos_results.ods