Semantic Terminology Management for Applications: Contextualized SKOS-XL

Andreas Thalhammer, Martin Romacker, and Joachim Rupp

Roche Pharma Research and Early Development Informatics, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland
{andreas.thalhammer, martin.romacker, joachim.rupp}@roche.com

Abstract. Terminology management is an important aspect for ensuring data quality in large organizations. To enable expert applications the use of agreed and curated terms enhances data quality while it significantly reduces the long-term cost for data integration. In this abstract, we outline our solution for two problems that occur in the context of terminology management for applications.

1 Introduction

In a large organization, like Roche, it is essential to ensure that different stakeholders are referring to the same thing when they mean the same thing. Instead of costly (and repeated) processing and mapping steps that are typically performed when "data integration is needed", a functioning organization-wide terminology management system preemptively supports users in selecting the right terms at the point of data creation. This leads to a situation where two independent applications can refer to `roche:ROX1305277804386` when they mean "Non-Small Cell Lung Cancer". Next to straight-forward integration, terminology management significantly improves data quality in the long term. For this, the Simple Knowledge Organization System (SKOS) not just enables to have a single identifier per concept but also enables to assign different labels with preferences (typically distinguished as preferred and alternative labels) and concept hierarchies. This supports consuming applications (that are designed for expert users) to select the acronym "NSCLC" as a label for the previously mentioned entity and group it in "Diseases→Lung→Lung Cancer". However, there are two main problems when applications consume SKOS terminologies:

Problem 1 Domain-specific schemes, like "Diseases" (organized via domain-centric polyhierarchies), would be used for a drop-down field called "Search cell line inventory—related disease)".[1] This means that an end user would need to select a single value from more than 4000 while, in the context of the drop-down field, far less entries are actually needed.

[1] Note that `skos:broader/skos:narrower` cannot be used in this case as the hierarchical organization of the content is domain-dependent rather than application-centric.

Problem 2 When an application uses a preferred or alternative label in its specific context, the label would need to be transferred to the application (as otherwise it would be unclear which label the application wants to use). Therefore, the consuming application becomes the "owner" of the term and potential changes (on the application side) might not be fed back to the terminology management system.

2 Contextualized SKOS-XL

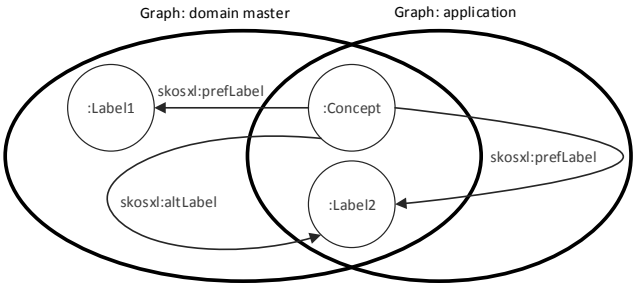

Fig. 1. Schematic overview of contextualized SKOS-XL.

In the pREDi-Roche Terminology System (RTS) we use RDF and SKOS-XL for specifying domain-centric terminologies. Consuming applications can rely on the defined semantics and the stable URIs that the terminology system provides. In order to address Problem 1, applications need access to subsets. However, in the design of the RTS system we decided not to maintain these subsets on the application side. This is due to two main reasons: **1) Semantic -** Applications use the same concepts with slightly different semantics (e.g., a disease can also be interpreted as an adverse event, genetic background, or pheno-type): an application naturally sets a context for a term. Data curators need to be aware of such contexts (which is enabled by maintaining the subsets at the side of the terminology system). **2) Organizational -** If subsets are maintained at the application side (i.e., via lists of URIs and according preferred terms), the technical infrastructure for storing and retrieving terms is readily in place. In combination with corporate structure (i.e., shortest paths) and the long-term orientation of central terminology management (i.e., no quick benefit) this can lead to unnoticed disconnection of the applications.

In RTS, we make use of named graphs to maintain application contexts/subsets. The URIs of the concepts and labels of a domain-master graph are reused and contextualized in application graphs. In order to address Problem 2, the domain master maintains all different types of SKOS-XL labels (preferred, alternative, or hidden label) and applications can choose one of these labels as their preferred label (see Fig. 1). This enables to provide acronyms like "NSCLC" in restricted, application-centric subsets (i.e., maximum flexibility). At any point of time, data curators can see which applications consume which terms and how they refer to them.