SKOS Tool: A Tool for Creating Knowledge Graphs to Support Semantic Text Classification

Farhad Ameri, Reid Yoder, Kimia Zandbiglari

SKOS Tool: A Tool for Creating Knowledge Graphs to Support Semantic Text Classification

Farhad Ameri, Reid Yoder, Kimia Zandbiglari

Engineering Informatics Lab, Texas State University, San Marcos, U.S.A
{ameri, rjy15, k_z54}@txstate.edu

Abstract. Knowledge graphs are being increasingly adopted in industry in order to add meaning to data and improve the intelligence of data analytics methods. Simple Knowledge Management System (SKOS) is a W3C standard for representation of knowledge graphs in a web-native and machine-understandable format. This paper introduces SKOS Tool; a web-based application developed at the Engineering Informatics Lab at Texas State University. It can be used for creating knowledge graphs and concept schemes based on the SKOS standard. The main feature and functions of SKOS Tool are described in this paper. Beyond creating knowledge graphs, SKOS Tool has additional features that can be used to support semantic document classification based on the Bag of Concepts technique. To demonstrate the utilities of SKOS Tool, a use case related to classifications of manufacturing suppliers with Medical Grade Polymer Tubing capabilities is presented.

Keywords: knowledge graph, semantic classifier, Natural Language Processing, Artificial Intelligence, SKOS

1 Introduction

Semantic Artificial Intelligence (AI) is a branch of AI that uses semantic models for supporting intelligent systems that mimic human-like cognitive functions such as learning, reasoning, and problem solving. Semantic models are intended to represent a model of the reality in a machine-understandable and logical fashion [1]. They can be used to represent the implicit meaning of data and add context to it. There are different types of semantic models ranging form simple controlled vocabularies and taxonomies to more sophisticated formal thesauri and ontologies that vary based on their expressivity and development cost and time. Most formal semantic models can be represented as graphs with nodes (concepts or entities) and edges (relationships). *Knowledge Graph* is a general term that can be applied to the semantic models that are represented as one or more connected graphs [2]. Knowledge graphs can serve as unifying models that can semantically connect and integrate disparate silos of structured and unstructured data. A knowledge graph can provide a strong foundation for various machine learning and cognitive computing projects as it adds a semantic layer on top of metadata and data layers in AI application [3].

There are multiple standards for representation of knowledge graph. The focus of this paper is on a specific type of knowledge graph that serves as a *concept scheme* or *thesaurus* and is represented using Simple Knowledge Organization System (SKOS)

formalism [4]. SKOS is a standard, published by World Wide Web Consortium (W3C), that provides a structured framework for building controlled vocabularies such as thesauri, concept schemes, and taxonomies to be used and understood by both human and machine agents. SKOS models are considered to be lightweight ontologies as they don't have the expressivity of heavyweight, axiomatic ontologies such as OWL models. However, for many applications that require basic semantics in terms of the structural and lexical relationships between various entities, SKOS models can be developed fairly easily without requiring to invest heavily on developing rich, logic-based ontologies.

This paper describes a web-based tool called INFONEER SKOS Tool (or SKOS Tool for short) that is developed for creation and extension of SKOS models. Beyond its core function, SKOS Tool provides some other useful services that can support supervised and unsupervised document classification applications. Although SKOS Tool was originally developed to support a particular application related to classification of manufacturing suppliers based on their website content, it can be used for any type of document classification and semantic similarity measurement applications.

The remainder of this paper is organized as follows. Section 2 provides an overview of the underlying Semantic Model of SKOS. Section 3 describes a use case related to ventilator supply chain. Section 4 presents different functions and features of SKOS Tool. Section 5 describes how SKOS Tool can be used for supporting a supplier classification task related to the ventilator use case.

2 SKOS Semantic Model

The building block of a SKOS knowledge graph is called a *Concept*. A SKOS Concept (skos:concept) is any unit of thought such as an idea, an object, or an event. SKOS concepts, as abstract notions in mind, are independent of the terms that are used in natural language to describe them. For example, the English terms *car* and *automobile* point to the same concept, or entity, which is basically an artifact that is used for transporting people. Separation of the concepts form their descriptors (labels) is a core feature of SKOS models. Humans can identify concepts through their labels and machines can identify concepts via their Uniform Resource Identifier (URI) [5].

Each concept in SKOS has exactly one *preferred label* (skos:prefLabel) and can have multiple *alternative labels* (skos:altLabel). Preferred Label is a SKOS element that makes it possible to assign an authorized name to a concept. For example, in the context of metal casting terminology, *Foundry Sand* is the alternative label for *Molding Sand* as it is used frequently for referring to the same concept (Figure 1). The broader concept of the *Molding Sand* is *Sand*, while *Silica Sand* and *Chromite Sand* are the narrower concepts; meaning that they are more specialized forms of *Molding Sand*. The concept that is semantically *related* to *Molding Sand* is *Mold*. While skos:broader and skos:narrower indicate a *hierarchical* link between two concepts, skos:related represents an *associative* relationships between concepts. Each SKOS concept can also have a definition provided in plain English or any other natural language. One major advantage of the SKOS thesauri is that they can be extended, enriched, and validated incrementally by community crowds and shared as linked open data due to their open and standard syntax and semantics. A SKOS thesaurus forms the nucleus of a knowledge graph that can be continuously enriched to support various data-driven and

knowledge-intensive application such as semantic search and reasoning, text mining, data integration and alignment, and data analytic.

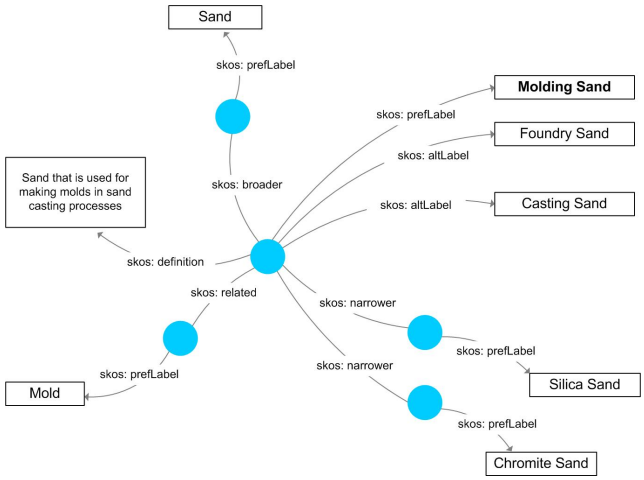

Fig. 1. The concept diagram of the Molding Sand based on SKOS terminology.

3 Use Case: COVID-19 Response Manufacturing Suppliers

COVID-19 pandemic caused a demand surge for certain medical equipment and supplies such as ventilators and face shields [6]. Supply chains have been slow in responding to this emergency mainly because finding the qualified suppliers with the required set of capability and capacities is a time-consuming process. Using keyword search method for finding suppliers is inefficient because online keyword search doesn't take into account the contextual semantics of the terms. Additionally, the contents of the websites of manufacturers vary significantly in term of quality and depth. Another issue arises from the heavy use of 'tribal knowledge' on the websites of contract manufacturers. The informal terminology that dominates this body of tribal knowledge causes a semantic discontinuity throughout the domain.

In presence of a knowledge graph that captures the important concepts (notions) in medical equipment manufacturing, supplier search can be conducted on a semantic level. For example, the manufacturers' websites can be annotated, or tagged, with the concepts coming from the knowledge graph. Another solution is to use a Semantic Classifier for classifying suppliers, represented by documents extracted from their websites, based on their capabilities.

4 INFONEER SKOS Tool

INFONEER SKOS Tool is developed for creating knowledge graphs. It also supports document classification applications by providing means for tokenizing and annotating documents using SKOS concepts. The SKOS Tool runs as a Django web application. Django is a free and open-source web framework that utilizes Python to realize a traditional model-template-view architecture. In addition to Django, various other libraries such as BeautifulSoup4 are bundled in a virtual environment to help carry out the tool's functions. While the back-end of the application is developed with Python, the application's front-end is presented with HTML and JavaScript. The latest stable release of the web application is deployed on a developmental virtual machine running Red Hat Enterprise Linux at Texas State University, providing accessibility for select users through Secure Shell (SSH). SKOS Tool has different gadgets such as Thesaurus Manager, Term Selector, Entity Extractor, Concept Model Builder, Concept Model Manager, and Capability Scorer that are describe in the following sections.

4.1 Thesaurus Manager

Thesaurus Manager (TM) (Figure 2) can be used for creating and extending a SKOS thesaurus.

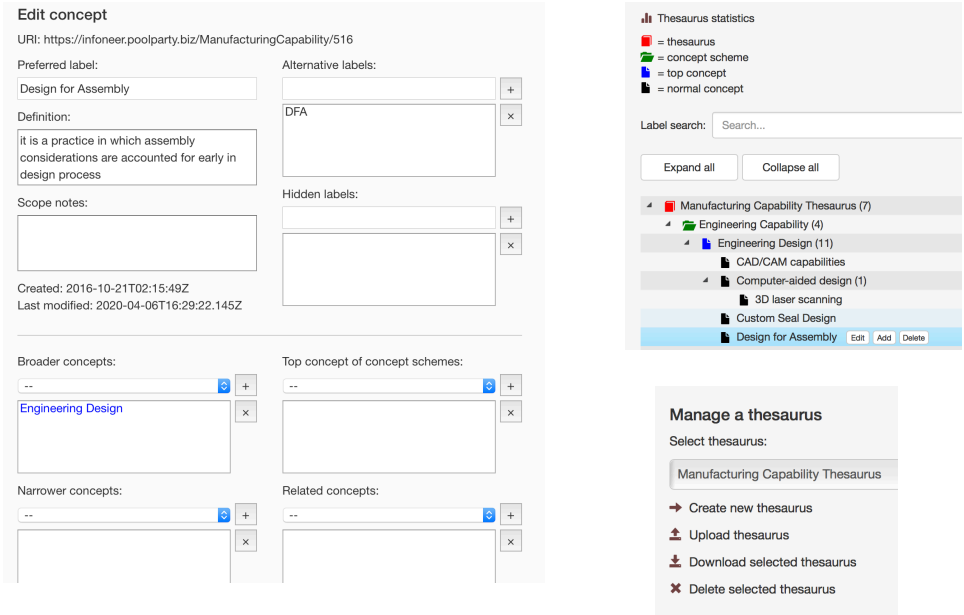

Fig. 2. The user interface for Thesaurus Manager and partial view of the Manufacturing Capability Thesaurus

The user can create a thesaurus from scratch by building a taxonomy of concepts, adding the necessary preferred and alternative labels and providing natural language definition for each concept, and relating them to one another. The final model can be exported in RFD/JSON format. For example, in the thesaurus partially shown in Figure 2, Design for Assembly is a narrower concept for Engineering Design under Engineering Capability concept scheme. Thesaurus imports are also allowed using the same format.

4.2 Term Selector

The thesaurus can be extended directly by adding concepts using the TM gadget. An alternative method is to select terms from inserted text and integrate the selected terms with the thesaurus. The Term Selector gadget allows the user to select the relevant terms from a given text (through copy&paste or entering the URL) and add them to the thesaurus directly or export the result as an intermediate CSV file to be integrated with the thesaurus after verification by domain experts.

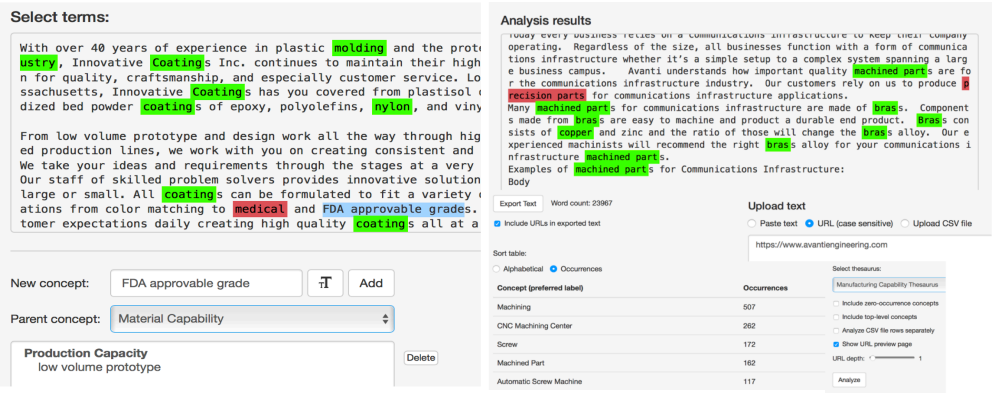

Fig. 3. Term Selector (left)User can highlight the terms that should be added to the thesaurus . Entity Extractor (right) identifies the occurrences of the thesaurus concepts in a text

As shown in Figure 3, the user needs to specify the parent (skos:broader) concept for each selected term. In the example shown in this figure, “FDA approvable grade” is selected as a new concept to be added to the thesaurus and placed under “Material Capability” is the broader concept.

4.3 Entity Extractor

Entity Extractor (Figure 3 -right) is used for tokenizing a text or document. The tokens are the concepts that exist in the treasures and appear in the inserted text through either their preferred labels (highlighted in green) or alternative labels (highlighted in red). The input text can be inserted directly or grabbed from a given URL. The number of occurrences of those concepts is also captured using this gadget. This will result in vectorization of the unstructured text. The resulting concept vector can be exported as

a CSV file. The concept vector for each document can be used for more advanced text analytics processes such as document classification and clustering.

4.4 Concept Model (CM) Builder and Manager

A Concept Model (CM) is a subset of the thesaurus that represents a class of interest in a document classification task (Figure 4 - left).

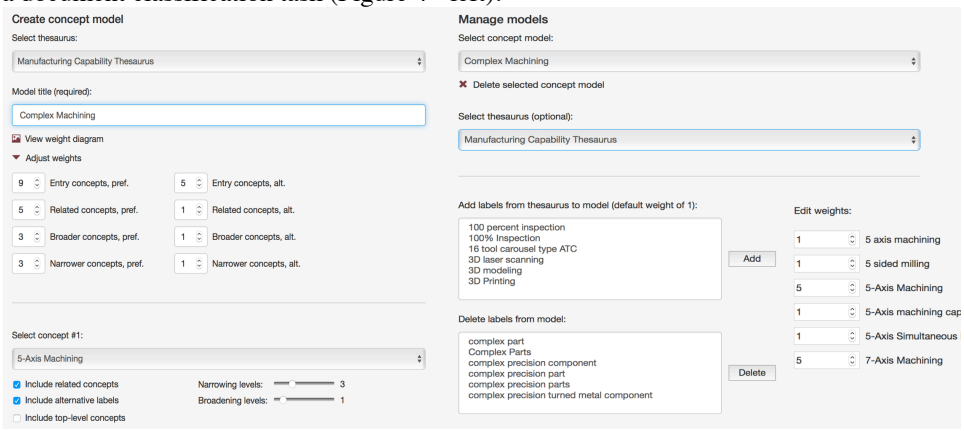

Fig. 4. Concept Model Builder (Left), CM Manager (Right) Domain expert can use CM builder to select the representative concepts for a given class of documents

For example, if the class of interest is Heavy Part Machining, then the CM related to this class include the labels for all processes and equipment that can be used in heavy part machining. CM Builder provides a user-friendly environment for domain experts to pick the relevant concepts from thesaurus and add them to the concept model for a specific class. The degree of importance of the concepts for a given class can be specified through assigning weights to the concepts. Concept Model is used as the input for document classification algorithms that use techniques such as Random Forests (RF) and Support Vector Machine (SVM). CM Manger can be used for modifying a concept model through adding or removing concepts and/or changing their weighting.

4.5 Capability Scorer

SKOS Tool was originally developed for evaluating the capabilities of manufacturing companies based on the textual description of their services provided on their websites. Capability Scores uses a scoring scheme that assigns a score to a given text based on the normalized frequencies of occurrences of terms that can be mapped to concepts in a given concept model. In the given example in Figure 5, the company's score with respect to complex machining and heavy machining capabilities is 0.133 and .053, respectively.

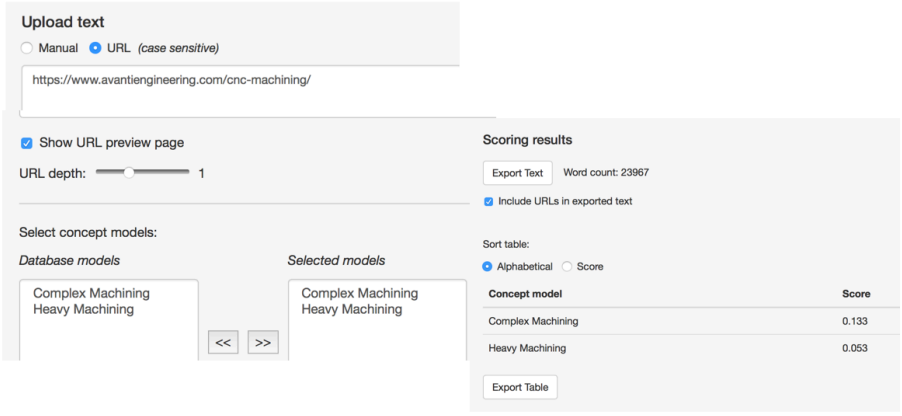

Fig. 5. A document (company website) scored based on complex machining and heavy machining capabilities

5 SKOS Tool for Classification of Manufacturing Suppliers

Going back to the COVID-19 use case discussed earlier, since most ventilators need some sort of silicone and polymer tubing, suppose we want to create a group of suppliers with specialization in "Medical Grade Polymer Tubing". We already have a knowledge graph named Manufacturing Capability Thesaurus (MCT). Through web crawling, all suppliers in North America can be screened and evaluated based on their websites information. If they meet the minimum membership strength threshold, they will be added as a member of this class. Alternatively, using the capability scorer gadget in SKOS Tool, a score can be assigned to each participating supplier. This sore can be used for ranking and initial screening before going through more rigorous capability analysis steps. Using the SKOS Tool, the Manufacturing Capability Thesaurus was extended with the concepts that are related to Medical Grade Polymer Tubing capability. Some of those concepts are shown in Table 1.

Table 1. A subset of concepts related to Medical Grade Tubing concept model

Materials	Methods	Finishing	Technologies
PTFE Tube	Extrusion	Laser Welding	Tri-TIE
Teflon Tube	Tip forming	Laser Machining	Taper-TIE
PEEK Tube	Flanging	Tipping	Variable Braid & Coil
FEP Tube	flaring	Precision Cutting	Polyimide Shaft Liners
TPU Tube	Coextrusion	Precision Machining	Coil & Braid Reinforced
PP Tube	Single Lumen	Hole Punching	Sheath Extrusion
PEBAX Tube	Multi-Lumen	RF Welding	Proximally Reinforced
PE Tube/polyethylene	Tri-Layer Extrusion	Overmolding	Marker Bands
Nylon Tube	Intermittent Extrusion	Printing	Super- Tri
ABS Tube	Braiding	Plasma Etching	PTFE Coated Mandrels
PETG tube	Coiling	Annealling	Multi-Lumen Tubing

To collect these concepts, the websites of about 100 suppliers with medical tubing capability was parsed. This step is equivalent to the training phase of the conventional text classification methods that results in an automatically-generated dictionary of terms (a.k.a Bag of Words) [7]. However, the Bag of Words method often creates a dictionary

which is cluttered with irrelevant terms that create a noisy environment for text classification. However, a curated thesaurus ensures that every term included in the Concept Model is terminologically and semantically relevant and meaningful. The collected concepts were then made *skso:related* to one another in order to capture the associative relationships among them. We refer to this semantically enhanced document classification method as Bag of Concepts (BoC) method. Using the Bag of Concepts method, new suppliers can be analyzed to check if they belong to different capability classes of interest. It was demonstrated previously that BoC method significantly improve the precisions of document classifiers [8].

6 Conclusion

In this paper, the main feature and functions of SKOS Tool were described and a use case related to supplier classification was discussed. In future, we will extend the medical equipment supplier classification use case by creating multiple capability classes. SKOS Tool will be extended in future to provide more sophisticated functionalities such as creating probabilistic Naïve Bayes networks from unstructured text. SKOS Tool is currently in its alpha test phase and it is being evaluated by a small group of researchers. The beta version will be released to larger group of domain experts for creation of knowledge graphs in various domains.

References:

[1] Hagedorn, T., Bone, M., Kruse, B., Grosse, I., and Blackburn, M., 2020, "Knowledge Representation with Ontologies and Semantic Web Technologies to Promote Augmented and Artificial Intelligence in Systems Engineering," INSIGHT, 23(1), pp. 15-20.
[2] Wilcke, X., Bloem, P., and De Boer, V., 2017, "The knowledge graph as the default data model for learning on heterogeneous knowledge," Data Science, 1(1-2), pp. 39-57.
[3] Ameri, F., Urbanovsky, C., and McArthur, C., "A systematic approach to developing ontologies for manufacturing service modeling," Proc. Workshop on Ontology and Semantic Web for Manufacturing, OSEMA 2012, July 24, 2012 - July 24, 2012, Sun SITE Central Europe CEUR-WS, pp. 1-14.
[4] Miles, A., and Bechhofer, S., 2009, "SKOS simple knowledge organization system reference," W3C.
[5] Blumauer, A., 2017, "PoolParty Technical White Paper," https://help.poolparty.biz/doc/white-papers-release-notes/poolparty-technical-white-paper.
[6] Ranney, M. L., Griffeth, V., and Jha, A. K., 2020, "Critical supply shortages—the need for ventilators and personal protective equipment during the Covid-19 pandemic," New England Journal of Medicine, 382(18), p. e41.
[7] Korde, V., 2012, "Text Classification and Classifiers: A Survey," International Journal of Artificial Intelligence & Applications, 3(2), pp. pp. 85-99.
[8] Sabbagh, R., Ameri, F., 2017, " A thesaursi-guided text analytics technique for capability based classification of manufacturing suppliers"" Proc. ASME International Design Engineering Technical Conferences / Computers and Information in Engineering Conference (IDETC/CIE 2017).