Incorporating New Concepts into the Scientific Variables Ontology

1st Maria Stoica
Institute of Arctic and Alpine Research
University of Colorado, Boulder
Boulder, USA
maria.stoica@colorado.edu

2nd Scott D. Peckham
Institute of Arctic and Alpine Research
University of Colorado, Boulder
Boulder, USA

***Abstract*—We present a preliminary methodology, currently in development, for automated generation of domain-specific, machine-readable representations of qualitative and quantitative scientific variable concepts. The method presented is based on the top level universal categories and modular design patterns declared within the Scientific Variables Ontology (v 1.0.0) blueprint. These scientific variable representations can be used to annotate electronic resources, such as data and models and, along with reasoning algorithms, can be used to provide explainable automated resource alignment capabilities in the assembly of scientific workflows.**

***Index Terms*—ontology, interoperability, data models, metadata**

I. Introduction

A data model is a blueprint that contains information about the conceptual, logical, and physical layout of a particular database. A data model is usually custom-created and optimized for a particular use-case, and thus can cause difficulties in the integration and interoperability of resources stored across disparate databases. In a world of growing data availability and a rise of interdisciplinary research, it has become increasingly desirable to create more generic data models that can be used to bring together disperse, heterogenous resources. Ontologies are a means of achieving this goal. Ontologies aim to provide a generic conceptual representation of information within a domain and should be designed to be as application or task independent as possible; ontologies are thus shareable and reusable [1], [2]. Furthermore, ontologies are explicitly represented and support inferencing and reasoning.

In the sciences in particular, it is desirable to have a universal, machine-readable language to use when storing information about scientific variables. In computational scientific workflows, variables comprise the flow of information from one workflow component to another. In order to address the need of expressing scientific variables in a domain-agnostic format, we developed the Scientific Variables Ontology (SVO) blueprint [3], [4]. SVO was designed not just to describe information about variables, but also to allow reasoning for the purpose of aligning variables across resources as well as generating new variable concepts. The latter functionality is particularly useful when there is no ontology given for a particular domain but there may be a list of curated variables that has been carefully compiled by domain experts. Such lists are usually very long, containing hundreds, if not thousands, of variables. The lists usually contain a short description of each variable along with other tabular information such as measurement units, and possibly a controlled vocabulary or standard naming scheme. Examples include the NWIS SRS parameter codes [5], the CF Standard Names [6] and the World Development Indicators [7]. When such resources are available, it is desirable to be able to quickly generate a meaningful, machine-interpretable ontology that can then be used to annotate data and model resources.

SVO is a framework for principled, machine-readable representation of quantitative and qualitative scientific variables. It consists of (a) a domain-independent upper ontology blueprint that declares the component class concepts and design patterns for mix-and-match customizable creation of scientific variables; (b) a set of rules for inferring spatiotemporal context, quantification reference frames, and roles of recorded observation components; (c) an extensible lower ontology currently populated with a wide range of variables from the earth science and socio- and agro- economic domains; (d) ontology guided tools for searching entries present in the ontology and grounding natural language phrases to variables; and (e) ontology structure based tools for creating and introducing new concepts into the lower ontology. SVO ensures interoperability among digital resources, such as models and data, by providing automated, rules-based variable alignment between elements in scientific workflows. The work presented here focuses specifically on the latter component of the SVO framework—the ontology augmentation tools.

II. Methods

The new variable generation tools have two main components, each addressing a different level of refinement. The first component performs a coarse mapping by ingesting a freeform, concise string that describes the desired variable, processing it against the SVO blueprint, and proposing possible term categorization as well as a basic variable mapping recommendation. Unlike a step-by-step guided variable cre-

The authors would like to acknowledge the funding sources for this work: DARPA award W911NF-18-1-0027 for MINT: Model Integration through Knowledge-Rich Data and Process Composition and NSF award 1740696 for BALTO: Brokered Alignment of Long-Tail Observations.

ation mechanism, this design choice allows the quick bulk processing and rough mapping of a long list of variables. The second component poses clarifying questions to the user to correct mistakes from the bulk categorization process and fill in missing information or disambiguate broad terms.

Similar to a human's decision making process, the new variable generation tools follow a set of decision-making guidelines with the following hierarchy: (1) a bias towards trusting the information already present in the ontology more than information ingested from an external resource; (2) leveraging crowd-sourced and expertly generated linked data and ontology resources to categorize unknown concepts; (3) relying on short, precise definitions of terms from a combination of curated, reliable resources to reduce noise and processing time; (4) using part-of-speech analysis for assigning roles to components and determining possible term categorization; (5) using n-gram language indexes (preferably from a domain corpus) to recommend disambiguation of terms and lastly, (6) using machine learning algorithms to determine context and organization of components within the design pattern rules by training on the verified content already stored in the ontology.

III. Preliminary Results

In order to populate an ontology of variables in a domain that is not currently in SVO, multiple approaches are possible. As mentioned previously, the first step is to look for atomistic concepts within what is already encoded and determine if variables can be pieced together with the already vetted information; again, this is similar to how the human brain learns new concepts by making associations to what it already knows. For information that cannot be found, one approach is to implement rules-based or statistical detection of word morphology in combination with part of speech tagging to identify to which ontological class a term is most likely to belong. A second approach is to leverage the wealth of information already publicly available as structured linked data across the web. Such resources include, but are not limited to, WordNet, Wiktionary, and Wikidata. Again, this is similar to how a human would perform preliminary research to help their understanding of a concept. In this step, a human would curate the resources they would consult based on trustworthiness as well as accessibility—if the resource is too advanced, it may initially be skipped.

A first pass approach that illustrates this latter step involves determining the mappings between the SVO classes and categories in the WordNet hierarchical tree. Conveniently, WordNet has categories for concepts like process, phenomenon, attribute, and state that map relatively well to ontological classes in SVO. The result of blindly categorizing previously categorized terms in SVO using categories from WordNet shows a relatively high success rate in most scenarios (see Table I), with the exception of complex and dynamic phenomena.

TABLE I
Result of automated categorization into ontological classes based on WordNet hierarchy. The terms used for this analysis were previously manually categorized.

Category	Type	Correct	Incorrect	Missing	Total
Process	All	1250	216	36	1502
Phenomenon	Natural Body	48	10	8	66
Phenomenon	Complex	46	38	23	107
Property	Qualitative	89	14	9	112

Once categorization of atomistic terms has been achieved, one can assemble suggested variable instantiations using the design patterns provided by SVO by filling in the known information. A knowledgeable user can then be asked to fill in missing information and correct any errors made by the automated categorization procedure.

IV. Summary

The purpose of this work is to describe preliminary work in automated ontology generation for scientific variables using the design pattern templates and universal categories defined in the Scientific Variables Ontology ver. 1.0.0. This is a multi-step process that is informed by and mimics the mental processes of the human mind when attempting to learn new concepts.

References

[1] Peter Spyns, Robert Meersman, and Mustafa Jarrar. Data modelling versus ontology engineering. *ACM SIGMod Record*, 31(4):12–17, 2002.

[2] Michael Uschold and Semantic Arts. Ontologies and database schema: What's the difference. In *Semantic Technology Conference, San Francisco, CA*, 2011.

[3] Maria Stoica and Scott D Peckham. An ontology blueprint for constructing qualitative and quantitative scientific variables. In *International Semantic Web Conference (P&D/Industry/BlueSky)*, 2018.

[4] Maria Stoica. Scientific variables ontology, 2019. [Online; accessed 30-August-2019].

[5] National Water Quality Monitoring Council. Nwis public srs names, 2019. [Online; accessed 29-August-2019].

[6] CF Conventions. Guidelines for construction of cf standard names, 2008. [Online; accessed 26-August-2019].

[7] The World Bank. World development indicators, 2019. [Online; accessed 30-August-2019].