The Scientific Variables Ontology: A Blueprint for Custom Manual and Automated Creation and Alignment of Machine-Interpretable Qualitative and Quantitative Variable Concepts

Maria Stoica[1] and Scott D. Peckham[1]

[1]Institute of Arctic and Alpine Research, University of Colorado, Boulder

Abstract

The integration of heterogenous, transdisciplinary digital resources—such as models and data—into customizable workflows is a key step in modeling complex systems. To achieve successful integration, it is necessary to have a uniform means of expressing the conceptual information exchanged between workflow components. The Scientific Variables Ontology (SVO) is a framework for expressing and aligning the flow of information in such complex systems. SVO is an ontological blueprint for creating unambiguous, machine-interpretable representations of quantitative and qualitative variable concepts. Created from the iterative process of inspecting more than 15,000 variables from various disciplines in the natural sciences, the ontology template contains a detailed yet manageable set of atomistic, modular elements as well as a set of principled rules for how to combine those elements to create complex variable concepts. Since their development, the ontological patterns expressed in SVO have been applied to aligning concepts detected in freeform text to variables present in the ontology, as well as to generating new variables in various domains, including rural economics and sustainable development. In this work, we outline the design of SVO as well as review its current applications to concept alignment and automated population of domain-specific variable ontologies.

1 Introduction

The evolution of the state and behavior of complex systems can be modeled by integrating diverse digital resources such as models and data into complex workflows. In such systems, it is critical to have a uniform, machine-interpretable language for identifying the conceptual information, or variables, exchanged between system components. When automating the assembly of workflows, it is critical to correctly link data sets or model outputs to model inputs. Furthermore, when extracting knowledge graphs from unstructured text for driving the creation of modeling workflows, it is important to be able to ground the phenomena and variables detected to the appropriate ontological concepts. The Scientific Variables Ontology (SVO) aims to address these needs.

SVO is a blueprint that provides a template for required and optional components that make up a variable concept. By breaking down the concepts within a variable to atomistic ontological classes and the relationships between them, SVO can be used in a range of applications, such as a template for detecting variables in freeform text, a guide for ontological alignment between concepts presented in different formats, or a blueprint for guided custom variable creation for tagging digital resources, such as data and models. The structure of SVO arose from the iterative, copious analysis of over 15,000 variables in a range of domains across the natural sciences, and has so far proven successful in creating principled variable concepts in other domains, such as economics and sustainable development.

Significance Complicated, interacting systems are not naturally restricted to a single domain; rather, the evolution of various subsystems making up a complex system is highly coupled and interdependent. This interdisciplinary coupling of models and data requires accurately linking the exchange of information via unambiguous variable concepts. Therefore, it is critical to be able to generate a generic, domain-independent template for representing system variables. The Scientific Variables Ontology aims to provide a reusable, modular template for generating uniform variable representations in a complicated modeling context.

2 Ontology Blueprint Description

The Scientific Variables Ontology (SVO) provides a blueprint for constructing qualitative and quantitative scientific variable concepts for labeling resources such as data, models, or knowledge graphs extracted from freeform text. SVO comprises a top-down hierarchy, or taxonomy, (shown in Figure 1), of high level concepts that represent the key components of a variable, as well as a "lateral" architecture, or relationship architecture, (illustrated in Figure 2), that allows the modular mixing and matching of atomic elements to compose complex concepts.

The foundational concepts of SVO are Phenomenon, Property, and Variable. A Phenomenon is anything that is or can be observed to exist or happen, an idea which comes (loosely) from Kant's philosophy that noumena (things as they are, in and of themselves) are perceived as phenomena[1]. Put a different way, a Phenomenon is the object of an observation. In order to observe something, one need not fixate on a specific aspect of it. However, once one desires to communicate information about the observation to others, one must identify which Property of the phenomenon one wants to record. The core principle of SVO is that, in order to communicate about a particular Variable, it is not sufficient to identify just a Phenomenon or just a Property; rather, both concepts are necessary. Furthermore, before a property can be qualitatively or quantitatively assessed, one must first identify the object (Phenomenon) to which a Property belongs. Therefore, a well-defined Variable must comprise both a Phenomenon (object of observation) and a Property (observable). In common parlance, there is a tendency to refer to variables using just a property, such as 'temperature' or 'price', or just a Phenomenon (object), such as 'carbon-dioxide' or 'crop production'. However, in order to automate the linking of data and models properly, a variable must comprise, at the very least, a Phenomenon and a Property. So, for example, 'air temperature' and 'car price' both satisfy the minimum requirements for identifying a variable, since each identifies a Phenomenon object—air and car, respectively—and a property—temperature and price, respectively. Furthermore, in order for a variable to be useful in modeling or analysis, the values its Property can take should be standardized in some way, whether qualitatively or quantitatively.

The examples of 'air temperature' and 'car price' are relatively simple, and it will be rare to use variables that are this elementary in complex modeling. For creating more complex variables, SVO provides modular templates for combining elementary concepts to create custom, complex concepts, including the identification of processes, assigning Phenomena to Roles within more complex Phenomena, labeling reference frames for measurement, and providing spatio-temporal context. More details on the other ontological concepts and constructs provided can be found at the SVO website[2]. Figure 3 shows an example of how modular components can be combined in a custom way to create new variable concepts.

SVO arose from the need to generalize a set of standard names[3]. Although intrinsically SVO is terminology agnostic, in some cases, especially for human accessibility, it may be necessary to identify a variable concept with a standard label that can provide a snapshot of the information associated

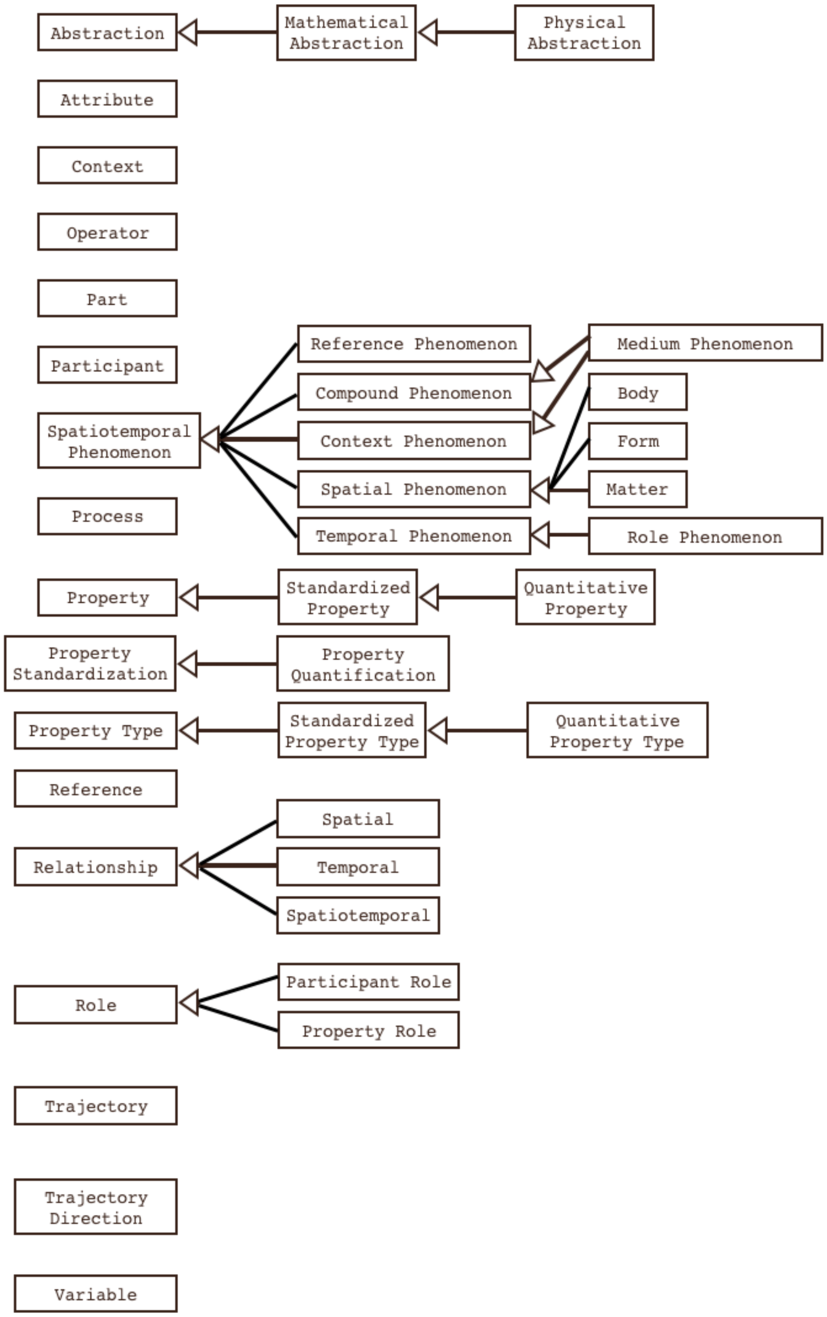


Figure 1: Hierarchical structure of SVO.

with a particular variable concept. To that end, SVO also comes equipped with a standard name generator that uses a specific set of rules to serialize particular key concepts into a one-dimensional, human-readable string (see Figure 4). Although SVO comes with a name construction engine based on the preferred controlled vocabulary and format of the CSDMS standard names[3], custom name generators can be created with a different controlled vocabulary and serialization algorithm. This

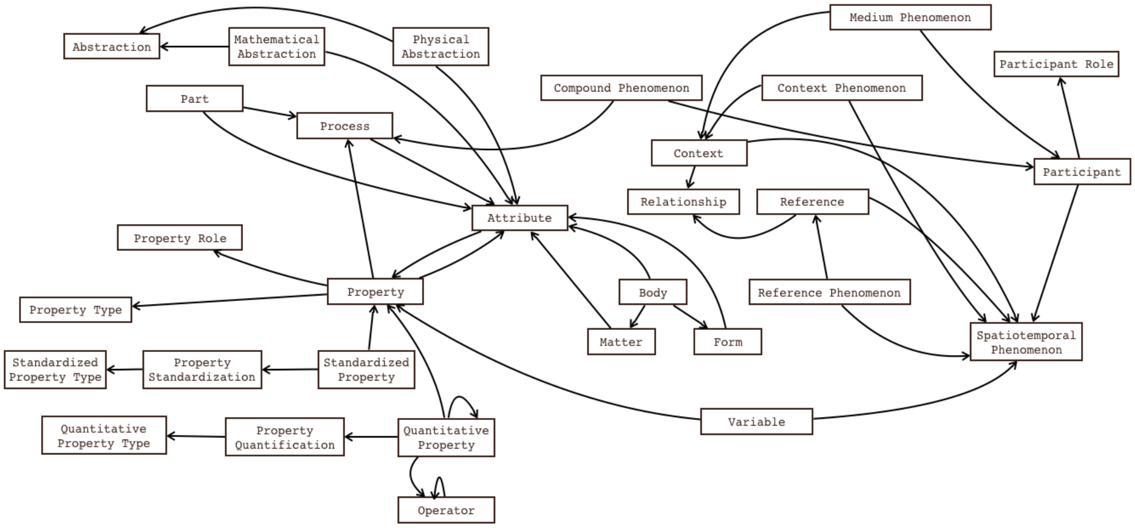
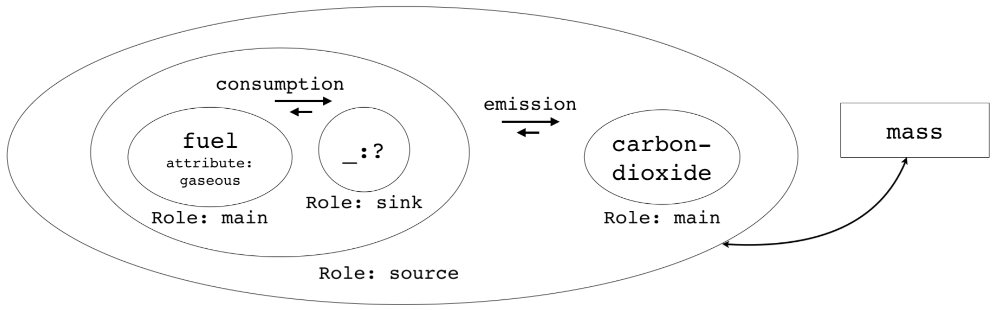


Figure 2: Lateral structure of SVO. This diagram is illustrative of the relationships between the ontological classes. For simplicity, it does not present all relations, nor does it label the relations or the restrictions placed on each relationship.



World development indicator: carbon dioxide emissions from gaseous fuel consumption

Figure 3: Example of how elementary concepts and templates are used to construct more complex variable concepts utilizing SVO. Ovals denote Phenomenon instances and rectangles denote Property instances. Compound Phenomena are created by combining simpler Phenomena, and, if a Process is also involved, assigning Roles to those Phenomena.

can be particularly useful when experts from different domain areas desire to search and scan terms in the ontology.

3 Ontology-guided Concept Creation

As described in the previous section, the core requirement for identifying a variable is the Phenomenon-Property pair. In order to populate an ontology of variables in a domain that is not currently in SVO, multiple approaches are possible. One approach is to implement rules-based detection of word morphology in combination with part of speech tagging to identify what ontological class a term is most likely to belong to. 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

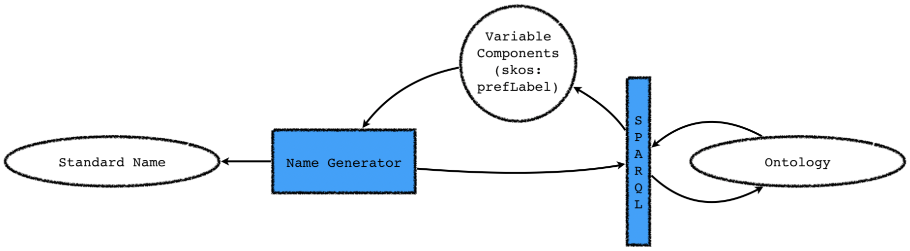


Figure 4: Generating standard names from SVO.

Table 1: Result of automated categorization into ontological classes based on WordNet hierarchy. The terms used for this analysis were previously manually categorized.

Ontological Category	Type	Correct	Incorrect	Not Found	Total
Process	All	1250	216	36	1502
Phenomenon	Natural Body	48	10	8	66
Phenomenon	Complex/Dynamic	46	38	23	107
Property	Qualitative	89	14	9	112

limited to, WordNet, Wiktionary, and Wikidata. In this work, we present the success rate of the latter approach and develop a reasoning algorithm for finding the likely category of a term.

A simple, first pass approach 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 1), with the exception of complex and dynamic phenomena. Current attempts to improve the accuracy of this approach include (a) automatically extending the WordNet graph using information from other linked data sources, (b) creating a definition-based "sentiment analysis"-type algorithm that attempts to categorize a word using its dictionary definition, and (c) extending the resource set to include information from other linked data databases.

In addition to categorizing terms into their ontological classes, the concept creation approach also requires a set of assumptions about commonly used patterns for presenting concept information. The most common patterns are Phenomenon + Process and Process + of + Phenomenon. Once terms in a phrase have been assigned to their ontological classes, a new concept can be created. Additionally, concept refinement can be improved through looking for attributes of components, as well as detecting context and reference frames.

4 Ontology-guided Concept Alignment

In addition to creating new custom domain ontologies following the SVO template, it is also possible to use the SVO structure to implement reasoning algorithms that can determine when concepts can be properly aligned. In order to understand this capability, it is first important to highlight that data and models indicate their variables at different levels of granularity. Descriptions of data tend to be much more descriptive than those of models. The reason for this is that models are designed to be generic and may apply to anything that belongs to a specific category. Data, on the

Table 2: Alignment of various search terms to concepts present in the ontology.

search term	first ontology match	explanation
cereal production	land_crop__production_cost-per-area	cereal can play the role of a crop (information found in Wikidata)
soil moisture	soil_water__volume_fraction	moisture is a state and is thus looked up and associated with the Matter phenomenon water
drought	atmosphere_water__rainfall_volume_flux	drought is a state and is associated with the phenomenon of rainfall
humidity	atmosphere_hydrometeor__volume	humidity is also a condition or state and is associated with water; however, the property alignment is not yet implemented so this part was not correctly identified by the matching algorithm

other hand, needs to be as specific as possible in order to allow for the application of that data in any way that could be useful. For example, if one is interested in the crop production in a region, one will want to track multiple aspects of growing crops. Data about which crops are grown will in general be specific, indicating the family or even species of crop. On the other hand, a model for crop production may be applicable to a wide range of crops, and thus its variables may indicate that it can take any crop data as an input. In such a scenario, it is important to determine what data sources indicate items that can be crops. In a separate scenario, one may wish to model food availability. Some crops are classified as food, but others are not. In order to determine whether this model is appropriate to use with a data set, we would need to determine whether the data pertains to an item that can take the role of food. Since variables in data sets are more granular than those of models, the ontology can guide us in determining whether the phenomenon described by a data set is a subclass of a phenomenon, or can take on a particular phenomenon role, so that it may be modeled with a particular model. For example, data about cereal production may be applicable to both crop production models and food availability models. Utilizing the SVO template as a guide, and leveraging information already stored in large databases such as WordNet and WikiData, we are currently creating and expanding the capability of generating such data to model alignments. Current alignment capabilities from our algorithms are shown in Table 2. An advantage of such an approach is that, as show in the 'explanation' column, alignments using these types of methods have high explainability, and so we can detect whether alignments are coincidental or whether they make sense. An in-development API for concept alignment is provided in the Resources section below.

In addition to the concern of categorical alignment, another concern is whether or not data needs to be transformed between resources. Such transformation may involve unit conversion or it may require a mathematical operation (such as a time derivative or average). To provide support for unit conversions, SVO associates Quantitative Properties with a unit dimension string that can be aligned with QUDT[4], an ontology with capabilities for unit conversion. For mathematical operations, SVO has a class of Operators that can be applied to Quantitative Properties to indicate a transformation.

5 Conclusion

SVO is an ontology template for constructing custom, unambiguous variables in a format that is compatible across domains. Such uniform construction of variables is desirable for the enabling of correct automated alignment of model inputs and outputs and data resources. In this work, we have presented the foundational concepts of SVO—the Phenomenon, Property, and Variable. Additionally, we have reviewed applications of SVO patterns in the automated detection and creation of new variables, as well as its application to variable alignment.

6 Acknowledgments

We 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.

7 Resources

The Scientific Variables Ontology can be found at `http://www.geoscienceontology.org` as well as `http://www.scienceontology.org`. The website provides a simple search tool for browsing the current content of the ontology. The namespace for the ontology is located at `http://www.geoscienceontology.org/svo` (SVO), `http://www.geoscienceontology.org/svo/svu` (SVU), and `http://www.geoscienceontology.org/svo/svl` (SVL). More documentation can be found at these locations when accessed with a web browser. Code related to the generation of the ontology can be found on GitHub at https://github.com/mariutzica/Scientific-Variables-Ontology. An API for alignment of concepts to current ontological entries (currently in development) can be found at `http://termmatch.geoscienceontology.org`.

SVO implementation follows Semantic Web best practices. We provide some references to great resources on ontologies, knowledge organization, and the Semantic Web in the references[5, 6, 7].

References

[1] Russell, Bertrand. The History of Western Philosophy. Simon and Shuster: New York, 1945.

[2] Scientific Variables Ontology. `http://www.geoscienceontology.org`. 2019.

[3] CSDMS Standard Names (CSN). `https://csdms.colorado.edu/wiki/CSDMS_Standard_Names`. 2015.

[4] Quantities, Units, Dimensions, and Types (QUDT) Ontology. `http://www.qudt.org`. 2019.

[5] Dean Allemang and Jim Hendler. The Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL. 2011.

[6] Grega Jakus, Veljko Milutinović, Sanida Omerević, Sašo Tomažič. Concepts, Ontologies, and Knowledge Representation. 2013.

[7] Smiraglia, Richard. The Elements of Knowledge Organization. 2014.