A Pattern for Representing Scientific Taxonomies*

Shirly Stephen[1,*], Cogan Shimizu[2], Mark Schildhauer[1], Rui Zhu[3], Krzysztof Janowicz[1,4] and Pascal Hitzler[5]

[1]University of California, Santa Barbara, USA
[2]Wright State University, Dayton, Ohio, USA
[3]University of Bristol, England
[4]University of Vienna, Austria
[5]Kansas State University, Manhattan, USA

Abstract

Standard taxonomies that are meant to serve as reference specifications of specific scientific domains are developed through extensive review by scientists and experts. For data integration and interoperability needs within Knowledge Graphs (KGs), these taxonomies must be translated into formal ontologies. In this paper we present an ontology design pattern for modeling a scientific taxonomy as an ontology. The focus of the pattern is to 1) capture temporal dynamics of concepts as taxonomies evolve, 2) model the provenance of concepts to add context and enable governance, 3) assist the translation of taxonomic relations to ontological relations appropriately that will empower their use within KGs, and 4) tag provenance and other metadata information to mappings or alignments of uncertainty between concepts in different ontologies.

Keywords

ontology design patterns, semantic web, linked data, scientific taxonomies, SKOS

1. Introductions

A *domain taxonomy* is a type of controlled vocabulary used to structurally organize concepts in a particular domain. Frequently, multiple taxonomies are developed to understand the relationships between concepts in a single domain. Many scientific domains have a *scientific* domain taxonomy that is developed as a *standard*, and meant to serve as a *reference* specification and to assist data interoperability needs. The classification or organization structure, concept definitions, and other metadata in such a taxonomy are generally developed with a certain level of consensus, curation, and peer-review by experts. Examples include the Hazard Information Profile (HIP) taxonomy [1] constructed by the United Nations Office for Disaster Risk Reduction (UNDRR) and the International Classification of Diseases constructed by the World Health Organization (WHO) [2]. However, these scientific taxonomies are largely documented in informal representation formats (e.g., plain text, UML diagrams) and as such cannot be robustly re-used as reference data. It is preferred to represent these taxonomies as formal ontologies that conform to World Wide Web Consortium (W3C) recommendations (e.g., the Web Ontology

WOP22: 13th Workshop on Ontology Design and Patterns, October 23-24, 2022, Hangzhou, China
*Corresponding author.
kulyabov-ds@rudn.ru (S. Stephen)

Language-OWL) for 1) both human and machine interpretability of their precise semantics; 2) modeling taxonomic relationships between concepts, such as subsumption, partonomy, membership, hypernymy; 3) creating mappings between concepts across different taxonomies; and 4) a more meaningful way to capture, store, navigate, discover, retrieve, and archive information in Knowledge Graphs (KGs).

The framework of standard taxonomies in their informal documents may vary for different domains - e.g., for hazards, one could say the metadata framework should capture cause and effect, spatial and temporal characteristics, co-occurring and cascading events, available proactive and recovery measures. But we found that overall, certain discourse elements overlap. Many of these taxonomies have at least one if not several definitions for every concept, each attributed from some reference sources or organizations, along with related and synonymous names. Other attributes that are common include scientific descriptions, units of measurement, vernacular names. Capturing *all* the rich information from the text document into the ontology is necessary for many purposes e.g., for ontology alignment tasks, where we need as much lexical information as possible for better matching. However, for many reasons (lack of time, and knowledge about ontology engineering, etc.), taxonomy-to-ontology formalization efforts mostly represent all the descriptive semantics either using Resource Description Framework (RDF) or Simple Knowledge Organization System (SKOS) annotation properties such as rdfs:comment, rdfs:label, skos:definition in a haphazard manner, and the structural semantics are altogether confused or neglected. Here below we highlight four characteristics of scientific taxonomies that have influenced the construction of the design pattern presented in this paper.

- Science evolves, and as such scientific concepts change over time. Therefore, the *temporal dynamics* of corresponding taxonomies (definitions, names, addition/removal of concepts) have to be captured in the ontology to assimilate new scientific information and model revisions.

- Scientific concepts have canonical names, but may also have many vernacular names in different contexts (e.g., based on language, geographic region, organization). Moreover, the definitions of these concepts are prescribed by subject matter experts within their respective domains (or context). Extending the contextual attributes of names and definitions using provenance, versus simply using annotation "labels" provides a richer description of a concept's profile. It could also enable *governance* to ensure reference data quality.

- Translating a scientific taxonomy into an ontology requires a well thought out approach that addresses not only how to designate classes and properties, but more importantly, how to translate taxonomic relations to structurally and semantically accurate *ontological relations*. This will enable the ontology to be efficiently used for access, navigation, and retrieval in a KG.

- Scientific taxonomies are essentially used for data representation and integration purposes, and therefore may have to be aligned to other analogous or related reference or dataset taxonomies. In this situation, where there is *uncertainty in alignment*, additional metadata information regarding the alignment (e.g., mapping relation, score, matching technique, matcher) will need to be modeled.

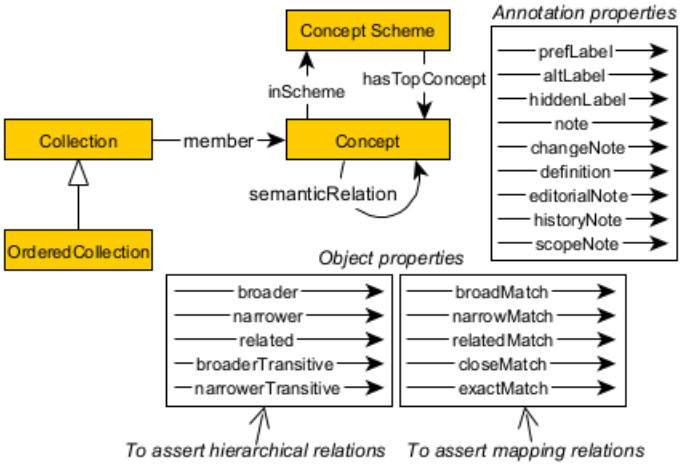

Figure 1: Core classes, their relationships, list of annotation and object properties in the Simple Knowledge Organization System (SKOS) framework-from [6].

Standards such as SKOS [3], PROV [4], and OWL-Time [5] can be re-used together in the context of the above-mentioned issues, but there is no template that domain scientists and experts can refer to, that will guide their taxonomy-to-ontology translation process. In this paper we present a conceptual pattern that can be implemented by linking or deferring to such standard ontologies for the development of formal ontologies from scientific taxonomies[1].

2. Background: Simple Knowledge Organization System

Simple Knowledge Organization System (SKOS) [3], a W3C standard is the most common data model being used to represent controlled vocabularies such as thesauri, and taxonomies in the Semantic Web. In this section we will briefly overview key elements and syntax of the SKOS RDF vocabulary - see Fig. 1.

The skos:Concept class is meant to denote any abstract entity such as an idea, an object, an event. Each unique concept in this class may have several synonymous or vernacular name(s) that can be denoted via their labels. The three annotation properties for representing concept labels are: skos:prefLabel to assign an authoritative name, skos:altLabel for unauthorized name(s) and synonyms, and skos:hiddenLabel for names to be hidden from text-search and visual interfaces. The class skos:Collection (disjoint with skos:Concept) is used to describe labeled or ordered groups of SKOS concepts. The object property skos:member is used to define concept members of a collection.

The object property skos:semanticRelation is used when there is an inherent meaning between two concepts. The two transitive (skos:broaderTransitive, skos:narrowTransitive) and two non-transitive (skos:broader, skos:narrower) sub-properties of skos:semanticRelation are intended to model hypernym-hyponym relations, i.e., one concept is broader or narrower in scope with

[1]The OWL file can be found at https://github.com/shirlysteph/taxonomy-odp

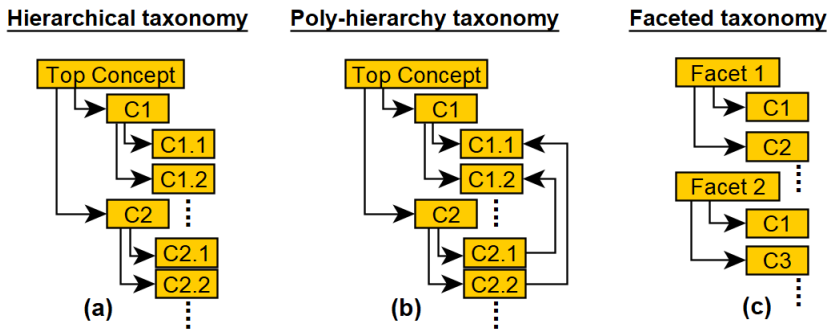

Figure 2: Typical taxonomy structures used for representing reference taxonomies.

another. The object property skos:mappingRelation is used to state mapping (or alignment) between concepts existing in different ontologies. Its five sub-properties include skos:closeMatch (symmetric), skos:exactMatch (symmetric and transitive), skos:broadMatch, skos:narrowmatch, and skos:relatedMatch (symmetric).

3. Overview of Different Taxonomy Structures

The structure of a taxonomy can consist of many levels and sub-levels, each representing a specific category of information. Here we discuss some of the common taxonomy organization structures.

A flat taxonomy, also known as an unlayered taxonomy, is simply a list of concepts, without a top-concept. An example of such a taxonomy is the list of hazards in the National Oceanic and Atmospheric Administration (NOAA) Storm Events Database[2].

A hierarchical taxonomy-see Fig. 2(a) is represented as a tree, where individual concepts are arranged in a hierarchy. A hierarchy is often thought of as subsumption, but in meta-modeling they indicate many kinds of semantic refinements. An example is the biological hierarchy ranking[3], i.e., Domain > Kingdom > Family > Species.

A poly-hierarchical taxonomy-see Fig. 2(b), represents hierarchies having one-to-many child-parent relationships. While constructing a formalization for such a taxonomy, one should not explicitly construct OWL subsumption relationships as a poly-hierarchy, but it is fine to infer concepts into a poly-hierarchy [7]. This helps avoid inconsistencies and unwanted inferences from OWL subsumption reasoning. As an example, Fig. 3 shows a subset of the UNDRR's HIP taxonomy constructed as a poly-hierarchical ontology, but results in an incorrect inference, i.e., *HydrogenCyanide* as an instance of BiologicalHazard. This is when one shifts towards using other forms of semantic relations that are not necessarily transitive e.g., hypernymy or membership relations, to organize concepts.

A faceted taxonomy allows a concept to be classified in multiple ways (sets of attributes), enabling the classification to be ordered in multiple ways, rather than in a single, predetermined

[2]https://www.ncdc.noaa.gov/stormevents/
[3]https://en.wikipedia.org/wiki/Taxonomic_rank

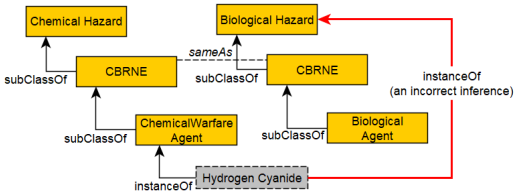

Figure 3: A subset of the UNDRR's HIP taxonomy that shows the poly-hierarchical arrangement of concepts.

order (as in a strict hierarchy). Each facet is a unique way of characterizing concepts, and therefore represents a collection of concepts. Concepts in a faceted taxonomy are either 1) classified by every facet, or 2) classified across different facets and are then related through specific relations, such as partonomy, membership, hypernymy. A faceted taxonomy allows users to locate and discover concepts based on the facets that are important to them, and therefore it does not necessarily have one top concept.

Finally, taxonomies can also be a complex combination, i.e., hierarchical and faceted. The SKOS models favors the formal representation of taxonomies that 1) have a hierarchical tree structure, where the skos:hasTopConcept relation can be used to represent the top-most entry point of a hierarchy; 2) have concepts related through subsumption, hyponym-hypernym, member-collection, or relatedness relations. But scientific taxonomies may interrelate concepts using relations beyond the scope of SKOS. For example, the relation between *StrongWind-Tornado* denotes functional-parthood [8], while the relation between *Earthquake-DebrisFlow* denotes cause-effect [9]. While the limited set of SKOS relations are meant to pragmatically represent taxonomic hierarchical relations, in an ontology using relations that closely resemble the domain will be useful for downstream applications.

4. The Scientific Taxonomy Pattern

The Scientific Taxonomy Pattern (STP) as shown in Fig. 4 and as we will describe below is primarily meant to assist *scientific* taxonomy-to-ontology translation tasks. It is therefore restricted, but meant to be useful for scientific purposes rather than being all-encompassing by using fuzzy names and vague reification.

Concept. The Concept class is intended to denote abstractions of entities, ideas, events, or processes in scientific domains. They have clear, referenced, intentional, and extensional semantics [10]–in the form of definitions and descriptions–that are obtained through scientific or scholarly work. As such, they are identified as the nodes in any taxonomy. Examples of the Concept class are *Hazard*, *Hurricane*, and *Disease*.

ConceptCollection. The ConceptCollection class is used to denote a "bag of concepts". When a set of concepts have certain similar characteristics, they are organized into a category in a taxonomy, and such a node represents a ConceptCollection. For example, *SpecificHazard* is a ConceptCollection that represents the set of all individual hazards (*Tornado*, *TropicalStorm* etc.). *SpecificHazard* is also a Concept with specific hazard properties or metrics such as spatial and

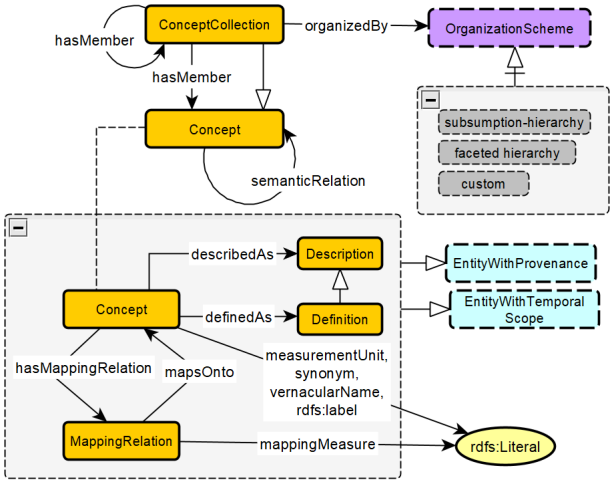

Figure 4: The schema diagram for the Scientific Taxonomy Pattern (STP). Gold boxes represent concepts central to this pattern. Purple boxes with a dashed border represent controlled vocabularies (i.e., classes that have been defined as a specific set of individuals). Yellow ellipses are datatypes. Blue boxes with dashed borders represent interfaces to external patterns or concepts (i.e., representing hidden or additional complexity not covered by this pattern). Black filled arrows are object or data properties and open arrows represent subclass relationships.

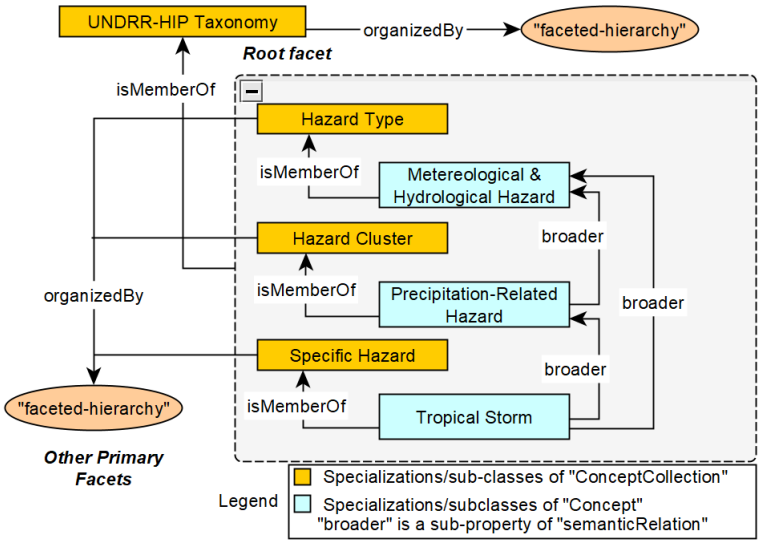

Figure 5: Structure of UNDRR's HIP represented using a subset of its concepts.

temporal characteristics, and human impacts. Thus, an entity that is a ConceptCollection is also simultaneously a Concept (Ax:1) having a prescribed name with characteristic attributes[4].

[4]This interpretation of a collection is slightly different from SKOS collections, which are intended to only represent groupings of "closely-related" concepts. skos:Collection and skos:Concept are therefore disjoint classes.

Membership of a Concept within a ConceptCollection is made using the isMemberOf and its inverse hasMember relations (Ax:2-5). Finally, Concept is also an EntityWithProvenance and an EntityWithTemporalScope (Ax:6-7).

$$\text{ConceptCollection} \sqsubseteq \text{Concept} \tag{1}$$
$$\text{ConceptCollection} \sqsubseteq \forall\ \text{hasMember.Concept} \tag{2}$$
$$\exists\ \text{hasMember.Concept} \sqsubseteq \text{ConceptCollection} \tag{3}$$
$$\text{ConceptCollection} \equiv \exists \text{hasMember.}\top \tag{4}$$
$$\text{isMemberOf} \equiv \text{hasMember}^{-} \tag{5}$$
$$\text{Concept} \equiv \text{EntityWithProvenance} \tag{6}$$
$$\text{Concept} \equiv \text{EntityWithTemporalScope} \tag{7}$$

OrganizationScheme. The OrganizationScheme class is meant to capture the taxonomic or organization structure of a ConceptCollection - represented through the organizedBy relation (Ax:8-10). The structures by which concepts and concept collections can be organized are subsumption-hierarchy, faceted-hierarchy, or a complex combination. Ideally each ontology developed from a taxonomy must have a "root" ConceptCollection entity that will indicate the source scheme from which the descendant concepts and concept collections are derived. This root entity is also the primary form of access into the ontology within a KG. A ConceptCollection may also be representative of a *facet*, which is a way of organizing a set of concepts (or instances) based on a perspective or purpose that makes more sense to a human or system. In that sense every Concept or ConceptCollection is a facet. When the root ConceptCollection of an ontology is organized according to a faceted organization scheme, each of its direct descendant or member concept collections will now have an organization scheme of its own. Fig. 5 shows the faceted organization scheme for the UNDRR's HIP taxonomy. The root ConceptCollection class is organized based on faceted-classification scheme, where the members at the next level are the three facets of ConceptCollection classes namely *HazardType*, *HazardCluster*, and *SpecificHazard* that are also organized following a faceted structure.

$$\text{ConceptCollection} \sqsubseteq \forall\ \text{organizedBy.OrganizationScheme} \tag{8}$$
$$\exists\ \text{organizedBy.OrganizationScheme} \sqsubseteq \text{ConceptCollection} \tag{9}$$
$$\text{ConceptCollection} \sqsubseteq =1\ \text{organizedBy.OrganizationScheme} \tag{10}$$

semanticRelation. The general semanticRelation is included here as a placeholder for any organizational structure specific-relations between concepts (Ax:11-12). Such relations could be explicit subsumption relations from RDF, semantic refinement (i.e., hyponym/hypernym) relations from SKOS, variations of partonomic or membership relations such as those mentioned in [8]. Fig. 5 shows concepts across the three facets related through the hyponym relation *broader*, a specialization of semanticRelation.

$$\text{Concept} \sqsubseteq \forall\ \text{semanticRelation.Concept} \tag{11}$$
$$\exists\ \text{semanticRelation.Concept} \sqsubseteq \text{Concept} \tag{12}$$

MappingRelation. Ontology alignment tasks determine mappings between terms in two ontologies. Automated alignment between two ontologies that are constructed from taxonomies rely mostly on their lexical information (concept names, labels, definitions, synonyms, descriptions, etc). Due to lexical ambiguity, the risk of matching two unrelated concepts increases, which may affect precision. Moreover the mappings generated by most alignment techniques represent is-a (i.e., subsumption) or equivalence relations by default. Our experience has shown that the generated mappings may also indirectly indicate other specific semantic relations that could be derived using machine learning approaches. To reliably represent the alignments between ontologies, we need to capture qualitative mappings of varying degrees of complexity, and corresponding quantitative mapping (or similarity) measures. The MappingRelation class, a reification of the mapsOnto property is introduced for this purpose i.e., to attach provenance, to denote different kinds of semantic mappings, and to capture uncertainty measure of alignment (Ax:13-17). We note in Ax:16 that the inverse filler of hasMappingRelation must exist and must be a Concept.

$$\mathsf{MappingRelation} \sqsubseteq \forall\ \mathsf{mapsOnto.Concept} \tag{13}$$

$$\exists\ \mathsf{mapsOnto.MappingRelation} \sqsubseteq \mathsf{Concept} \tag{14}$$

$$\mathsf{MappingRelation} \sqsubseteq \exists\ \mathsf{mapsOnto.Concept} \tag{15}$$

$$\mathsf{MappingRelation} \sqsubseteq \exists\ \mathsf{hasMappingRelation}^{-}\mathsf{.Concept} \tag{16}$$

$$\mathsf{MappingRelation} \sqsubseteq \mathsf{EntityWithProvenance} \tag{17}$$

Definition. We specify *Definition* as the description of a scientific concept having some level of consensus amongst scientists or experts in the domain. They are obtained from an authoritative source such as international or government agencies, or academic, or other scientific sources with literature reference(s). As reducing *ambiguity* is a goal of standard taxonomies, there is a need for ontology patterns that can reconcile established meanings in different sub-disciplines. For example, the definition of an *EnvironmentalDisaster* in disaster management tends to focus on causes, whereas from the perspective of climate change or sustainability the focus is on effects [11]. This is why standard taxonomies emphasize or mention alternative definitions to capture a better conceptual understanding.

On the other hand, conceptual landscape of domains change over time resulting in taxonomies that evolve to allow for more precision, representation, coverage, and better consensus. These changes may be reflected in the form of the addition/replacement/removal of definitions and concepts, or updated classification structure. For all these reasons, it is necessary that the formal representations are capable of capturing temporal trends of the taxonomy, but also empower their functionality as *reference* specifications by ascribing provenance to concept names, definitions, and descriptions.

$$\mathsf{Description} \sqsubseteq \mathsf{Definition} \tag{18}$$

$$\mathsf{Concept} \sqsubseteq \forall\ \mathsf{definedAs.Definition} \tag{19}$$

$$\exists\ \mathsf{definedAs.Concept} \sqsubseteq \mathsf{Definition} \tag{20}$$

$$\mathsf{Concept} \sqsubseteq \forall\ \mathsf{describedAs.Description} \tag{21}$$

$$\exists\ \mathsf{describedAs.Concept} \sqsubseteq \mathsf{Description} \tag{22}$$

$$\text{Definition} \sqsubseteq \text{EntityWithProvenance} \tag{23}$$

$$\text{Definition} \sqsubseteq \text{EntityWithTemporalScope} \tag{24}$$

EntityWithProvenance. This concept is left as a placeholder for assigning provenance. Detailed modeling of provenance is already done in W3C standard, the PROV Ontology [4] which can be adopted as needed. Thus, by explicating ConceptCollection as an entity with provenance one can also tag additional metadata information providing context.

EntityWithTemporalScope. This concept is left as a placeholder for capturing temporal information. Detailed modeling of temporal concepts is already done in the OWL-Time ontology [5] which can be adopted as needed.

Other Metadata Attributes.

$$\text{Concept} \sqsubseteq \forall\ \text{measurementUnit.rdfs:Literal} \tag{25}$$

$$\text{Concept} \sqsubseteq \forall\ \text{synonym.rdfs:Literal} \tag{26}$$

$$\text{Concept} \sqsubseteq \forall\ \text{vernacularName.rdfs:Literal} \tag{27}$$

5. A Worked Example for UNDRR's HIP Taxonomy

UNDRR's HIP taxonomy [1] is a harmonized hazard typology containing definitions and descriptions to aid disaster risk management efforts. While the taxonomy is "not a scientific product", it is the most up-to date, extensively consulted, and consolidated hazard nomenclature developed by a large group of domain scientists and inter-government experts.

The template used in this taxonomy is developed both from a scientific perspective (i.e., reviewed by experts for robustness and scientific consensus), and user-driven perspective (i.e., acknowledging the different contexts and useful ways by which hazards can be organized). As a result of the latter approach, the resulting HIP taxonomy is not a single hierarchy of concepts that can be organized into a neat subsumption hierarchy, but rather is multi-dimensional (or multi-faceted), with concepts across three different facets semantically and meaningfully interlinked–see facets in Fig. 5. In addition to structural information, each HIP captures the following metadata information: hazard name, hazard reference number, definition(s) with reference, synonyms, scientific description, globally used metrics and numeric limits, references to key relevant UN conventions or multilateral treaties, examples, key references to support facts and statements made for each hazard, coordination agency or organisation that provided technical guidance on the hazard.

Fig. 6 shows a subset of the HIP for *Extra-TropicalStorm* that populates a portion of the pattern, emphasizing how concepts may be connected by reusing relations from SKOS, PROV, and OWL-Time.

6. Conclusion

Modeling scientific taxonomies as formal ontologies is crucial in their utilization for data integration, discovery, and exploration purposes within KGs. Scientific taxonomies are inherently

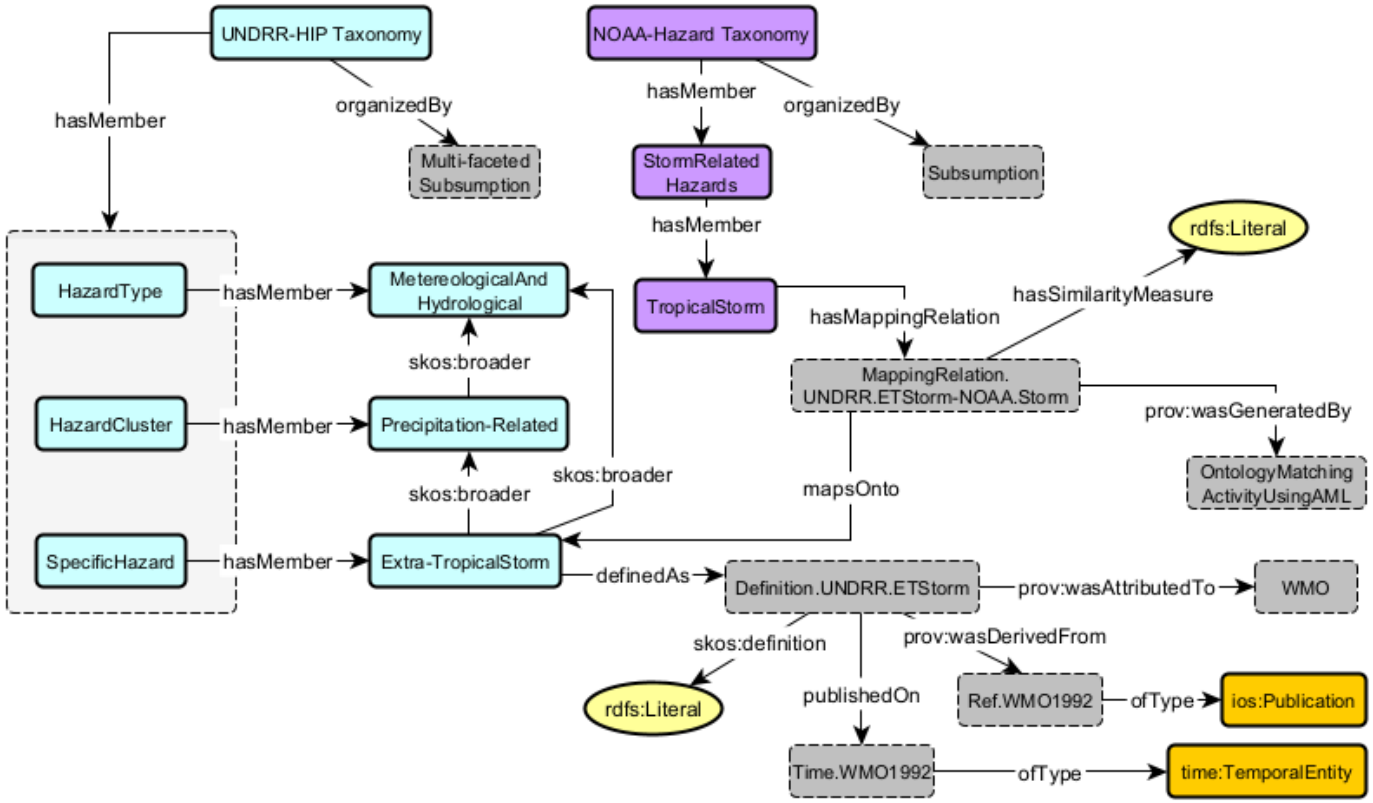

Figure 6: Boxes in yellow indicate general concepts from the Taxonomy Alignment Pattern (they could also indicate concepts from external vocabularies, e.g. SKOS, PROV-O, IOS, OWL-Time etc); boxes in blue indicate concepts from UNDRR; boxes in grey indicate instances.

different from non-scientific taxonomies in striving for clarity, consistency, and universality in the use of terminologies, and hence there is a need to appropriate model in their ontology evolving conceptual trends, provenance of the standardized semantics, meaningful structural relations, and other scientific metadata. The most commonly used SKOS standard is too informal to optimally represent all these aspects, but can be expanded and used in combination with other W3 standards for more precise representations of domain taxonomies as is desirable in scientific inquiry. As such, we have developed the Scientific Taxonomy Pattern (STP) to augment taxonomy-to-ontology translation tasks by demonstrating how to 1) model temporal dynamics of concepts, their definitions, names, 2) enable semantic governance of reference concepts and associated information, 3) translate taxonomic hierarchical relations to useful ontology relations that extend the scope of simple subsumption and SKOS, 4) capture provenance and mapping metadata for alignment with uncertainty. Additionally, we have demonstrated using an example, how the pattern can be used to develop an ontology for the UNDRR's HIP taxonomy by deferring to existing W3C standards such as SKOS, PROV, and OWL-Time.

Future work will focus on expanding the semanticRelation property for more refined modeling of structural relations that will be useful for domain modeling needs.

Acknowledgement. This material is based upon work supported by the National Science Foundation under Grant No. 2033521: "KnowWhereGraph: Enriching and Linking Cross-Domain Knowledge Graphs using Spatially-Explicit AI Technologies". Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Hazard Information Profiles: Supplement to UNDRR-ISC Hazard Definition & Classification Review - Technical Report, https://www.undrr.org/publication/hazard-information-profiles-supplement-undrr-isc-hazard-definition-classification, 2021.

[2] WHO, International classification of diseases for mortality and morbidity statistics (11th revision), 2018.

[3] A. Miles, S. Bechhofer, SKOS - Simple Knowledge Organization System reference, W3C recommendation 18 (2009) W3C.

[4] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, PROV-O: The PROV ontology (2013).

[5] J. R. Hobbs, F. Pan, Time ontology in OWL, W3C working draft 27 (2006) 3–36.

[6] T. Baker, S. Bechhofer, A. Isaac, A. Miles, G. Schreiber, E. Summers, Key choices in the design of Simple Knowledge Organization System (SKOS), Journal of Web Semantics 20 (2013) 35–49.

[7] A. Rector, N. Drummond, M. Horridge, J. Rogers, H. Knublauch, R. Stevens, H. Wang, C. Wroe, OWL pizzas: Practical experience of teaching OWL-DL: Common errors & common patterns, in: International Conference on Knowledge Engineering and Knowledge Management, Springer, 2004, pp. 63–81.

[8] M. E. Winston, R. Chaffin, D. Herrmann, A taxonomy of part-whole relations, Cognitive science 11 (1987) 417–444.

[9] B. Rehder, Categorization as causal reasoning, Cognitive Science 27 (2003) 709–748.
[10] W. G. Stock, Concepts and semantic relations in information science, Journal of the American Society for Information Science and Technology 61 (2010) 1951–1969.
[11] P. R. Mulvihill, The ambiguity of environmental disasters, Journal of environmental studies and sciences 11 (2021) 1–5.