Harmonizing GCW Cryosphere Vocabularies with ENVO and SWEET. Towards a General Model for Semantic Harmonization

RESEARCH PAPER

RUTH DUERR
PIER LUIGI BUTTIGIEG
GARY BERG CROSS
KAI LEWIS BLUMBERG
BRANDON WHITEHEAD
NANCY WIEGAND
KATE ROSE

*Author affiliations can be found in the back matter of this article

ABSTRACT

This paper presents the specific process used by members of the Earth Science Information Partners (ESIP) Semantic Harmonization Cluster, to harmonize cryospheric terms gathered by the Global Cryosphere Watch (GCW) with two leading semantic resources used in the Earth and Environmental science communities—the Semantic Web for Earth and Environmental Terminology (SWEET) and the Environment Ontology (ENVO). This process led to updates to both ENVO and SWEET as well as the development of an alignment file relating cryospheric terms in ENVO to those in SWEET. In addition, we summarize several leading practices which may be applied to other projects/realms within Earth and Environmental science and perhaps beyond, as well as suggest a generalized process for doing so. This paper describes the history of the effort, the technical and decision-making processes used to resolve differences between semantic resources, and describes several issues encountered, with a focus on those that were addressed during the effort. Lessons learned, examples of the problems encountered and a summary of resulting leading practices growing out of this work is provided.

CORRESPONDING AUTHOR:
Ruth Duerr
Ronin Institute for Independent Scholarship, United States
ruth.duerr@ronininstitute.org

KEYWORDS:
semantic resource; semantic harmonization; cryosphere; lessons learned; FAIR data; leading practices

TO CITE THIS ARTICLE:
Duerr, R, Buttigieg, PL, Cross, GB, Blumberg, KL, Whitehead, B, Wiegand, N and Rose, K. 2024. Harmonizing GCW Cryosphere Vocabularies with ENVO and SWEET. Towards a General Model for Semantic Harmonization. *Data Science Journal*, 23: 26, pp. 1–22. DOI: https://doi.org/10.5334/dsj-2024-026

DOI: 10.5334/dsj-2024-026

INTRODUCTION

Over the last decades it has become apparent that to solve any of humanity's pressing issues, inter- and trans-disciplinary research is needed. This requires that data that are collected, developed, and described for one community become readily accessible and understandable by other communities, that the data become globally FAIR (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al. 2016).

What is often not understood by researchers is that for data to be FAIR, both the data and its metadata must be amenable to reasoning by both humans and computers ('FAIR Principles' 2015). This implies that formally defined language be used to describe the structure and content of both the data and its metadata ('FAIR Principle I1' 2015). *Consequently, understanding and harmonizing disciplinary semantic resources with those in other fields is necessary* (Gil et al. 2018).

Historically, the data systems used by the research community were independently developed and customized to suit their requirements. Underpinning these systems are a variety of semantically heterogeneous resources, including controlled vocabularies, glossaries, thesauri, and ontologies (see Figure 1 and section Types of Semantic Resources). Moreover, these underlying resources come in a wide variety of formats, including spreadsheets, documents, programming languages, and schemas, which are typically embedded with a non-trivial amount of tacit domain knowledge. Consequently, these data systems, which may support large, well-established user communities such as those of the Global Cryosphere Watch, are unlikely to naturally merge with those of other disciplines without a great deal of effort. *In light of this problem, it is increasingly clear there is a pressing need for a sound and sustainable way to align and harmonize these underlying semantic resources in order to allow for inter-, cross- and trans-disciplinary data discovery and use.*

The World Meteorological Organization's (WMO) Global Cryosphere Watch (GCW) supports many historical, or legacy, discipline-specific research data. The term 'cryosphere' refers collectively to the portions of the earth where water is in solid form, including snow and ice cover, sea ice, river ice, lake ice, glaciers, ice caps, ice sheets, and seasonally and perennially frozen ground (permafrost). Given the geographic scope of the cryosphere, its data comprise several scientific and sociological disciplines and is thus extremely heterogeneous. A few examples include remotely sensed data acquired by satellites, airplanes, and drones; long-term time-series data gathered at stations such as permafrost borehole temperature profiles and ship-born sea ice and ocean temperature profiles; 'in-situ' sample data such as snow depth, density and water equivalent, ice cores, sea ice, or permafrost soil samples; laboratory measurements and experimentally derived data; and computational environmental models.

The cryosphere is an integral part of the global climate system. The presence or absence of snow and ice affects heating and cooling over the Earth's surface, influencing the entire planet's energy balance. Indeed, as the 2023 Global Tipping Points Report (Lenton et al. 2023) notes, of the five major systems currently at risk of crossing tipping points, four of them—the Greenland and West Antarctic ice sheets, the North Atlantic Subpolar Gyre circulation and permafrost regions—all have cryospheric components. *Thus, harmonizing the semantic resources underlying data systems holding cryospheric data is critical to enabling the inter-, cross-, and trans-disciplinary research needed to understand the impacts of and to mitigate climate change.*

The Earth Science Information Partners (ESIP) is a non-profit organization with a mission to 'empower innovative use and stewardship of Earth Science data to solve our planet's greatest challenges' (ESIP 2023). Supported by the US National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), and the US Geological Survey (USGS), and with more than 130 member organizations, ESIP provides a neutral, open, and welcoming space for collaboration between researchers, educators, industry, and government agencies to accomplish these goals.

In 2009, ESIP convened a Semantic Web Cluster to help its community adopt a wide range of technologies to digitally represent knowledge from diverse scientific domains and bridge between them. As the popularity and importance of semantic technologies grew, this cluster

was promoted to become the Semantic Technologies Committee in 2016 to address needs in this operational space. In ESIP, Committees can convene their own clusters, and as recognition of the substantial expertise and domain knowledge present within the ESIP community, several subsidiary clusters were formed to address specific aspects of semantics.

One of these clusters was the ESIP Semantic Harmonization Cluster which was formed in 2018 to *propose a route towards sustainably bridging terminologies across the Earth Sciences to other domains, as well as to disseminate best practices for harmonizing semantic resources*. Successful bridges need to be usable across implementation scenarios and user communities, as well as applicable across the spectrum of semantic resource types—that is, from resources with weak expressivity such as controlled vocabularies and glossaries (see Figure 1), through those that support best practices for publishing structured scientific data on the Web (Shepherd et al. 2022), and to those that enable computational reasoning—that is, ontologies.

In this paper, we describe the methods used to harmonize cryosphere terms from the 27 semantic resources in the Global Cryosphere Watch (GCW) glossary compilation with two major Earth science ontologies, ENVO and SWEET, and propose a general process for harmonizing semantic resources across the semantic ladder. This work was done as a project through ESIP *to fulfill the mandate of the ESIP Semantic Harmonization Cluster*.

BACKGROUND: TYPES OF SEMANTIC RESOURCES

In the Earth Sciences there is no single semantic resource or semantic resource type to rule them all. The phrase *semantic resource* typically refers to a *spectrum* of artifacts ranging from simple controlled vocabularies (e.g., term lists) to complex, logically consistent, and formally rigorous structures (e.g., ontologies), each providing a level of interoperability to innumerable applications (see Figure 1). The terminology describing semantic resources varies significantly depending on the community with which it is employed. As such, the following are the types of semantic resources considered during this work along with our definitions for each.

- Controlled vocabulary (e.g., term list): Limited set of terms in a sequential order without definition (Zeng 2008).
    - Example: AGU Index of terms (AGU 2021)
- Glossary: Alphabetical list of terms with definitions (Zeng 2008)
    - Example: Glossary of Geology (Neuendorf, Mehl, Jr., and Jackson 2011)
- Thesaurus: sets of terms representing concepts and the relationships connecting them (Zeng 2008).
    - Example: The USGS Thesaurus ("USGS Thesaurus" 2023)
- Taxonomy: Divisions of terms into ordered, hierarchical groups, or categories based on particular characteristics (Zeng 2008).
    - Example: The classification of living organisms by their Kingdom, Phylum, Class, Order, Family, Genus, and Species
- Ontology: More than a taxonomy in that an ontology is a structured vocabulary in which 1) terms (classes) are related by logically consistent axioms (defined in a formal language), primarily formal subclass/superclass relations where subclasses inherit all the properties of their superclass(es) and 2) terms are associated with consistently written, human-readable definitions (such as from a controlled vocabulary), which are aligned to their logical axioms.
    - Example: ENVO (Buttigieg et al. 2023)

Each type of semantic resource defined above has been placed on the semantic ladder depicted in Figure 1 along with the three resources used in this work (GCW glossaries, SWEET, and ENVO).

METHODS

As previously described, the ESIP Semantic Harmonization Cluster was formed to develop processes for sustainably bridging terminologies across the Earth Sciences and to other

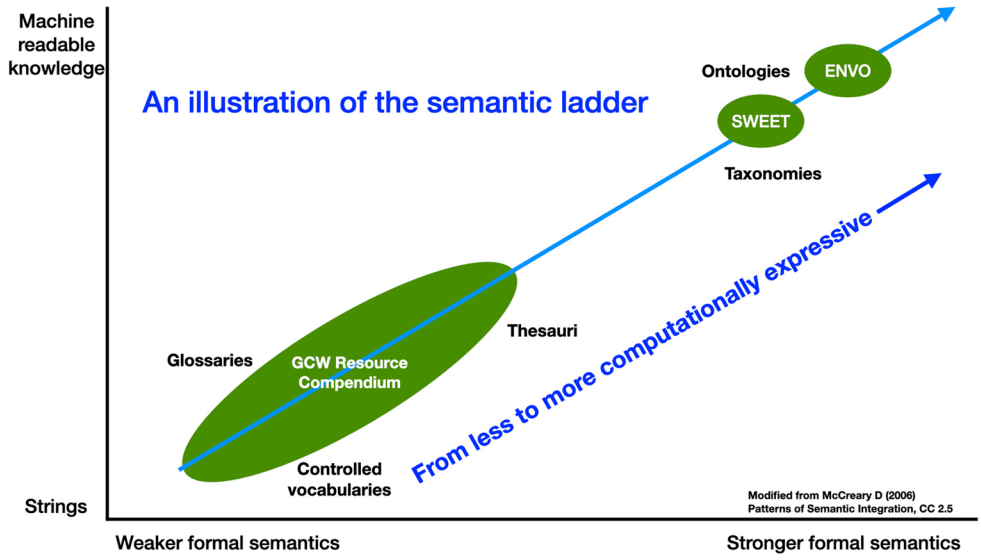

Figure 1 A depiction of the semantic ladder illustrating the extent of machine-aided interoperability of semantic resources, loosely based on Dan McCreary's 2006 presentation (McCreary 2006).

related domains, as well as to disseminate best practices for harmonizing semantic resources. Figure 2 depicts the general process used here, which is reproducible across other projects and disciplines.

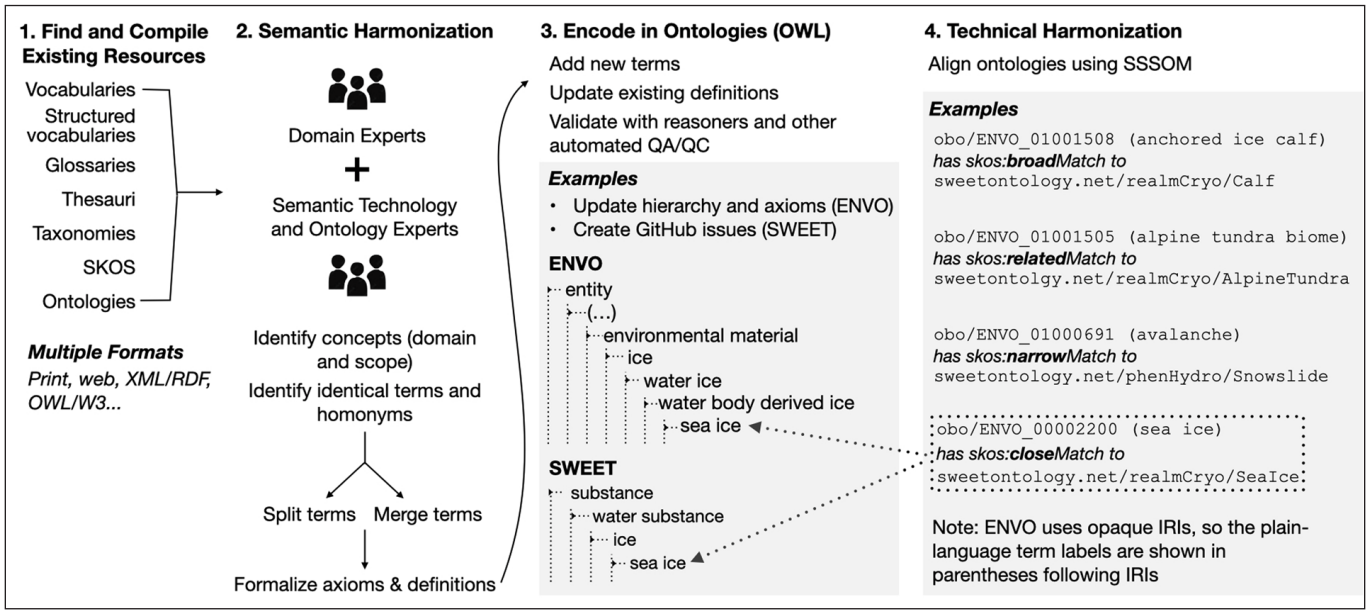

Figure 2 Overview of the harmonization process used in the project and described below.

STEP 1: FIND AND COMPILE EXISTING RESOURCES

The first task was to select the set of semantic resources to harmonize and the discipline to cover. Given the expertise within the group and the critical importance of the cryosphere to climate change impacts, we agreed that cryospheric terminology would be our focus.

This task was greatly aided by previous work commissioned by the GCW to analyze the 27 cryospheric semantic resources they had gathered (Duerr 2018b; 2018a). One result of that work are tables containing terms:

- that are not problematic from a semantic standpoint
    - In this case usually because only one of the glossaries defined the term or where multiple definitions are exact copies of each other and therefore do not conflict.
- where multiple definitions could be coalesced into a single definition,
    - For example, the term *adfreezing* is defined as 'the process by which two objects are bonded together by ice formed between them' in the International Permafrost Association's glossary (van Everdingen 2005) and as 'the process by which one object becomes adhered to another by the binding action of ice' in the AMS Glossary of Meteorology (American Meteorological Society 2024). These two definitions were combined into the ENVO definition 'a freezing process during which two objects adhere to each other via ice.'

- where the terminology was inconsistent and therefore problematic from a semantic standpoint, that is, where definitions are in conflict
    - For example, the term 'blizzard' is defined as 'violent and very cold wind which is loaded with snow, some of which has been raised from snow covered ground' by the Australian Bureau of Meteorology (Australian Government, n.d.); as 'a severe weather condition characterized by reduced visibility from falling and/or blowing snow and strong winds that may be accompanied by low temperatures' by Canada (Government of Canada n.d.); and having 'sustained wind or frequent gusts of 16 m per second (30 kt or 35 mi per hour) or greater, accompanied by falling and/or blowing snow, frequently reducing visibility to less than 400 m (0.25 mi) for 3 hours or longer' by the US National Weather Service (NOAA/NWS 2009). There are additional definitions for other regions such as France, England, and Russia as well, each with some distinguishing set of criteria that usually differs in some way from the examples given here. This example is discussed further below.
- where community resolution was needed to either agree on a definition or to split the terms up into separate entities.
    - See the detailed discussion of the term *calving* and the *calving process* in the Results section and in Figure 5.

Terms from the categories above formed the initial scope of this project.

A recent survey identified both the Semantic Web for Earth and Environmental Terminology (SWEET) and the Environment Ontology (ENVO) as amongst the five most important semantic resources within the community (Whitehead 2022). Of the other three resources in the group, neither QUDT (FAIRsharing Team 2015) nor the Sensor, Observation, Sampler, and Actuator/Semantic Sensor Network (SOSA/SSN) (Haller et al. 2019; Janowicz et al. 2019) contains cryospheric terminology. The last member, the UK Natural Environment Research Council (NERC Vocabularies) (British Oceanographic Data Centre 2023), is focused on marine science but not the cryosphere. Moreover, Wolodkin, Welland, and Grieb explicitly mention the need to bridge between SWEET and ENVO in order to facilitate reuse of biodiversity data (Wolodkin, Weiland & Grieb 2023). The previously mentioned survey also noted that SWEET should be harmonized with other semantic resources. Consequently, the cluster agreed to harmonize GCW terminology within and between both SWEET and ENVO.

SWEET (McGibbney et al. 2022) organizes over 11,000 Earth and Environmental concepts into roughly 200 separate ontology modules based on nine top-level categories (below), some of which contain subcategories with cryosphere-related terms (Table 1):

- Representation – Math, Space, Science, Time, Data,
- Realm – Ocean, Land Surface, Terrestrial Hydrosphere, Atmosphere, Heliosphere, Cryosphere, Geosphere,
- Phenomena (macro-scale) – Ecological, Physical,
- Process (micro-scale) – Physical, Biological, Chemical, and Mathematical,
- Matter – Living Thing, Material Thing, Chemical,
- Human Activities – Decision, Commerce, Jurisdiction, Environmental, Research,
- Property (observation) – Binary Property, Quantity, Categorical Property, Ordinal Property
- State (adjective, adverb) – Role, Biological, Physical, Space, Chemical, and
- Relation (verb) – Human, Chemical, Physical, Space, Time

Initially developed at NASA's Jet Propulsion Lab (Raskin & Pan 2005) and originally based on the Global Change Master Directory (GCMD) keywords (Nagendra et al. 2001), SWEET is now officially under the governance of the ESIP federation. Despite the broad coverage, historically, SWEET did not include terminology definitions or their equivalent machine readable axioms, so despite routinely being referred to as a set of ontologies in relation to the semantic spectrum, in many areas SWEET is more along the lines of a taxonomy or *lightweight* ontology (Giunchiglia & Zaihrayeu 2017).

ENVO was initially created to represent environmental characteristics in which biological entities are found. ENVO includes, for example, descriptions of physical environments such as geological, ecological, or astronomical (Buttigieg et al. 2013; 2016). As such, expanding ENVO to include cryospheric terms enhances ENVO's coverage of physical environments.

In relation to the semantic ladder (see Figure 1), ENVO is an ontology with both human and machine-readable axiomatic definitions. It is being developed following the recommendations and principles of the Open Biological and Biomedical Ontologies (OBO) Foundry and Library (OBO Technical Working Group 2022) and can be formally represented in the Ontology Web Language (OWL) or OBO formats. ENVO is aligned with the Basic Formal Ontology (Arp, Smith & Spear 2015; Brochhausen et al. 2019) at an *upper* level, so that ENVO is interoperable with other OBO ontologies. Compared to SWEET, ENVO has numerous defining axioms and overall is a more formally rigorous ontology.

STEP 2: SEMANTIC HARMONIZATION

Work proceeded by identifying SWEET terms that were cryospheric from within that subset of SWEET files whose name indicated that they were likely to contain relevant terms (see the list of files addressed in Table 1). A Google sheet containing the relevant SWEET terms was created for each SWEET file addressed (Semantic Harmonization Cluster 2023).

For each SWEET term on the spreadsheets, the team determined whether there were equivalent terms in the GCW compilation. If not, the term was not addressed further. If the SWEET term was found in the GCW compilation, then we searched for the term in the ENVO ontology. If found, we paid attention to which hierarchy, that is, superclass, it was under compared to SWEET's hierarchies to be sure we had a match. Then additions or updates to ENVO were made using guidelines developed by Seppälä et al. (Seppälä, Ruttenberg & Smith 2017) and extended for ENVO (Buttigieg 2021). This included creating minimal but robust definitions following the genus-differentia model which produces definitions of the form 'X is a Y that Zs' and numbering each discrete differentia in the definition (see Figure 4 for an example) as well as ensuring that the axioms for the term reflect the differentia in the definition (see Figure 3 for an example).

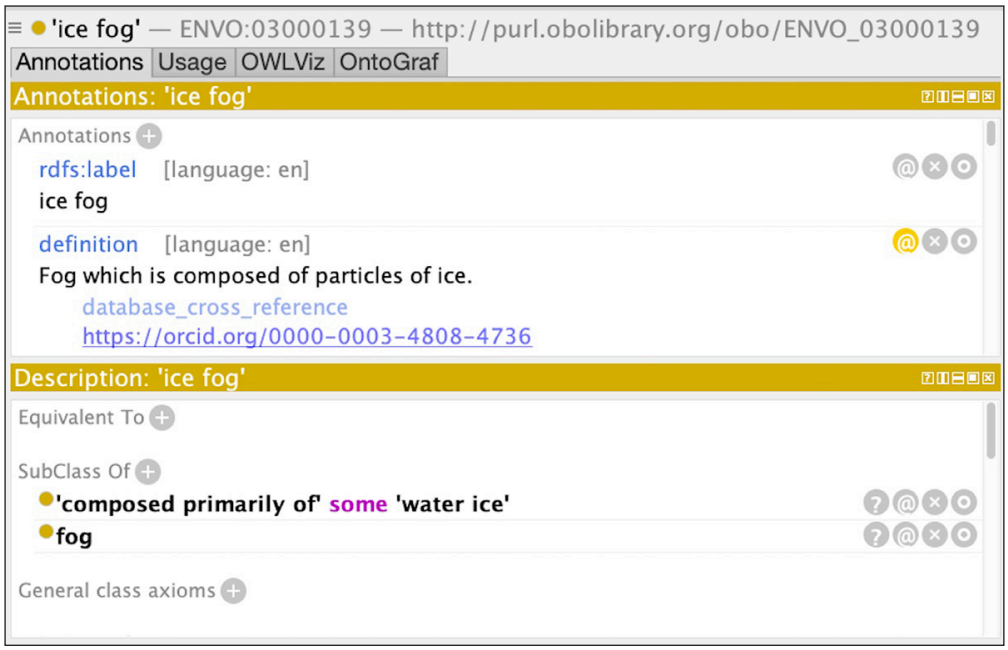

Figure 3 Term *ice fog* added to the ENVO ontology using a GCW derived definition showing parallel definition and axioms.

Many of the terms in the GCW compilation included additional information that went well beyond a definition. These extra materials were not included as part of the ENVO definition, but instead kept as separate annotating comments on the ENVO term (see Figure 4 for an example). When revising definitions or adding terms to ENVO, we paid special attention to the taxonomically inherited axioms of each class, correcting issues higher in the ontology hierarchy or adding additional levels to the hierarchy as needed.

We initially intended to update SWEET directly as well—adding definitions and relationships to the equivalent terms from ENVO directly into SWEET. However, during the project a SWEET roadmap was debated within the larger ESIP Semantic Technologies Committee which might have invalidated our work. Instead, we opted to create GitHub Issues for anything related to

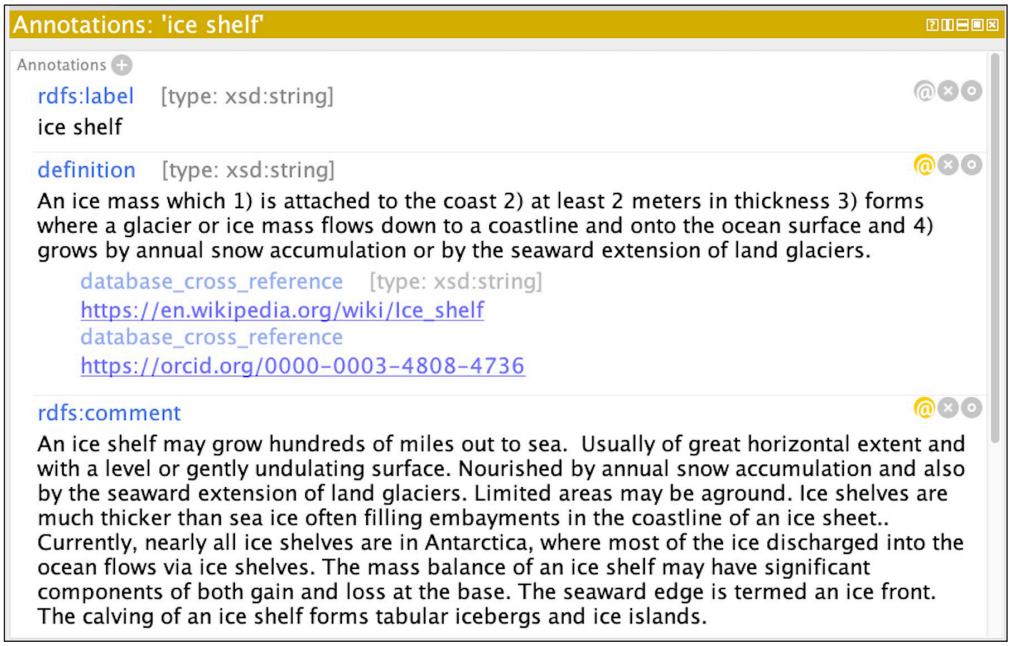

Figure 4 Term *ice shelf* added to ENVO with numbered differentia and added GCW comments.

SWEET, to defer the addition of definitions to after completion of the roadmap, and to record SWEET and ENVO term relationships using the recently developed Simple Standard for Sharing Ontological Mappings (SSSOM) (Matentzoglu et al. 2022) (see Step 4 below).

STEP 3: ENCODE IN ONTOLOGIES (OWL)

Initially, the examination of terms in ENVO occurred using the Protégé ontology editor (Musen 2015) and the development branch of ENVO available from the ENVO GitHub repository. We were editing/updating ENVO one term at a time. However, later in the project, after having worked through many terms using this process, we switched to using a ROBOT spreadsheet (Jackson et al. 2019; Overton et al. 2015) to automate the process of updating ENVO in bulk.

ROBOT is a general-purpose command-line tool for working with ontologies and is used by many projects contributing to the OBO Foundry. It provides commands for merging ontologies, extracting subsets, filtering for selected axioms, running reasoners, and converting between file formats. ROBOT commands can be chained together to form powerful, repeatable workflows.

In this work, we created the ENVO ROBOT template and merge workflow, which allowed us to update existing as well as to add new terms to ENVO. The workflow enables the use of collaborative spreadsheets to add information into ENVO. A generalized version of the workflow is available from the ENVO wiki (Blumberg & Duerr 2022) involving the following steps:

1. Creating a GitHub issue detailing the material to be added.
2. Making a copy of the template spreadsheet formatted with headers necessary to compile a ROBOT template.
3. Preparing new terminology by filling out the spreadsheet following the best documented practices (Blumberg, Chong & Buttigieg 2021).
4. Compiling the ROBOT template spreadsheet into OWL code.
5. Using a GitHub pull request to merge the OWL code into the main ENVO codebase.

The spreadsheets we created while using the new workflow for this material added to ENVO discussed in this paper are available from our GitHub site (Duerr 2023). Once finalized, the new information added to ENVO though the ROBOT template and merge workflow was made publicly available within a new release of ENVO using the standard ENVO release process.

Using ROBOT improved overall efficiency as well as decreased the conceptual workload for those team members without a great deal of ontology engineering experience, though it did not decrease the time required to assess the GCW definitions or any existing ENVO definitions and axioms.

STEP 4: TECHNICAL HARMONIZATION

Finally, to formally record the relation between ENVO and SWEET terms, we used the recently developed Simple Standard for Sharing Ontological Mappings (SSSOM) (Matentzoglu et al. 2022) to document the relationships between the identified SWEET terms and their related ENVO terms.

To use SSSOM, we first populated a spreadsheet with our newly entered ENVO terms alongside potential matching terms in SWEET. For each term, we determined a potential relationship that we expressed using Simple Knowledge Organization System (SKOS) predicates (Miles & Bechhofer 2009), by analyzing the placement of the SWEET and ENVO terms in their class hierarchies, and comparing any available definitions and axioms (see the last column of Figure 2 for examples). While time consuming, this human curated approach proved to be much more accurate than other approaches which generally ignore both differences in the organization of the hierarchies of different resources as well as the richness of the subclasses and axioms underlying the mapped terms (see Results section below).

In addition to the SKOS relationship between terms, such as skos:broadMatch or skos:relatedMatch, we recorded a comment explaining the reasoning behind the type of match assigned. In many cases, these comments also include suggestions for future work and/or conditions for changing the type of match if either ontology is updated. For example, for the term Arete we recorded a comment to the effect that in SWEET an arete is a type of plain, but in ENVO an arete is a kind of ridge; so the SWEET hierarchy needed to be changed. The SSSOM file generated was added to the ENVO repository on GitHub and the ESIP Community Ontology Repository (ESIPFed 2023).

RESULTS

Of the 626 terms currently in the polar subset of ENVO, a total of 302 terms were added or updated as a part of this work. This represents roughly 15% of the unique terms in the GCW compilation; though it should be noted that many of the other GCW terms had been addressed in ENVO prior to this project. Of these terms, 151 were mapped from ENVO to SWEET using the SSSOM mapping standard, mapping available in the ENVO GitHub repository (Buttigieg et al. 2023).

Table 1 contains a list of the SWEET ontology files addressed during this work, the number of cryosphere terms identified in each file, the number of these that were also present in the GCW compilation and the number that were common between all three sources.

Table 1 SWEET files addressed during this work.

SWEET FILE	TOTAL TERMS IN SWEET FILE	CRYOSPHERIC TERMS IN SWEET FILE	GCW TERMS OVERLAPPING WITH SWEET TERMS	GCW + ENVO TERMS OVERLAPPING WITH SWEET TERMS
realmCryo.ttl	32	32	12	11
phenCryo.ttl	17	17	14	11
mtrWater.ttl	41	14	10	9
phenAtmoFog.ttl	32	3	3	3
realmClimateZone.ttl	24	3	3	1
realm.ttl	20	1	1	1
realmOcean.ttl	26	1	1	1
realmSoil.ttl	34	5	5	5
propTime.ttl	41	5	0	0
propSpaceThickness.ttl	32	3	3	3
phenHydro.ttl	33	2	1	1
phenAtmoPrecipitation.ttl	58	15	13	13
realmLandGlacial.ttl	18	16	11	8
phenSolid.ttl	63	7	4	3
Total Terms assessed	**471**	**124**	**81**	**70**

Of the almost 500 terms in the 12 SWEET files identified as containing some cryospheric terms, 124 or 26% of those were cryospheric terms. And, of those 124 cryosphere terms, 81 or 65% were also found in the Global Cryosphere Watch, and 70 or 56% were found among all three resources. Again, this overlap of similar terms found in multiple resources as well as the lack of comprehensiveness of terms relevant for a domain in any one resource shows the need and value of our work.

Figure 5 provides a graphic representation of the results of harmonizing ENVO terms related to the 'ice calving process.' This has the advantage of showing terms and relationships that are not immediately obvious when looking at one term at a time. In ENVO, 'ice calving process' is represented as a form of (subclass of) mass wasting. The subclasses of ice calving process captured differentia noted during our glossary review, in particular, 'where' the ice was calved, either into water or upon land, and 'from' which entity it was calved, that is, an iceberg or glacier. The definitions of these terms often reveal semantics which are implicitly obvious for domain scientists, but not apparent from their commonly used labels. Similarly, *Land ice*, is a term used to refer to ice formed over land masses, rather than present upon them, thus allowing marine icebergs to be a valid (sub)subclass. That is, by definition, icebergs come from land ice versus ice floes which are an expanse of sea ice. So, a marine iceberg is an iceberg which is a type of land ice mass, even though it's no longer on land. Relationships between terms (i.e., axioms such as 'has participant') come from another OBO Foundry ontology, the Relations Ontology (Huntley et al. 2014; Mungall et al. 2020), which supports reasoning and verifies logical coherence.

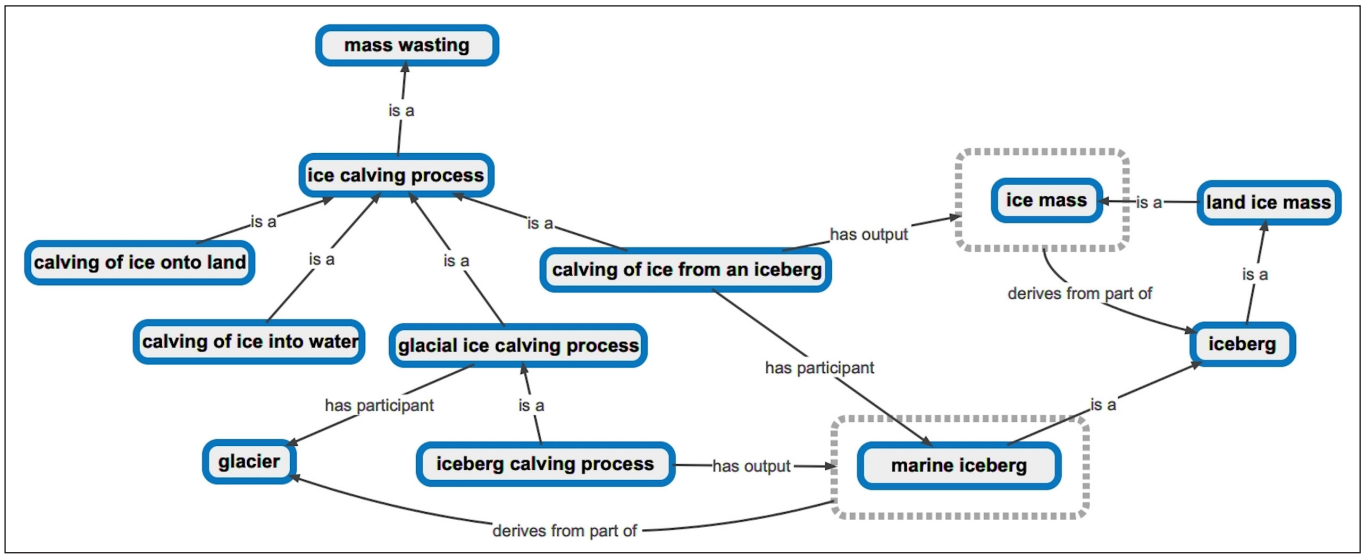

Figure 5 A partial ENVO representation of harmonized 'ice calving process' terms. Blue boxes represent terms within the ontology, the lines indicate subclass (i.e., is a) and other relationships between terms, while dotted gray boxes indicate that the enclosed terms inherit the relationships from other levels within the ontology.

As mentioned earlier, SSSOM was used to document the relationship between cryospheric terms in SWEET and ENVO. In total, 151 relationships between terms were developed. As you can see from Figure 6, roughly 40% of the terms were categorized as being a skos:closeMatch which typically implies that positioning within each hierarchy is comparable but that SWEET's lack of definitions inhibited assumptions of exact equivalence. An additional 40% of the terms were categorized as being related matches, which typically implies that while the terms are in some way related, that positioning within each hierarchy is sufficiently different to eliminate there being any possibility that the terms are equivalent. For example, if a term was considered to be a process in ENVO and a landform in SWEET, the match was deemed a related match. The remaining 20% of the terms were either categorized as being skos:broad or skos:narrow matches indicating that one of the terms is less specific than the other. skos:broad matches provided the bulk of these types of matches indicating that the ENVO term was more specific than the SWEET term.

It is quite common in the field for folks to attempt lexical matching of concepts from multiple ontologies (Euzenat & Shvaiko 2013; X. Liu et al. 2021), that is, matching based on similarity of the un-defined concept label only (or where the concept label is the most heavily weighted feature of the matching algorithm). To investigate the impact that this would have had on the ontology term relationships developed here, the match types assigned to the 61 lexically equivalent strings in the SSSOM file were examined. Figure 7 provides a summary of the match

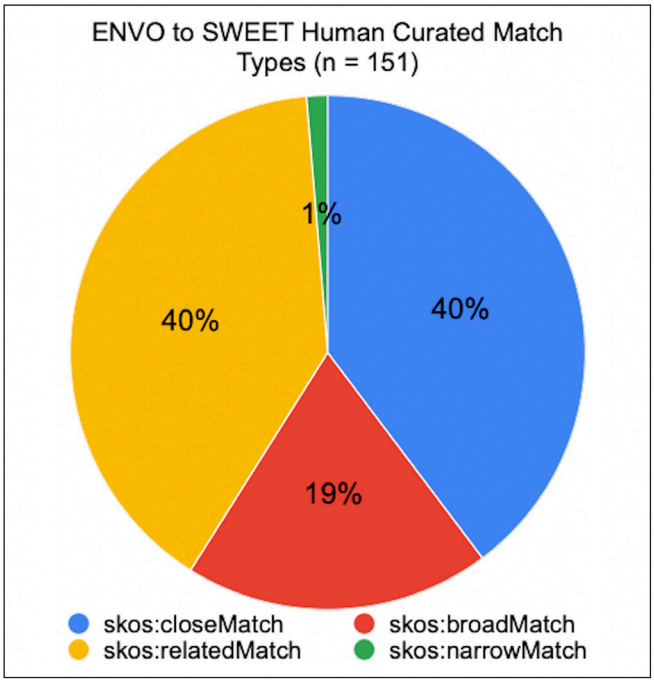

Figure 6 Match types in the SSSOM created for ENVO and SWEET.

types found. Roughly half of the terms matched closely; while the other half did not; indicating that a purely lexical match would be wrong in our case roughly half the time. Moreover, we note that the majority of the terms for which we assigned a relationship could not be matched based on their labels, since they had little or no lexical similarity.

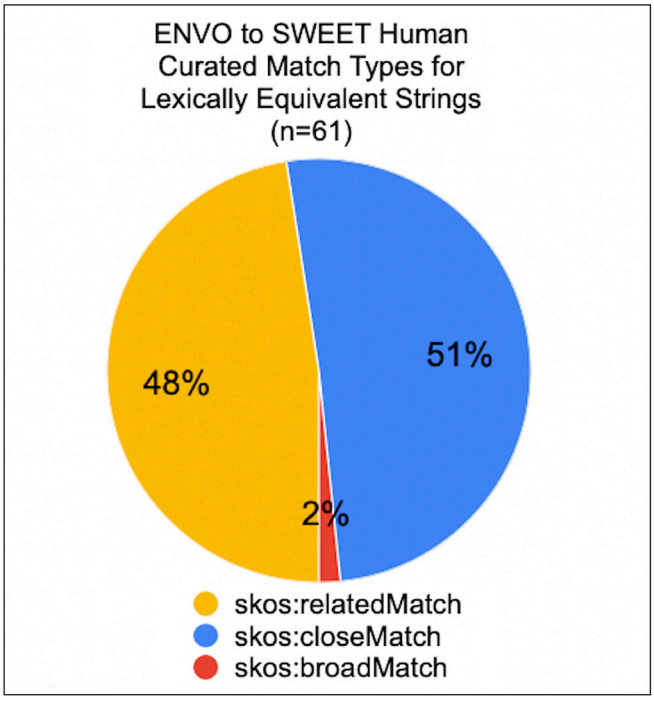

Figure 7 Match types for Lexically Equivalent Strings.

As summarized in Figure 8, we also characterized the reasons for the match types chosen for those 61 lexically equivalent strings. While these characterizations are subjective and the number of terms addressed is small, the results are still instructive. As you might expect, most of the lexically equivalent terms rated as being close matches did not have definitions in SWEET (25 terms). However, there were six such terms where it also was not clear that the placement of the term in each hierarchy was equivalent. For example, SWEET considers fiords to be a type of estuary, while ENVO doesn't. Similarly, ENVO considers rime to be a type of frost; while in SWEET frost and rime are parallel concepts placed in different parts of the overall hierarchy. In addition, there were 21 terms where the type of the term in

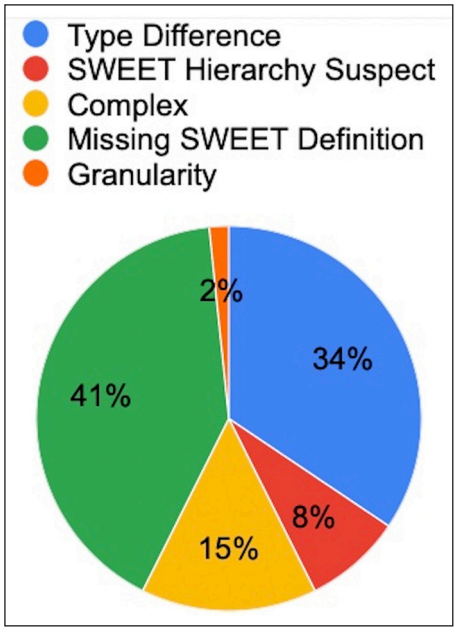

Figure 8 Reasons why lexically equivalent terms were not said to be semantically equivalent.

each ontology was different. For example, in SWEET, terms such as permafrost are three-dimensional geometric objects, while in ENVO they are environmental materials. Moreover, in nine cases, the reasons for not equating the SWEET and ENVO terms were complex, typically involving both definitional and structural differences between the two resources. In one such case, the term had been deprecated in SWEET. In another such case, SWEET had two identical terms defined in different branches of the SWEET hierarchy. In five cases, the existing SWEET hierarchy was called into question. GitHub Issues have been created to address the concerns identified from these cases.

Lastly, over the last year interactions with other communities, both within ESIP and beyond, spurred us to generalize the harmonization process so that it could be tailored to the needs of other communities. Figure 2 depicts this general process using the GCW glossaries, ENVO and SWEET purely as examples of the types of resources that can be harmonized. A summary of the general process we developed follows:

1. Existing thematic semantic resources in a variety of formats of term-definition pairs are identified by domain experts, who work together with semantic technology and ontology experts.
2. Domain experts identify source/target terms for harmonization; usually those required to advance their work. If definitions, comments, or provenance do not accompany terms, more work will be needed to understand and describe each term. Semantic technology and ontology experts work with the domain experts to reduce ambiguity by comparing terms and definitions, splitting, or merging terms, and updating targets and formalizing definitions where necessary (see Discussion).
3. The resulting terms and definitions can then be encoded in one or more semantic resources (including their provenance). To allow machine-actionable search and understanding of terms, formal axioms need to be written. This is best done by a collaboration of domain experts who know the field along with semantic technology and ontology experts who know the logic and technology. The result is a domain-correct and machine-readable final set of terms described and expressed with formal axioms. If OWL is used, reasoners can be used for quality assurance and control (QA/QC) and other logical analyses.
4. Lastly, multiple semantic/ontology resources can be formally aligned, in our case documented with SSSOM.

DISCUSSION

Here, we discuss issues found regarding harmonizing terminology and definitions, harmonizing across different ontology hierarchies, and finally sociotechnical issues.

HARMONIZING GLOSSARIES AND ONTOLOGIES

Harmonizing semantic resources developed by different groups over different periods of time is fraught with issues. However, using analysis methods such as those promulgated by the semantics community (Seppälä, Ruttenberg & Smith 2017) can help clarify, simplify, and resolve many issues. Broadly over the course of this project two major kinds of glossary inconsistencies were encountered: terminology incoherence and imprecise definitions. How we dealt with each is described in the following sections.

Terminology differences

First, we need to simply acknowledge the fact that language is fluid, in some sense alive. Terminology meaning and usage varies and drifts over time, place, and community. Consequently, there may be multiple meanings for a term depending on the exact discipline or subdiscipline defining it. For example, in the permafrost community *hummocks* are 'small lumps of soil pushed up by frost action, often found uniformly spaced in large groups' (NSIDC n.d.), while in the sea ice community a *hummock* is 'a hillock of broken ice that has been forced upwards by pressure' (WMO/OMM/BMO 1970). Both definitions are equally valid but specific to usage within a particular community. It would be pointless to argue about which of these is the right definition, since both clearly are 'right' and useful in their specific community. However, semantically speaking, these are two distinct terms that can each have their own unique identifier. For example, ENVO handles this by including the terms *sea ice hummock* (ENVO:01001537) and *frost-formed hummock* (ENVO:01001538) both under its elevated landforms branch.

Similarly, it is often the case that a term's meaning depends either on the organization providing the definition or the region of the world from which the definition came. In either case, arguing over who is right is still pointless; simply acknowledging and understanding the differences and generating multiple terms in an ontology appropriately is sufficient. For example, there are differences in the definition of the term 'blizzard' depending on which country or continent the definition came from. Thus, in the US the Weather Service definition is not the same as that of the Australian Bureau of Meteorology. The real issue here becomes simply ensuring that there is a superclass concept able to account for all the variation and nuance of the more precise local variations as subclasses (in this case for any differences in the definition of the term blizzard from other meteorological services around the world).

Another case that often occurs is where the definitions of a term are not parallel concepts but are completely different but still related. For example, the term *thermokarst* can either be a type of landform or the process that results in those kinds of landforms. In these types of cases, resolution is simple – define multiple terms accordingly! In the case of thermokarst, the ENVO ontology includes the term *thermokarst* (ENVO:03000085) as 'an irregular land surface which consists of marshy hollows, hummocks, thermokarst depressions and thermokarst lakes formed from the erosion of ice-rich thawing permafrost areas' and the term *thermokarst formation process* (ENVO:01001498), which is 'a process by which landforms are formed from the thawing of ice-rich permafrost or the melting of massive ground ice.' The thing to remember here is that the labels *thermokarst* and *thermokarst formation process* are just that—labels—and as such are easily changed without impacting in any way the organization or structure of the ontology. The only reason why the label for the term ENVO:03000085 is not something like *thermokarst landform* is simply that it was inserted into the ontology first and the label wasn't updated when the formation process was added to the ontology later on.

The situation when a term's meaning changes over time is more complicated, for example, semantic drift. For example, when discussing snow and ice processes prior to 1980, the term *ablation* did not include mechanical removal of either snow or ice by processes such as wind erosion, avalanches, or calving. Now it does. While semantic technologies and languages such as OWL can deal with temporal and numeric constraints, their inclusion in ontologies such as those within the OBO Foundry has not yet been standardized. Even if such usage were standardized, it isn't clear how such a temporal constraint could be operationalized without explicitly capturing the date the term was used wherever that term was used. For example, in natural language applications, associating the date when a particular text including that term was written, would be needed, and there would always be edge cases where it would be unclear which definition was used (e.g., papers written during or near 1980).

Worst yet are cases where there are disagreements over concepts. Unfortunately, ontology modeling cannot resolve disputes in the domain of discourse. In these situations, resolution will ultimately require discussion within the various communities involved. For example, within the cryospheric community as a whole there are disagreements about whether an ice sheet is a glacier, a glacier is an ice sheet or whether these are parallel concepts (A more complex case of *calving* is discussed in the next subsection). In these cases, there are two courses of action, with only one being considered practical. The practical alternative involves 1) acknowledging the problem, 2) include terms in ontologies wherever their inclusion is absolutely required and 3) include a note with the term itself, possibly as a skos:scopeNote, as well as to the editor of the ontology, about the problem and the likelihood that the term's placement, axiomatization, and/or inclusion may need to change in the future. The other option would be to create a separate ontology capturing the alternate world view, but this option is often considered wildly impractical.

Precise definitions and their axiomatization

While scientists are often accused of using jargon and trying to be very precise, sometimes inhumanely so, it is surprising that many of the definitions in the various disciplinary glossaries and other vocabulary resources developed are often not semantically consistent or complete. This is one reason why formal semantics calls for 1) the careful creation of definitions using analysis methods such the genus-differentia definitional form (that is, dividing terms into classes and subclasses differentiated by properties) complemented by 2) machine-actionable axiomatization which uses a logical language to formally specify the vocabulary of concepts and the relationships among them and 3) by ensuring that the human-readable definition and the corresponding machine-actionable axioms are equivalent (Seppälä, Ruttenberg & Smith 2017). Doing so can both call out and/or fix problems with existing glossaries. Inconsistencies between axioms represented in OWL, for example, can be shown by theorem provers available in tools like Protégé (Musen 2015). However, it is up to the ontology developer(s) to ensure that the human readable definitions and their machine-actionable counterparts actually are equivalent, so that any machine made logical inferences are as expected by humans.

Let's return to our example in Figure 5. The term *calving* is an ablative process where chunks of ice fall off a parent body (e.g., a calving glacier). There is ambiguity in the existing dozen definitions in the GCW compilation for both the process and the resulting chunks of ice. Some definitions assume that the calving process can only happen going into water while others allow calving on land. Also some definitions allow calving to occur from any form of ice of land origin (e.g., ice sheets, ice caps, ice shelves), while others restrict it to glaciers or some other subset of all of the types of ice of land origin. What ice calved onto land would be called is not obvious, especially since the only definition of calved ice in the GCW compilation excludes ice falling onto land. To resolve the ambiguity with process terminology, we defined four subclasses in ENVO under the class 'ice calving process': calving of ice from an iceberg, calving of ice into water, calving of ice onto land (i.e., dry calving or terrestrial calving), and glacial ice calving process. While it is unlikely that there will ever be a need for other terms for what ice is falling onto (can ice fall onto or into anything other than water or land?), there may well be the need to add terms for other sources of the falling ice (e.g., ice sheet, ice cap, thick permafrost embedded in an eroding cliff, etc.) in the future, provided of course that there are use cases where such distinctions are important.

As an example of the genus-differentia definitional form, the definition of the term *calving of ice into water* is 'An ice calving process during which a mass of ice falls from a larger mass into a body of water' where *ice calving process* is the parent, more general class. The rest of the sentence describes how this term is specialized from its parent. In terms of the machine-actionable axiomatization of the term, the only difference in axiomatization of the term and its parent is the addition of a water body as a participant in the process (i.e., 'has participant' some 'water body').

Another example of axiomatization of an ENVO term is *permafrost*. We created formally defined axioms that specify that *permafrost* is a type of 'environmental material' which 'has quality some decreased temperature' and is 'composed primarily of some (sediment or soil or rock).' One of its sub-types is 'ice-bearing permafrost' which 'has part some water ice.' Permafrost also

has a human-readable definition of 'Soil or rock and included ice or organic material at or below the freezing point of water (0 degrees Celsius or 32 degrees Fahrenheit) for two or more years.' This is a case where the human readable definition is more precise than the axiomatization. Clearly, when or if the larger semantic community promulgates a standard way of including numeric constraints into axioms, these axioms will need to be updated, perhaps as 'has quality some freezing years' >=2; where 'freezing years' axioms are something like 'has quality maximum temperature < 0C' and 'has quality minimum duration.'

MAPPING ACROSS INCONSISTENT ONTOLOGY HIERARCHIES

Given the issues with harmonizing terms in glossaries as discussed above, and the vast number of glossaries, it would be surprising if two ontologies created by different groups, for different purposes at possibly different times had internal hierarchies that were the same. Yet, that doesn't mean that it is impossible to harmonize across such resources; it is just not as straightforward as simply mapping lexically equivalent terms.

Consequently, when adding terms to an existing ontology the resulting contextual structure/hierarchy for the added terms may not necessarily be the same as would occur if adding to a different ontology or if creating a new and independent ontology, say a stand-alone cryosphere ontology. But, even when creating a new ontology, the order of adding classes can result in a functionally similar but different ontology structure. That is, which terms were added first can influence where later terms are placed. So, as we added cryosphere terms one by one to ENVO, the terms were subclassed into the most relevant existing classes. This scattered some terms that, on later inspection, could have been more closely related, and the initial result may eventually be slightly changed. The piecemeal process of adding terms and creating a new whole that makes sense is difficult regardless of creating a new ontology or adding to an existing ontology and is probably non-deterministic regarding the exact same hierarchical result. Accuracy can be retained, however. A few examples follow.

For example, the concept 'greenhouse gas' encompasses both a role and a material entity. In ENVO there is no material entity that is a 'greenhouse' gas, but certain gasses can bear this role. So in ENVO, *greenhouse gas* is a term from the Chemical Entities of Biological Interest (ChEBI) ontology (i.e., CHEBI_76413) and not a term under 'gas molecular entity.' However, in SWEET, *greenhouse gas* is a both a subclass of 'chemical substance' and a subclass of 'chemical.'

As another example, in the ENVO ontology, 'cryosol' is a subclass of frozen soil, and 'part_of' is its relationship to 'permafrost'; but in SWEET 'cryosol' is a 'categorical property,' specifically a subclass of 'soil order.' Also, in SWEET, 'gelisol' is listed as a sibling of 'cryosol,' whereas 'gelisol' is a synonym of 'cryosol' in ENVO.

'Snowpack' is a subclass under 'thickness' in SWEET, although immediately under 'snow cover.' In ENVO, 'snowpack' is under 'snow mass,' which is under 'mass of compounded environmental materials.' Given that SWEET considers the term to be a thickness and ENVO currently considers it a mass of snow, there is a mismatch. The definition in ENVO does refer to size, however, as in being large enough and persisting long enough to form layers under its own weight. Overall, the GCW analysis found eight definitions of snowpack over multiple glossaries, with many commonalities but also disagreements.

In ENVO, *proglacial* (ENVO:01001853) is a 'positional quality which inheres in a bearer by virtue of the bearer in being in physical contact with, or close to, a glacial margin.' But, in SWEET, 'proglacial' is not a concept that refers to being, say, in front of a glacier, but instead is a process, that is, found under 'glacial process' along with other processes such as 'accumulation,' 'calving,' and 'glacial retreat'.

In each of the examples above, it was possible to generate a SSSOM relationship between the terms despite their differences.

In summary, the definitions and uses of terms can vary across ontologies such that hierarchies and conceptualizations differ. This makes alignment or harmonization imprecise. Delving into these differences, however, can expand one's knowledge across disciplines and perspectives and may help the expert community reassess and standardize its definitions.

SOCIOTECHNICAL ISSUES

In addition to issues related to the often ambiguous or incomplete definitions, difficulties with inconsistent ontology structures and current limitations in axiomatization, we encountered several issues that were more on the social side of the sociotechnical spectrum that needed to be resolved.

First, many GCW terms are entirely missing from either ENVO or SWEET or both. Simply put, the GCW provides a much more comprehensive compilation of terms in use within this discipline. The question then becomes one of scoping—how much coverage of the terms in the GCW would be appropriate for this work? We decided to limit ourselves to terms that were present in SWEET or ENVO and to add related terms to ENVO as were judged relevant to the existing ENVO community. For example, several compaction and erosion related terms were added to ENVO because material transformation processes having inputs and outputs are an important branch of the ENVO ontology. This decision constrained the work to the limited bandwidth available within the ESIP harmonization cluster membership.

Second, this work reflects the understanding that practical and resource limitations mean that collaborative development of a single encompassing semantic resource for a domain is likely to be impossible. A better target is harmonizing semantic resources within a defined scope of work, the scope of work that participants in the harmonization process care about. This can start at the lower end of the semantic spectrum by harvesting well-established and well-defined terminologies as was done in this work. Agreement on the meaning of termed concepts is a first step toward alignment across the semantic spectrum and its impact on the overall ontological structure can be judged as work continues. A degree of interoperability, though minimal, is the reward.

In practice, what this also means is that it is likely that semantic modeling of any term in any ontology will only be as deep as is necessary to satisfy current use cases. For example, the term *snow water equivalent* describes the output of a method used to determine how much water is present in a given volume of snow. Snow covering a defined area is collected and then melted. The depth of the resulting snowmelt is measured after it has been transferred to a standardized container. A value for snow water equivalent (SWE) can also be inferred via remote sensing technologies. Complete semantic modeling of this term would require that the processes of identifying, collecting, and melting a volume of snow and subsequently measuring the volume of the resulting water be modeled for ground-based methods and the algorithms used to infer SWE from remote sensing observations also be modeled. Neither SWEET nor ENVO currently model this term or many comparable terms to that level of detail; though either could be updated to include deeper modeling if and when new use cases surfaced that require it.

In general, semantic resources of any type are living objects, subject to change over time, just as all languages in use (i.e., living languages) change over time. Both ENVO and SWEET have existed for more than a decade and some of the glossaries compiled by the GCW are well over 60 years old. What this meant in practical terms was that we needed to review the history of each term and its placement within the ENVO and SWEET hierarchies for every term addressed. In some cases this meant we needed to change an ontology to use better and more recently defined terms. For example, we switched to using the Chemical Entities of Biological Interest Ontology term for water, CHEBI:water, rather than the original ENVO term for water to handle issues of the hydrological precipitation process that arose when revising *hailfall* and *snowfall* in a systematic way.

As a corollary to these last several issues and given the hierarchy inconsistencies evident in comparing ontologies such as ENVO and SWEET, it should be noted that the need for semantic harmonization will only grow as long as people continue to reinvent the wheel each and every time they need to use semantic resources within their work. Currently the norm within the Earth and Environmental sciences is for folks who need to use semantics to invent their own semantic resources no matter how many resources either partially or totally covering that topic already exist. A better use of these people's time would be for them to collaborate with the communities currently maintaining existing semantic resources and determining what extensions, refactoring, and so on of those resources are needed and contributing their efforts

to the larger community. Having a well-maintained repository and ontology/term discovery resource for the Earth sciences, akin to the OBO Foundry and BioPortal resources in the Biomedical community, might go a long way to helping resolve this problem which is currently inhibiting uptake of semantics in our field.

LESSONS LEARNED AND RECOMMENDATIONS

Many lessons were learned along the way, with some noted as part of the previous discussion. The following are some of the main lessons along with recommendations for managing semantic harmonization.

PROPER SCOPE AND INTERDISCIPLINARY TEAMS ARE NEEDED

From a project perspective, starting with the right scope and an adequate, interdisciplinary team is important. Selecting a proper set of terms is important as is the value of building a coalition of interested parties around the selected set of concepts to harmonize. This starts with clearly identifying the conceptual space you are trying to describe and define. With the help of definitions one can analyze the conceptual space to understand the key concepts and relationships that are contained in a core subset of the terms looked at. Next is to evaluate the feasibility of a preliminary scope based on factors such as available resources and time constraints and prioritize a final set of semantic resources that need to be included in the scope based on the targeted conceptual space, stakeholders' needs, domain analysis, and feasibility considerations. It is also important to identify the stakeholders who will likely use the harmonized vocabulary and ensure that the team has a good balance of domain and semantic technology experts with good communication skills for effective collaboration and resolving any conflicts that may arise.

GLOSSARY HARMONIZATION IS FOUNDATIONAL

Merging and splitting of glossary terms at lower levels of the semantic ladder (as well as identification of sub meanings) is needed before the more difficult alignment at higher levels of the semantic ladder because many terms can have a variety of synonyms and closely related terms that make them similar. For example, the term *tabular iceberg* can be found in glossaries under the synonyms *tabular berg* and *table iceberg*, and it was formerly called a *barrier iceberg*. Similarly, ensuring that the same label is not re-used for another term within an ontology is important for minimizing confusion. This problem can be easily prevented simply by adding disambiguating phrases to the term, for example, *thermokarst landscape* and *thermokarst process*, as discussed earlier. Once mapped, the alignment of textual definitions with axiomized representations in ontologies can be performed. For all these reasons and to make the sequence of changes to the ontology clear (i.e., its provenance), there should be an item by item commit to updates and documentation of the changes made.

USE TOOLS WHENEVER POSSIBLE

The well-documented ROBOT Templates (Jackson et al. 2019) and their supporting scripts, described in Step 3 above, allow shared best practices with spreadsheet-like editing modality for more inclusivity. These tools help cross the domain expert to ontologist divide by allowing routine, asynchronous work within domain communities without relying on a trained ontology engineer.

HUMAN EXPERTISE IS IMPORTANT

A central lesson is that while automation, such as simple label matching and tools like ROBOT can help with routine tasks, a human-in-the-loop for things like ontology curation was needed. While time consuming, this human curated approach proved to be much more accurate than other approaches which generally ignore both differences in the organization of the hierarchies of different resources as well as the richness of the subclasses and axioms underlying the mapped terms.

As seen in the Discussion, there were many lessons learned in assigning the type of SKOS match between terms, especially when there is not an adequate definition in one of the ontologies. The most important lesson is that when alternate definitions exist from different points of view, arguing over who is right is less useful than simply acknowledging, understanding, and documenting the differences by appropriately generating multiple terms in an ontology.

FUTURE WORK

Based on the results of this work, the ESIP semantic community expects to continue working in three areas: 1) pushing the greater OBO Foundry and general semantics community to formalize the handling of numeric values and ranges in ontologies; 2) evolving the SWEET ontology in support of harmonization and 3) pursuing related semantic harmonization work in several other ESIP clusters. These topics are described in more detail in the following paragraphs.

FORMALIZING THE HANDLING OF NUMERIC VALUES AND RANGES IN ONTOLOGIES

As has been mentioned previously it is often the case in science that the definition of a concept will include numeric values. For example, the composite definition for the term *ice pellet* from the 27 glossaries in the GCW compilation and included in the ENVO ontology is 'An ice mass which is 1) transparent or translucent, 2) rounded, spherically, or cylindrically shaped, and 3) less than 5 millimeters in diameter.' Similarly, nearly all of the terms in the WMO Sea Ice Nomenclature (WMO/OMM/BMO 1970) include numeric criteria related to the age of the ice, the size of the floe, and so on. Currently, there is no agreement as to a uniform way of adding numeric values, with units, as an axiom. This is critical if ontologies are to be useful for characterizing and understanding scientific data. In particular, for this project it would have been very useful if the OBO Foundry consortium had agreed to a convention for this, since, as is, many terms within ENVO currently have incomplete axiomatization where the human readable definition is more accurate and complete than the computer processable axiomatization.

SWEET

The SWEET ontology suite is a long-standing community resource and continues to evolve. Pursuant to the work described here, the harmonized GCW definitions now in ENVO are also being added to SWEET. As such, SWEET developers and the broader community of practice will soon be able to utilize SSSOM mappings to cross-reference back to ENVO and/or add further definition annotations which include the provenance available from that resource.

In addition to the SSSOM mappings, updates to the curation process, creation and enhancement of domain and observational concepts and properties, as well as the underlying technology stack supporting the resource, it was determined by the community that SWEET could fill a current gap by housing textual concept definitions from disparate Earth and Environmental science resources, thus making SWEET a *hub* for domain relevant concepts including, potentially, multiple independently sourced definitions which are not semantically equivalent. In this context, resources for definitions could be established vocabularies—for example, GCMD, USGS Thesaurus, and so on—as well as resources which currently exist in a static, unstructured format—for example, Dictionary of Geologic Terms (Bates & Jackson 1984) or Glossary of Geology (Neuendorf, Mehl, Jr., & Jackson 2011) currently available in hard copy format, or other resources perhaps only available as a PDF. Each candidate definition is to be added using annotation properties (i.e., it will not affect any axioms in the initial investigation) with proper citation and contributor information (i.e., creator and reviewer) attached to each recorded textual definition.

It is the hope that using SWEET as hub for concept definitions will highlight similarities and gaps in Earth science conceptual descriptions and knowledge as well as provide the groundwork for making concepts more precise and increasing their expressivity. This latter point will be crucial for the future development of the resource.

FUTURE HARMONIZATION WORK

We believe that semantic harmonization is an important and often missing ingredient to help find, make sense of, and usefully employ digital data as well as being critical to making data FAIR. Our outcomes and progress with the cryosphere have motivated us to begin work with other ESIP clusters in harmonizing key terminological resources in the following domains (see Table 2).

Table 2 Future harmonization work by Earth and Environmental science domain.

DOMAIN	DESCRIPTION	WORK DONE TO DATE
Wildfires	Initial topic under the ESIP Disaster Lifecycle Cluster	Initial vocabulary (boundary, fuels, water sources, causes, wildfire behavior etc.) terms were identified from expert narratives and a conceptual model drafted from a work session.
Soils	Work in the Soil ontology and Informatics Cluster	Source glossaries identified. Topics include: • Geolocation: surface location, sample time, depth of sample • Soil organic carbon: bulk density, coarse fraction, organic fraction • Metals, salts, and acids: pH, elemental analysis, and ionic exchange • Nutrients: phosphorus and nitrogen • Gas flux: field respiration and incubation • Fractions: texture (sand/silt/clay) and sample subsetting (physicochemical fractionation) • Isotopes: Radiocarbon and other isotopes
Coastal and Marine Ecological Classification Standard (CMECS)	Attempt to extend existing harmonization with ENVO	• Assess domains where CMECS and ENVO can contribute additional terms to each other • Harmonize like terms in ENVO and CMECS
Heliophysics	Long term goal is to create a knowledge commons for heliophysics and Earth sciences	Several sessions have been held at ESIP meetings. Initial target glossaries and terms have been identified and are being loaded into YAMZ.net. A workshop to kick off the glossary harmonization effort is being planned.
Earth and Environmental science domains	Adding definitions from other semantic resources (electronic and hardcopy) to SWEET	Match candidates from GCMD, USGS Thesaurus, USGS Lithology terms, CMECS, Marine Planning Data(MPD), and GEneral Multilingual Environmental Thesaurus (GEMET) are currently under review with several others scheduled.

CONCLUSION

Alignment and semantic harmonization across the growing types of semantic resources is important for data interoperability and reuse, thus satisfying FAIR principles. In this work we have shown how a focused interdisciplinary team of domain experts and semantic technology developers can effectively harmonize semantic resources using a standard method. The process developed is to review and synthesize content in a stepwise fashion from a collection of thematic glossaries into a harmonized collection and then to align these and further document them along with richer, more machine-actionable resources higher on the semantic ladder (i.e., here, SWEET, and ENVO).

In piloting this process we encountered several issues and documented the lessons learned from these experiences. This includes many examples that we hope will help other communities attempting to perform similar activities.

DATA ACCESSIBILITY STATEMENTS

All of the data associated with this work is publicly available on Zenodo (Semantic Harmonization Cluster 2023) as well as on the ESIP github (Duerr 2023). The ENVO ontology can be found on the OBO Foundry (OBO Technical Working Group 2024); while the SWEET ontology can be found on the ESIP Community Ontology Repository (COR) (ESIPFed 2023).

ACKNOWLEDGEMENTS

This work is based on materials, programs, collaboration platform, and meeting spaces provided by the ESIP Community with support from the National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), and the United States Geological Survey (USGS). We'd especially like to thank Chantelle Verhey for her very helpful comments on drafts of this work and Mark Schildhauer and Anne Thessen for their insights and suggestions in defining the challenges and approaches early in the project. Some of the work included here was conducted using Protégé.

FUNDING INFORAMTION

Pier Luigi-Buttigieg was supported by the Helmholtz Metadata Collaboration and the Frontiers in Arctic Marine Monitoring Programme of the Alfred Wegener Institute, Helmholtz Institute for Polar and Marine Research. Brandon Whitehead was supported by New Zealand's Ministry of Business Innovation and Employment (MBIE) Infrastructure Platform. Kate Rose was supported by the Northern Gulf Institute under a grant from NOAA National Centers for Environmental Information.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Conceptualization: Ruth Duerr, Gary Berg-Cross, Brandon Whitehead, Mark Schildhauer, Anne Thessen, Pier Luigi Buttigieg

Data curation: Ruth Duerr

Formal analysis: Ruth Duerr, Gary Berg-Cross, Nancy Wiegand, Brandon Whitehead, Pier Luigi Buttigieg

Investigation: Ruth Duerr, Gary Berg-Cross, Nancy Wiegand, Brandon Whitehead, Anne Thessen, Pier Luigi Buttigieg

Methodology: Ruth Duerr, Gary Berg-Cross, Brandon Whitehead, Pier Luigi Buttigieg

Resources: Ruth Duerr, Pier Luigi-Buttigieg, ESIP

Software: Kai Blumberg, Pier Luigi-Buttigieg, Ruth Duerr, Brandon Whitehead

Validation: Ruth Duerr

Visualization: Ruth Duerr, Kate Rose, Pier Luigi Buttigieg

Writing, original draft: Gary Berg-Cross, Nancy Wiegand, Brandon Whitehead, Kate Rose, Ruth Duerr, Pier Luigi Buttigieg

Writing, review and editing: Gary Berg-Cross, Nancy Wiegand, Brandon Whitehead, Kate Rose, Ruth Duerr, Chantelle Verhey

AUTHOR AFFILIATIONS

Ruth Duerr orcid.org/0000-0003-4808-4736
Ronin Institute for Independent Scholarship, United States

Pier Luigi Buttigieg orcid.org/0000-0002-4366-3088
Helmholtz Metadata Collaboration, Alfred Wegener Institute, Germany

Gary Berg Cross orcid.org/0000-0002-2282-7215
Ontolog Forum Board, United States

Kai Lewis Blumberg orcid.org/0000-0002-3410-4655
Department of Biosystems Engineering, University of Arizona, Tucson, AZ, United States

Brandon Whitehead orcid.org/0000-0002-0337-8610
Manaaki Whenua – Landcare Research, Ltd., Palmerston North, New Zealand

Nancy Wiegand orcid.org/0000-0001-7543-6173
University of Wisconsin-Madison (emeritus), United States
Kate Rose orcid.org/0000-0002-8454-8481
Northern Gulf Institute/Mississippi State University, NOAA National Centers for Environmental Information, United States

REFERENCES

AGU. 2021. Index of terms. American Geophysical Union. https://www.agu.org/Publish-with-AGU/Publish/Author-Resources/Index-terms.

American Meteorological Society. 2024. Glossary of meteorology. AMS. https://www.ametsoc.org/index.cfm/ams/publications/glossary-of-meteorology/.

Arp, R, Smith, B and **Spear, AD.** 2015. *Building ontologies with basic formal ontology*. London, UK: The MIT Press. DOI: https://doi.org/10.7551/mitpress/9780262527811.001.0001

Australian Government. n.d. Bureau of Meteorology Glossary. Bureau – Glossary of Terms. corporateName=Bureau of Meteorology. Accessed February 16, 2024. http://www.bom.gov.au/lam/glossary/.

Bates, RL and **Jackson, JA.** (eds) 1984. *Dictionary of geological terms*. Third. Garden City, New York: Anchor Press/Doubleday.

Blumberg, K, Chong, S and **Buttigieg, PL.** 2021. Adding classes to ENVO. *GitHub*. February 5, 2021. https://github.com/EnvironmentOntology/envo/wiki/Adding-classes-to-ENVO.

Blumberg, K and **Duerr, R.** 2022. ENVO Robot template and merge workflow. *GitHub*. August 11, 2022. https://github.com/EnvironmentOntology/envo/wiki/ENVO-Robot-template-and-merge-workflow.

British Oceanographic Data Centre. 2023. The NERC vocabulary server. *Natural Environment Research Council*. https://vocab.nerc.ac.uk.

Brochhausen, M, Ceusters, W, Courtot, M, Dipert, R, Hastings, J, Mungall, C, Natale, D, et al. 2019. Basic formal ontology. https://github.com/BFO-ontology/BFO.

Buttigieg, PL. 2021. Creating good definitions. *EnvironmentOntology/Envo Wiki*. March 1, 2021. https://github.com/EnvironmentOntology/envo/wiki/Creating-good-definitions.

Buttigieg, PL, Morrison, N, Smith, B, Mungall, CJ, Lewis, SE and **the ENVO Consortium.** 2013. The environment ontology: Contextualising biological and biomedical entities. *Journal of Biomedical Semantics,* 4(1): 43. DOI: https://doi.org/10.1186/2041-1480-4-43

Buttigieg, PL, Mungall, C, Duncan, B, Blumberg, K, Wilkie, I, meichen-liu, Meyer, R, et al. 2023. EnvironmentOntology/Envo: 2023-02-13 release. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.7636736

Buttigieg, PL, Pafilis, E, Lewis, SE, Schildhauer, MP, Walls, RL and **Mungall, CJ.** 2016. The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics,* 7(1): 57–57. DOI: https://doi.org/10.1186/s13326-016-0097-6

Duerr, R. 2018a. WMO GCW glossary terms and their alignment with terms in leading semantic resources. DOI: https://doi.org/10.13140/RG.2.2.26770.12489

Duerr, R. 2018b. WMO GCW nomenclature assessment. DOI: https://doi.org/10.13140/RG.2.2.35158.73284

Duerr, R. 2023. ESIPFed/CRYO-Harmonization: Repository for ESIP semantic harmonization work on cryosphere terms. ESIPFed/CRYO-Harmonization. 2023. https://github.com/ESIPFed/CRYO-Harmonization.

ESIP. 2023. About ESIP. 2023. https://www.esipfed.org/about.

ESIPFed. 2023. ESIP federation community ontology repository services. ESIP Community Ontology Repository. March 8, 2023. http://cor.esipfed.org/.

Euzenat, J and **Shvaiko, P.** 2013. Ontology matching. 2nd ed. Heidelberg: Springer Berlin. DOI: https://doi.org/10.1007/978-3-642-38721-0

Everdingen, Rv. (ed.) 2005. Multi-Language glossary of permafrost and related ground-ice terms. NSIDC/WDC-A Glaciology. https://globalcryospherewatch.org/reference/glossary_docs/Glossary_of_Permafrost_and_Ground-Ice_IPA_2005.pdf.

FAIR Principle I1. 2015 I1: (Meta)Data use a formal, accessible, shared, and broadly applicable language for knowledge representation. March 1, 2015. https://www.go-fair.org/fair-principles/.

'FAIR Principles'. 2015. FAIR Principles – GO FAIR. March 1, 2015. https://www.go-fair.org/fair-principles/.

FAIRsharing Team. 2015. FAIRsharing record for: Quantities, units, dimensions and types. *FAIRsharing*. DOI: https://doi.org/10.25504/FAIRSHARING.D3PQW7

Gil, Y, Pierce, SA, Babaie, H, Banerjee, A, Borne, K, Bust, G, Cheatham, M, et al. 2018. Intelligent systems for geosciences: An essential research agenda. *Commun. ACM,* 62(1): 76–84. DOI: https://doi.org/10.1145/3192335

Giunchiglia, F and **Zaihrayeu, I.** 2017. Lightweight ontologies. In Liu, L and Tamer Özsu, M (eds.), *Encyclopedia of database systems*. New York, NY, USA: Springer. pp. 1613–19. DOI: https://doi.org/10.1007/978-1-4899-7993-3_1314-2

Government of Canada. n.d. Glossary – Climate – Environment and Climate Change Canada. Accessed February 16, 2024. https://climate.weather.gc.ca/glossary_e.html.

Haller, A, Janowicz, K, Cox, SJD, Lefrançois, M, Taylor, K, Le Phuoc, D, Lieberman, J, García-Castro, R, Atkinson, R and **Stadler, C.** 2019. The modular SSN ontology: A joint W3C and OGC standard specifying the semantics of sensors, observations, sampling, and actuation. *Semantic Web*, 10(1): 9–32. DOI: https://doi.org/10.3233/SW-180320

Huntley, RP, Harris, MA, Alam-Faruque, Y, Blake, JA, Carbon, S, Dietze, H, Dimmer, EC, et al. 2014. A method for increasing expressivity of gene ontology annotations using a compositional approach. *BMC Bioinformatics*, 15(May): 155. DOI: https://doi.org/10.1186/1471-2105-15-155

Jackson, RC, Balhoff, JP, Douglass, E, Harris, NL, Mungall, CJ and **Overton, JA.** 2019. ROBOT: A tool for automating ontology workflows. *BMC Bioinformatics*, 20(1): 407. DOI: https://doi.org/10.1186/s12859-019-3002-3

Janowicz, K, Haller, A, Cox, SJD, Le Phuoc, D and **Lefrançois, M.** 2019. SOSA: A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics*, 56(May): 1–10. DOI: https://doi.org/10.1016/j.websem.2018.06.003

Lenton, TM, Armstrong McKay, DI, Loriani, S, Abrams, JF, Lade, SJ, Donges, JF, Milkoreit, M, Powell, T, Smith, SR and **Zimm, C.** 2023. The global tipping points report 2023. https://global-tipping-points.org/.

Liu, X, Tong, Q, Liu, X and **Qin, Z.** 2021. Ontology matching: State of the art, future challenges, and thinking based on utilized information. *IEEE Access*, 9: 91235–43. DOI: https://doi.org/10.1109/ACCESS.2021.3057081

Matentzoglu, N, Balhoff, JP, Bello, SM, Bizon, C, Brush, M, Callahan, TJ, Chute, CG, et al. 2022. A Simple Standard for Sharing Ontological Mappings (SSSOM). *Database*, 2022(January): baac035. DOI: https://doi.org/10.1093/database/baac035

McCreary, D. 2006. Patterns of semantic integration. April. https://www.slideshare.net/dmccreary/semint.

McGibbney, LJ, Whitehead, B, Rueda-Velásquez, CA, Duerr, R, Keil, JM, Berg-Cross, G, Rose, K, et al. 2022. Semantic Web for Earth and Environmental Terminology (SWEET). http://sweetontology.net.

Miles, A and **Bechhofer, S.** 2009. SKOS Simple Knowledge Organization System Reference. 2009. *World Wide Web Consortium*. August 18, 2009. https://www.escholar.manchester.ac.uk/uk-ac-man-scw:66505.

Mungall, C, Osumi-Sutherland, D, Overton, JA, Balhoff, J, Clare72, pgaudet, Brush, M, et al. 2020. Oborel/Obo-Relations: 2020-07-21. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.3955125

Musen, MA. 2015. The protégé project: A look back and a look forward. *AI Matters*, 1(4): 4–12. DOI: https://doi.org/10.1145/2757001.2757003

Nagendra, KS, Bukhres, O, Sikkupparbathyam, S, Areal, M, Miled, ZB, Olsen, L, Gokey, C, et al. 2001. NASA global change master directory: An implementation of asynchronous management protocol in a heterogeneous distributed environment. In *Proceedings 3rd International Symposium on Distributed Objects and Applications*, 136–45. DOI: https://doi.org/10.1109/DOA.2001.954079

Neuendorf, KKE, Mehl, JP, Jr. and **Jackson, JA.** 2011. *Glossary of geology*. (Revised) Fifth; Version 1.9. American Geosciences Institute.

NOAA/NWS. 2009. Glossary – NOAA's National Weather Service. 2009. https://w1.weather.gov/glossary/.

NSIDC. n.d. Hummock. National Snow and Ice Data Center. Accessed December 19, 2023. https://nsidc.org/learn/cryosphere-glossary/hummock.

OBO Technical Working Group. 2022. OBO Foundry Principles. Principles: Overview. 2022. https://obofoundry.org/principles/fp-000-summary.html.

OBO Technical Working Group. 2024. OBO Foundry. https://obofoundry.org.

Overton, JA, Dietze, H, Essaid, S, Osumi-Sutherland, D and **Mungall, CJ.** 2015. ROBOT: A command-line tool for ontology development. In Couto, FM and Hastings, J (eds.), *Proceedings of the International Conference on Biomedical Ontology (ICBO 2015)*, 1515(2). Lisbon, Portugal: CEUR. https://ceur-ws.org/Vol-1515/demo6.pdf.

Raskin, RG and **Pan, MJ.** 2005. Knowledge representation in the Semantic Web for Earth and Environmental Terminology (SWEET). *Computers & Geosciences*, 31(9): 1119–25. DOI: https://doi.org/10.1016/j.cageo.2004.12.004

Semantic Harmonization Cluster. 2023. ESIPFed/CR.YO-Harmonization: Data for DSJ Article. *Zenodo*. DOI: https://doi.org/10.1038/sdata.2016.18

Seppälä, S, Ruttenberg, A and **Smith, B.** 2017. Guidelines for writing definitions in ontologies. *Ciência Da Informação*, 46(1): 73–88. DOI: https://doi.org/10.18225/ci.inf..v46i1.4015

Shepherd, A, Jones, M, Richard, S, Jarboe, N, Vieglais, D, Fils, D, Duerr, R, et al. 2022. Science-on-Schema.Org v1.3.1. *Zenodo*. DOI: https://doi.org/10.5281/ZENODO.7872383

USGS Thesaurus. 2023. USGS. https://apps.usgs.gov/thesaurus/thesaurus-full.php.

Whitehead, B. 2022. SWEET Governance and Roadmapping: Survey Results. Presentation, July 12. DOI: https://doi.org/10.6084/m9.figshare.20293860.v1

Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1): 160018. DOI: https://doi.org/10.1038/sdata.2016.18

WMO/OMM/BMO. 1970. WMO Sea-Ice Nomenclature – No. 259. WMO.

Wolodkin, A, Weiland, C and **Grieb, J.** 2023. Mapping. Bio: Piloting FAIR semantic mappings for biodiversity digital twins. *Biodiversity Information Science and Standards*, 7: e111979. https://biss.pensoft.net/article/111979/download/pdf/. DOI: https://doi.org/10.3897/biss.7.111979

Zeng, ML. 2008. Knowledge Organization Systems (KOS). *Knowledge Organization*, 35(3/2): 160–82. DOI: https://doi.org/10.5771/0943-7444-2008-2-3-160

TO CITE THIS ARTICLE:
Duerr, R, Buttigieg, PL, Cross, GB, Blumberg, KL, Whitehead, B, Wiegand, N and Rose, K. 2024. Harmonizing GCW Cryosphere Vocabularies with ENVO and SWEET. Towards a General Model for Semantic Harmonization. *Data Science Journal*, 23: 26, pp. 1–22. DOI: https://doi.org/10.5334/dsj-2024-026

Submitted: 17 May 2023
Accepted: 17 May 2023
Published: 07 May 2024

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.