

Ontology Summit 2021 Communiqué: Ontology Generation and Harmonization

First Draft: 2 June 2021

Abstract: This Communiqué surveys the landscape of ontologies throughout their lifecycle, including generation and development, harmonization with both human and machine environments, and sustainability. [The abstract will be completed later.]

Keywords: Ontology, Machine Learning, Definitions, Sustainability

1. Introduction

Ontologies are proliferating, producing a complex landscape of many types, roles and uses for many purposes. Ontologies can be extracted, learned, modularized, interrelated, transformed, analyzed, and harmonized as well as developed in a formal process; and there are now many ways that ontologies interact with other technologies, including but not limited to, generation by machine learning tools, serving as the basis for machine learning so as to improve the quality of the results, integration into deep learning architectures. The variety has made it difficult to communicate about them and between them. There is a need for harmonization and better definitions. Furthermore, it isn't enough to construct ontologies, they must be sustained over time.

The Ontology Summit 2021 examined the topic of ontology generation and harmonization at a series of virtual sessions from February to May 2021. This Communiqué synthesizes and summarizes the findings of this series. The Ontology Summit 2021 was organized into four tracks and this Communiqué is also organized based on these tracks: The Ontological Landscape (Section 2), Definitions (Section 3), Neuro-Symbolic Learning Ontologies (Section 4), and Sustainability of Ontologies (Section 5). Section 6 lists some of the problems and challenges that were identified by the Ontology Summit. The Communiqué ends with a conclusion and acknowledgments.

2. The Ontological Landscape

This section surveys the many aspects and dimensions of ontologies and their purposes. The most common classification of ontologies is by their level of generality. Section 2.1 discusses this dimension of the landscape. One of the primary purposes of ontologies is for communication. Section

28 2.2 surveys this dimension of the landscape. Section 2.3 discusses the different approaches to what
29 entities of an ontology represent. Sections 2.4 and 2.5 survey two more dimensions: attitudes toward
30 realism and representation of uncertainty. This section ends with a short discussion of ontological
31 commitment in Section 2.6. However, there are still other dimensions in the landscape, and these are
32 surveyed later on in Sections 3, 4 and 5.

33 **2.1 Generality**

34 Ontologies are most often classified by their level of generality. The most generic ontologies are called
35 foundation ontologies, top-level ontologies, and upper ontologies. The most specific ontologies are
36 called application ontologies because they are generally associated with a specific application or
37 narrow range of applications. In between these two extremes, there are reference ontologies and
38 domain ontologies. Reference ontologies are more specific than foundation ontologies but are not
39 limited to a particular domain. Domain ontologies are limited to a single domain, but domains can
40 form hierarchies with many levels of generality and the domain ontologies may also have many levels
41 of generality. [Schneider:21]

42 **2.2 Communication**

43 One of the main purposes of ontology is to improve communication. Communication can be between
44 people, between people and machines, and between machines. Machines operate best when the terms
45 are precise and logical so that there is no ambiguity. Human language, on the other hand, is much
46 richer with large numbers of types of figure of speech and inherent ambiguities. [Baclawski:21w] Far
47 from being a flaw of human language, this richness is its strength. Humans deal with the ambiguity of
48 language by engaging in dialogs that clarify the ambiguities. [Sowa:21] Accordingly, ontologies
49 should recognize the distinctions between the needs of humans and machines. One means of
50 accomplishing this is to have a "language interface" that mediates between the human and machine
51 terminology. The language interface is an important feature of modern software engineering processes,
52 and ontology development can also benefit from these practices. [Bennett:21], [Woods.Low:21]
53 Another effective means for communication is the use of events and narratives. [Westerinen:21]

54 **2.3 Ontology Styles**

55 There are many ontology styles that can be employed to better match ontologies to their purposes.
56 Some of the issues that must be addressed include [Bennett:21]:

- 57 • Truth makers versus data surrogates
 - 58 ◦ What does it mean to be a thing?
 - 59 ◦ What data is available about that thing or its distinguishing properties?
- 60 • Digital Fingerprints
 - 61 ◦ When there is no data directly reflecting a real-world truth maker
 - 62 ◦ What data can be seen to reliably exist as a 'fingerprint' indicating the presence of the
 - 63 relevant truth-makers?

- Data typing considerations
 - Kinds of information in the world versus computational datatypes
 - Typing decisions as a kind of design decision

2.4 Realism

Another dimension of ontology is the degree of "realism". Philosophical realism is the thesis that there is a reality that exists independently of people. An ontology is realist if it is based on this thesis. One consequence of realism is that one can only have classes that have at least one instance. Ontologies do not necessarily require realism; indeed, computer scientists focus are generally more pluralistic with respect to what entities may be relevant, whether they are ontological (in the philosophical sense), conceptual, cognitive, hypothetical, etc. From this point of view in knowledge representation and conceptual modeling there is no absolute reality. This allows for many different "realities" and views. This has benefits for dealing with natural language, common sense, and other human capabilities. Another consequence of a pluralistic attitude toward realism is that there is no absolute truth. There are different notions of truth that underly different reasoning mechanisms and that depend on context and situation. [Masolo:21] One of the most important examples of a different reasoning mechanism is probabilistic reasoning that is discussed in the next section.

2.5 Uncertainty

Human understanding is probabilistic. To reflect our best understanding of the world, ontologies must be, too. However, there is no generally agreed upon way to specify uncertainty in an ontology. Perhaps this is an advantage. It may be that the notion of uncertainty, like other aspects of an ontology, will depend on the particular domain, community and purpose. Rather than attempt to create a single reference ontology for uncertainty, there should be a meta-theory for uncertainty that is modularized and customizable for the particular task and purpose. [Breiner:21]

2.6 Commitment

Ontology development is similar to software development; indeed, ontologies may be regarded as a kind of software. Consequently, many of the same issues arise. In particular, during the development process, many design decisions must be made. These decisions vary with respect to how much of the ontology is affected. While modularity can help limit the scope of a decision, there will nevertheless be decisions that have major consequences for the process and the artifacts that are developed. These design decisions are commonly referred to as "ontological commitments". When using this term in the context of ontology development, one must be careful not to confuse the term with the philosophical notion of ontological commitment, such as Quine's Criterion.

96 3. Definitions

97 A definition is a formal statement of the meaning or significance of an entity, including words, phrases,
98 classes and properties. Accordingly, definitions can serve as links between formal ontologies and
99 informal ontologies as well as between different formal ontologies. Historical attempts to standardize
100 terms included creating core metadata models and common conceptual models for combining data into
101 a single representation. These however have largely failed to be adopted because of flawed
102 conceptualizations, lack of community agreement, and inadequate representation; and thus they have
103 resulted in silos. Some progress has been made leveraging best practices including the use of
104 ontological analysis and design. This section surveys the different notions and levels of formality of
105 definitions, with emphasis on practical methods to harmonize a variety of semantic resources.

106 Domain vocabularies vary significantly in quality and scope, often with alternative definitions for the
107 same term and definitions that have varying degrees of formality. These problems have been
108 recognized for a long time. It is said that when Confucius was asked what he would do if he was a
109 governor, he replied that he would "rectify the names" to make words correspond to reality.
110 Standardization of term meanings is challenging since there are many conflicting and overlapping
111 glossaries or data models that define domain terms in idiosyncratic ways. Completely "rectifying the
112 names" may be too ambitious a goal, but harmonization can be achieved, albeit with some effort.
113 Harmonizing terminology is underway in some domains, such as the cryosphere, which is concerned
114 with ice fields and glaciers.

115 Writing a good definition is not as easy as one would expect. It is even more difficult when one must
116 harmonize many terms. Indeed, one of the main problems with harmonizing definitions for ontologies
117 is precisely the common assumption that anyone can write definitions and that they will then be
118 harmonious. In fact, it requires training and experience. Experiments have shown that even for well
119 understood domains, the results of manual classification tasks performed by domain experts are highly
120 inconsistent. [Westerinen:21] So the first step in writing definitions is to accept that one does not know
121 how to do it. The next step is to start learning how. Guidelines for writing definitions in ontologies are
122 available at [Seppälä.Ruttenberg.Smith:17]. The rest of this section gives examples and lessons learned
123 about developing and harmonizing definitions.

124 3.1 Definitions in the Environmental Sciences

125 In this section, a specific example is presented in detail of a domain that has addressed semantic. The
126 domain is the environmental sciences, and the ontology is ENVO. [ENVO] Envo is a widely cited
127 semantic resource for semantically controlled descriptions of environmental entities. For example, the
128 Darwin Core glossary uses ENVO in its habitat descriptions and was developed by applying text-
129 mining approaches to extract habitat information from the Encyclopedia of Life and automatically
130 create experimental habitat classes within ENVO.

131 Here is an example of an ENVO definition: A habitat is "An environmental system which can sustain
132 and allow the growth of an ecological population."

133 ENVO's initial focus was to represent biomes, environmental features, and environmental materials,
134 and the initial purpose was for genomic and microbiome-related investigations. However, the need for
135 environmental semantics is common to a multitude of fields, and ENVO's scope has steadily grown
136 since its initial description. As the scope has expanded, the ontology has been enhanced and
137 generalized to support its increasingly diverse applications such as the Cryo (glaciers and ice fields)
138 and the Marine (ocean) realms that is now examined in more detail.

139 The global cryosphere watch (GCW) has sponsored the effort to harmonize term definitions that are
140 already in use in the Cryo domain. Sometimes a single term will have dozens of different term
141 definitions. Consider the term "snow cover". The GCW harmonized definition is the following:

142 "An area density which inheres in snow distributed over an area of a landmass or other substrate."

143 Additional information about this term is provided by annotations. For example, one annotation states
144 that, in general, snow cover is a layer of snow on the ground surface and can be compared to the related
145 terms of snowfield and snowpack. Annotations can also be expressed using logical axioms. One such
146 annotation is the following:

147 "has part some (surface layer and (composed primarily of some snow))"

148 ENVO definitions conform to the Minimum Information for the Reporting of an Ontology (MIRO)
149 guidelines. [Matentzoglou et al:18]

150 The ENVO harmonization effort has had a number of accomplishments.

- 151 1. The GCW glossary analysis results were harmonized with ENVO and aligned with
152 corresponding terms in the Semantic Web for Earth and Environment Technology (SWEET)
153 ontology.
- 154 2. The alignment with SWEET has improved the definitions of SWEET terms.
- 155 3. ENVO and SWEET terms have been aligned with other OBO Foundry ontology terms.
- 156 4. A special envopolar subset of ENVO has been crafted with relevant terms and axioms.
- 157 5. Attribution for all of these updates have been documented to the people or groups responsible
158 for these changes.
- 159 6. Best practices for documentation were instituted, including annotating the time the definition
160 was added, the orcid.org of the individual who created the update, and provenance information
161 about definition derivation from ENVO to SWEET.
- 162 7. Templating methods were employed to accelerate class creation, and spreadsheets were used to
163 help update ontologies with definitions. Specifically, the OBO Robot tools was used to
164 facilitate ENVO collaboration. This tool guides users through the process of creating new
165 terms and is intended to be used by non-ontologists. Robot organizes new term requests in a
166 standardized google sheet template, and users can follow a step by step process to fill out the
167 appropriate spreadsheet columns.

3.2 Lessons Learned and Best Practices

In this section, some of the best practices for harmonization of definitions are presented. For general advice about developing definitions, see the Guidelines for Writing Definitions in Ontologies. [Seppala.et.al:17] The guidelines include general advice such as: be brief, align, re-use, extend, and revise semantic resources. The following are other best practices and advice [Berg-Cross:21]:

1. Organize the terms into a concept system (network of concepts), and indicate the position of each term in the concept system. Somewhat informal conceptual models may help organize the networks during the early phases of ontology development.
2. To ensure that definitions are brief, put other, more encyclopedic information in annotation notes.
3. To structure taxonomies use the nearest generically super-ordinate concept, adding one or a few constraining characteristics.
4. Consider systematizing the role of lexical modifiers as well as nouns and verbs. For example, words like alpine or liquid.
5. Reuse broad vocabularies such as Schema.org, DCAT2, VIVO, DDI, etc. However, doing so will usually involve specializing the more general term.
6. Eliminate (conceptual) ambiguity by explicitly showing relations between the terms.
7. Be sensitive to issues of granularity. Metonymy is a the naming of a thing by something related to it. For example, using the name of the whole for a part or vice versa. Figures of speech such as metonymy are so commonplace that one may not be aware that one is using a figure of speech at all.
8. Avoid vague comparisons, such as "similar". There will usually be many similar terms and terms with overlapping meanings. Perform semantic analysis to distinguish similar terms. This involves establishing relations among terms where appropriate, such as subtypes, part-whole relationships, roles, influences, production (output) relationships, etc.
9. The interface between the business (natural) language and the technical language (FOL) helps human-machine communication by making language more precise and less ambiguous. [Woods.Low:21] However, when communicating with natural language, one must be aware of how people categorize the world. Unlike the classes and properties of ontologies, human categories "shimmer". [Baclawski:21w]
10. Terms may be standardized but the meaning has to be too. [BergCross:21]
11. Definitions will vary according to context of use and target audience. [BergCross:21]

4. Neuro-Symbolic Learning Ontologies

Symbolic reasoning has a long history, and continues to be an active area of research. Machine learning, also known as subsymbolic methods, is also a very active area of research. Although both are

part of AI, they two areas have been developed under clearly distinct technical foundations and by separate research communities. The two areas have complementary strengths and weaknesses. As a result, finding ways for bridging the gap between symbolic and subsymbolic approaches to AI is a long-standing unresolved challenge, and integrating these two areas is now the subject of growing research interest in AI. Neuro-symbolic learning aims to integrate neural learning with symbolic approaches typically used in computational logic and knowledge representation in AI. One benefit of such an integration is the development of effective knowledge extraction methods towards explainable AI, which was examined in the Ontology Summit 2019 [Lamb:21],[Ontolog:19b], but there are many other advantages. While there are significant benefits for tighter integration of neural and symbolic paradigms, it not known how best to integrate them, and many integration architectures have been proposed. Symbolic models can be the result of, or the basis for, different stages of a neural process. The following are some of the architectures of techniques that integrate symbolic and subsymbolic methods [Kautz:21]:

1. The simplest and most common architecture is one in which symbolic data (e.g., documents) are processed with symbolic techniques to produce vectors that are input to a subsymbolic module (e.g., a neural network). The vector output of the subsymbolic module is then interpreted in symbolic form using symbolic techniques.
2. Another architecture is a symbolic system that can invoke subsymbolic submodules. As far as the symbolic system is concerned, the submodules are just subroutines like any others. This is the architecture commonly used by self-driving vehicles.
3. One could, in principle, invert the roles of symbolic and subsymbolic in the second architecture to get an architecture in which it is the subsymbolic system that is invoking the symbolic one. This architecture would be useful for dealing with complex decision making, since symbolic reasoners can perform combinatorial reasoning much more scalably and efficiently than subsymbolic systems.
4. The architectures discussed so far do not learn during normal processing. The symbolic and the subsymbolic modules have already been programmed and trained, respectively. Some recently developed architectures incorporate symbolic reasoning in the subsymbolic system by organizing the subsymbolic system according to symbolic rules. Examples of this architecture are tensor product representations and logic tensor networks. This architecture is useful for finding concept hierarchies such as generalization and part-whole hierarchies. In other words, this architecture can generate ontologies, or at least some aspects of ontologies. The forms of reasoning that can be incorporated include temporal logic, description logic, and first-order predicate logic. [Hitzler:21]
5. Another learning architecture is to input instances of logical inferences, expressed as input-output pairs, to train a subsymbolic model. This is primarily useful for mathematical problems, and it is surprisingly effective, although it will make mistakes sometimes. [Kapanipathi:21]
6. Another architecture uses subsymbolic methods, such as back-propagation, to train a symbolic system. This can be useful for question-answering systems. The back-propagation is invoked whenever the system makes a mistake.

Another reason why subsymbolic methods need knowledge based methods is the intuition, based on human behavior, that intelligence necessarily involves learning, knowledge from experience, and reasoning, which could be expressed as an equation like this:

Intelligence = Learning from Data + Knowledge/Experience + Reasoning [Sheth:21]

This intuition has been used as the basis for Knowledge-Infused Learning (KIL). The KIL notion is a collection of architectures that range from shallow to deep infusion of knowledge, allowing one to match an architecture with a particular application. KIL has the potential to have impacts on robotics, cognitive science, autonomous vehicles, and personal assistants. [Sheth:21]

The examples discussed so far have been primarily concerned with single-input decision making. In other words, while the architectures may vary, the overall structure is essentially the same as the first architecture. Machine learning methods have also been applied to signal processing tasks. One example of such a task is to construct an ontology from a graph signal. [Majumdar:21],[Sowa:21]

The survey given in this section has attempted to give a rough organization of the large variety of neuro-symbolic architectures that are now being developed, but one can expect new architectures to be developed by researchers in this very active field.

5. Sustainability of Ontologies

Many organizations, including government agencies, standards bodies and commercial firms, use ontologies and have developed tools for various ontological activities, such as creation, evolution, mapping and other forms of harmonization. This section discusses the issues involved in sustainability of ontologies.

Sustainability involves addressing much more than simply ensuring sufficient funding. Achieving sustainability fully requires building on a firm foundation. The most important aspects of such a foundation are the following three "pillars" [Dickerson:21]

1. Economic Viability. This pillar is the most obvious. One must ensure there is sufficient funding for maintaining the ontology for as long as its purpose remains relevant.
2. Social Equity. Ontologies can have subtle biases, especially when ontologies are generated using machine learning. The problem is that data invariably has biases to a greater or lesser extent, and machine learning techniques cannot find or correct them on its own. Combining symbolic knowledge with machine learning can help to discover and mitigate bias, but the most important problem is accepting that correcting bias is an important part of ontology development and maintenance.
3. Environmental Protection. The term "environment" in this pillar refers to the human environment that surrounds the specific community that developed an ontology. It is important to recognize that communities and their ontologies do not exist in isolation. One must maintain avenues of communication and cooperation with adjacent and other related communities.

Well designed definitions, documentation and harmonization, as discussed in Sections 3 and 4 above, can contribute both to the Social Equity and the Environmental Protection pillars of sustainability.

280 The pillars of sustainability illustrate that there is much more to ontology development than simply
281 creating the ontology. One must put the ontology into production. This is similar to the problem of
282 putting software into production. [Need more content here.]

283 **6. Problems and Challenges**

284 In this section, some of the problems and challenges of ontology harmonization are presented.

- 285 • Harmonization can be very time-consuming, so tools are important to simplify the effort
286 required as well as to manage the effort over time. The challenge is to develop better tools for
287 harmonization.
- 288 • Community agreement is needed for extensions and revisions. Good mechanisms for
289 community discussion are important as are partner agreements with groups with domain
290 vocabularies. A variety of tools are used for coordination and harmonization, such as Slack and
291 Github, but community members are not necessarily skilled in the use of these tools. The
292 challenge is to develop state-of-the-art tools and techniques for reaching agreement efficiently
293 with a diverse community.
- 294 • Another challenge is maintaining access to the vocabulary for reuse or for alignment with
295 ontologies. Some preliminary mechanisms have been established for this, but improving these
296 mechanisms is needed.

297 **7. Conclusion**

298 **8. Acknowledgments**

299 Certain commercial software systems are identified in this paper. Such identification does not imply
300 recommendation or endorsement by the National Institute of Standards and Technology (NIST) or by
301 the organizations of the authors or the endorsers of this Communiqué; nor does it imply that the
302 products identified are necessarily the best available for the purpose. Further, any opinions, findings,
303 conclusions or recommendations expressed in this material are those of the authors and do not
304 necessarily reflect the views of NIST or any other supporting U.S. government or corporate
305 organizations.

306 We wish to acknowledge the support of the ontology community, especially the invited speakers and
307 participants who contributed to the Ontology Summit. There were many invited speakers, some of
308 whom gave presentations at more than one session. The complete list of sessions, speakers, and links
309 to presentation slides and video recordings is available at <https://bit.ly/3gKukvC>

310 [References are in a separate document.]

