Environmental Health Language Collaborative (EHLC): a route to environmental health science data harmonization

Anna Maria Masci [1], Stephanie Holmgren [1], Charles Schmitt[1], Rima Habre2, Anne E Thessen[3], Rebecca Boyles[4], Carmen Marsit[5].

[1] *Office of Data Science, National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, North Carolina, USA*

[2] *University of Southern California, Los Angeles, CA, USA*

[3] *University of Colorado Anschutz Medical Campus, Center for Health AI, Aurora CO, USA*

[4] *Center for Data Modernization Solutions, RTI International. Durham, NC. USA*

[5] *Gangarose Department of Environmental Health, Emory University Rollins School of Public Health, Atlanta, GA, USA*

Abstract

Standard language is critical for helping scientists share, compare, and reanalyze data. The increased use of automatization and AI technologies has made the adoption of machine interpretable language essential. Due to the broadness of the domains that are under the environmental health umbrella there is not yet a set of common standard terminologies. To address this lack of standardized language, NIEHS has launched the Environmental Health Language Collaborative (EHLC) https://www.niehs.nih.gov/research/programs/ehlc/index.cfm. This is a new initiative to advance community development and application of a harmonized language for describing Environmental Health Science (EHS) research.

As a first step toward the development of standard terminology, a working group of environmental health researchers and NIEHS program officers established an initial set of four general use cases. Here we present one of the initial use cases on place-based exposures. This preliminary work is intended to be expanded as the community develops. EHLC is seeking larger community involvement as well as additional use cases. NIEHS encourages anyone interested in advancing this mission to engage in this community.

Keywords

Ontology; controlled vocabulary; data reuse; FAIR data metadata; taxonomy; standards; semantic; environmental health; toxicology; community of practice; community driven, geospatial, place, location, exposure.

1. Introduction

Environmental health (EH) is a science that studies the effect of exposure to environmental factors on human health. The definition of environment is wide and includes the "totality of exposures we face throughout our lives, e.g., the food we ingest, the air we breathe, the objects we touch, the psychological stresses we face, the activities in which we engage" [1] EH research is not just focused on external exposures, but also considers the molecules in our body that derive from external exposures, the environmental influences we receive through our parents, the socio-economic factors that play into disparities in health as well as research that seeks to remediate and reduce the impact of these factors, e.g., by engineering plants that can remove or reduce pollutants.

The EH field covers a diversity of domains and methodologies, such as environmental epidemiology, toxicology, clinical and translational research, immunology, microbiology, exposure science, social science, and environmental engineering.

Progress in EH research depends on the ability to compare, contrast, and integrate data from across the field, which requires adoption of the principles of Findable, Accessible, Integrable, and Reusable (FAIR) [2, 3]data. FAIR requires the use of either common or comparable language in describing scientific data, metadata, and findings.

The breadth of the EH field challenges the use of a common language, not only because there are inconsistencies and gaps in the terminologies and ontologies used within subfields, but scientific language is often domain-specific and standardizing or even harmonizing language across subfields is especially challenging.

To address the need for common language, NIEHS has launched the Environmental Health Language Collaborative (EHLC)[4] https://www.niehs.nih.gov/research/programs/ehlc/index.cfm. This is a new initiative to advance community development and application of a harmonized language for describing Environmental Health Science (EHS) research.

The proposed mission of this community is to:

- Apply language standards and best practices for accurate environmental health data and knowledge representation
- Cultivate a vocabulary aware environmental health community through training and education
- Foster community-based development of harmonized vocabularies, terminologies, and ontologies
- Identify use cases for applying knowledge organization systems in research
- Promote and develop methods and tools for applying harmonized language in research

ICBO 2022, September 25–28, 2022, Ann Arbor, MI, USA
EMAIL: mascia2@niehs.nih.gov (Anna Maria masci.); holmgre1@niehs.nih.gov (Stephanie Holmgren.); charles.schmitt@nih.gov (Charles Schmitt.); habre@usc.edu (Rima Habre.); annethessen@gmail.com (Anne E. Thessen); rboyles@rti.org (Rebecca Boyles); carmen.j.marsit@emory.edu (Carmen Marsit).
ORCID: 0000-0003-1940-6740 (Anna Maria Masci.); 0000-0002-3148-2263 (Charles Schmitt); 0000-0002-2908-3327 (Anne E. Thessen); 0000-0003-0073-6854 (Rebecca Boyles); 0000-0003-4566-150X (Carmen Marsit).

A first step towards a practical approach has been asking what scientific questions would benefit most from development and adoption of a harmonized language standard? A working group of environmental health researchers and NIEHS program officers developed an initial set of five general use cases examples as starting points for community discussion. Working groups led by use case champions were formed from the community to address each use case. The working group teams decided to focus all the use cases on the effect of Particulate Matter (PM) main component of the air pollutants on Asthma as a common theme to unite their work. '*Asthma is a disease of the respiratory tract which is caused by a combination of environmental and genetic factors*' [5, 6]. '*Particulate matter is an environmental material which is composed of microscopic portions of solid or liquid material suspended in another environmental material*'[7, 8]. PM derives from multiple different sources, such as vehicle and industrial emissions from fossil fuel combustion, cigarette smoke, and burning organic matter, such as wildfires, as well as chemical reactions that can form PM from precursors.

WHO has estimated that 4.2 million deaths occur as a result of exposure to ambient (outdoor) air pollution. (https://www.who.int/health-topics/air-pollution - tab=tab_2).

Although all the five use cases focus on Asthma and PM, each of them is trying to answer different questions:

1. What data exists for a given chemical/endpoint/exposure scenario?
2. How best to combine data from multiple independent studies?
3. Given measures of biological responses to one or more exposures, what are the biological processes that might be related to the observed changes?
4. What are the biomarkers, phenotypes, and/or outcomes that can be measured and used as an indicator of exposure?
5. What do my unique exposure conditions based on where I live and work (E.g., Geographical Location, Occupation, Regulations, Hobbies) indicate about potential risks to my health?

Due to space limitation, we present the very preliminary work done for use case five to begin building the ontological representation of environmental exposures and social stressors or factors assessed based on place and geospatial information.

2. Methods and Results

Geospatial data are composed of three general components Object, Event, Location, and each of these components has specific characteristics that are time related Fig (1).

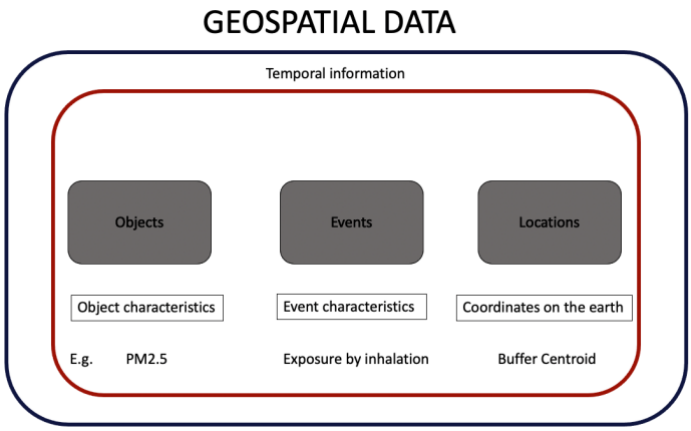

Figure 1. The main components associated with geospatial data.

Members of the Geospatial Working Group started by looking at an available data set from the Personalized Environment and Genes Study (PEGS) (https://www.niehs.nih.gov/research/clinical/studies/pegs/index.cfm).

The study's panel of experts had already identified the initial set of essential components to represent the geospatial data.

Because we are using an existing list of data elements, the first step was to look at the OBO Foundry ontologies [9] to see if those data elements were already captured in existing ontologies. Figure 2 shows a representation of

the data elements that were captured and their relations. The different box color represents the different ontologies from which the terms were imported: Exposure Ontology (ExO) **[10]**, Gazetteer (GZ) (http://environmentontology.github.io/gaz/), Ontology of Biomedical Investigations (OBI) [11, 12], phenotype and trait ontology (PATO) [13]. In italic are the relations from the relation Ontology (RO) **[14]** that we have used to link the terms. In addition to the imported terms, new terms have been identified as '*stressor detection assay*' and '*stressor* detector'. For the stressor detection assay we are proposing the following definition: '*an assay that aims to detect exposure stressor'*. We are proposing the stressor detection to be a child of a more general term assay defined in OBI.

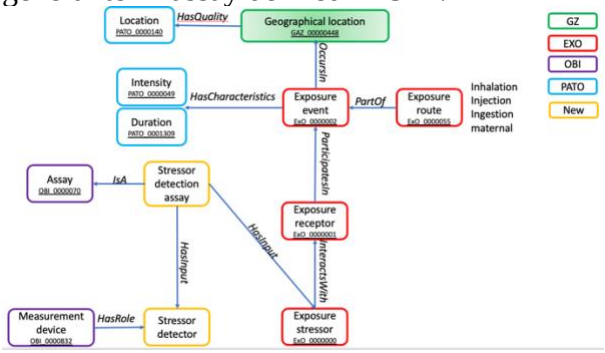

Figure 2. Ontological representation of an exposure event. The different box colors indicate the different ontologies from which the terms were imported. The gray boxes indicate the term has not been found in any ontology. The green filled box highlights the term that is present in Figure 2 as well as Figure 3.

The second additional term is a *stressor detector.*

We are proposing the following definition: *Is a role that inheres in a material entity, and which is realized through a process of exposure stressor detection.*

These two new terms as well as their definitions have been proposed to the ontology community.

The red triangle in Figure 2 represents the term that is the linking node between Figures 2 and 3.

We then explored the ability of ontology to capture more specific geographical types of information. Figure 3 shows a list of terms related to geographic location. There are terms like *latitude measurement datum, longitude measurement datum* that have been already described in Ontology of Biomedical Investigations. Other terms like *geographical identifier (GEO ID)* and *Buffer zone*, which are commonly used in geospatial studies, are not present in any ontology.

The Census Bureau and other state and federal agencies are responsible for assigning geographic identifiers, or GEOIDs, to geographic entities to facilitate the organization, presentation, and exchange of geographic and statistical data. (https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html)

We have classified the *GEO ID* term as *identifier* class defined in the IAO (https://obofoundry.org/ontology/iao.html).

We have modified the Census Bureau definition for the *GEO ID* to be 'is an identifier composed by numeric codes that uniquely identify all administrative/legal and statistical geographic areas for which the Census Bureau tabulates data.'

Another term that we needed to represent is a *Buffer Zone.* This is a very common term used to define a zone and its characteristics, that are the object of the study. Although there is this term in ENVO its classification under *administrative region* does not fit with our usage of the term. In our use case the buffer zone is used to define a zone from which collecting data (point, line, area) that is equidistant from the stressor.

As is shown in Figure 3 classification of this term is still under discussion as well as how to relate it to a specific geographic location.

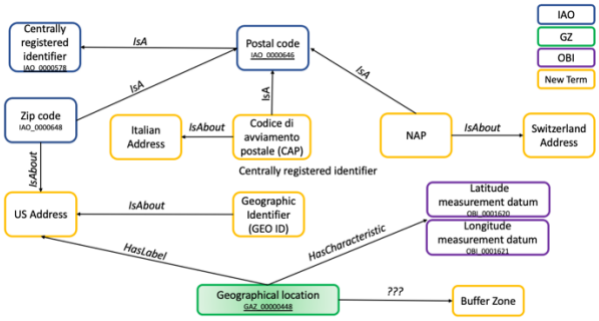

Figure 3. Ontological representation of the geographical specific entities. The different box colors indicate the different ontologies from which the terms were imported. The gray boxes indicate terms that have not been found in any ontology. The green filled box highlights the term that links Figure 3 and Figure 2.

3. Discussion

Geospatial studies use several heterogeneous data types. Although some efforts have been developed to create a common language like the Open Geospatial Consortium (OGC) (https://www.ogc.org/), there is still a lack of standardized machine-readable terminology.
As ontology good practice we are reusing several terms from other OBO Foundry ontologies. We are creating new relations between these terms to better represent geospatial information. For the terms that were not found in other ontologies we have created a new term and definitions.
This is a preliminary attempt to use an ontological representation of the geospatial data for an exposure. *Our near future goal for the use cases is to extract and represent the minimal information necessary for capturing use cases data. That could serve as reference for the environmental data*
These efforts are being developed under the community-driven Environmental Health Language Collaborative to ensure an open and broad community participation and development of harmonized language

4. Acknowledgements

Environmental Health Language Collaborative. Members of Environmental Health Language Collaborative Geospatial working group

5. References

Uncategorized References

1. Miller, G., *The Exposome: A primer* 2014, https://doi.org/10.1016/C2013-0-06870-3: Elseview Inc.
2. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Sci Data, 2016. **3**: p. 160018.
3. Wilkinson, M.D., et al., *Addendum: The FAIR Guiding Principles for scientific data management and stewardship.* Sci Data, 2019. **6**(1): p. 6.
4. Holmgren, S.D., et al., *Catalyzing Knowledge-Driven Discovery in Environmental Health Sciences through a Community-Driven Harmonized Language.* Int J Environ Res Public Health, 2021. **18**(17).
5. Schriml, L.M., et al., *The Human Disease Ontology 2022 update.* Nucleic Acids Res, 2022. **50**(D1): p. D1255-D1261.
6. Schriml, L.M., et al., *Human Disease Ontology 2018 update: classification, content and workflow expansion.* Nucleic Acids Res, 2019. **47**(D1): p. D955-D962.
7. Buttigieg, P.L., et al., *The environment ontology: contextualising biological and biomedical entities.* J Biomed Semantics, 2013. **4**(1): p. 43.
8. Buttigieg, P.L., et al., *The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation.* J Biomed Semantics, 2016. **7**(1): p. 57.

9. Smith, B., et al., *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nat Biotechnol, 2007. **25**(11): p. 1251-5.
10. Mattingly, C.J., et al., *Providing the missing link: the exposure science ontology ExO.* Environ Sci Technol, 2012. **46**(6): p. 3046-53.
11. Bandrowski, A., et al., *The Ontology for Biomedical Investigations.* PLoS One, 2016. **11**(4): p. e0154556.
12. Vita, R., et al., *Standardization of assay representation in the Ontology for Biomedical Investigations.* Database (Oxford), 2021. **2021**.
13. Gkoutos, G.V., P.N. Schofield, and R. Hoehndorf, *The anatomy of phenotype ontologies: principles, properties and applications.* Brief Bioinform, 2018. **19**(5): p. 1008-1021.
14. Smith, B., et al., *Relations in biomedical ontologies.* Genome Biol, 2005. **6**(5): p. R46.