Environmental Health Language Collaborative

Catalyzing Knowledge-Driven Discovery in Environmental Health Sciences through a Harmonized Language

Virtual Workshop | September 9-10, 2021

Contents

Workshop Purpose

The goal of this workshop is to advance the development and adoption of harmonized environmental health language approaches through the formation of a sustained community effort, the Environmental Health Language Collaborative. We will spend time at the workshop obtaining your input and achieving community agreement on the proposed elements for the Collaborative.

In this highly interactive workshop, participants will be charged to "think together" on two themes:

1. **Building a Sustainable Community:** Obtain agreement on the proposed Environmental Health Language Collaborative's vision, mission, community model, and strategy to build an impactful community.
2. **Developing Sustainable Semantic Solutions:** Define use cases in environmental health sciences research and begin identifying semantic needs, gaps, and strategies for implementing solutions.

Pre-Workshop Preparation

In keeping with the spirit of the Collaborative being a community-driven initiative, the workshop is less presentation and more participant engagement. For community building, we will be polling attendees to gauge initial reaction to the proposed community model and using Mural, an online collaboration tool, to solicit more detailed input.

To promote productive discussions during the workshop, we encourage attendees to:

- Review the information about the Collaborative and use cases in the workshop program.
- Reflect on the Questions to Ponder (see p. 6-10) in advance of the workshop.
- Review the use case preparatory materials if you will be participating in a use case.
- Take time to review the recommended resources listed later in the program.
  Watch the pre-workshop webinar on The Value of Creating Language and Community in Catalyzing Knowledge-Driven Discovery in Environmental Health Research, held on June 24.
- Watch the pre-workshop webinar on A Primer on Using Terminologies, Vocabularies, and Ontologies for Knowledge Organization, held on July 20.
- Become familiar with Mural, a platform being used to obtain participant input throughout the workshop.
  - It is best to display Mural on a desktop monitor in a Chrome browser (do not open in Internet Explorer). While laptop monitors will work, you will need to do more scrolling.
  - If you use VPN, you will need to turn off VPN during the session to use Mural successfully.
  - Consider watching this tutorial for Mural beginners (1-min video) or learn how to navigate Mural (3.5-min video).

Workshop Agenda

Thursday, September 9, 2021

10:00am EDT	Welcome and Background *Setting the stage for the goals and outputs of the workshop* Stephanie Holmgren, NIEHS Charles Schmitt, NIEHS	
11:00am EDT	Developing Sustainable Semantic Solutions *10-minute presentations on the defined use cases, followed by discussion*	
12:30pm EDT	Break	
1:00pm EDT	Use Case Work-a-Thon *Continue to define the use case, develop action plan for next steps, and/or begin working on next steps*	
	Use Case: Discovery of exposure data Facilitator: Michelle Angrish, EPA	Use Case: Place-based exposures Facilitator: Carmen Marsit, Emory
2:30pm EDT	Break	
2:45pm EDT	Use Case: Integration of Exposure Data Facilitator: Jeanette Stingone, Columbia	Community Input – Semantic Solutions *Mural session – see Questions to Ponder (p. 10)*
4:15pm EDT	Day 1 Recap *Participants share highlights and key takeaways of the day* Facilitator: Charles Schmitt, NIEHS	
4:30pm EDT	Adjourn	

Friday, September 10, 2021

10:00am EDT	Introduction *Review of Day 1 and Goals for Day 2* Stephanie Holmgren, NIEHS	
10:10am EDT	Building a Sustainable Community, Part 1 *Goals: obtain agreement on proposed community vision, mission, goals, and activities* Stephanie Holmgren, NIEHS	
11:30am EDT	Break	
12:00pm EDT	Use Case: Bridging Exposure and Biomarkers of Exposure Facilitators: Stephen Edwards, RTI and Chirag Patel, Harvard	Community Input – Collaborative Community *Mural session – see Questions to Ponder (pp. 6-9)*
1:30pm EDT	Break	
1:40pm EDT	Building a Sustainable Community, Part 2 *Goals: obtain agreement on proposed governance and infrastructure model* Stephanie Holmgren, NIEHS	
2:40pm EDT	Report Out: Use Case Work-a-Thons *10-minute presentations on the status of work, followed by discussion* Michelle Angrish, Carmen Marsit, Jeanette Stingone, Steve Edwards, and Chirag Patel	
3:40pm EDT	Moving Forward *Outline action plan for follow-up, including timeline, people, and next steps* Stephanie Holmgren, NIEHS	
4:00pm EDT	Adjourn	

Background

The generation of Environmental Health Sciences (EHS) data continues to increase at a rapid rate. An essential component to leveraging this data to answer large-scale complex questions is to describe data with a harmonized language to promote data sharing, reuse, and reanalysis. Applying a harmonized language to EHS data enhances its value by increasing the findability of data, facilitating consistent interpretation of data and metadata, permitting integration and promoting interoperability of data and databases, and enabling the assembly of datasets for computational modeling and knowledge discovery. The need for a harmonized language to express the knowledge gained from EHS research is equally critical to allow for transfer of knowledge between scientific communities, to support development of tools and aggregation of knowledge, and to convey findings to the broader population.

Increasing the use of a common, or even harmonized set of languages, is a grand challenge problem. While many biomedical terminologies and ontologies exist, often these terms do not exist in an EHS context or it is unclear which of the synonymous terms to use as the standard. To start addressing this issue, several programs and projects have created application-specific EHS-focused terminologies and ontologies. However, a major question arises as to how we can achieve consensus and coordinate, map, and harmonize those efforts without impeding research. Furthermore, researchers have limited automated tools to describe study designs and study findings using common terminologies and to subsequently link those terminologies to ontologies to further aid in data analysis, integration, and interpretation.

To address this challenge in the environmental health field, the NIEHS and partners are working to establish an Environmental Health Language Collaborative, a community-driven initiative to advance the development and adoption of harmonized language approaches within environmental health and toxicology.

The Collaborative is being organized around addressing use case problems (see pg. 10 for further information). Use cases provide a means of communicating between different scientific communities on enabling practices and technologies and allow for collective examination of:

- terminology and ontology gaps that impede research goals,
- specific challenges in advancing harmonized languages, and
- opportunities for advancing the creation and adoption of terminologies and ontologies.

This initiative is based on work begun at an NIEHS-sponsored event, “Workshop for the Development of a Framework for Environmental Health Science Language” (workshop proceedings) held in September 2014 and will leverage additional work achieved at the Computable Exposures Workshop.

Building a Sustainable Community

What is the Environmental Health Language Collaborative?

The Collaborative is a new initiative to advance community development and application of harmonized language approaches for describing Environmental Health Science (EHS) research.

In 2021, we strive to define the Collaborative and begin working together. We welcome diverse representation of expertise, needs, and scientific interests to make this a successful and sustainable community.

To start the process of community building, several working groups have drafted a community name, Vision, Mission, Goals, Roles and Activities, Community Model, and Use Case Profiles. We will spend time at the workshop obtaining your input and achieving community agreement on these proposed elements for the Collaborative.

Questions to Ponder

On a scale of 1 (low) – 5 (high), do you endorse/agree with the proposed a) vision, b) mission, c) goals, and d) activities?

What changes do you recommend being made to the proposed statement(s) that would lead to your endorsement?

For you to become engaged, what would you like to see the community work on/accomplish in the next year?

Vision

The **vision** of the Environmental Health Language Collaborative is to leverage community-driven environmental health language standards to catalyze knowledge-driven discovery and improve public health.

Mission

The Collaborative's **mission** is to advance integrative environmental health research by developing and promoting adoption of a harmonized language.

Goals

To achieve the Collaborative's mission, the community will:

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

Roles and Activities

We envision the community would be composed of three elements:

- **Community of Practice**. A community of practice to exchange information, ideas, and expertise. In addition, the community will be a place to advance the appreciation for and adoption of semantic and language approaches through education and training.

- **Forum to Coordinate**. The community serves as a hub to coordinate and prioritize harmonization activities, define use cases and gaps, and describe the language strategies or approaches to enable data querying, sharing, and interoperability.
- **Platform to Collaborate**. Based on identified gaps in use cases, the community serves to support and promote the development and application of harmonized language solutions to address the use case needs.

Questions to Ponder

How can the Collaborative support creating a more "vocabulary aware" EHS research community?

What are the barriers to adoption and what can the Collaborative do to promote adoption of harmonized language approaches?

How do we want to define success for the Collaborative? How can we measure it?

Community Model

Based on interviews with four community organizations (AOP Wiki, Earth Sciences Information Partnership, OBO Foundry, and Research Data Alliance) and discussions within a community-model working group, the community model shown in Figure 1 is proposed.

A key aspect of any community is having an infrastructure for communications, hosting meetings, and other ongoing operational activities. One of the recurring messages from stakeholders and community interviews was to not reinvent the wheel in this regard. As such, the Research Data Alliance (RDA) is proposed to provide structure for the Environmental Health Language Collaborative.

RDA is an international community-driven organization with the mission to "build the social and technical bridges to enable open sharing and re-use of data to accelerate data-driven innovation." Through Interest Groups and Working Groups, its members exchange knowledge, discuss barriers and potential solutions, explore and define policies, and harmonize standards to facilitate global data sharing and re-use. As such, RDA's activities and outputs strongly align with NIH's interest to improve data management, data sharing, and data interoperability to maximize the value of NIH data.

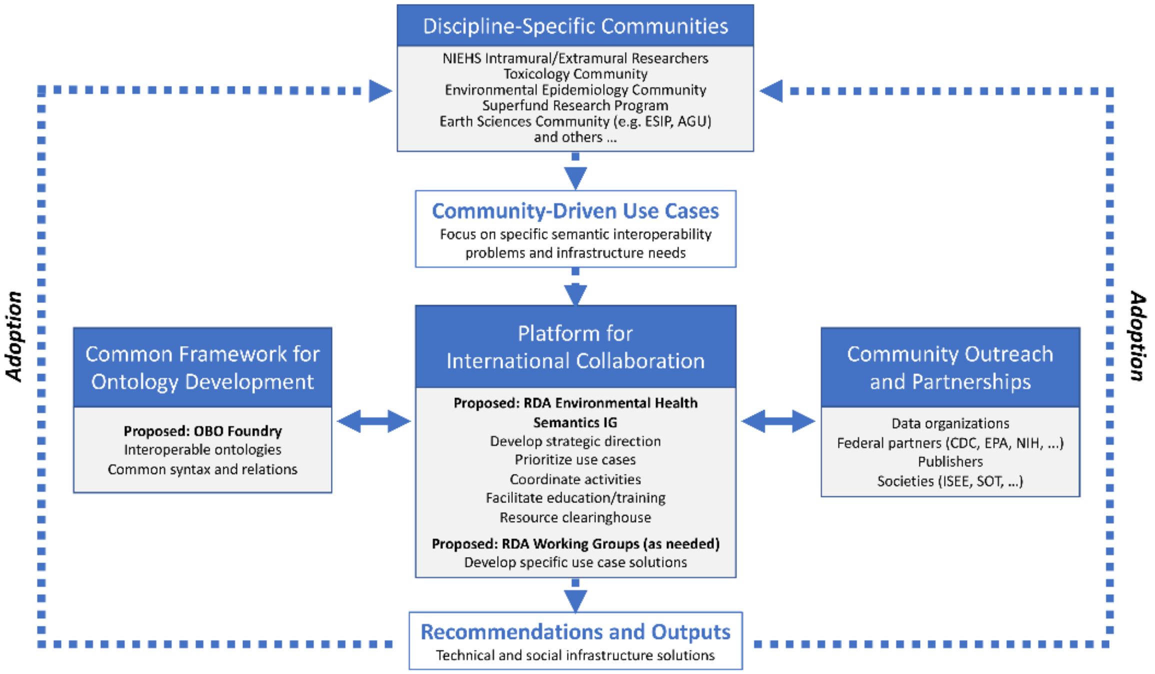

Discipline-Specific Communities
NIEHS Intramural/Extramural Researchers
Toxicology Community
Environmental Epidemiology Community
Superfund Research Program
Earth Sciences Community (e.g. ESIP, AGU)
and others ...
Community-Driven Use Cases
Focus on specific semantic interoperability problems and infrastructure needs
Adoption
Common Framework for Ontology Development
Proposed: OBO Foundry
Interoperable ontologies
Common syntax and relations
Platform for International Collaboration
Proposed: RDA Environmental Health Semantics IG
Develop strategic direction
Prioritize use cases
Coordinate activities
Facilitate education/training
Resource clearinghouse
Proposed: RDA Working Groups (as needed)
Develop specific use case solutions
Community Outreach and Partnerships
Data organizations
Federal partners (CDC, EPA, NIH, ...)
Publishers
Societies (ISEE, SOT, ...)
Adoption
Recommendations and Outputs
Technical and social infrastructure solutions

Figure 1. Proposed Community Model

Discipline-specific communities. The proposed model begins with individuals and/or groups from discipline-specific communities generating use cases based on research questions of interest to them.

Community-driven use cases. These use cases represent needs for harmonized language solutions that will enhance the findability, sharing, and interoperability of environmental health sciences data.

Platform for international collaboration. The use cases will be brought to a proposed RDA Environmental Health Semantics Interest Group (IG). This IG will provide a platform for overall coordination and collaboration among interested members. Its goal will be to design a strategic direction for developing and adopting language solutions, identify and prioritize use cases, coordinate activities, and be a Community of Practice space for exchanging information, offering a resource clearinghouse, and fostering education and training. An RDA Working Group (WG) could be formed whenever a specific work product needs to be developed.

Common framework for ontology development. If the product is an ontology, then ideally its development would follow the OBO Foundry framework to be interoperable with other ontologies.

Community outreach and partnerships. The IG and WG will work in concert with other relevant communities or partner organizations towards the development and implementation of any recommendations and outputs. Those products will be communicated back to the discipline specific communities with the anticipation of adoption.

How would this model work in practice?

The intent of the model is to provide support to those developing and applying language approaches, as outlined in Figure 2.

The example begins with an investigator (or someone else) who has a use case that can benefit from a semantic solution. At this stage, the investigator can work with the RDA Interest Group to raise awareness of the need, tap into expertise, and identify potential collaborators to work on a team. The investigator may choose to form a working group outside of RDA, but they can also decide that creating an RDA Working Group will assist in gaining broader community input and perspectives.

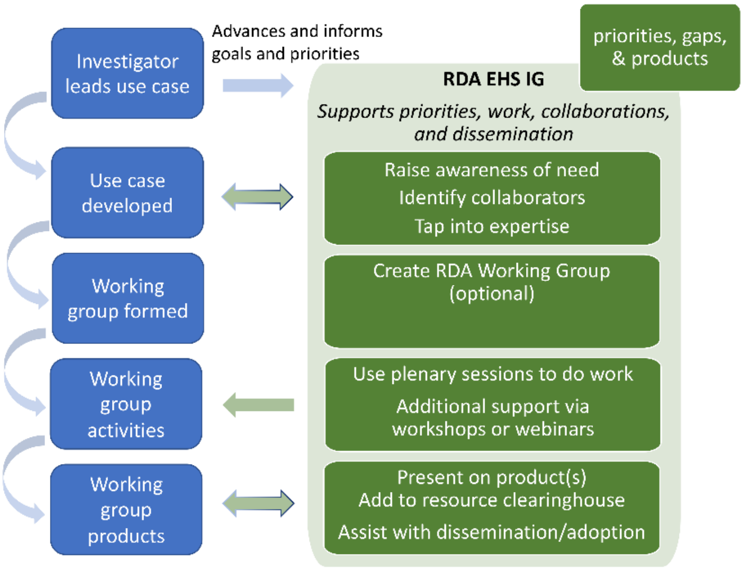

Investigator leads use case
Advances and informs goals and priorities
priorities, gaps, & products
RDA EHS IG
Supports priorities, work, collaborations, and dissemination
Use case developed
Raise awareness of need
Identify collaborators
Tap into expertise
Working group formed
Create RDA Working Group (optional)
Working group activities
Use plenary sessions to do work
Additional support via workshops or webinars
Working group products
Present on product(s)
Add to resource clearinghouse
Assist with dissemination/adoption

Figure 2: Community Model in Practice

Whether the activities are done within or outside an RDA WG, the IG can support the working group's activities by offering time at the IG's plenary sessions to do work and/or providing additional support in the form of workshop activities, presentation time, and webinars.

Any developed product(s) from the working group would be brought to RDA and shared with the broader community, as well as added to a resource clearinghouse. In addition, the RDA IG can assist with disseminating and promoting adoption of the product if needed. Finally, the RDA IG will maintain the catalogue of existing use cases which will aid other investigators in recognizing and prioritizing gaps and issues to which they can provide solutions

How do we sustain the community model?

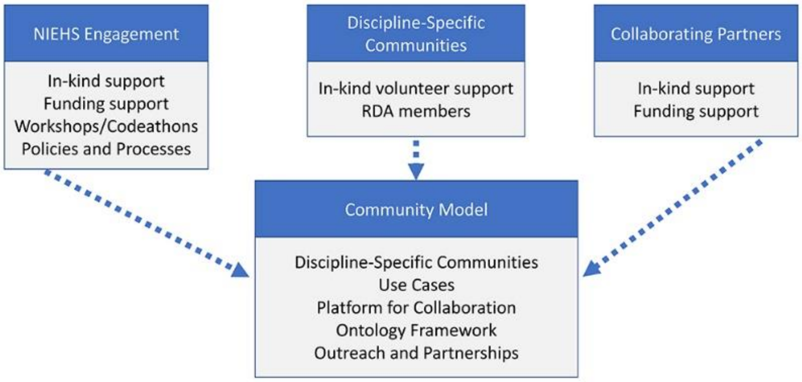


Figure 3: Sustaining the Community

Sustaining the proposed community model requires three supporting players as shown in Figure 3. NIEHS proposes to engage by providing in-kind volunteer support to the IG and WGs and working to promote funding support for relevant efforts. NIEHS will work to establish workshops, or other events such as codeathons, as well as develop policies and processes based on RDA or other's recommendations that would advance the community's goals. In-kind volunteer support will be provided through discipline specific communities, primarily through serving on the IG and WG. Finally, collaborating partners in academic, federal, and industry sectors will be identified and involved to provide both in-kind contributions and support for funding community activities.

Questions to Ponder

On a scale of 1 (low) – 5 (high), do you endorse/agree with the proposed community structure?

What changes do you recommend being made to the proposed structure that would lead to your endorsement?

What additional governance is needed?

What other communities have you participated in that we could use as an alternative model to RDA?

What would be the best technical platform for creating community and fostering collaborations? e.g., listserv, Teams, other?

Developing Sustainable Semantic Solutions

Environmental Health Use Cases

What scientific questions would benefit most from development and adoption of harmonized language approaches?

To kick-start the initiative, a working group of environmental health researchers and staff from across NIEHS developed an initial set of five general use cases, along with sub use case examples. In several cases, the use cases require not only advances in standardized vocabularies, but also in statistical and modeling approaches, which represents opportunities to engage with those communities.

These initial use cases focus on 1) discovery of exposure-specific data sets, 2) integration of exposure data from multiple studies, 3) bridging exposure to biology, 4) identification of biomarkers of exposures, and 5) relating exposures to place-based living and work locations.

Below are the use cases that are being put forward for community development.

For each of these use cases, a small group of subject matter experts drawn from research institutions and federal agencies drafted Use Case Profiles that will provide the basis for discussion at the workshop. The profiles include

- a definition of the use case research question,
- why we are exploring the use case,
- the benefit of developing solutions around the use case,
- the intended output of the use case,
- the workshop goal for the use case and the proposed approach to achieve that goal,
- who should attend discussions on that use case, and
- any background and preparatory materials.

Questions to Ponder

What other use cases/research questions are of interest that you would want to participate in?

What gaps/pain points/challenges would you like to propose be worked on in the Collaborative?

What data/terminology standards and/or tools are you currently using for data query and aggregation?

Where do terminologies need to be harmonized? What terminology gaps exist? Which terminologies should be endorsed for EHS-related use?

Use Case Profiles

What data exists for a given chemical/endpoint/exposure scenario?

Facilitator: Michelle Angrish, EPA

Why we are exploring this use case

Understanding the health effects of environmental exposure requires finding and integrating relevant information. Finding that information can be a challenge because (assuming the information exists) one must 1) know where to look and how to find it, 2) have the resources to collect, screen, and curate the information, and 3) assimilate that information so that it is accessible and usable. Such a workflow is further complicated because study reports are the typical form of information. These reports can be open access, paid for, or confidential business information and generally all are stored in databases with output formats that do not readily map to each other. Therefore, the purpose of this use case is to develop solutions toward identifying, connecting, and making use of environmental health science resources.

Benefit of developing solutions around this use case

Solutions to this use case will enhance the ability to:

1. find existing data
2. understand where there are data gaps
3. develop workflows for screening and curating data
4. increase usability and adoption of existing datasets
5. prospectively consider database interoperability

Intended final output of this use case

We will aim to develop tools and strategies to facilitate interoperability of existing databases.

Workshop goal(s) for this use case

We will aim to identify and define concepts and features that are common across representative environmental health datasets that are needed to achieve resource interoperability

Proposed approach to achieve workshop goal(s)

1. Develop an accurate understanding of a subset of existing environmental health datasets that leverage ontologies or controlled vocabularies, and for each dataset describe relevant data elements and terminology
    a. Example data: HAWC, CDISC-SEND CV as distributed through NCI, ECOTOX knowledgebase
2. Use the reference resources from goal 1 to define “minimal data models” that include the features and metadata to allow for interoperability.”
3. Outline needs and barriers related to aligning existing datasets with the minimal data model.
4. Develop strategies to increase the adoption of “minimal data models” and terminologies within relevant study and data repositories.

Who should attend discussions on this use case

Individuals with expertise in

- Bioinformatics
- Ontology, controlled vocabulary, semantics

- Exposure assessment
- Health hazard assessment
- Data science

Background materials and preparation (PDFs available on TEAMS site; you will need to request access)

- Review above noted example databases
- Familiarity with data schemas
- Familiarity with extracting information into a database
- Familiarity with data cleaning and normalization
- Consideration of workflows that include the elements above

Davis AP, Wiegers J, Weigers TC, and Mattingly CJ (2019). Public data sources to support systems toxicology applications. *Curr Opin Toxicol* 16, 17-24. https://doi.org/10.1016/j.cotox.2019.03.002

Vinken M, Benfenati E, Busquet F, et al. (2021). Safer chemicals using less animals: kick-off of the European ONTOX project. *Toxicology* 458:152846. https://dx.doi.org/10.1016/j.tox.2021.152846

The use of AI in evidence management: EFSA's vision (webinar) - https://www.efsa.europa.eu/en/events/webinar-use-ai-evidence-management-efsas-vision

Data and tools needed to harmonize place-based health research

Facilitator: Carmen Marsit, Emory

Why we are exploring this use case

Place-based research has been used extensively in Environmental Health to examine exposures such as air pollution, soil and water contaminants, industrial facilities, and radiation exposures and is a major contributor to climate-change related research. In addition, data sources beyond those examining chemical or physical exposures, including neighborhood and built environment characteristics and historical information present opportunities to integrate structural and societal factors that underlie exposures and characterize environmental injustice. Opportunities to integrate data from multiple place-based exposures and data from different studies across varied geographic locations is important to further understand how place influences health, and having a unified language to describe the data can improve its interpretation and utility. As a starting point, in this use case discussion, we will focus specifically on data harmonization and the initial development of a shared language to achieve that purpose.

Benefit of developing solutions around this use case

Creation of harmonized language for use in place-based environmental health research will:

1. Increase opportunities for data harmonization across studies
2. Improve rigor and reproducibility of place-based research
3. Increase usability and adoption of existing datasets
4. Allow for increased geographic variation in studies by combining datasets; thereby improving risk estimates and generalizability of findings
5. Improve communication within and beyond the EH community

Intended final output of this use case

To develop tools and strategies for a shared vocabulary and semantic ontologies that could improve the rigor and interoperability of place-based research to increase the impact of the research to improve public health and inform prevention and policy efforts.

Workshop goal(s) for this use case

To develop a model set of minimum information and tools needed to harmonize place-based health research.

Proposed approach to achieve workshop goal(s)

Place-based health research is extremely broad and diverse. In this use case, we will specifically discuss a sub-use case of air pollution exposure and asthma risk. This area of research was chosen because it is a relatively mature area of research and incorporates a wide variety of approaches and data sources.

1. We will review example studies and consider the question, *"How feasible would it be to link data sources to address these kinds of air pollution – asthma questions (e.g., exposure and health linkages needed within each study), and what would that achieve?"*
2. Through discussion, we will identify what information is needed about the studies and what procedures are necessary to harmonize the datasets to harmonize the exposure, health, and other data needed to address the study questions.

3. We will use this information to define a minimum data model that would provide the necessary elements for effective harmonization of place-based data
4. We will identify needs and barriers related to aligning and harmonizing existing datasets
5. Develop strategies to promote the adoption of minimum data models and ontologies to assure interoperability of datasets

Who should attend discussions on this use case

Individuals with expertise in

- Geographic-based tools and analyses
- Bioinformatics and ontology
- Environmental epidemiology
- Exposure assessment
- Climate change research
- Structural determinants of health and disease
- Biostatistics and Data Science

Background materials and preparation

Readings (PDFs available on TEAMS site; you will need to request access)

Hua J, Yin Y, Peng L, et al. (2014). Acute effects of black carbon and PM2.5 on children asthma admissions: a time-series study in a Chinese city. *Sci Total Environ* 481:433-8. https://doi.org/10.1016/j.scitotenv.2014.02.070

Mölter A, Simpson A, Berdel D, et al. (2015). A multicentre study of air pollution exposure and childhood asthma prevalence: the ESCAPE project. *Eur Respir J* 45(3):610-24. https://doi.org/10.1183/09031936.00083614

Rosenquist NA, Metcalf WJ, Ryu SY, et al (2020). Acute associations between PM2.5 and ozone concentrations and asthma exacerbations among patients with and without allergic comorbidities. *J Expo Sci Environ Epidemiol* 30, 795-804. https://doi.org/10.1038/s41370-020-0213-7

Combine individual-level data from multiple independent studies (heterogeneous study designs and data collection protocols) to understand (with increased statistical power) how exposures X + Y impact health outcome Z

Facilitator: Jeanette Stingone, Columbia

Why we are exploring this use case

Integration of data and knowledge across multiple studies, sources and platforms can facilitate the acceleration of scientific discovery and ignite the generation of new knowledge to improve health. Tools, such as ontologies and knowledge graphs, can facilitate the enhanced use and application of scientific data across studies in collaborative efforts. These tools enable the systematic representation of data, metadata, and exposure-disease relationships generated by scientific studies, promoting our ability to leverage algorithms and other machine-based analytics for subsequent data-driven discovery. They also facilitate harmonization across studies and platforms, ensuring that pooling of research data across studies is both accurate and appropriate. Yet, these tools for systematic knowledge representation have often been underutilized in the environmental health sciences. Therefore, the purpose of this use case is to address **the feasibility of using harmonized language for combining data across independent research studies.** While combining data across studies can take many forms, we choose to focus on combining individual-level data across multiple independent studies to pool data and apply analytic techniques to understand how exposures impact health outcomes.

Benefit of developing solutions around this use case

Creation of a harmonized language and reporting structure for harmonization of study data will:

- Promote broader usability and adoption of existing datasets across the environmental health community
- Develop infrastructure to guide future data collection to promote harmonization and integration
- Increase linkages and interoperability between datasets across disparate studies and research initiatives
- Enable the use of machine-learning and artificial intelligence-enabled technologies to analyze existing data and generate new knowledge

Intended final output of this use case

We will aim to develop tools and strategies to facilitate data sharing and harmonization through use of data and metadata standards and annotation of existing datasets.

Workshop goal(s) for this use case

We will identify gaps in metadata reporting and knowledge representation that hinder the harmonization of data across studies and prevent the use of semantic and other technologies. This will be accomplished through the mock harmonization of 2-3 existing environmental health studies that cover demographic, biomarker-based and external exposure data.

Proposed approach to achieve workshop goal(s)

Conduct mock data harmonization of 2-3 existing environmental health studies. Some features within datasets may be easily harmonizable while others may lack needed information. Identifying these gaps will help us target solutions in future work. The mock harmonization process will include the following tasks:

1. Review existing metadata templates including data dictionary templates to assess completeness of what is currently reported by existing datasets including both variable and study information.
2. Identify minimum amount of metadata needed to harmonize data with focus on descriptors of study design and potential differences across demographic, biomarker-based and external environmental exposure features.
3. Discuss sources of data standards that could be used within metadata templates to facilitate harmonization, sharing and application of semantic-enabled technologies such as ontologies
4. Develop strategies to increase the adoption of metadata and data standards when describing existing datasets to facilitate harmonization and sharing.

Who should attend discussions on this use case

- Individuals working to harmonize data across disparate human health studies including epidemiologists and environmental health scientists within consortia, research collectives, data centers or other collaborative science efforts
- Individuals developing data and metadata standards to represent common terms in environmental health and epidemiology studies including ontologists and other semantic scientists
- Individuals designing and/or implementing data management systems to harmonize human health and exposure data across disparate scientific studies including computer scientists and semantic scientists
- Individuals working to prepare existing environmental health data to be used by the broader scientific community, including activities related to preparation for the application of ML/AI-enabled technologies, systematic organization of study data and improvement to data and metadata reporting

Background materials and preparation

Review Study Summaries in TEAMS folder (to be posted Sept. 1; you will need to request access)

What are the biological processes and biomarkers associated with exposure and how do they relate to the potential for an adverse outcome associated with a given exposure?

Facilitators: Stephen Edwards, RTI and Chirag Patel, Harvard

Why we are exploring this use case

An exposure event occurs when a stressor from the environment interacts with a biological receptor (i.e., human, animal, etc.). Because of the transient nature of an exposure event, they must either be predicted based on modeling of different exposure pathways leading to the exposure or estimated based on the changes in the biological receptor following the exposure. The latter consist of both the presence of the stressor and its metabolites as well as the biological changes that result from the exposure event. The biological changes occur when the exposure event alters the organism's normal biological processes. *Processes* are defined as functions that are vital for a human to live, or, "operations or sets of molecular events with a defined beginning or end, pertinent to the functioning of lying units along a hierarchy, such as cells, tissues, organs, and the human" (according to *Gene Ontology*). *Biomarkers*, on the other hand, are measurable phenotypes that are indicative of biological processes or direct measurements of an environmental stressor or its metabolites.

The adverse outcome pathway (AOP) framework connects the perturbation of biological processes by an environmental stressor to adverse outcomes. By definition, however, the AOP framework excludes the environmental stressors that cause the perturbation of biological processes to initiate the sequence of events described by the AOP. As a result, it is impossible to connect databases containing information regarding the potential for chemical exposure and databases containing measurements of chemicals and metabolites with potential downstream effects. Ontologies describing exposure events exist and could easily be extended to capture the information needed to make those connections. Therefore, the purpose of our discussion is to address **how a semantic description of exposure events that incorporates the associated biomarkers and biological processes would support the integration of existing data resources to connect measured biomarkers to exposure-response relationships, using pre-documented sub use-cases.**

Benefit of developing solutions around this use case

This use case focuses on the ultimate application of the technology solutions developed under the other use cases.

Intended final output of this use case discussion

The specific deliverables for this session will be:

- Clear definition of the use case and refinement of sub-use-cases (e.g., see below)
- Existing data resources that contain exposures, biomarkers, and/or biological processes.
- Identification of proof-of-concept scientific examples.
- Semantic gaps that need to be resolved and current challenges in computationally querying across exposures, biomarkers, and processes.

Workshop goal(s) for this use case

To address how a semantic description of exposure events that incorporates the associated biomarkers and biological processes would support the integration of existing data resources to connect measured biomarkers to exposure-response relationships, using pre-documented sub use-cases.

Proposed approach to achieve workshop goal(s)

We will consider the exposure event from two perspectives. One breakout group will consider the interpretation of the biomarkers that are indicative of the exposure event, and a second breakout group will consider how AOPs can be used to connect the measured biomarkers with adverse outcomes that could result from an exposure event. Following the breakout discussions, we will reconvene to define a unified use case that allows both perspectives to be addressed via a common topic.

Sub-use cases

1. **What biomarkers are directly indicative of exposure to a given chemical?** Biomarkers can include direct measurement of the chemical or its metabolites and can be identified associatively or experimentally through epidemiological or experimental approaches, respectively.
2. **What are the exposures that are associated with the observed biomarkers in an epidemiological study?** One may observationally or experimentally find biomarkers associated with health and disease – what are potential exposures that may also induce changes in the biomarkers?
3. **Map signatures of 'omic changes to chemical exposure:** Query for organ-specific signatures of 'omic biomarkers, across the metabolome or the transcriptome, that are indirectly or directly associated with exposure.
4. **What biological processes are linked to biomarkers that are indicative of the exposure?** If an exposure is causal for a change in state, their biomarkers must also be directly or indirectly associated with biological processes. Given biomarkers that are indicative of exposure to a chemical or class of mechanistically related chemicals, query for all biological processes that are associated with changes in the biomarker(s).

Who should attend discussions on this use case

- Individuals responsible for implementation of other use cases will provide valuable input into the definition of this use case and can utilize their understanding of this practical application of the tools to inform the implementation of those use cases.
- Individuals interested in developing a practical application that utilizes a harmonized environmental health science vocabulary can participate in post-workshop activities focused on this use case.

Background materials and preparation

Thessen, Anne E., Cynthia J. Grondin, Resham D. Kulkarni, Susanne Brander, Lisa Truong, Nicole A. Vasilevsky, Tiffany J. Callahan, et al. 2020. "Community Approaches for Integrating Environmental Exposures into Human Models of Disease." *Environmental Health Perspectives* 128 (12): 125002. https://doi.org/10.1289/EHP7215.

Watford, Sean, Stephen Edwards, Michelle Angrish, Richard S. Judson, and Katie Paul Friedman. 2019. "Progress in Data Interoperability to Support Computational Toxicology and Chemical Safety Evaluation." *Toxicology and Applied Pharmacology* 380 (October): UNSP 114707. https://doi.org/10.1016/j.taap.2019.114707.

Boyles, R.R., A.E. Thessen, A. Waldrop, and M.A. Haendel. 2019. "Ontology-Based Data Integration for Advancing Toxicological Knowledge." *Current Opinion in Toxicology* 16 (August): 67–74. https://doi.org/10.1016/j.cotox.2019.05.005.

Environmental Health Language Collaborative

Harmonizing data. Connecting knowledge. Improving health.

Semantically Speaking

Glossary

This glossary is being put forward as a starting point for the EHS language community. Having an agreed upon meaning for a term reduces problems in understanding. If you have suggestions for new terms, revisions to the proposed descriptions, etc. please send them to Stephanie (holmgren@niehs.nih.gov). The glossary will eventually be posted to the Ontology Resource webpage.

Term	Description	Examples
Annotation	An explanatory or critical comment, or other in-context information (e.g., pattern, motif, link), that has been associated with data or other types of information. [Source: NCIt_C44272]	A GO annotation is a statement about the function of a particular gene. Annotations associate a gene/gene product with a GO term. Source: Introduction to GO annotations
Common data element (CDE)	A piece of data common to multiple data sets across different studies (may be universal or domain-specific). Development and use of CDEs supports standardization of terms and facilitates data sharing so that data can be compared and combined across studies. [Source: Glossary, NIH Strategic Plan for Data Science]	NIH Common Date Elements Repository offers access to CDEs recommended or required by NIH Institutes and others.
Controlled vocabulary	A controlled vocabulary, also called an authority file or term list, is an authoritative set of terms selected and defined based on the requirements set out by the user group. used to ensure consistent indexing (human or automated) or description of data or information. Controlled vocabularies do not necessarily have any structure or relationships between terms within the list. [Source: NCIt C48697 and About Taxonomies & Controlled Vocabularies]	Some definitions of controlled vocabulary are more expansive and include taxonomy, thesaurus, ontology, and other systems as types of controlled vocabulary. For our purposes, it is being considered only as a term list typically encountered as drop-down pick list, index list of terms, etc.
Data curation	A managed process, throughout the data lifecycle, by which data & data collections are cleansed, documented, standardized, formatted and inter-related. Such processes ensure the value of the data is preserved over time and available for discovery and reuse. A second meaning of the phrase is used in the context of extracting information from research articles and storing that information in a database. [Source: CASRAI and Wikipedia]	

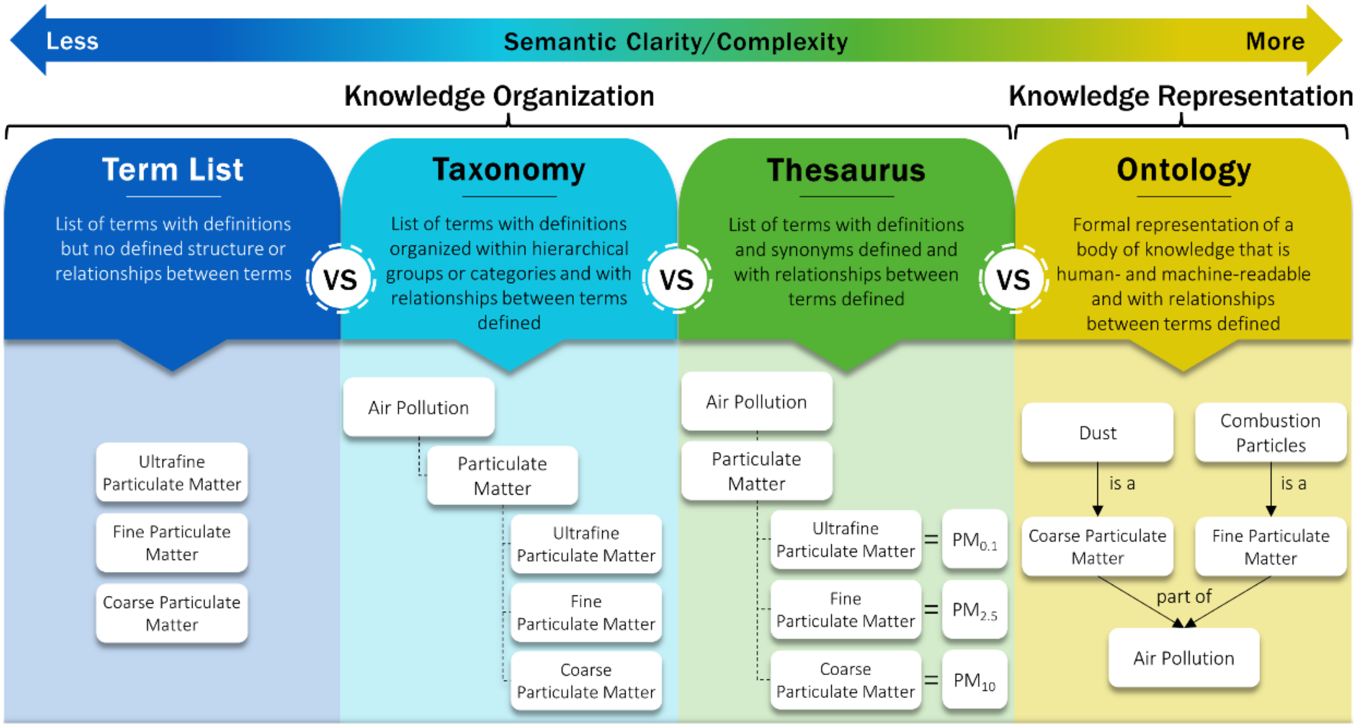
<table>
<tr><th>Term</th><th>Description</th><th>Examples</th></tr>
<tr><td>Data dictionary</td><td>A “super catalog” that provides for each data field or element, a list of information describing the field, where the data originates, edits or rules that apply to that field, type and width of field, description of codes used (if any), what applications or reports use that data element, etc. Use of a data dictionary minimizes inconsistencies related to collection and use of data by members within a team as well as across projects. The resulting consistency makes data easier to analyze.
[What is a data dictionary and Why use a data dictionary?]</td><td><table>
<tr><th>Variable Name</th><th>Data Type</th><th>Data Format</th><th>Field Size</th></tr>
<tr><td>Birthdate</td><td>Integer</td><td>DD/MM /YY</td><td>8</td></tr>
<tr><td>Last Name</td><td>Text</td><td></td><td>Unlimited</td></tr>
<tr><td>Symptoms</td><td>Text</td><td></td><td>unlimited</td></tr>
</table>Additional items include description, required values, among others.</td></tr>
<tr><td>Data elements</td><td>Information that describes a piece of data to be collected in a study. The description includes a data element name, definition, permissible values, and other attributes
[Source: CDE Glossary and NCIT_C41002]</td><td></td></tr>
<tr><td>Data harmonization</td><td>Data harmonization is an extension of data integration. The harmonization process combines data from different sources and reorganizes it according to a single schema to provide users with a comparable view of data from different studies. Data is combined by either identifying equivalent data elements between the sources or by developing unequivocable transformations between the elements, to create a view of the unified data. In some cases, transformations can lead to loss of information or subtle changes in meaning within the unified view.
[Adapted from ICPSR]</td><td>Lear more about HHEAR’s data harmonization, NCI’s Quest for Harmonized Data and role of data harmonization in a molecularly driven health system</td></tr>
<tr><td>Data integration</td><td>The practice of consolidating data from disparate sources into a single dataset with the goal to provide a unified, single view of the data. (Source: Omnisci)</td><td>Repositories integrate data by bringing disparate sources and collating them in a single database to improve findability</td></tr>
<tr><td>Harmonized language</td><td>A harmonized language combines multiple languages into a single comparable view building from the components of each language
[Source: Modified from ICPSR]</td><td></td></tr>
</table>

Term	Description	Examples
Interoperability	Interoperability refers to the ability of two or more systems or components to exchange information and to use the information that has been exchanged. There are four types of issues that may impede interoperability: system-level (incompatibilities between hardware and operating systems), syntactic (differences in encodings and representation), structural (variance in data models, data structures, and schema), and semantic (inconsistencies in terminology and meanings). [Source: ISKO] Semantic interoperability is a requirement to enable machine computable logic, inferencing, knowledge discovery, and data federation between information systems. Semantic interoperability is achieved when the information transferred has, in its communicated form, all of the meaning required for the receiving system to interpret it correctly [Source: CASRAI]	
Knowledge base	In general, a knowledge base is a database that holds statements about our knowledge in a particular domain instead of actual data points. More specifically, biomedical knowledgebases have the primary function to extract, accumulate, organize, annotate, and link growing bodies of information related to core datasets, in compliance with the FAIR Data Principles. [Source: NIH ODSS]	Database: Organism X was observed at lat/lon on datetime Knowledgebase: Species X lives in cypress swamps Comparative Toxicogenomics Database (CTD)
Knowledge graph	A method for representing knowledge as entities (nodes) and the relationship between them (edges) in a way that enables large-scale computing and takes advantage of our knowledge of those relationships. [Source: based on https://towardsdatascience.com/an-introduction-to-knowledge-graphs-841bbc0e796e]	See graphic
Knowledge organization	A term applied to all types of schemes (controlled vocabulary, taxonomy, etc.) used to	

Term	Description	Examples
	organize, represent, and manage a set of information. [Source: https://www.isko.org/cyclo/kos]	
Knowledge representation	The field of computer science devoted to representing information about the world in a form that a computer system can utilize to solve complex tasks.	
Metadata	Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage data. Metadata is often called data about data or information about information. It ensures that the context for how your data was created, analyzed and stored, is clear, detailed and therefore, more usable and reusable in for the future. Metadata can be descriptive, administrative, or technical in nature. [Source: Adapted from NISO]	Descriptive: title, author, study date Technical: file type, file size, creation date Administrative: license terms, checksum,
Minimum information standards	A specification of a minimum amount of information needed to reproduce or fully interpret a scientific result. The standard is typically composed of two parts: a table or checklist of reporting requirements and a data format. [Source: Ontobee and Wikipedia]	Numerous research methods use minimum information standards; e.g., MIATE (in vivo animal toxicology), MIAME (gene expression), MIBBI (biological and biomedical investigations. Find more at FAIRsharing.org.
Ontology	A formal representation of a body of knowledge within a given domain. Ontologies usually consist of a set of classes (or terms or concepts) with relations that operate between them. Ontologies are used to provide the underlying semantic structure for knowledge graphs to ensure shared meaning and understanding of the data both by humans and machine. [Source: About Taxonomies & Controlled Vocabularies and Ontotext]	Human Health Exposure and Analysis Resource (HHEAR) Ontology, AOP Ontology, and others can be found by searching the following ontology portals: BioPortal, OBO Foundry, OntoBee, and Ontology Lookup Service
Metadata standard	A standard that specifies what types of metadata should be collected and how for any given datum, what format the metadata should be in, what units and terms should be used, and the file format the metadata should be in.	Cancer Data Standards Registry (caDSR), Crystallographic Information Framework

Term	Description	Examples
Semantics	The meaning of a string (e.g., words, phrases, sentences) in a language; of or relating to the study of meaning and changes of meaning. [Source: NCIt C54194]	
Syntax	The rules (word order, punctuation, sentence structure, etc.) for writing a language. As applied in computer science, it refers to the structure needed for a computer to read and understand the coded instructions or information to perform a task. [Source: Wikipedia]	Programming (Java, Python, …), mark-up (HMTL, JSON, …), and knowledge representation (OWL, RDF, …) languages each have their own syntax for coding.
Taxonomy	A taxonomy (or taxonomical classification) is a scheme of classification with a tree-based hierarchical structure showing the relationships (parent/child or broader/narrow) of terms with each other within the taxonomy. Taxonomies typically lack the more complex relationships found in thesauri or ontologies. [Source: About Taxonomies & Controlled Vocabularies]	Integrated Taxonomic Information System is based on the Linnaean taxonomy for classification of organisms. Other biomedical examples include the International Classification of Disease and NCBI Taxonomy
Thesaurus	A thesaurus is an extension of a taxonomy. At its base is a standard hierarchical structure showing broader/narrower term relationships. In addition, a thesaurus also shows associative (*see also*), and equivalent (*use/used from* or *see/seen from*) term relationships. It is common in thesauri that some or all terms have scope notes, which are brief explanations of how the term should be used. [Source: About Taxonomies & Controlled Vocabularies]	NCBI's Medical Subject Headings (MeSH) and NCI Thesaurus

The Classification Continuum



Term List

List of terms with definitions but no defined structure or relationships between terms

Taxonomy

List of terms with definitions organized within hierarchical groups or categories and with relationships between terms defined

Thesaurus

List of terms with definitions and synonyms defined and with relationships between terms defined

Ontology

Formal representation of a body of knowledge that is human- and machine-readable and with relationships between terms defined

Why Use It?

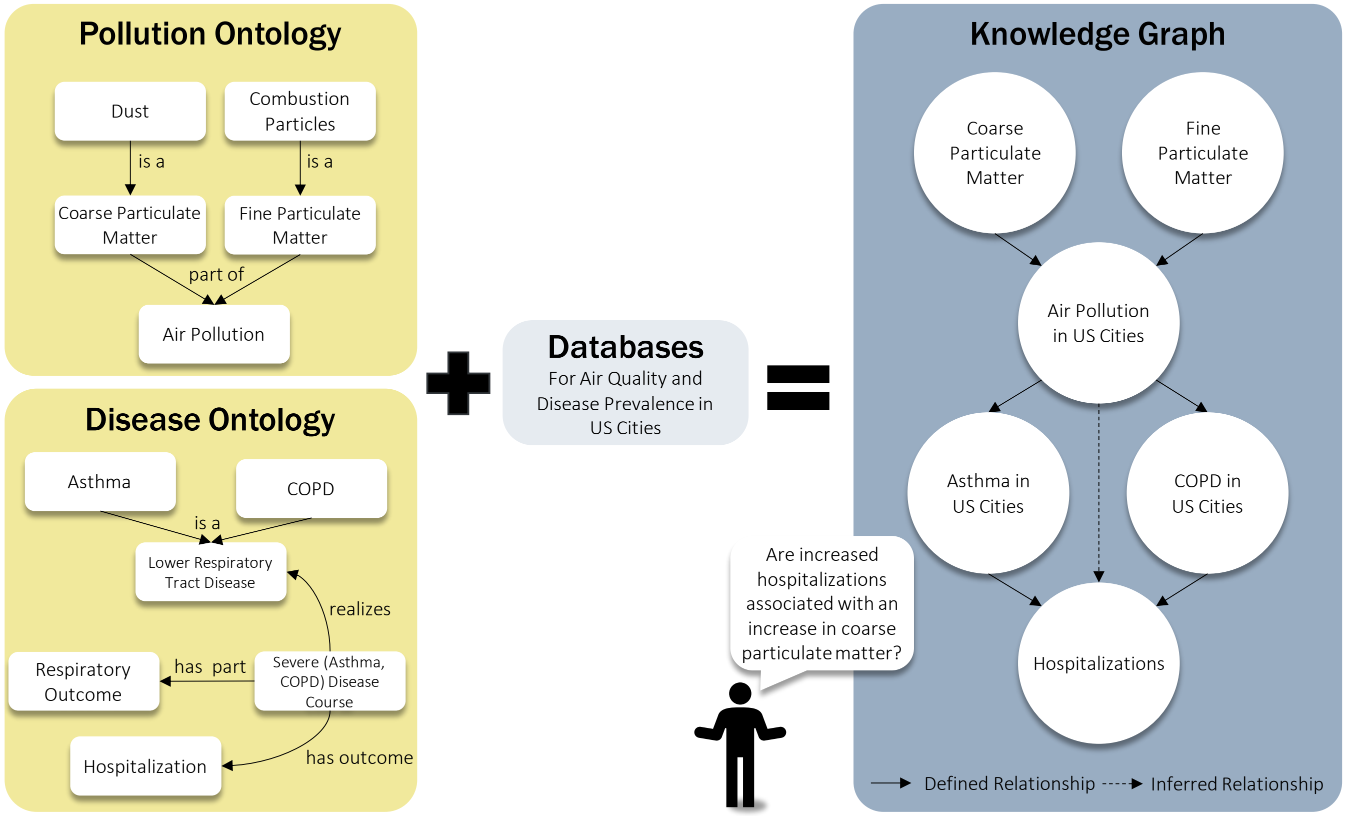
Consistent	Hierarchical	Associative	Inferential
Collaboration and search are easier when everyone is using a set of terms with an agreed upon definition.	The hierarchical structure helps us integrate data collected at different levels of granularity and search for data using high level categories.	In addition to the hierarchy, the presence of synonyms further improves integration and search over heterogeneous data.	The combination of formal logic and persistent, unique identifiers enables inferencing, makes data computable, and reveals novel connections.

Examples

HAWC EHV; LTER Controlled Vocabulary	NCBITaxon	NCIThesaurus; MeSH	OBO Foundry Ontologies; UBERON

Knowledge Representation

The knowledge graph is built using ontologies AND data. Building a knowledge graph involves combining the relationships described in ontologies and in the data. By integrating heterogeneous data and applying a formal, machine-readable representation of the data, users can query the knowledge graph to answer more complex questions; e.g., Are increased hospitalizations associated with an increase in coarse particulate matter?



Frequently Asked Questions

When should I use an ontology instead of a controlled vocabulary?

The answer to this question will depend on your research question and goals. Controlled vocabularies are easy to create, maintain, update, and understand. Ontologies are more computationally and semantically complex and require more effort to maintain and update. If your goal is to annotate and make heterogenous data consistent with each other and retrievable, a controlled vocabulary might be sufficient for your needs. However, if your goal is to infer new knowledge or postulate new relationships across heterogenous data, then an ontology might be warranted.

How do I know which ontology to use?

Picking an ontology is like picking any other scientific instrument or method. The exact ontology that will be best for your research question will depend on your specific use case. However, consider the following elements when choosing an ontology:

- Coverage of the topic/domain of interest
- If the ontology is actively maintained and updated
- If the ontology is likely to be sustained over time
- If the ontology is open source and interoperable
- If the ontology is widely adopted and used

What do I use a knowledge graph for?

While an ontology represents information that is ALWAYS true (i.e., every instance of a femur is part of some leg), a knowledge graph can include information about specific instances of a class (i.e., Patient X has Y Gene and a shortened femur). While ontologies are more authoritative, they take longer to build and are more difficult to maintain. A knowledge graph can be created much faster and for a specific research purpose. There are far more tools and services available to build and query knowledge graphs. Knowledge graphs can be used to represent a combination of an ontology (a generalized data model with defined classes and relationships) and a specific dataset (that aligns with the ontology).

[Learn more]

Resources

Recommended Reading

Boyles RR, Thessen AE, and Haendel MA. (2019). **Ontology-based data integration for advancing toxicological knowledge**. *Current Opinion in Toxicology*. 16: 67-74. https://doi.org/10.1016/j.cotox.2019.05.005

Holmgren SD, Boyles RR, Cronk RD, et al. (2021). **Catalyzing knowledge-driven discovery in environmental health sciences through a community-driven harmonized language**. *International Journal of Environmental Research and Public Health* 18(17), 8985. https://doi.org/10.3390/ijerph18178985

Mattingly CJ, Boyles R, Lawler CP, et al. (2016). **Laying a community-based foundation for data-driven semantic standards in environmental health sciences**. *Environmental Health Perspectives*. 124: 1136-1140. https://ehp.niehs.nih.gov/doi/10.1289/ehp.1510438

Thessen AE, Grondin CJ, Kulkarni RD, et al. (2020). **Community approaches for integrating environmental exposures into human models of disease**. *Environmental Health Perspectives*. 128(12): 125002. https://ehp.niehs.nih.gov/doi/10.1289/EHP7215

Whaley P, Edwards SW, Kraft A, et al. (2020). **Knowledge organization systems for systematic chemical assessments**. *Environmental Health Perspectives*.128(12): 125001. https://ehp.niehs.nih.gov/doi/10.1289/EHP6994

Recommended Viewing

What can ontologies do for you? Perspectives from an environmental epidemiologist (21:52 min.)
https://www.youtube.com/watch?v=eVQAKPZz3Jo

How to use ontologies for Superfund Research Center data when you've never used an ontology before (33:18 min.)
https://www.youtube.com/watch?v=siMVfWcb-XI

Environmental Health Collaborative Web Pages

Collaborative Email Listserv
https://list.nih.gov/cgi-bin/wa.exe?SUBED1=EHSCOMMONLANGUAGE&A=1
Sign up for our email distribution list and join the community of researchers, ontologists, informaticists, and engineers working together on environmental health common language standards

Environmental Health Language Collaborative
https://www.niehs.nih.gov/research/programs/ehlc/index.cfm

> Proposed Collaborative vision, mission, and goals
> https://www.niehs.nih.gov/research/programs/ehlc/purpose/index.cfm
>
> Proposed Collaborative community model
> https://www.niehs.nih.gov/research/programs/ehlc/model/index.cfm

Workshop Web Page: *Catalyzing Knowledge-Driven Discovery in Environmental Health Sciences Through a Harmonized Language*
https://tools.niehs.nih.gov/conference/ehslanguage/

Additional Resources

Ontology Resource Toolbox

https://www.niehs.nih.gov/research/programs/ehlc/resources/index.cfm

Check out a compilation of organizations, recommended readings, ontologies/terminologies, and tools useful to harmonizing environmental health research.