



# ENVIRONMENTAL HEALTH PERSPECTIVES

<http://www.ehponline.org>

## Laying a Community-Based Foundation for Data-Driven Semantic Standards in Environmental Health Sciences

Carolyn J. Mattingly, Rebecca Boyles, Cindy P. Lawler, Astrid C. Haugen, Allen Dearry, and Melissa Haendel

<http://dx.doi.org/10.1289/ehp.1510438>

Received: 7 July 2015

Accepted: 3 February 2016

Advance Publication: 12 February 2016

**Note to readers with disabilities:** *EHP* will provide a [508-conformant](#) version of this article upon final publication. If you require a 508-conformant version before then, please contact [ehp508@niehs.nih.gov](mailto:ehp508@niehs.nih.gov). Our staff will work with you to assess and meet your accessibility needs within 3 working days.



## **Laying a Community-Based Foundation for Data-Driven Semantic Standards in Environmental Health Sciences**

Carolyn J. Mattingly<sup>1</sup>, Rebecca Boyles<sup>2</sup>, Cindy P. Lawler<sup>2</sup>, Astrid C. Haugen<sup>2</sup>, Allen Dearry<sup>2</sup>, and Melissa Haendel<sup>3</sup>

<sup>1</sup>Department of Biological Sciences and the Center for Human Health and the Environment, North Carolina State University, Raleigh, North Carolina, USA; <sup>2</sup>National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA; <sup>3</sup>Library and Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA

**Address correspondence to** Carolyn Mattingly, Department of Biological Sciences, North Carolina State University, Campus Box 7633, Raleigh, NC 27695-7617 USA. Telephone: (919) 515-1509. E-mail: [cjmattin@ncsu.edu](mailto:cjmattin@ncsu.edu)

**Running title:** A phased data- and community-driven framework

**Acknowledgments:** The authors, who comprised the core planning committee and user community, thank NC State's James B. Hunt Library for Webcast support; the Institute for Emerging Issues for the venue and associated support; Jennifer Solomon, Beth Anderson, Jennifer Collins, Kerri Moran, Whitney Freberg, and Lesley Skalla (NIEHS SaRPS contract) for meeting support, such as logistics, AV support, note-taking, editing, and travel reimbursements to workshop participants; Julie McMurry for figure content; and the Environmental Health Science Language Workshop Working Group members-- Yuxia Cui, Stephanie Holmgren, Lisa Chadwick, and Kimberly Thigpen-Tart for help planning the workshop. Finally, the authors

would like to acknowledge the dedication and enthusiasm of the attendees who collectively helped to clarify needs, stimulated discussion, and expressed their commitment to participating in future efforts to develop environmental health semantic standards. This work was supported by NIH, National Institute of Environmental Health Sciences (NIEHS) and the Office of the Associate Director for Data Science.

**Competing financial interests:** The authors declare they have no actual or potential competing financial interests.

## **Abstract**

**Background:** Despite increasing availability of environmental health science (EHS) data, development and implementation of relevant semantic standards, such as ontologies or hierarchical vocabularies, has lagged. Consequently, integration and analysis of information needed to better model environmental influences on human health remains a significant challenge.

**Objectives:** Identify a committed community and mechanisms needed to develop EHS semantic standards that will advance understanding about the impacts of environmental exposures on human disease.

**Methods:** The National Institute of Environmental Health Sciences (NIEHS) sponsored a *Workshop for the Development of a Framework for Environmental Health Science Language* hosted at North Carolina State University on September 15-16, 2014. Through the assembly of data generators, users, publishers and funders, we aimed to develop a foundation for enabling community-based and data driven standards development that will ultimately improve standardization, sharing, and interoperability of EHS information.

**Discussion:** Creating and maintaining an EHS common language is a continuous and iterative process, requiring community building around research interests and needs, enabling integration and reuse of existing data, and providing a low barrier of access for researchers needing to use or extend such a resource.

**Conclusions:** Recommendations included developing a community-supported Web-based toolkit that would enable: 1) collaborative development of EHS research questions and use cases; 2) construction of user-friendly tools for searching and extending existing semantic resources; 3) education and guidance about standards and their implementation; and 4) creation of a plan for governance and sustainability.

## Introduction

This review is derived from a workshop held at North Carolina State University, Raleigh, North Carolina, USA, on September 15-16 2014. Sharing, analysis and integration of environmental health science (EHS) data is limited by a lack of data standards, in particular, common language standards. Language standards are shared vocabularies that are used for data annotation and common data elements specification to aid interoperability. They may be as complex as an ontology, whereby the terms and the relations between them are defined using logic and are expressed in computable languages such as the Web Ontology Language (OWL)(OWL 2016), or, they may be as simple as a hierarchical vocabulary. This workshop aimed to: a) articulate research areas that would be advanced by EHS language standards and data interoperability; b) identify a community to initiate the creation and champion the extension of EHS language standards; and c) develop guidelines for development of EHS standards.

Exposure to environmental factors significantly impacts human health. The environment, broadly defined, can range from everyday products (e.g., toothpaste) to hazardous materials (e.g., open pit mining sites) and socioeconomic stressors. Consideration of this spectrum is needed to better understand *how*, *when*, and *to whom* exposures pose health risks. There is an enormity of available data that, if structured and integrated, could be leveraged to inform mechanistic hypotheses, therapeutic approaches, and policy-making. However, a lack of semantic standards has been a major barrier to data sharing and integration (van Panhuis et al. 2014). This need for semantic standards is being recognized in many areas of biomedical research. For example, the National Research Council's *Toward Precision Medicine* report called for clinical and research advancements based upon systems that would be enabled by a new language standard (National Research Council 2011). The authors of this report (Committee on A Framework for Developing

a New Taxonomy of Disease, Board on Life Sciences, and Division on Earth and Life Studies) determined that, “The rise of data-intensive biology, advances in information technology, and changes in the way health care is delivered have created a compelling opportunity to improve the diagnosis and treatment of disease by developing a Knowledge Network, and associated New Taxonomy, that would integrate biological, patient, and outcomes data on a scale hitherto beyond our reach” (National Research Council 2011).

Development of semantic standards, such as logically constructed ontologies, EHS data and integration of this effort within the broader biomedical context through crosscutting research programs, such as the Exposome (Wild 2005) and Big Data to Knowledge (BD2K) (Margolis et al. 2014), will enhance the capacity to inform disease research with environmental data while also improving understanding of environmental impacts on human disease. The lack of language standards and their consistent implementation affects not only the capacity to analyze across diverse data sets, but even hinders the ability to identify available data sets, limiting the value of potentially important scientific findings. A query of microbiome samples using PubMed from the National Center for Biotechnology Information (NCBI) (NCBI 2016) illustrates the variability in results that stem from a lack of harmonized language standards and annotation of data using such standards (Table I). Standardization has the potential to benefit many areas of biomedical science by augmenting discovery and reuse (Richesson and Nadkarni 2011; Tenopir et al. 2015; Zimmerman 2008).

A few projects have specifically demonstrated the potential of adopting standards to advance EHS data integration, research, and discovery. For example, the Oceans and Human Health program (supported by NIEHS and the National Science Foundation) links oceanographic and metagenomics data sets (NCBI’s Sequence Read Archive, Metagenomic Rapid Annotations

using Subsystems Technology)(NCBI-SRA 2015; Youngblood et al. 2014), and custom public health databases (Antibiotic Resistance Database, Computer Access to Research on Dietary Supplements Database)(ARDB 2015; CARDS 2015) using ontologies to provide an innovative, health-based framing for oceanographic observatories (microbial diversity and antibiotic resistance of ocean ecosystems)(Port et al. 2012; Port et al. 2014). The Comparative Toxicogenomics Database (CTD)(Davis et al. 2014) provides integrated information about chemicals, genes/proteins, phenotypes, diseases and exposures to provide mechanistic insights into the effects of the environment on human health (Davis et al. 2014). Data are annotated and integrated using public ontologies describing chemicals (MeSH)(MeSH 2015), genes and proteins (Entrez Gene)(Entrez-Gene 2015), diseases (MeSH), and interactions (CTD interaction ontology)(Davis et al. 2014). Consequently, users may query cross-species mechanistic data for specific or broad classes of chemicals and identify associated diseases or disease models. Broader development and adoption of EHS standards will be necessary to ensure access, reuse, innovative integration, and ongoing reanalysis of data that describe the complex interactions between the environment and human health.

## **Discussion**

### Gaps in EHS semantic standards

The data standardization needs within EHS are diverse and include genomics, metabolomics, chemistry, toxicology, epidemiology, exposure science, phenotypes, geospatial data, and clinical health records among others. While some of these components are better standardized than others (e.g., genomics) and not necessarily specific to EHS, it is the need for integration *across* these diverse entities in order to better model the complexity of environmental health interactions that is unique. In apparent contradiction there are a large number of existing

standards (Tenenbaum et al. 2014), yet often the needed content is missing, occurs redundantly in more than one context, or cannot be found. Although there are several public resources that have centralized some publicly available semantic vocabulary standards and ontologies (OboFoundry, NCBO BioPortal, Biosharing.org, Ontobee)(Biosharing 2015; NCBO 2015; Smith et al. 2007; Xiang et al. 2011), there is still limited capacity for the community to identify the concepts they need across the spectrum, contribute in such a way that reduces redundancy and enhances existing standards, and easily compare the content between selected standards. In addition, few of these resources are associated with the data that are annotated using the ontologies or vocabularies. This disconnect leads to semantic standards that are not necessarily built fit-for-purpose and lacking examples that would help users determine which standards would be most appropriate for their needs. There is a need for a tool in this space to inform decision-making about the incorporation of an existing standard, the need to extend such resources, or create and coordinate new standards. Critical to this decision-making is the need to link to existing datasets in which semantic standards have been applied and understand the impacts of standards use and evolution on downstream data analyses. Further, EHS needs to incorporate emerging biomedical concepts (e.g., the exposome) that are not adequately represented among existing vocabulary resources. Consequently, there is a need for tools that allow community-based development of new standards, such as in cases of emerging research areas.

A critical component of development and adoption of semantic standards is community agreement on the meaning of terms and their use in different contexts. Gaining agreement is often difficult and imperfect, and consideration should be given to achieving agreement where there is a natural propensity, whether at a specific level of detail or around specific concepts.

Semantic disagreements can be due to community diversity, over-specification of terms, or changes in the meaning of terms over time. In cases where agreement cannot be achieved, community-specific synonyms must be incorporated to avoid limiting the utility of the standard or stalling future development. Furthermore, once a standard is available, its value is largely determined by the datasets and projects that adopt it. Wide adoption of standards is best achieved when diverse constituencies, such as data generators, data users, standards developers, publishers, and government agencies are involved and incentivized to participate in community education, participation and tool building. New tools are needed to cultivate a greater degree of collaborative development.

### Lessons learned

The Gene Ontology (GO) (Ashburner et al. 2000) is often referenced as a gold standard for ontology-based initiatives by virtue of its global community participation and implementation, development of tools to browse and access content, and its impact on data integration and analysis; however, it had humble beginnings and there is much to be learned from its early roots and subsequent path. Developed with input from an international consortium to represent how genes encode biological functions at the molecular, cellular, and tissue system levels across diverse species, GO now describes more than 40,000 biological concepts (GO 2015). GO annotations are incorporated into countless biological resources and it has been cited in over 100,000 peer-reviewed articles (GO 2015). GO has enabled integrative analyses that are now common in genomic experiments, such as gene set enrichment. Drawing upon GO, the following successful features of a semantic standards development process were identified:

- Start simple and practical;
- Utilize a modular, building block approach to allow for flexibility and reuse;

- Leverage and interoperate with existing standards where possible;
- Accommodate fuzziness: language standards need to work with scientific uncertainty;
- Find balance between logic engineering and easier-to-use vocabulary editing;
- Develop standards in close contact with the data and specific scientific need;
- Focus on capturing scientific findings (i.e., durable facts);
- Facilitate community-based collaborative curation of term definition and annotation;
- Provide stable unique identifiers;
- Incorporate significant time for community engagement and debate;
- Provide accessible user interfaces for ongoing development.

### Guiding principles

In order to ensure buy-in and use of EHS standards, we provide the following eight recommendations for establishing a community willing to participate in the development of an EHS ontology and the resources needed to accomplish this development.

1. **Engage a broad community.** Consider a broad community of stakeholders including researchers and clinicians (data generators and data users), publishers, and government agencies. Engagement can be achieved through standalone workshops, events that are embedded within broader yet aligned programs (e.g., BD2K) (Margolis et al. 2014), and Web-based resources where users can participate in discussions or add to data sets.
2. **Facilitate collaboration.** Proactively enable collaborations by planning for redundancy or inconsistency across terms within standard resources. A Web-based, automated method of identifying incongruences between concepts would provide the user a valuable “status check” across specified terms. By highlighting these inconsistencies, they may be collaboratively resolved.

3. **Enable navigation of existing language standards.** Current inventories of standards resources lack descriptive details about the standards themselves as well as applications for which they have been used. There is a need for details that allow a user an accessible glimpse of what “coverage” exists, perhaps by terms or functions of standards as well as how they have been used to standardize data. The EHS research community should be able to easily find and evaluate standards for use with their own data.
4. **Support citation and attribution of semantic standards.** A language standard used within a project, data resource, report, publication, or other scholarly product needs to be a citable entity. Standards contributions must be recognized scholarly endeavors to incentivize development. Small contributions to languages (e.g. creation of classes in ontologies) can be tracked with contributor IDs (e.g., ORCID IDs)(ORCID 2015). Attribution to funding entities (e.g., grant award) may also be included to fully capture the roles within the standard lifecycle.
5. **Adopt software development best practices.** Development should address a need in the context of real data. Break the work into modularized portions, and provide descriptors for each module. Utilize robust version control and attribution for each module as routinely practiced in software development. Publish each module to enable testing, reuse, and integration by others.
6. **Assist early development.** One challenge is that early standards development is rarely funded, but the initial stages of projects involving standards are crucial for establishing effective collaborations. Small funding sources for collaborative exchanges can help. National Science Foundation Research Coordination Networks (RCN 2015) are one mechanism for this, but there could be a more general funding mechanism.

7. **Be sustainable and flexible.** A successful framework must allow for continuous iteration of standards, be extensible in the face of evolving technologies, be driven by the data and community needs, and enable community participation/crowdsourcing.
8. **Capitalize on opportunistic development.** Seek existing projects or use cases that require or are developing language standards. Utilize these opportunities to pilot a framework approach.

These guiding principles should be operationalized to serve as a resource for the EHS research community. A Web-based toolkit could enable navigation of relevant standards from existing sources and serve as a collaborative infrastructure for community-based participatory research. Such a resource could include navigation not only of existing standards, but also the data within resources that leverage those standards. This connection would facilitate crowdsourcing approaches and tool development such as trackers, forum pages for the community to contribute use cases, and success stories. The intention of such a toolkit would be to complement and work synergistically to achieve an environmental health “slice” of existing standards efforts and technologies. For example, a project investigating the microbiome population and its response to different dietary and environmental exposures needed to standardize a) the microbiome species, b) the source from which the microbiome sample was taken (e.g. stool, mouth, etc.), c) a set of key nutrients, d) environmental contaminants, and e) disease and phenotypic characteristics at the time of sampling. The EHS toolkit could potentially go to the Human Microbiome Project (Group et al. 2009) to uniquely identify microbial strains, collect anatomical terms from the Uberon anatomy ontology (Haendel et al. 2014), foods from Wikipedia (Wikipedia 2015), target chemicals from MeSH (MeSH 2015), diseases from the Disease Ontology (Schriml and Mitraka 2015), and phenotypes from the Human Phenotype

ontology (Kohler et al. 2014). In choosing the terms, the user would want to see what data were already associated – for example, which phenotypes had been associated with the candidate disease? Which toxicants were found in the groundwater near certain population(s)? The output would be a logically constructed collection of vocabulary terms that could be used in the project, edited, and contributed back to the source resources, while maintaining provenance.

Development of an EHS toolkit would require expertise in technical standards development processes, such as software engineering that leverages the Web Ontology Language (OWL) (OWL 2016). It would also require close collaboration with the various sources of vocabulary standards to support interoperability and coordination of community contributions, and environmental health related data resource developers. Finally, tools such as Web Protégé (WebProtege 2015) or Semantic Media Wiki (SMW 2015), if enhanced with functionality to meet the above needs, may potentially be utilized as Web-based locations for collaborative editing, reviewing, and sharing the “slices” of the vocabulary standards.

#### Phased Approach to EHS Semantic Standards Development

There are several current challenges to development and broad adoption of EHS semantic standards including identification of an invested community, accessibility of semantic standards and development resources, and availability of funding to ensure ongoing support and sustainability. A major accomplishment of this workshop was identification of a community, comprised of the workshop participants, who are committed to initiating and participating in a collaborative effort to develop EHS semantic standards. This community strongly recommended a) federal funding to ensure augmentation and adoption of these standards and b) interdependent and iterative phases of development described below.

### **Phase I: Identify EHS research questions and use cases**

To ensure currency and immediate application, the EHS community should focus semantic development efforts related to current research questions. Refinement through development of detailed use cases within the community forum would serve to engage the multidisciplinary EHS community and help to prioritize standards development. Use cases describe minimally, a research question; the data, standards, and resources required to address the question including gaps; and competency questions (essentially questions that are used to test adequacy of the standard) that enable clear and focused communication around the research need. To facilitate progress, we developed a use case template and applied it to a sample research question (see “Use Case Template” in Supplemental Material). Initial research questions and use cases should attempt to use existing or ongoing data input streams. By embracing a needs-based approach and working openly to provide solutions on a focused, well-understood project, development efforts are more likely to evolve and address real research needs. Currently, through a listserv mechanism (see below), our community has begun the process of identifying research areas for use case development. Development of use cases would be an ongoing activity that serves to expand EHS data representation and the capacity for integration and reuse over time.

### **Phase II: Develop a Web-based, EHS standards toolkit**

We propose development of a Web-based toolkit that will support navigation of existing standards, knowledge, and data sources; and allow users to extract vocabulary “slices” for a given research project and enable extension of these standards (Figure 1). The overall goal is to provide a) navigation of environmental health relevant vocabulary standards and concepts that can be found in a broad diversity of locations on the web; b) allow custom term set creation

(“slices”) in a logically consistent, shareable, and reusable manner; c) allow perusal of existing data to inform term selection and enable quality assurance; and d) provide a brokering mechanism to contribute terms and edits back to the source vocabularies and knowledgebases. This tool would therefore facilitate crowdsourcing vocabulary development. Group sharing of the “slices” could potentially increase the EHS community’s adoption and extension of existing standards, and provide a mechanism for broadening the collection of research questions, use cases, and success stories described in Phase 1.

To facilitate participation, data entry and automated validation tools for quality control assessment were recommended as part of the toolkit. One example of a validation tool is the Annotation Sufficiency Meter (Phenoday 2014) provided by the Monarch Initiative (Monarch 2015), which leverages diverse large-scale semantically integrated data. This validation tool allows clinicians or model organism researchers to enter phenotypic data at the point of care or in the lab, and then get back quality assurance metrics on their phenotype ontology annotations. It will be critical for those experienced in developing such resources to help develop tools that leverage language standards and data stores together. This integration will ensure that researchers benefit during the process of data creation, analysis, and publication from the use of language standards while simultaneously contributing to them.

### **Phase III: Develop a plan for governance and sustainability**

A governance model is essential for maintaining a coordinated suite of semantic standards, sustaining community efforts in keeping with scientific and technical advances, and championing public availability of standards for EHS data to ensure continued relevancy. Governance should involve representation from the full spectrum of data-generating labs, funders, domain scientists, informaticians, and publicly available resources. Modern open

source software development environments, such as GitHub (GitHub 2015), have become more accessible to the layperson and have been extremely successful in coordinating distributed vocabulary development projects. We recommend coordinating with such efforts as well as emerging funding mechanisms (e.g., BD2K, Children's Health Exposure Analysis Resource, and other Exposome initiatives at NIEHS) (CHEAR 2016; Margolis et al. 2014) for which standards development is an expressed need in the interest of establishing best practices and avoiding redundancy.

A common problem for resource development projects such as databases or ontologies is the lack of dedicated and sustainable funding mechanisms. A paradigm shift by funding agencies and reviewers is needed such that development of data resources is not evaluated through the same lens of traditional hypothesis-driven research projects. Effective and broadly used semantic standards require a high level of scholarship and community involvement, result in major capacity-building impacts on research, are increasingly recognized for their integral role in data analysis and integration, yet there are virtually no dedicated funding mechanisms for their development or sustainability. Dedicated funding mechanisms are needed as standalone or as part of ongoing research programs. For either mechanism, funding agencies should consider upfront how developed resources will be sustained long term and integrated into other ongoing research projects. To justify continued funding, metrics that reflect scientific value must be incorporated to track use (e.g., numbers of citations where semantic resources were used). Although seemingly straightforward, such metrics are challenging to compile because infrastructure and standards are generally not well cited, Web-based tracking is not uniformly defined and can be wildly misleading, and new metrics are needed to properly credit infrastructure developers and collaborative teams that are not based solely on publications

(National Institutes of Health 2014). Many of these issues are not unique to EHS; however, the lack of semantic standards for EHS-specific areas (e.g., exposure related contexts, chemicals) and the need for improved integration within the broader biomedical research landscape will only be rectified by the EHS community and associated funding.

## **Conclusions**

It is an opportune time for the EHS community to help catalyze development of standards given the increasing quantity and diversity of data that is poised to advance our understanding about environmental impacts on human health. Lessons from previous language standard development efforts emphasize the long-term nature of such endeavors, and that persistence and endurance are critical characteristics of successful efforts. Toward this end, sustaining community engagement is critical and a phased approach is recommended as follows: 1) EHS research questions and use cases; 2) identify existing language resources, build navigational tools to encourage adoption and extension; and 3) determine a plan for governance and sustainability.

Clearly such advances will require dedication of resources, must address real needs, remain close to the data, and follow a sustained, but phased approach. In the coming months, NIEHS will pursue an engagement and outreach strategy providing a listserv for discussion and dissemination of materials, a research question and use case template, and a sample semantic standard inventory to be used in a community forum to give shape to the recommendations that have been described in this report. To contribute to this community, please register with the listserv at [EHSCOMMONLANGUAGE@LIST.NIH.GOV](mailto:EHSCOMMONLANGUAGE@LIST.NIH.GOV).

## References

- ARDB (Antibiotic Resistance Genes Database). 2015. Available: <http://ardb.cbc.umd.edu> [accessed 10 October 2015].
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25:25-29.
- BioSharing 2015. Available: <https://http://www.biosharing.org> [accessed 1 February 2016].
- CARDS (Computer Access to Research on Dietary Supplements). 2015. Available: [http://ods.od.nih.gov/Research/CARDS\\_Database.aspx](http://ods.od.nih.gov/Research/CARDS_Database.aspx) [accessed 10 October 2015].
- CHEAR (Children's Health Exposure Analysis Resource). 2016. Available: <http://www.niehs.nih.gov/research/supported/exposure/chea/> [accessed 1 February 2016].
- Davis AP, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, et al. 2014. The Comparative Toxicogenomics Database's 10th year anniversary: update 2015 *Nucleic Acids Res.*
- Entrez-Gene (NCBI Entrez-Gene). 2015. Available: <http://www.ncbi.nlm.nih.gov/entrez> [accessed 10 October 2015].
- GitHub 2015. Available: <https://github.com> [accessed 10 October 2015].
- GO (Gene Ontology). 2015. Available: <http://geneontology.org/page/about> [accessed 10 October 2015].
- Group NHW, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, et al. 2009. The NIH Human Microbiome Project. *Genome Res* 19:2317-2323.
- Haendel MA, Balhoff JP, Bastian FB, Blackburn DC, Blake JA, Bradford Y, et al. 2014. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *Journal of biomedical semantics* 5:21.
- Kohler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. 2014. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res* 42:D966-974.
- Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. 2014. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *Journal of the American Medical Informatics Association* : JAMIA 21:957-958.

- MeSH (NLM Medical Subject Headings). 2015. Available: <http://www.nlm.nih.gov/mesh> [accessed 10 October 2015].
- Monarch (The Monarch Initiative). 2015. Available: <http://monarchinitiative.org> [accessed 10 October 2015].
- National Institutes of Health. 2014. Software Discovery Index Meeting Report ([https://nciphub.org/resources/889/download/Software\\_Discovery\\_Index\\_Workshop\\_Report.pdf](https://nciphub.org/resources/889/download/Software_Discovery_Index_Workshop_Report.pdf)).
- National Research Council. 2011. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease (2011). Washington, D.C.:National Academies Press.
- NCBI (National Center for Biotechnology Information). 2016. Available: <http://www.ncbi.nlm.nih.gov> [accessed 1 February 2016].
- NCBI-SRA (National Center for Biotechnology Information Sequence Read Archive (SRA)). 2015. Available: <http://www.ncbi.nlm.nih.gov/sra> [accessed 10 October 2015].
- NCBO (National Center for Biomedical Ontology (NCBO) BioPortal home page). 2015. Available: <http://bioportal.bioontology.org> [accessed 10 October 2015].
- ORCID 2015. Available: <http://orcid.org> [accessed 10 October 2015].
- OWL (Web Ontology Language). 2016. Available: <http://www.w3.org/TR/owl2-overview/> [accessed 1 February 2016].
- Phenoday (Phenotype Day). 2014. Available: <http://phenoday2014.bio-lark.org> [accessed 1 February 2016].
- Port JA, Wallace JC, Griffith WC, Faustman EM. 2012. Metagenomic profiling of microbial composition and antibiotic resistance determinants in Puget Sound. *PLoS One* 7:e48000.
- Port JA, Cullen AC, Wallace JC, Smith MN, Faustman EM. 2014. Metagenomic frameworks for monitoring antibiotic resistance in aquatic environments. *Environ Health Perspect* 122:222-228.
- RCN (National Science Foundation, Research Coordination Networks). 2015. Available: [http://www.nsf.gov/funding/pgm\\_summ.jsp?pims\\_id=11691](http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=11691) [accessed October 10].
- Richesson RL, Nadkarni P. 2011. Data standards for clinical research data collection forms: current status and challenges. *Journal of the American Medical Informatics Association : JAMIA* 18:341-346.

- Schriml LM, Mitra E. 2015. The Disease Ontology: fostering interoperability between biological and clinical human disease-related data. *Mammalian genome : official journal of the International Mammalian Genome Society*.
- Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25:1251-1255.
- SMW (Semantic MediaWiki). 2015. Available: <https://semantic-mediawiki.org> [accessed 1 February 2016].
- Tenenbaum JD, Sansone SA, Haendel M. 2014. A sea of standards for omics data: sink or swim? *Journal of the American Medical Informatics Association : JAMIA* 21:200-203.
- Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. 2015. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLoS One* 10:e0134826.
- van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. 2014. A systematic review of barriers to data sharing in public health. *BMC public health* 14:1144.
- WebProtege 2015. Available: <http://protege.stanford.edu> [accessed 1 February 2016].
- Wikipedia 2015. Available: [https://en.wikipedia.org/wiki/Main\\_Page](https://en.wikipedia.org/wiki/Main_Page) [accessed 1 February 2016].
- Wild CP. 2005. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* 14:1847-1850.
- Xiang Z, Mungall C, Ruttenberg A, He Y. 2011. Ontobee: A Linked Data Server and Browser for Ontology Terms. In: *ICBO: International Conference on Biomedical Ontology*. Buffalo, NY.
- Youngblood J, Wallance J, Port JA, Cullen AC, Faustman EM. 2014. Metagenomic Applications for Environmental Health Surveillance: A One Health Case Study from the Pacific Northwest Ecosystem. *Planet@Risk* 2:281-285.
- Zimmerman A. 2008. New Knowledge from Old Data. The role of standards in the sharing and reuse of ecological data. *Science, Technology, & Human Values* 33:631-652.

**Table 1.** Variable results from a PubMed query of microbiome samples illustrates the consequences of lacking semantic standards and implementation (NCBI 2016).

| <b>Query</b>     | <b>Number of results</b> |
|------------------|--------------------------|
| Feces            | 22,592                   |
| Faeces           | 1,750                    |
| Ordure           | 2                        |
| Dung             | 19                       |
| Manure           | 154                      |
| Excreta          | 153                      |
| Stool            | 22,756                   |
| Stool NOT faeces | 21,798                   |
| Stool NOT feces  | 18,314                   |

## Figure Legend

**Figure 1.** An EHS Semantic Toolkit (phase II). We recommend establishing a Web-based toolkit to facilitate exchange, extension, and adoption of EHS data standards. Priority areas of research and associated use cases (phase I) will drive use of the toolkit, which will allow users to: search broadly for EHS concepts and related existing terms, and evaluate the context of terms through associated annotations in knowledge bases; develop custom sets of terms to address their use cases and detect gaps in available standards; extend existing ontologies and enrich new terms with associated annotations; and continually expand the search capability of identifying and reusing data standards. This workflow will inform the development of a governance and sustainability plan to ensure ongoing use and expansion in increasingly broader and cross-disciplinary contexts (phase III). This cycle will iterate as more research questions are identified and the community becomes more involved.

Figure 1.

