Concept Clearance

Branch: GEHB, HSRB, PHB, ERTB

Council Period: 202210

Concept Title: Data and metadata standards efforts to promote development and adoption of a harmonized environmental health language

Introduction

The use of harmonized language approaches to describe scientific data, methods, and knowledge is critical for a variety of needs, including helping researchers to find, procure, and integrate data and knowledge, conduct comparative analysis, and promote common understanding. Over the past decade, NIEHS has engaged the Environmental Health Sciences (EHS) community around the topic of advancing environmental health research through the development and adoption of a harmonized language. Several workshops have been held to mobilize the EHS and biomedical communities around language standards development, including the 2013 EPA-NIEHS Workshop for Advancing Environmental Health Data Sharing and Analysis: Finding a Common Language [1], the 2014 Workshop for the Development of a Framework for Environmental Health Science Language [2,3], and the 2015 NIH BD2K Workshop on Community-Based Data & Metadata Standards [4,5]. Additional community-driven efforts, including the 2019 Computable Exposures Workshop and the 2021 Workshop on Integrating Multiscale Geospatial Environmental Data into Large Population Health Studies, reflect a desire by the exposure science, epidemiology, and toxicology communities to foster the development of data reporting standards and models and to use informatics approaches to improve their research workflow, gain new insights, and increase data reuse [6,7].

Building on these and other efforts, NIEHS launched the Environmental Health Language Collaborative (EHLC) [8,9] in 2021 to coordinate an ongoing community discussion around advancing integrative environmental health research by promoting access, use, and harmonization of data through interoperable terminologies and best practices. EHLC sponsors several ongoing community forums, including hosting a public webinar series, facilitating use case-based working groups, and hosting an annual workshop series, which launched with the 2021 Virtual Workshop on Catalyzing Knowledge-driven Discovery in Environmental Health Sciences Through a Harmonized Language [10]. Through these efforts, the EHLC initiative aims to build a community of practice to exchange information, ideas, and expertise. EHLC events provide a forum to coordinate harmonization activities and collaborate on defining use cases and gaps, prioritizing activities, and describing the language strategies or approaches to enable data querying, sharing, and interoperability. Participation in EHLC is volunteer-based, and the Collaborative does not provide funding support for participation, technical development, or implementation of recommendations in domain communities.

Collectively, these efforts and ongoing discussions point to several key gap areas in the environmental health language and a critical need for development of community-driven data and metadata standards across many EHS subdomain areas as well as tools, software, and workflows to implement standards. The lack of common, harmonized language approaches for the EHS field is a well-recognized and persistent barrier for research and policy decisions. Gaps in standard terminologies, vocabularies, ontologies, and related schemas and tools hamper the capabilities to address large-scale, complex EHS research questions that require the integration of disparate data and knowledge sources [9]. The data and metadata standardization needs within EHS are diverse, and standards gaps persist in EHS focus areas including multi-omics, chemistry, toxicology, epidemiology, exposure science, phenotypes, geospatial data, and clinical health records among others [3]. This lack of shared standards creates an inefficient research practice of having to continuously recreate models and workflows, spend huge amounts of time looking for data, and recollect data [6].

Research Goals and Scope

To address these challenges, we propose a new concept to advance community-driven standards development efforts in key gap areas of the environmental health language. The purpose of the proposed initiative is to support resource-focused projects to enable EHS domain and/or subdomain communities to openly develop, extend, adapt, or refine data and metadata standards and associated tools to implement standards. These projects are intended to support activities at any point in the data standards lifecycle.

For the purposes of the proposed initiative, ‘data and metadata standards’ refer to documented agreements on representation, format, definition, structuring, tagging, transmission, manipulation, use, and management of data [11]. Standards may specify what types of metadata should be collected for any given dataset or data type, what format the metadata should be in, what units and terminologies should be used, and the file format to be used for the metadata. The 'data standards lifecycle' includes the (a) development (including initiation, establishment of key collaborators/contributors/user-communities and working groups, establishment of requirements/use cases, design, test, and approval), (b) dissemination and distribution, (c) adoption and use, (d) evaluation, and (e) maintenance (including review, revision, and retirement) of a standard [5].

For the proposed program, anticipated activities and outcomes may include:

- **Open standards for data and metadata:** Developing, extending, adapting, or refining data and metadata standards in key gap areas of the environmental health language. These standards include but are not limited to standard terminologies, taxonomies, controlled vocabularies, ontologies, common data elements, minimal information standards, and related schemas that link together semantic meaning and data.
- **Tools for standards implementation:** Developing, extending, adapting, or refining software tools to implement data and metadata standards for the EHS community, including but not limited to templates and software for metadata capture, software tools for terminology mapping or harmonization, tools for data transformation into common formats, data submission and curation workflows, and related Application Programming Interfaces (APIs).
- **Collaborator, contributor, and user-community engagement:** Engaging relevant persons, groups, and organizations throughout the data standards lifecycle by convening expert panels, working groups, workshops, codeathons, trainings, and other outreach mechanisms. Within this context, the 'community' for a given standard encompasses a broad and diverse range of individuals or groups involved at various points throughout the lifecycle including, but not limited to science domain experts, researchers, ontologists, librarians, data scientists, data stewards, data engineers, software developers, vendors, repositories, societies, publishers, advocacy groups, and other end-users. Data scientists, engineers, or related experts in domains outside of environmental health are also important community members to consider.

Mechanism and Justification

There is currently a need to stimulate the development and adoption of consensus-based standards for environmental health data and metadata. The proposed program is intended to provide catalytic support for a diverse array of EHS-focused standards development activities that address unmet needs within the NIEHS strategic mission and the broader biomedical data ecosystem. This program is expected to generate standards, software tools, and best practices that will be openly disseminated for broad adoption by the relevant biomedical communities, following best practices for sharing research software [12] as well as FAIR principles for associated data and documentation [13]. The R24 Resource-Related Research Project mechanism will likely be used for this initiative, with the intent to support projects that provide open resources to enhance research and data infrastructure. Activities supported under this concept would likely include an aspect of NIEHS coordination. Ultimately, the resource projects should support a vision where our scientific community can rapidly bring together the data, knowledge, and tools needed to make new discoveries and inform solutions to environmental health concerns. In addition, this program will support cross-cutting themes of the NIH Strategic Plan for Data Science [11] and the NIEHS 2018-2023 Strategic Plan goals #1.7 Data Science and Big Data; #2.1 Creating Knowledge From Data; #3.3 Promotion of Collaborative Science; and #3.5 Scientific Research and Data Infrastructure [14].

References:

[1] **EPA-NIEHS Advancing Environmental Health Data Sharing and Analysis: Finding a Common Language.** June 25, 2013. https://www.niehs.nih.gov/news/events/pastmtg/2013/epa-niehs-health-data/index.cfm

[2] **Workshop for the Development of a Framework for Environmental Health Science Language.** September 15 – 16, 2014. https://www.niehs.nih.gov/news/events/pastmtg/2014/language/index.cfm

[3] Mattingly, C.J.; Boyles, R.; Lawler, C.P.; Haugen, A.C.; Dearry, A.; Haendel, M. **Laying a Community-Based Foundation for Data-Driven Semantic Standards in Environmental Health Sciences.** Environ. Health. Perspect. 2016, 124, 1136–1140. https://doi.org/10.1289/ehp.1510438

[4] **NIH BD2K Workshop on Community-Based Data & Metadata Standards.** February 25 – 26, 2015. https://www.niehs.nih.gov/news/events/pastmtg/2015/community-based_standards/index.cfm

[5] **Executive Summary for the NIH BD2K Workshop on Community-based Data and Metadata Standards Development: Best practices to support healthy development and maximize Impact.** 2015. https://datascience.nih.gov/sites/default/files/bd2k/docs/ExecSumm_CBDMSworkshopFEB2015.pdf

[6] Thessen, A.E.; Grondin, C.J.; Kulkarni, R.D.; Brander, S.; Truong, L.; Vasilevsky, N.A.; Callahan, T.J.; Chan, L.E.; Westra, B.; Willis, M.; et al. **Community Approaches for Integrating Environmental Exposures into Human Models of Disease.** Environ. Health. Perspect. 2020, 128, 125002. https://doi.org/10.1289/EHP7215

[7] **Workshop on Integrating Multiscale Geospatial Environmental Data into Large Population Health Studies**. June 15 -16, 2021. https://www.niehs.nih.gov/news/events/pastmtg/2021/dert_geospatial_2021/index.cfm.

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

[9] Holmgren, S.D.; Boyles, R.R.; Cronk, R.D.; Duncan, C.G.; Kwok, R.K.; Lunn, R.M.; Osborn, K.C.; Thessen, A.E.; Schmitt, C.P. **Catalyzing Knowledge-Driven Discovery in Environmental Health Sciences through a Community-Driven Harmonized Language.** Int. J. Environ. Res. Public Health 2021, 18, 8985. https://doi.org/10.3390/ijerph18178985

[10] **Catalyzing Knowledge-driven Discovery in Environmental Health Sciences Through a Harmonized Language.** September 9-10, 2021. https://www.niehs.nih.gov/news/events/pastmtg/2021/ehslanguage/index.cfm

[11] **NIH Strategic Plan for Data Science.** NIH Office of Data Science Strategy. https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf

[12] **Best Practices for Sharing Research Software.** NIH Office of Data Science Strategy. https://datascience.nih.gov/tools-and-analytics/best-practices-for-sharing-research-software-faq

[13] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. **The FAIR Guiding Principles for scientific data management and stewardship.** Sci. Data. 2016, 3, 160018. https://doi.org/10.1038/sdata.2016.18

[14] **NIEHS Strategic Plan 2018-2023: Advancing Environmental Health Sciences, Improving Health.** National Institute of Environmental Health Sciences. https://www.niehs.nih.gov/about/strategicplan/strategicplan20182023_508.pdf