Environmental Health Language Collaborative (EHLC)

2023 Virtual Workshop Report
Sharing Your Environmental Health Sciences (EHS) Data: Metadata, Standards, and Tools

January 13, 19, and February 1, 2023

Contents

Background and Objectives

In January and February 2023, the Environmental Health Language Collaborative (EHLC) hosted a three-day virtual workshop, *Sharing Your Environmental Health Sciences (EHS) Data: Metadata, Standards, and Tools*, to raise awareness of and encourage use of metadata, standards, and tools that researchers can use to comply with the NIH Data Management and Sharing Policies and to promote effective management, sharing, and reuse of EHS data.

The National Institutes of Health (NIH) Data Management and Sharing Policies require that applicants submit a Data Management and Sharing (DMS) Plan for any NIH-conducted or funded research that will generate scientific data. The workshop aimed to prepare the EHS research community for the creation and implementation of their plans.

Presentation recordings and materials for the three-day workshop are available on the workshop website. In addition, a compilation of data management and sharing resources is available that includes resources mentioned by the speakers, submitted to the workshop chat, or added to the workshop Mural boards.

EHLC is a community-driven initiative to improve standardization, sharing, and interoperability of EHS information. In keeping with EHLC's mission, the workshop focused on elements of the DMS Plan associated with data/metadata description and standards.

The objectives of the virtual workshop were to promote the ability of attendees to:

- understand the individual parts and the overall importance of a DMS Plan,
- appreciate how a DMS Plan can aid in research,
- understand the basics of standards in the context of data management and sharing, including the value of using community-based standards, and
- become familiar with resources to aid in development and implementation of DMS Plans.

EHLC 2023 Workshop Overview

Workshop Summary

The virtual three-day workshop, *Sharing Your Environmental Health Sciences (EHS) Data: Metadata, Standards, and Tools*, showcased speakers from academic, governmental, and non-governmental organizations highlighting topics, tools, and tips related to DMS.

The three-day workshop had a total of 481 registrants (Figure 1) from 41 States in the U.S. and from 17 other countries.

Researchers (295 registrants, 208 attendees) were the most represented self-reported roles among the workshop registrants, followed by data stewards/managers/curators (45 registrants, 37 attendees) (Figure 2). Registrant affiliations were most represented by academic institutions (321 registrants, 89 attendees), followed by consulting, research, medical, and laboratory organizations (76 registrants, 55 attendees), NIH/National Institute of Environmental Health Sciences (NIEHS) (57 registrants, 48 attendees), and non-NIH/NIEHS U.S. governmental agencies (23 registrants, 15 attendees) (Figure 3). Of the attendees, about an equal number of attendees attended one day as did those who attended two or three days (Figure 4). Of the 481 workshop registrants, 197 reported they anticipated being involved in writing or implementing the NIH 2023 DMS Plan for their organization, and 199 reported being unsure (Figure 5). Of the 352 workshop attendees, 148 reported they anticipated being involved in writing or implementing the NIH 2023 DMS Plan for their organization, and 154 reported being unsure (Figure 5).

Detailed agendas from each workshop event can be found in Appendix A. Summary details on workshop participants are illustrated in Figures 1-5.

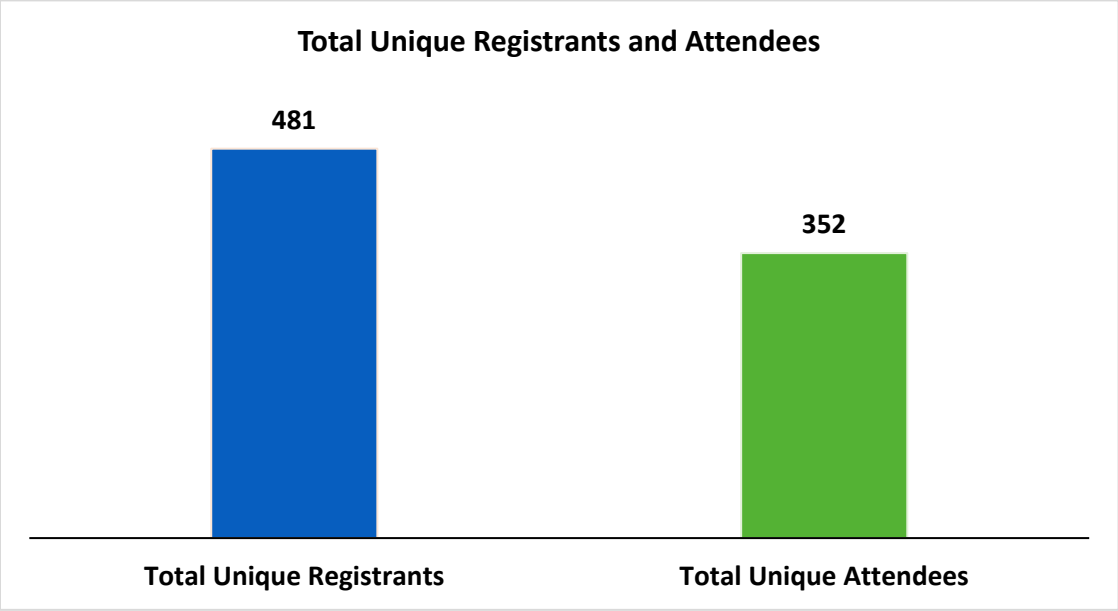

Figure 1 Total Unique Registrants and Attendees

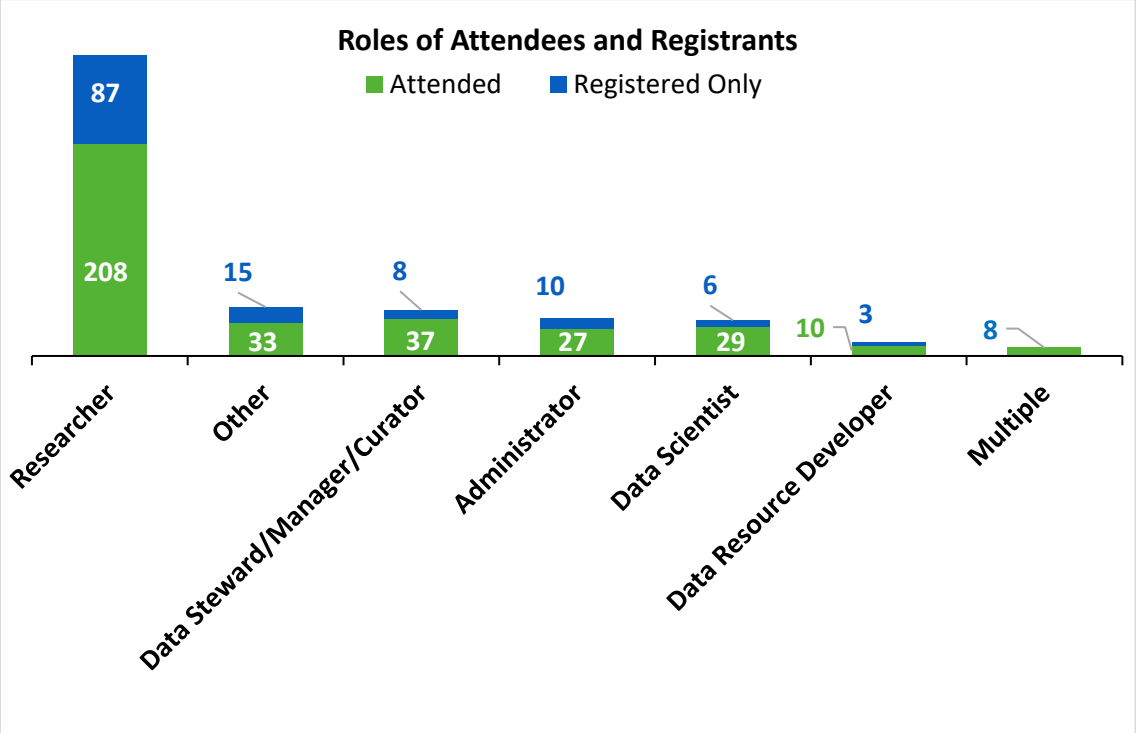

Figure 2 Roles of Attendees and Registrants

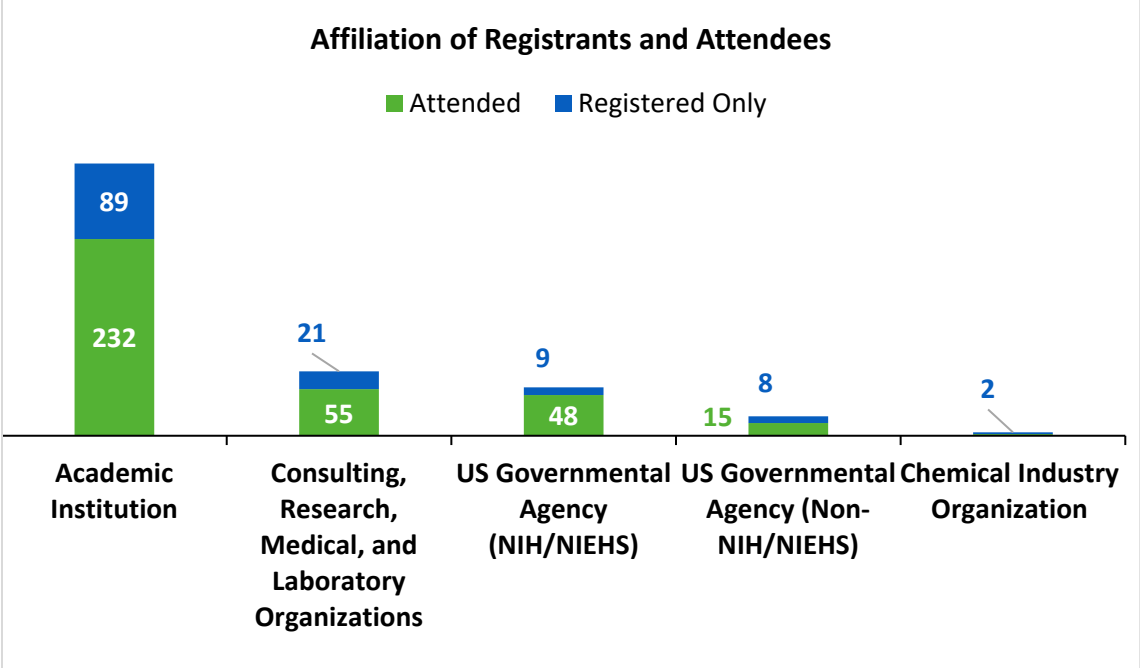

Figure 3 Affiliation of Registrants and Attendees

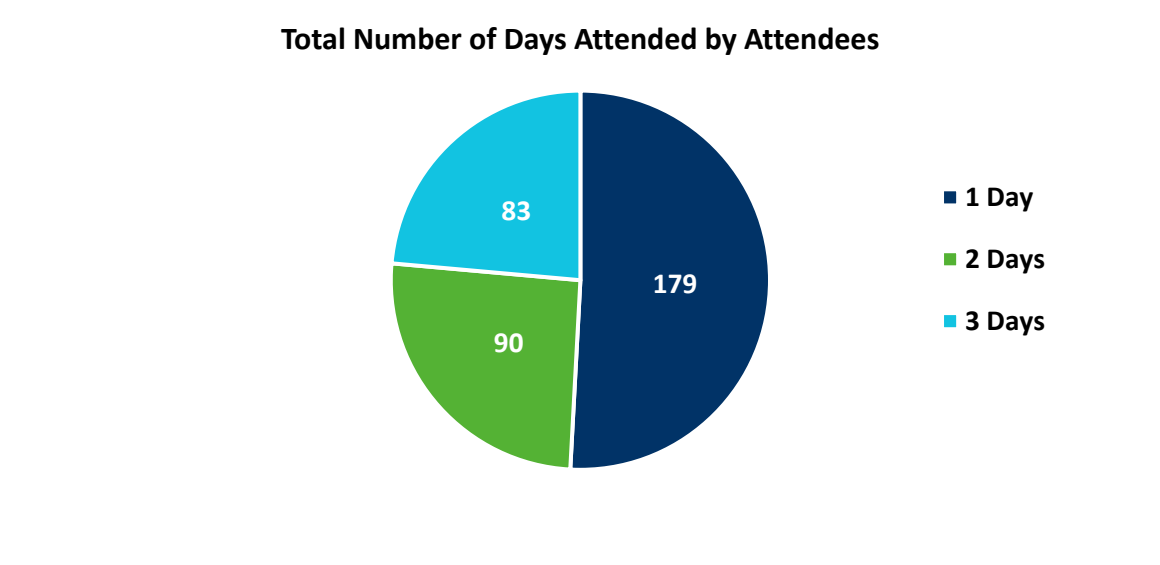

Figure 4 Total Number of Days Attended by Attendees

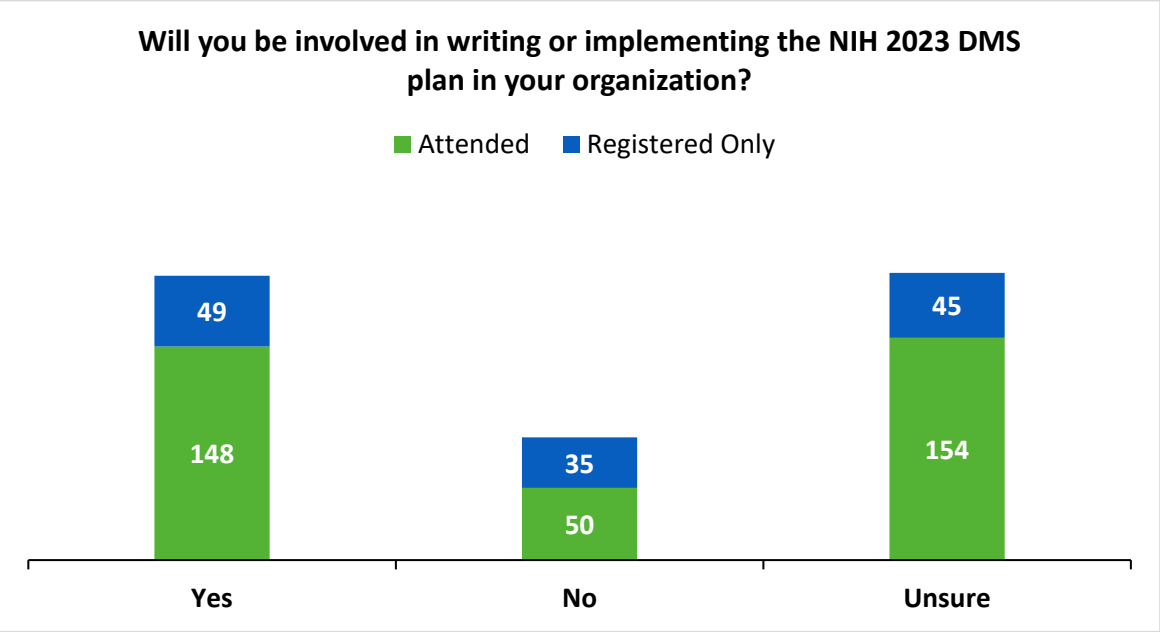

Figure 5 Will you be involved in writing or implementing the NIH 2023 DMS Plan in your organization?

Mural, a collaborative workspace, was used throughout the workshop to capture attendee input on three specific topic areas (see Appendix B for more details on Mural boards).

- **The Data Management and Sharing Plan (DMSP) Input Mural Board** was created to capture the community's input, understanding, and preparedness to implement the new policy. Registrants were invited to record answers to five questions on the DMSP Mural board from January 13th to February 8th.
- **The Unconference Session Mural Board** was open from January 13th to January 26th for attendees to submit topics they would like to facilitate for group discussion at the Unconference session held on February 1st.
- **The Data Harmonization Use Case Feedback Mural Board** was created to obtain feedback on work-in-progress to compile a list of terminologies and ontologies relevant for the EHS field. EHLC workshop participant feedback was requested on the scope and gaps of the list and its usefulness during the February 1st session.

Workshop Content

Key takeaways and showcased materials are summarized below for each of the three days of the workshop. Detailed agendas from each workshop event can be found in Appendix A. Mural board content, attendee demographic information, and additional resources are summarized in Appendix B, and Appendix C, respectively. The presentations spurred more questions than could be answered during the workshop and Q&A periods. As a result, a summary of questions and answers can be found in Appendix D. Full recordings and materials are available on the workshop website.

Day 1: January 13, 2023, 12:30–4:30 pm ET

Day 1 of the EHLC 2023 Workshop included 299 participants and featured nine talks covering the purpose of the NIH Data Management and Sharing Plan, an introduction to resources that can assist with developing and implementing DMS Plans, and the value of applying ontologies and metadata for data sharing.

Key takeaways were:

- Data sharing requirements are becoming more commonplace.
- Applying standards and metadata to EHS data makes data more Findable, Accessible, Interoperable, and Reusable (FAIR).
- Numerous services and resources are available to assist with DMS Plan development and implementation.

Welcome – *Rick Woychik, Ph.D., Director of the National Toxicology Program (NTP) and National Institute of Environmental Health Sciences (NIEHS)*

Dr. Rick Woychik acknowledged that environmental health research is complex due to the diverse spectrum of exposures and wide range of health outcomes. NIEHS is working to make advances in several areas including precision environmental medicine, exposomics, climate change, and health. Progress in these areas is dependent upon the research community working together to manage and share scientific data and metadata in such a way that the data are accessible, understandable, and interoperable. NIEHS is committed to promoting effective data management, sharing, and reuse of EHS data and works closely with NIH to implement the new NIH DMS Policy.

Introduction – *Charles Schmitt, Ph.D., National Institute of Environmental Health Sciences (NIEHS)*

Dr. Charles Schmitt noted that the EHLC was established to address the challenges of developing and adopting a common language for EHS research data. The value of using a common language makes it easier to find research data, combine and reuse the data, and interpret the results. In addition, a

common language enables software, tools, and databases to be interoperable. Finally, a common language ensures our communication with the public is accurately consistent. EHLC is evolving and working to ensure it serves the goals of common language and the needs of the EHS community.

Session 1: NIH Data Management and Sharing: Your Plan to Comply with Policy

Session 1 Introduction – *Chris Duncan, Ph.D., National Institute of Environmental Health Sciences (NIEHS)*

Dr. Chris Duncan explained that the January 25th policy implementation date was quickly approaching and conducted a DMS Plan preparedness poll. He highlighted the range of reported levels of preparedness (**Figure 6**) for DMS planning activities:

- 11% of participants reported they were "very prepared" to develop and implement a DMS Plan, and
- 15% reported they were "not at all prepared".

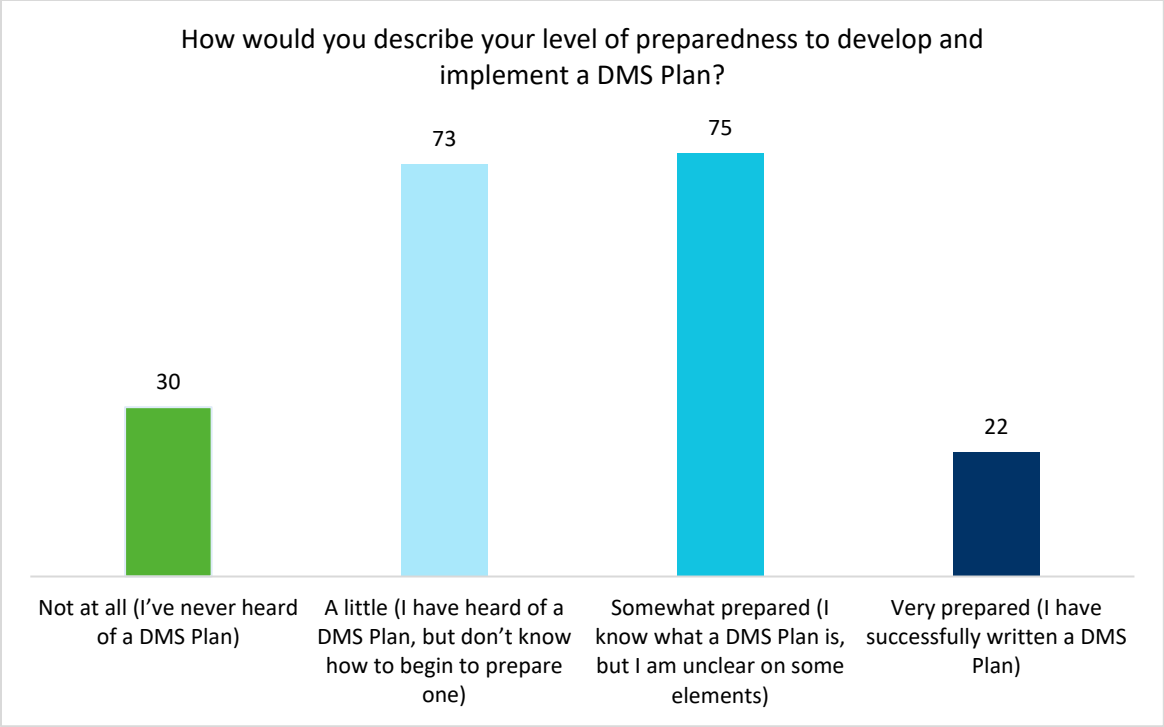

Figure 6 Workshop Day 1 Poll: Preparedness to Develop and Implement DMS Plan

Dr. Duncan shared several actions that could be taken to prepare for the upcoming policy. Dr. Duncan referenced NIEHS' new website, Scientific Data @ NIEHS, which will be continually updated and include data sharing policies and activities, as well as information on data funding related to environmental health.

To learn more, view the NIH Data Management and Sharing Plan website and the additional resources in Appendix C.

The NIH Data Management and Sharing Policy: Overview, Implementation, and Resources – *Taunton Paine, MA, National Institutes of Health (NIH) and Cindy Danielson, Ph.D., NIH*

Taunton Paine gave an overview of the NIH DMS Policy and highlighted the two basic requirements of the policy: submission of a DMS Plan for all NIH funded research and compliance with an Institutes, Centers, and Offices (ICOs)-approved plan. Mr. Paine explained that the policy would become effective on January 25, 2023, and would replace the 2003 policy. He offered additional details on the scope of the policy including awards, scientific data, timeline, format, data repository selection, and DMS. The NIH DMS Policy requires submission of and compliance with a DMS Plan, and applicants are expected to maximize appropriate data sharing, utilizing repositories as the preferred method of sharing.

> The NIH Data Sharing Policy aims to advance the rigor and reproducibility of research, promote public trust in research, and reaffirm the appropriate protections for research participants' data.
>
> *– Taunton Paine, National Institutes of Health*

Dr. Cindy Danielson gave an overview of how to plan for the submission, assessment, and compliance portions of the grant review process and emphasized that sharing data has allowable costs. In support of compliance with the policy, NIH published a set of recommended elements of a DMS Plan, guidance on selecting a data repository, and details on the allowable DMS costs. The presenters referenced the Scientific Data website and reviewed some of the available pages, resources, and tools the data sharing website provides.

> Creating metadata is key to data management, and implementing standards ensures that data and metadata are collected and stored consistently.
> *– Dr. Cindy Danielson, NIH*

Getting Started with Data Management – *Nicole Contaxis, M.L.I.S, New York University*

Nicole Contaxis started by providing an overview of the National Center for Data Services (NCDS), a program developed by the Network of the National Library of Medicine (NNLM) to train librarians on data services. Attendees were encouraged to reach out to their organization's library to see if they offer data management services. She emphasized the importance of data management and its role throughout the entire research lifecycle. Although researchers already do these data management tasks, the policy ensures researchers can think ahead about how to process and document these tasks. Critical to data management is the creation of metadata. Ms. Contaxis referenced the NCDS data glossary definition of metadata as "information that describes, explains, locates, classifies, contextualizes, or documents an information resource" and described the use and scope of metadata in the research lifecycle, including the distinction between metadata for discovery versus metadata for reuse and reproducibility. Ms. Contaxis defined a standard as "an established, community agreed-upon way to collect, organize, and document data," and highlighted examples of the diverse types of metadata, all of which may be subject to standards. Controlled vocabularies and ontologies offer the means for standardizing the metadata, and she provided examples of the appropriate use of each. The presentation closed with her highlighting a publication for researchers to learn more: "Support your data: a research data management guide for researchers."

Session 2: Resources to Get You Started

FAIRsharing.org – *Allyson Lister, Ph.D., Oxford e-Research*

Dr. Allyson Lister presented on FAIRsharing, a registry of research data standards, repositories, and policies that powers data management, FAIR evaluation, and FAIR-supporting tools. Every record is curated and the resources are cross-linked. Dr. Lister gave an overview of FAIRsharing by presenting a hypothetical use case that explained how the registry is used to identify databases to submit data, identify standards to describe data, and discover funder/publisher data policies that may apply to the research data. She closed by highlighting the FAIR Cookbook resource that offers "recipes" on how to make and keep data FAIR.

Introduction to the NIH Common Data Element Repository – *Robin Taylor, M.L.I.S, National Institute of Health (NIH)*

> Using Common Data Elements (CDEs) saves time and labor, making data interoperable.
>
> *– Robin Taylor, NIH*

Robin Taylor defined a common data element (CDE) as a standardized, precisely defined question paired with a set of allowable responses and used systematically across different sites, studies, or clinical trials to ensure consistent data collection. CDEs standardize what question is being asked, how it is being asked, and the allowable responses. The NIH CDE Repository is publicly available and contains about 29,000 CDEs across 18 collections. Users can search for CDEs that have been endorsed, recommended, or required. NIH endorses CDEs that meet

specific criteria. Ms. Taylor gave a live demonstration of the NIH CDE Repository and an in-depth walkthrough of a selected CDE.

Resources to Get You Started: DMPTool – *Maria Praetzellis, M.A., California Digital Library*

Maria Praetzellis presented on the DMPTool, an open-source, community-supported platform for data management plan (DMP) guidance and creation with templates for U.S. funders (NSF, NIH, DOE, DOT, etc.) and many international funders. Users create an account and log in to use the tool. Organizations can also join and customize the plan templates and guidance for their needs. In addition to being able to search for specific funder DMS Plan templates, the tool offers example plans that researchers chose to make public. The DMP Tool template guides users through each element of the NIH DMS Plan. The plan can be downloaded as a .doc or .pdf file. Researchers can also add research outputs (data, publications, etc.) to the plan as desired. Work is underway to create good structured interoperable metadata that is machine-actionable so that the DMS Plans are living documents that can be updated and followed up over time.

Session 3: Value of Applying Ontologies/Metadata for Data Sharing and Reuse

Introduction to NHLBI BioData Catalyst® (BDC) – *Rebecca Becky Boyles, M.S.P.H, RTI International*

Rebecca Boyles explained that the BioData Catalyst® is an advanced cyberinfrastructure with cutting-edge community tools to support FAIR data for use by the research community. BDC manages the computing environment, providing easier access to many high value datasets, tooling, and community and peer interactions. Ms. Boyles discussed how the Dug Semantic search engine is designed to enable hypothesis generation. Dug puts a biological lens on data and identifies similar terms so the user does not need to know what they are looking for in a search. Ms. Boyles detailed the Dug Annotation pipeline and showed how users can navigate to the BDC-PIC-SURE sort tool. The tool facilitates approachable research for all skill levels and produces data frames for the user.

Developing Semantic Technology for High-Throughput Zebrafish Studies – *Anne Thessen, Ph.D., University of Colorado at Anschutz*

> Ontology harmonization is a way to standardize endpoint measures across labs. Additionally, experimental plans that use a controlled vocabulary can increase consistency in endpoint reporting across laboratories.
>
> *– Dr. Anne Thessen, University of Colorado at Anschutz*

Dr. Anne Thessen opened by explaining that the Monarch Initiative uses an underlying knowledge graph to bridge the divide between laboratory and clinical data by integrating cross-species, genotype, and phenotype data. Currently, despite all the data sources and data types included, environmental data are largely missing. A project on high-throughput zebrafish studies examining microcephaly endpoints was undertaken to assess how to incorporate these types of data into Monarch. Integrating the data into Monarch requires the use of standardized and harmonized data. The challenge is that different labs use different terms for the same endpoints

and report endpoints with different levels of granularity. An experiment was conducted to determine whether a controlled vocabulary will increase consistency and endpoint reporting across laboratories. Results indicated slight improvement in annotator agreement on general terms, showing the potential to improve consistent endpoint reporting. Less agreement was recorded for more granular terms due to overloading of the terms. The next step includes development of a zebrafish endpoint atlas, and the group is seeking interested collaborators.

Monarch Initiative: Fuzzy Phenotype Matching – *Kevin Schaper, University of Colorado Anschutz Medical Campus*

> Ontologies are about more than keeping your data organized. Semantic connections help generate new insights and improve science.
>
> *– Kevin Schaper, University of Colorado Anschutz Medical Campus*

Kevin Schaper discussed how disease and phenotype associations are at the core of the Monarch Knowledge Graph. Model organisms are important to patients, meaning more species lead to more coverage. Including five model organism species boosts coverage. Human phenotype ontology (HPO) is a fundamental part of Monarch. Mr. Schaper described an example of a clinical case study of phenotype matching and the phenotype profile search site on Monarch.

KnowWhereGraph in a Nutshell – *Krzysztof Janowicz, Ph.D., University of California at Santa Barbara*

Dr. Krzysztof Janowicz described the KnowWhereGraph, a spatially enabled cross-domain knowledge graph for environmental intelligence applications with 12–15 billion statements. It seeks to tackle the key problem of researchers spending most of their time on data wrangling tasks. His presentation focused on the pros and cons of GeoEnrichment. The pros shared were that GeoEnrichment offers data on-demand, and that the data are well-curated. The data are also apportioned, which means they are tailored to the study of interest and are GIS-ready. The cons shared were that GeoEnrichment has predefined categories, closed data silos, flat tabular data, and limited support for automated integration. Additionally, GeoEnrichment is not always up to date and does not scale.

Day 2: January 19, 2023, 12:00–5:00 pm ET

Day 2 of the EHLC 2023 Workshop included 200 participants and featured five presentations highlighting specific resources and tools (e.g., CEDAR Workbench, ISA Framework, and OBO Foundry) that can assist researchers in implementing the metadata and standards elements of the NIH Data Management and Sharing Plan. The focus was on metadata annotation, learning more about best practices for ontologies and controlled vocabularies, and what to do if there are no standards for a specific domain of interest.

Key takeaways were:

- CEDAR Workbench, ISA Framework, and OBO Foundry are tools that can help implement metadata standards required for a data management and sharing plan.
- There was a focus on the need for FAIR data and following FAIR data standards.
- Collaboration is required to be a good member of the ontology community and is vital for the betterment of ontologies.

Welcome – *Stephanie Holmgren, M.L.I.S., MBA, Program Manager of the Office of Data Science at the National Institute of Environmental Health Sciences (NIEHS)*

Stephanie Holmgren welcomed the workshop participants and summarized the logistics of the EHLC workshop.

Introduction – *Charles Schmitt, Ph.D., National Institute of Environmental Health Sciences (NIEHS)*

Dr. Charles Schmitt reviewed the events from Day 1 of the workshop (January 13, 2023) and the agenda for Day 2 of the EHLC 2023 workshop, emphasizing the day would focus on two key elements of the DMS Plan: metadata and standards. These elements are relevant to the development of a common environmental health language and probably the most difficult parts of the data sharing plan to apply.

Session 1: Resource Sharing Session – DMS Plan Element 1c: Metadata

The CEDAR Workbench – *Mark Musen, MD, Ph.D., Stanford University*

Dr. Mark Musen introduced the Center for Expanded Data Annotation and Retrieval (CEDAR) tool, which supports FAIR by enabling researchers to annotate their data with standards-compliant metadata. If researchers want to share data in a way that guarantees other investigators can make the most use of the data, they need 1) *ontologies* to provide controlled terms so other investigators can make secondary use of the data, 2) *reporting guidelines* that standardize the types of information needed to know about an experiment, 3) *technology* like CEDAR to make it easy to apply metadata, and 4) *community of practice* procedures to create those metadata standards. Dr. Musen provided the Minimum Information About a Microarray Experiment (MIAME) as an example to show the value of creating minimum reporting guidelines for microarray experiments. The last half of the presentation focused on describing CEDAR, which translates reporting guidelines into a template, making it easy for researchers to assign metadata using ontology terms. He stepped through an example using the biosample template. Dr. Musen also highlighted two other FAIR data organizations: GO FAIR and ZonMw.

> The only way that investigators can ensure that their data are compliant with the FAIR Guiding Principles is to create *metadata* that adhere to appropriate community standards for both *reporting guidelines* and *ontologies*.
>
> *– Dr. Mark Musen, Stanford University*

10 years of ISA: Lessons Learned and Recent Developments – *Philippe Rocca-Serra, Ph.D., Oxford e-Research Centre*

Dr. Philippe Rocca-Serra introduced the Investigation Study Assay (ISA) Framework and spoke on the evolution of ISA. For the past 10 years, the ISA project has been growing a community of users, developing closer integration with public repositories, and improving Python support and documentation. Dr. Rocca-Serra discussed common curation practices for ISA configurations and highlighted the adoption of ISA by different communities, as well as the control guidelines and specifications that ISA follows.

Session 2: Resource Sharing Session – DMS Plan Element 3: Standards

Standard Terminology: Ontology Lookup Services. OBO Foundry. Specific Ontologies – *James Overton, Ph.D., Knocean Inc*

Dr. James Overton explained the different types of standardized terminologies including controlled vocabulary, taxonomy, and ontology. He detailed the pros and cons of each and noted they vary in the cost to build and maintain. Researchers should use the simplest type of terminology for their needs. Dr. Overton then gave an example of using an ontology term. Dr. Overton recommended that the

community not build new ontologies unless necessary due to the large amount of work required to create them.

> Standard terminology is important to use for several reasons: comply with the data sharing mandate, support FAIR principles, and especially offer clear communication not only among humans but interoperability of machines.
>
> *– Dr. James Overton, Knocean Inc*

In the last half of the presentation, Dr. Overton discussed ontology browsers like the National Library of Medicine's Unified Medical Language System (UMLS) and MetaThesauraus, BioPortal, the EMBL-EBI Ontology Lookup Service (OLS) and the Open Biological and Biomedical Ontology (OBO) Foundry. Dr. Overton provided an overview of the Open Biological and Biomedical Ontology (OBO) Foundry and its community of open scientific ontology projects. Dr. Overton listed the specific OBO Foundry principles and best practices. The core value of standards is in community consensus: breadth, depth, and detail. Ontologies provide standard terminologies with rich annotations and axioms. The OBO Foundry is an open community of interoperable ontology projects.

Session 3: Getting Hands-on with Ontologies

Using Ontologies: Tutorial on Finding and Requesting Ontology Terms – *Nicole Vasilevsky, Ph.D., Critical Path Institute*

Dr. Nicole Vasilevsky reviewed how to find and choose the right ontology terms and described how they can be used to standardize data. Dr. Vasilevsky mentioned several ontology websites to use including Ontology Lookup Service, BioPortal, and Ontobee. She described what to look for when assessing ontologies for use: licensing, quality/quality control, active/inactive status, community involvement, scientific soundness, etc. Dr. Vasilevsky then described how to use ontologies for annotations. For the annotation of biomedical data, she recommended identifying the best ontology terms and making new ontology term requests. Dr. Vasilevsky concluded by highlighting the importance of being collaborative team members within the ontology community. The information she covered as well as additional tutorials and materials on building and using ontologies can be found at the OBO Academy.

Session 4: But Standards Don't Exist for My Domain!

But Standards Don't Exist for My Domain – *Sierra Moxon, Lawrence Berkeley National Laboratory*

Sierra Moxon described the importance of data consistency and how this is currently lacking within the research community. To address the inoperability of databases, multiple models can be mapped together. Ms. Moxon described an example of a biosample dataset. When standards do not exist, best practices include: contribute to and reuse existing ontologies, separate annotation models from the technology that implements them, and map annotations to existing models and vocabularies. Ms. Moxon then discussed another example using LinkML as a modeling framework. LinkML is an open community with monthly meetings.

> The ideal way to standardize different datasets is through finding common vocabulary. Researchers can use ontology look up services like OBO Foundry.
>
> *– Sierra Moxon, Lawrence Berkeley National Laboratory*

Day 3: February 1, 2023, 12:00–5:00 pm ET

Day 3 of the EHLC 2023 Workshop included 115 participants and featured four presentations highlighting domain-specific resources (PhenX, Chemical Identifiers, MIATE, and Molgenis), a work-in-progress update of the Data Harmonization Use Case Working Group's draft list of ontologies useful for environmental health sciences research, and an open discussion on five topics of interest to attendees during the EHLC Unconference session.

Key takeaways were:

- Working together to follow the FAIR guiding principles.
- Making metadata accessible and useable for other researchers.
- Creating long term solutions for the research community to be able to foster metadata sharing.

Welcome and Introduction – *Charles Schmitt, Ph.D., National Institutes of Environmental Health Science (NIEHS)*

Dr. Charles Schmitt reviewed the events from Day 2 of the workshop (January 19, 2023) and the agenda for Day 3 of the EHLC 2023 Workshop including the Unconference session, introduced presenters, and reviewed the structure of the question and answers session.

Session 1: Spotlight on Domain-Specific Resources

PhenX Toolkit – *Carol Hamilton, Ph.D., RTI International*

Dr. Carol Hamilton presented an overview of the PhenX Toolkit, which is a free to use, web-based catalog of recommended measurement protocols for **Phe**notypes and e**X**posures developed by the scientific community via a consensus-based process. Dr. Hamilton explained that these protocols are standard measures for consistent data collection related to the study of common and complex diseases and are suitable for clinical and translational research. Using standard measures makes it easier to compare or combine data from different studies, increasing the impact of an individual study. Dr. Hamilton demonstrated how to use PhenX and walked through an example of a protocol.

Chemical Identifiers – Capabilities, Connection, and Contradictions – *Antony Williams, Ph.D., Environmental Protection Agency (EPA)*

Dr. Antony Williams provided an overview of various "flavors" of chemical identifiers (registry numbers, systematic names, InChI strings and keys, database identifiers, chemical structures, and structure formats) and highlighted their strengths and weaknesses. EPA has generated its own identifier, DTXSID, used in its DSSTox Database, CompTox Chemical Dashboard, and other platforms. These systems incorporate content from various sources by integrating the different flavors of chemical identifiers. He stated that despite the existence of chemical identifiers, identifying a substance is still challenging because identifiers can be ambiguous and are complex to manage and integrate. Human annotation and curation are still necessary.

MIATE: Supporting Standardized Collection of Metadata for *In Vivo* Toxicology Research – *Rance Nault, Ph.D., Michigan State University*

Dr. Rance Nault presented on Minimum Information about Animal Toxicology Experiments (MIATE), a minimum requirement checklist/reporting standard to ensure that the essential metadata for an *in vivo* study is recorded from the start and collected using a controlled language. MIATE aims to facilitate the interoperability and reuse of in vivo EHS datasets and uses existing tools and user bases to improve adoption. Dr. Nault identified previously proposed standards and then discussed how MIATE was created as an update to those standards. Dr. Nault concluded his presentation by sharing how his group is currently working on ToxDataCommons to move data from the lab bench to the public.

> MIATE is based on the ISA (Investigation, Study, Assay) Framework and can encourage researchers to collect metadata using a controlled language through use of MIATE standards and templates.
> *– Dr. Rance Nault, Michigan State University*

MOLGENIS Catalogues: For Multi-Center Cohort Studies and Beyond – *Morris Swertz, Ph.D., University of Groningen*

Dr. Morris Swertz introduced MOLGENIS, an open-source web application to collect, manage, analyze, visualize, and share large and complex biomedical datasets. Dr. Swertz described how the MOLGENIS application is being used to catalog multi-center cohort studies. He focused on the Europe and Canada (EUCAN) example to integrate observation data by treating projects as one family of projects. He showed an overview of the cohort metadata and secondary use metadata followed by the data model showing how data sources are linked and the elements used to harmonize the data. Dr. Swertz remarked that the MOLGENIS software is well suited to the creation of other types of domain-specific FAIR data catalogues including best practice models to document cohort studies, real-world evidence (RWE) data sources, and data harmonization.

Session 2: EHS Ontology Discussion

Work In Progress: The Development of a Semantic Resource Listing for EHS Data Harmonization Use Case – *Jeanette Stingone, Ph.D., Columbia University*

Dr. Jeanette Stingone presented the draft "Recommended" Semantic Resources list from the EHLC Data Harmonization Use Case Group. The list was developed to address the use case interest in combining data across independent environmental epidemiology research studies. The first step involved taking data from two Human Health Exposure Analysis Resource (HHEAR) studies related to air pollution and childhood asthma and trying to map the data to metadata standards. The goal being to determine what terminologies exist and where are the gaps. The list started with identifying domain areas and then populating resources available within those domains. During the process, the group identified significant overlap of terminologies for human endpoints and outcomes. This led to creating criteria to assess what makes quality terminology and whether to use it. In addition, the process highlighted gaps in the areas of positive outcomes and wellness, study context and quality, historical roots, barriers, and prospective trends. In the long-term, the group plans to use this information to create a minimum information template to facilitate data harmonization.

> The EHLC Data Harmonization Use Case's draft "Recommended" Semantic Resources list resource is not static, but rather will reflect what the community thinks is necessary. The Data Harmonization Use Case is seeking feedback on the Semantic Resources list, regarding subdomains, ontologies, and any items that may have been missed.
>
> *– Dr. Jeanette Stingone, Columbia University*

Data Harmonization Use Case Feedback Mural Board

The Data Harmonization Use Case Feedback Mural Board was created to present the group's work-in-progress related to compiling a list of terminologies and ontologies relevant for the EHS field. EHLC workshop participant feedback was requested on the scope and gaps of the list and its usefulness during the February 1st workshop event.

Feedback provided included:

Questions	Feedback
What gaps (if any) do you see in the spreadsheet?	• Epidemiology ontologies are under development • Is there a need to specify the level of granularity of the terminology to help users identify what would be appropriate • Need fields for every chemical studied to indicate if study reported only exogenous, endogenous, or both

Questions	Feedback
Do you think all types of standard terminologies should be included, or only ontologies? Why?	• Some important standards (SNOMED, CT, LOINC, UMLS, MeSH) are important. However, support to map those to ontologies is needed. • Not all standard terminologies, but if there are no formal ontologies and these terminologies are critical to filling gaps they should be included. • Other terminology types could be included if they are most relevant for specific types of use cases.
How would this spreadsheet benefit your work or research?	• Helpful to see standards and gaps from the use case • Helpful for a quick view of what ontologies are included in the NIEHS domains • This sheet would make it much easier to narrow the scope for folks getting started
What additional feedback do you have?	• Great to have a Google spreadsheet for accessible sharing of resources • Would be nice to have a static version that could be referenced. • What's the next stage of coming to "consensus" on the terminologies to be included for the domains

Session 3: Unconference Session Breakout Rooms

Based on contributions to the Unconference Mural Board, attendees split into five breakout rooms to discuss topics of interest. Attendees who would not be able to attend any breakout session had the opportunity to add comments to a topic on the Mural board (See EHLC 2023 Workshop Unconference Session Mural

in Appendix B).

Unconference Breakout Room One: Mechanisms to Incentivize Adoption and Adherence to CDE Collections among Clinical Studies in the NIH Extramural Community

Convener(s): Dr. Oswaldo Alonso Lozoya, RTI International

Key takeaway: Attendees discussed the layers of uncertainty that lie in the DMS Plan, such as 1) who decides what data are included in data sets and 2) who is tasked with the degree of selection/vetting of different analytical tools (peer-review process versus vetting by NIH).

Unconference Breakout Room Two: Data sharing, Privacy, and Geospatial/Spatiotemporal Data in Environmental Health

Convener(s): Dr. Allan Just, Icahn School of Medicine at Mount Sinai and Dr. Charles Schmitt, NIEHS

Key takeaway: Attendees discussed the balance of privacy risks and data sharing interests as it pertains to spatiotemporally linked data. Opportunities for action discussed included: 1) using a firewall like DataShield, 2) using a practice dataset so users do not have access to the raw data but still get outputs, and 3) getting participants to opt into sharing their data that may contain identifiers and having the NIH control who gets to see or use that data. Some obstacles discussed included: 1) small research groups being able to handle requests for their data and 2) gaining an understanding of the boundary and scope of data security problems.

Unconference Breakout Room Three: What is Involved in Making a Robust Submission Package for a Generalist Repository?

Convener(s): Dr. Jennifer Fostel, NIEHS

Key takeaway: Attendees discussed challenges for researchers to find and interpret data. Suggested next steps for the generalist repository included minimizing data capture initially, adding searchable indexed terms, developing a standardized list of environmental health variables or endpoints, facilitating the use of templates and a web-based platform, harmonizing data, and providing guidance to reduce the burden and complexity of terminology, technology, and tools. The group was interested in continuing this discussion as part of ongoing EHLC activities.

Unconference Breakout Room Four: Using General-Purpose Study Protocols to Automate Standards-Compliant Reporting of Environmental Health Research.

Convener(s): Dr. Paul Whaley, Whaley Research UK

Key takeaway: Attendees discussed the need for ethnographic research to understand what happens in laboratories. While scientific processes and standards do exist, they are not implemented or standardized related to the outputs of research. Due to this, attendees emphasized that it is important to incorporate these processes and standards related to outputs of research from the beginning of the research process. The group highlighted the benefits of using a democratic approach to develop these processes (i.e., bottom-up versus top-down).

Unconference Breakout Room Five: Moving EHLC Forward in the Next 6 to 12 Months.

Convener(s): Steven Black, ICF

Key takeaway: Attendees discussed potential activities for future EHLC efforts, including more interactive and collaborative opportunities that can be done asynchronously for simpler topics and focusing large group meetings on only the most important topics. Attendees were invited to continue to provide ideas for future EHLC activities and sessions by contacting Stephanie Holmgren, Office of Data Science.

> Data management happens throughout the research lifecycle. The DMS Plan aids researchers in thinking about data management and data sharing issues up front before they conduct a study.
>
> *– Stephanie Holmgren, National Institute of Environmental Health Sciences (NIEHS)*

Appendix A. Workshop Agendas

EHLC 2023 Workshop Day 1 | January 13, 2023, 12:30–4:30 pm ET

The objective of day 1 is to set the stage for understanding the purpose and elements of the DMS Plan.

Welcome and Opening Remarks

12:30–12:35 | Welcome and Introduction *Rick Woychik, NIEHS*

12:35–12:45 | Introduction to EHLC *Charles Schmitt, NIEHS*

Session 1: NIH Data Management and Sharing: Your Plan to Comply with Policy

Moderator: Chris Duncan, NIEHS

12:45–12:55 | Introductory Remarks *Chris Duncan, NIEHS*

12:55–1:15 | The NIH Data Management and Sharing Policy: Overview, Implementation, and Resources *Taunton Paine, NIH and Cindy Danielson, NIH*

1:15–1:35 | Getting Started with Data Management *Nicole Contaxis, NYU Health Sciences Library*

1:35–1:45 | Questions

Session 2: Resources to Get You Started

Moderator: Stephanie Holmgren, NIEHS

1:45–2:00 | FAIRsharing.org *Allyson Lister, University of Oxford e-Research Centre*

2:00–2:15 | Introduction to the NIH Common Data Element Repository *Robin Taylor, NIH/National Library of Medicine (NLM)*

2:15–2:30 | Resources to Get You Started: DMPTool *Maria Praetzellis, California Digital Library*

2:30–2:45 | Questions

[Break 2:45 pm–3:00 pm]

Session 3: Value of Applying Ontologies/Metadata for Data Sharing and Reuse

Moderator: Stephanie Holmgren, NIEHS

3:00–3:15 | Introduction to NHLBI BioData Catalyst® (BDC) *Rebecca Boyles, RTI International*

3:15–3:30 | Developing Semantic Technology for High-Throughput Zebrafish Studies *Anne Thessen, University of Colorado at Anschutz*

3:30–3:45 | Monarch Initiative: Fuzzy Phenotype Matching *Kevin Schaper, University of Colorado at Anschutz Medical Campus*

3:45–4:00 | KnowWhereGraph in a Nutshell *Krzysztof Janowicz, University of California at Santa Barbara*

4:00–4:15 | Questions

Closing and Adjourn

EHLC 2023 Workshop Day 2 | January 19, 2023, 12:00–5:00 pm ET

The objective of day 2 is to highlight specific resources and tools that support compliance with elements of the DMS Plan.

Welcome and Opening Remarks

12:00–12:15 | Welcome and Introduction *Charles Schmitt, NIEHS*

Session 1: Resource Sharing Session – DMS Plan Element 1c: Metadata

Moderator: Charles Schmitt, NIEHS

12:15–12:40 | The CEDAR Workbench *Mark Musen, Stanford University*

12:40–1:05 | 10 years of ISA: Lessons Learned and Recent Developments *Philippe Rocca-Serra, University of Oxford e-Research Centre*

1:05–1:15 | Questions

[Break 1:15 pm–1:30 pm]

Session 2: Resource Sharing Session – DMS Plan Element 3: Standards

Moderator: Anna Maria Masci, NIEHS

1:30–2:20 | Standard Terminology: Ontology Lookup Services. OBO Foundry. Specific Ontologies *James Overton, Knocean Inc*

2:20–2:30 | Questions

[Break 2:30–2:45 pm]

Session 3: Getting Hands-on with Ontologies

Moderator: Anna Maria Masci, NIEHS

2:45–3:35 | Using Ontologies: Tutorial on Finding and Requesting Ontology Terms *Nicole Vasilevsky, Critical Path Institute*

3:35–3:45 | Questions

Session 4: But Standards Don't Exist for My Domain!

Moderator: Anna Maria Masci, NIEHS

3:45–4:35 | But Standards Don't Exist for My Domain *Sierra Moxon, Lawrence Berkeley National Laboratory*

4:35–4:45 | Questions

Closing and Adjourn

EHLC 2023 Workshop Day 3 | February 1, 2023, 12:00–5:00 pm ET

The objective of day 3 is to continue to raise awareness of domain-specific resources as well as offer the opportunity to have more open discussions on topics of interest to attendees.

Opening Remarks

12:00–12:15 | Welcome and Introduction *Charles Schmitt, NIEHS*

Session 1: Spotlight on Domain-Specific Resources

Moderator: Charles Schmitt, NIEHS

12:15–12:35 | PhenX Toolkit *Carol Hamilton, RTI International*

12:35–12:55 | Chemical Identifiers – Capabilities, Connections and Contradictions *Antony Williams, U.S. EPA*

12:55–1:15 | MIATE: Supporting Standardized Collection of Metadata for *In Vivo* Toxicology Research *Rance Nault, Michigan State University*

1:15–1:35 | MOLGENIS Catalogues: For Multi-Center Cohort Studies and Beyond *Morris Swertz, University Medical Center Groningen*

1:35–1:45 | Questions

[Break 1:45 pm–2:00 pm]

Session 2: EHS Ontology Discussion

Moderator: Stephanie Holmgren, NIEHS

2:00–2:45 | Work in Progress: The Development of a Semantic Resource Listing for EHS Data Harmonization Use Case *Jeanette Stingone, Columbia University*

2:45–3:00 | Questions

[Break 3:00 pm–3:15 pm]

Session 3: Unconference Session

Moderator: Stephanie Holmgren, NIEHS

3:15–3:30 | Introduction to Unconference and Mural

3:30–4:30 | Breakout Discussions

4:30–4:45 | Report Outs

Closing and Adjourn

Appendix B. Workshop Mural Board Content

EHLC 2023 Workshop Data Management and Sharing Plan Input Mural

The DMSP Mural was created to gauge the EHS community's understanding, preparedness, and ability to implement the Data Management Sharing Plan. The option to submit answers to posted questions was open from January 13th to February 8th.

Responses provided by workshop attendees are shown in the tables below. Please note that some answers were edited for spelling and clarity.

Question 1: What data types do you generate or use?	
• Animal data (date of birth, age, weight)	• Metabolomics
• ATAC-Seq	• Metagenomics and microbial communities
• Biochemical assays	• Microscopy
• Biomarker data	• Modeling data (physical simulations)
• Chemical analysis data (e.g., concentrations of POPs, PBTs)	• Mutation spectra in mice and cells
• Chemical structures	• NGS data (multi-omics, molecular)
• Clinical outcomes	• Phenotypic information (participant-level)
• Comet-chip (image based)	• Phosphoproteomics
• Cytokine analysis	• Plasma and blood, urine levels of contaminants
• Data dictionaries	• Program-specific progress reports and training numbers
• Electronic health records	• QRT-PCR
• Environmental monitoring	• RNA Seq
• Environmental samples	• Secondary data from existing cohort studies
• Flow cytometry	• Sensor data (time series concentrations of contaminants)
• GC/MS	• Spatiotemporal exposures
• Geospatial information	• Survey data (study participants)

• Ion mobility spectrometry and mass spectrometry	• Tissue chip
• LC-MS	• Transcriptional analysis (RNASeq)
• Light emission (sensor data for water samples)	• Vital statistics data (birth, fetal death records)
• Medical imaging	• Western Blot

Question 2: What repositories do you currently or plan to use to comply with the DMS Policy?	
• Amazon	• HRSA
• Biodata Catalyst®	• Metabolomics Workbench
• Bioimage archive	• PNNL Data Hub
• Commercial non-profit portals (e.g., HCA, DNAnexus)	• Pride
• Dataverse	• Program-specific and developed database
• DbGaP	• SEEK or NExtSEEK
• Dryad	• SRA
• Figshare	• Synapse
• GEO	• Zenodo
• GitHub	

Question 3: Which standards are relevant to your work?	
• DICOM	• OBO Foundry-recognized ontologies
• FHIR	• OBO ontologies
• ICH	• OECD harmonized templates
• IHEC	• Reporting templates
• NIH CDEs	• UMLS

Question 4: What resources or tools do you find useful to help with data management?	
• Bioportal	• NIH templates for data management plants
• Cloud environments	• NLP tools (SciSpacey, INDRA, PubTator)
• Database-embedded software	• Ontobee
• Dropbox	• Personnel-dedicated data manager
• ezDMP and DMPTool	• RedCap
• GitHub	• Seek/NExtSEEK
• NIH ECHO Standards	• Versioned releases of code/data

Question 5: What pain points or challenges do you anticipate or have you experienced trying to complete a DMS Plan?	
• Timely data input from multiple sources (Element 1, 4)	• Data access requests (Element 5)
• Questions about long-term longitudinal data (Element 1, 4, 5)	• Indigenous data sovereignty (Element 5)
• Coding standards of environmental health data (Element 3)	• Sharing data in compliance with informed consent (Element 5)
• Harmonize data from different sub-projects (Element 3)	• Sharing data that do not belong to me (secondary data) (Element 5)
• Lack of rigor in adherence to standards (difficulties with harmonization, interoperability) (Element 3)	• State laws prohibiting sharing of vital statistics data (Element 5)
• Lacking cross-referencing, connectivity across extant databases that hold different types of data for same studies (Element 3, 4)	• Ability to budget for necessary costs in smaller grants (Element 6) • Expertise on our study team to deposit data correctly (Element 6)
• Data versioning and consent groups in dbGaP (Element 3, 4, 5)	• Staffing management requirements after funding ends (Element 6)
• Choosing NIH data repository (Element 4)	• Time investment to accommodate all sharing requests (Element 6)
• Identifying relevant data repositories (Element 4)	• Cultural resistance to FAIR tasks (General/Administrative)
• Timing of data sharing in relation to publishing primary results (protected time) (Element 4)	• Advising DIR PIs and facilitating scientific review processes (General/Administrative)
• Lack of long-term plan and maintenance of the data management (Element 4, 6)	• People don't understand what data management means (General/Administrative)

EHLC 2023 Workshop Unconference Session Mural

This Unconference Mural was created for attendees to submit their suggested topics for the February 1, 2023, Unconference session from January 13th to January 26th. Four Unconference topics were provided by participants and the fifth topic was added by the EHLC planning committee.

Topic 1: Mechanisms to Incentivize adoption and adherence to CDE collections among clinical studies in the NIH extramural community. Added by Oswaldo A. Lozoya

Comments left on Mural:

- Theme 1: Education about benefits for CDE use (meta-analysis) and considering any additional incentives and support.
- Consilience: I don't know how to achieve it technically but it would be great to introduce tools that check data integrity in the process of analysis.
- Theme 3: By creating layers of CDE requirements: required, optional, and others. That would allow all data producers to participate in the data lake. Also, data users would be able to capitalize on the available data.
- Do some retrospective analysis? How many "unique" terms are there, really? Sounds like a lot of work.
- Group 1 mechanisms: Expanding mutually supported semantic search of data elements across resources (e.g. REDCap capacity to search NIH CDE Repository via API, and prepopulation of REDCap Forms).
- Group 1 mechanisms: have large repositories provide templates for data submissions that includes CDEs.

Topic 2: Data sharing, privacy, and geospatial/spatiotemporal data in environmental health. Added by Allan Just

Comments left on Mural:

- Support needs for small research groups: Supplying data to NIH securely, so that NIH could supply data management expertise.
- Bounds and extent of the problem are not well-defined.
- Tools like DataSHIELD can be a useful part of a data management process to preserve privacy.
- How much can data protection be automated? How much of data management for privacy requires general versus highly specific work?

Topic 3: What is involved in making a robust submission package for a generalist repo? Added by Jennifer Fostel

- No comments were left for Unconference topic 3 on the Mural board.

Topic 4: Using general-purpose study protocols to automate standards-compliant reporting of environmental health research. Added by Paul Whaley

Comments left on Mural:

- Fork Gource
- ADAPT: Use as a visualization tool for PROCEDURE DEVELOPMENT, with version control, contributions, generate and parse logs, etc.
- See GOURCE.io and "Gource in Bloom" https://www.youtube.com/watch?v=NjUuAuBcoqs. GOURCE is a visualization tool for the process of a program's creation (source control repositories). Gource will visualize whenever a file is added or changed. The repository is displayed as a tree, with the root of the repository central. Directories (folders) are branches. Files are leaves (dots). Contributors to the source code appear and disappear as they contribute to specific files and directories.

Topic 5: Moving EHLC forward in the next 6 to 12 months. Added by EHLC planning committee.

Comments left on Mural:

- Have a workshop on case studies of DMSPs.
- Mural sessions on specific topics.
- Clinical Imaging - identifiable information?

Appendix C. Additional Resources

This list reflects resources that were discussed at the workshop and that will assist researchers in developing or implementing their DMS Plans.

General Resources	
NIH Data Management and Sharing Policies	Guidance on planning and budgeting for DMS and methods for sharing data
NIH Data Sharing Resources	Links to webinars, slide sets, and webpages to learn more about the DMS Policy
NIH ODSS – Data Curation Network Event	Three-hour webinar - Four presentations cover development of DMS Plans, addressing sensitive data, budget development, and how to think like a curator
Scientific Data at NIEHS	NIEHS website providing information specific to DMS policies and resources
EHLC Resources Catalog	A compilation of organizations, ontologies, standards, and tools useful for harmonizing environmental health research
EHLC Workshop 2023 Recordings	Three-day workshop focused on raising awareness of and encouraging the use of metadata, standards, and tools that researchers can use to comply with the NIH DMS Policy and reuse of EHS data
Elements of an NIH Data Management and Sharing Plan	Detailed overview of the elements of an NIH DMS Plan and links to related recent issued announcements
NIH Scientific Data Sharing News and Events	Latest news, upcoming events, and past events
Funding Opportunity Announcement: Accelerating Data and Metadata Standards in the Environmental Health Sciences (FOA #: RFA-ES-23-002)	Funding Opportunity Announcement (FOA) to support resource projects to enable EHS communities to openly develop, extend, adapt, or refine data and metadata standards as well as associated tools to implement standards

DMS Plan Development	
DMP Tool	A free, open-source, online application that helps researchers create data management plans (DMPs)
Research Data Management (RDM) Kit	Offering resources, best practices, guides, and tools to help make data FAIR
Instructional Resources	
Metadata Standards (Susanna Sansone)	1-hour webinar - Learn about metadata and content standards, their value, different types, and application
Ontology Training (OBO Foundry)	Free, online course to teach beginners how to use ontologies
A Primer on Ontologies for Toxicology and Environmental Health (EBTC)	1.5-hour webinar - Three speakers present on the value and use of ontologies in the field
NIH Data Management and Sharing Policy Webinar Series	NIH resource with links to webinars related to the implementation of the NIH DMS Policy
Metadata/Standards Resources	
Metadata Standards Catalog	A directory of metadata standards applicable to describing either data generated or collected for the purpose of research
Ontologies/Terminologies	
Chemical Entities of Biological Interest (ChEBI)	A free online dictionary of molecular entities focused on “small” chemical compounds
Open Biological and Biomedical Ontology (OBO) Foundry	A family of community-developed interoperable biological ontologies that follow specific development principles
Ontology Lookup Services (OLS)	A repository for biomedical ontologies
Repositories	
Generalist Repository List	NIH list of generalist repositories that accept data regardless of data type, format, content, or disciplinary focus
NIH Guidance on Repositories	Provides links to NIH-supported repositories and guidance on selecting a repository based on data type and discipline
Registry of Research Data Repositories	A global registry of research data repositories for different academic disciplines

Resources from Day 1 of the EHLC 2023 Workshop	
Session 1: NIH Data Management and Sharing: Your Plan to Comply with Policy **The NIH Data Management and Sharing Policy Overview, Implementation, and Resources** **Taunton Paine, NIH and Cindy Danielson, NIH**	
NIH Policy for Data Management and Sharing (NOT-OD-21-013)	Final NIH Policy for DMS Policy), effective January 25, 2023
2023 NIH Data Management and Sharing Policy	NIH announcement and information related to the NIH DMS Policy to promote the management and sharing of scientific data generated from NIH-funded or conducted research
NIH Data Management and Sharing Policy FAQs	Frequently asked questions about the 2023 DMS Policy
NIH Activity Codes	A comprehensive listing of NIH activity codes that require applicants to submit a DMS Plan
Requesting and Justifying Costs for Data Management and Sharing	Outline on how to request and justify costs for DMS
Writing a Data Management and Sharing Plan	Guide for writing and submitting a DMS Plan, including elements in the plan and sample plans
Getting Started with Data Management **Nicole Contaxis, NYU Health Sciences Library**	
Support Your Data: A Research Data Management Guide for Researchers	A journal article on research data management and the "Support Your Data" tools, which include a self-assessment and a series of short guides
Session 2: Resources to Get You Started	
FAIRsharing.org **Allyson Lister, University of Oxford e-Research Centre**	
FAIRsharing.org	FAIRsharing is a community-driven FAIR-supporting resource that provides an informative and educational registry on data standards, databases, repositories, and policy, alongside search and visualization tools and services that interoperate with other FAIR-enabling resources
Unsure Where to Start with FAIRsharing?	Read about tips on discovering the resources you need with FAIRsharing

FAIRsharing Community Curator Programme	Put your expertise into action and get credited by joining the Fairsharing.org curation programme
Introduction to the NIH Common Data Element Repository Robin Taylor, NIH/NLM	
NIH Common Data Element (CDE) Repository	A CDE is a standardized, precisely defined question, paired with a set of allowable responses, used systematically across different sites, studies, or clinical trials to ensure consistent data collection
Standardizing Data Collection	Self-paced tutorial: Common Data Elements: Standardizing Data Collection
Common Data Element (CDE) Training	On-demand class recording: Standardize Your Research Data with the NIH CDE Repository
Resources to Get You Started: DMPTool Maria Praetzellis, California Digital Library	
DMPTool_NIHTemplate	Link to DMP Tool sample plan templates for NIH
DMP Tool Sample Plans	Links to public DMPs, which were created using the DMPTool service and shared publicly by their owners
Materials to Promote the DMPTool	Landing page for the Working Group on NIH DMSP Guidance and links to available materials and resources to promote the DMPTool
DMPTool Blog	Updates on guidance and resources for your data management plan

Session 3: Value of Applying Ontologies/Metadata for Data Sharing and Reuse	
Intro to NHLBI BioData Catalyst® (BDC) Rebecca Boyles, RTI International	
BioData Catalyst® Services	A list of platforms and services offered by BioData Catalyst
Getting Started with BioData Catalyst®	Documentation for getting started on the NHLBI BioData Catalyst® ecosystem
Biodata Catalyst® Forum	NHLBI BioData Catalyst® (BDC) Forums
Join The Biodata Catalyst® Community	Join the NHLBI BioData Catalyst® Community here
Monarch Initiative: Fuzzy Phenotype Matching Kevin Schaper, University of Colorado at Anschutz Medical Campus	
The Monarch Initiative	The Monarch Initiative integrates, aligns, and re-distributes cross-species gene, genotype, variant, disease, and phenotype data
Human Phenotype Ontology	HPO provides a standardized vocabulary of phenotypic abnormalities encountered in human disease

EHLC 2023 Workshop Day 2	
Session 1: Resource Sharing Session – DMS Plan Element 1c: Metadata	
The CEDAR Workbench - Mark Musen, Stanford University and 10 years of ISA: Lessons Learned and Recent Developments. Philippe Rocca-Serra, University of Oxford e-Research Centre	
Center for Expanded Data Annotation and Retrieval (CEDAR)	Suite of tools that help make authoring metadata more manageable to help improve discovery, comparison, and analysis of data
ISA framework tools	The open-source Investigation, Study, and Assay (ISA) framework and tools help to manage an increasingly diverse set of life science, environmental, and biomedical experiments that employ one or a combination of technologies
ISA tools (GitHub)	GitHub webpage with ISA tools
Session 2: Resource Sharing Session – DMS Plan Element 3: Standards	
Standard Terminology: Ontology Lookup Services. OBO Foundry. Specific Ontologies	

James Overton, Knocean Inc	
Open Biological and Biomedical Ontology (OBO) Foundry	OBO Foundry: Community development of interoperable ontologies for the biological sciences
Ontology for Biomedical Investigations (OBI)	For OBI specifically, the metadata registry on the project page includes a publication field for this purpose
BioPortal	A comprehensive repository of biomedical ontologies
EMBL-EBI OLS	European Molecular Biology Laboratory and European Bioinformatic Institute (EMBL-EBI) EMBL-EBI OLS
Ontobee	A linked data server that facilitates ontology data sharing, visualization, query, integration, and analysis
Ontology Development Kit (ODK)	Manage your ontology's life cycle with the Ontology Development Kit (ODK)
OBO Foundry Dashboard	The OBO-Dashboard automatically checks ontologies for adherence to OBO Foundry principles
OBOOK OBO Training	Open Biological and Biomedical Ontologies Organized Knowledge (OBOOK) and OBO Semantic Engineering Training
Session 3: Getting Hands-on with Ontologies	
Using Ontologies: Tutorial on Finding and Requesting Ontology Terms Nicole Vasilevsky, Critical Path Institute	
OBO lesson- Getting Hands-on with Ontologies	Lesson trains biomedical researchers on how to find a term, what to do if they find too many terms, how to decide on which term to use, and what to do if no term is found
How to Guide for GitHub	Contributing to OBO ontologies
Session 4: But Standards Don't Exist for My Domain.	
But Standards Don't Exist for My Domain Sierra Moxon, Lawrence Berkeley National Laboratory	
LinkML	LinkML is a general-purpose modeling language that can be used with linked data, JSON, and other formalisms
SSSOM: Simple Standard for Sharing Semantic Mappings	The SSSOM TSV format in particular is geared towards the needs of the wider bioinformatics community as a way to safely

	exchange mappings in an easily readable yet semantically well-specified manner

EHLC 2023 Workshop Day 3 Session 1: Spotlight on Domain-Specific Resources	
PhenX **Carol Hamilton, RTI**	
Phenotypes and eXposures (PhenX) Toolkit	A web-based catalog of curated, recommended measurement protocols for genomic, epidemiologic, clinical, and translational research
MIATE **Rance Nault, Michigan State University**	
Minimum Information about Animal Toxicology Experiments (MIATE)	A community-developed set of minimal metadata requirements to promote making in vivo animal toxicology experiment data FAIR
MOLGENIS **Morris Swertz, University of Groningen**	
MOLGENIS	A free, open-source data platform to help researchers find, capture, exchange, manage, and analyze scientific data

Appendix D. Presentation Q&A

EHLC Workshop Day 1: Questions and Answers	
Question	**Response**
The NIH Data Management and Sharing Policy: Overview, Implementation, and Resources Taunton Paine, NIH and Cindy Danielson, NIH	
How does this new DMS Plan correspond to or differ from the Resource Sharing Plan required in the Application Submission System & Interface for Submission (ASSIST) when submitting a new grant application?	**Taunton Paine:** Resource Sharing Plans relate to sharing as expected by the model organism sharing policy, the research tools policy, and other NIH Institute and Center or program-specific requirements as applicable. Resource Sharing Plans are submitted in a different section of the application and are reviewed differently. For more information on the model organisms or research tools policy, see: https://sharing.nih.gov/other-sharing-policies.
Are you going to monitor and audit compliance with intramural plans?	**Taunton Paine:** Yes, NIH will monitor intramural compliance with DMS Plans, as stated in the DMS Policy.
Does it apply to the 'K' awards?	**Cindy Danielson:** It applies to some K awards - please refer to the list of NIH activity codes subject to the DMS Policy at https://sharing.nih.gov/sites/default/files/flmngr/List-of-Activity-Codes-Applicable-to-DMS-Policy.pdf. This will also be outlined in each FOA.
You are planning a publication but don't submit it prior to the end of your funding. Do you need to make those data available before publication?	**Taunton Paine:** The data should be shared by the time of peer reviewed publication or by the end of the period of performance but we may offer some additional guidance in the future.

EHLC Workshop Day 1: Questions and Answers	
Can we speculate that whatever behemoth these policy making efforts add up to in the U.S. will make things easier for small and medium-sized enterprise (SME) companies in countries adopting Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)-like regulations, which are extortionately expensive to comply with because of the access fees charged by data owners?	**Taunton Paine:** I'm not certain what REACH-like regulations are, but I want to point out that the DMS Policy prefers sharing data through established repositories, and provides certain desirable characteristics for these repositories, including "Free and Easy Access: Provides broad, equitable, and maximally open access to datasets and their metadata free of charge in a timely manner after submission, consistent with legal and ethical limits required to maintain privacy and confidentiality, Tribal sovereignty, and protection of other sensitive data." See: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html
With the exception of what I write in my data management plan, how do I incorporate standards when sharing data? Is this something the repository or journal asks for?	**Taunton Paine**: Some repositories, journals, or NIH programs may set specific expectations for the data standards to be used. Some examples of how standards may be incorporated in DMS Plans are provided in NIH sample DMS Plans: https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/writing-a-data-management-and-sharing-plan#sample-plans.
Do we need a separate budget item for data management/sharing or can we just address the budgeting in the justification? (For example, if the same person will manage data and be project manager)	**Cindy Danielson**: Please see https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/budgeting-for-data-management-sharing#requesting-&-justifying-costs-for-data-management-and-sharing, which outlines how to request and justify costs for DMS. For detailed budgets, you will include a line item in the budget form for DMS costs and also justify those costs in the Budget Justification Attachment. The NIH Application Guide contains more details on this.
Any consideration from NIH to increase budget caps to allow for these additional costs related to DMS activities (to preserve funds for research activities)?	**Cindy Danielson**: That is a question we heard and understand this may be new. There is no plan to raise the cap, but we are keeping our eyes and mind on things that can be done.

EHLC Workshop Day 1: Questions and Answers	
How does the Genomic Data Sharing Policy affect the new Sharing Policy? Does Genomic Data get specifically called out in the DMS Plan?	**Cindy Danielson:** The scope of that attachment is provided in the one plan and would outline any genomic sharing. There are other NIH wide sharing policies that may apply, and you may be subject to. You will need to look carefully at the specific FOA, but most applications will now be submitting the data sharing and management plan. **Taunton Paine**: The GDS Policy will continue to apply to awards that propose to generate large-scale genomic data. The GDS Policy sets expectations that may go beyond the DMS Policy, such as identifying earlier timelines for sharing human genomic data and indicating the use of specific repositories. The DMS Policy will also apply to awards that are subject to the GDS Policy, and applicants that are subject to the GDS Policy will describe their plans for sharing data under the GDS Policy in the DMS Plan. For more detail, see: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-22-198.html
Can you advise on how to justify DMS effort that will be performed by key personnel with roles in study conduct as well as DMS tasks (in budget and in budget justification). Does the effort need to be broken out into separate categories (Reg budget justification and DMS budget justification) or is the intent for the DMS budget justification section to include only personnel or other expenses specific to DMS). For example, if the principal investigator (PI) also plans to handle data curation, etc., does the PI effort for those tasks need to be included separately in the section of the justification for the DMS budget?	**Cindy Danielson:** Due to the specificity of this question, we cannot offer a straightforward answer - we encourage you to reach out to sharing@nih.gov so that we can confer with others as needed to answer this budget-specific question for your project.

EHLC Workshop Day 1: Questions and Answers	
What if consent (for human population studies) occurred before the policy takes effect? Can this be justification for not sharing human data? Or am I required to re-consent?	**Taunton Paine:** NIH has provided Frequently Asked Questions (FAQs) to the DMS Policy, including examples of potentially justifiable ethical, legal, and technical factors for limiting sharing to some degree. This includes reasons related to informed consent. See: https://sharing.nih.gov/faqs#/data-management-and-sharing-policy.htm?anchor=56549
Is the DMS Plan part of the Research Performance Progress Report (RPPR)?	**Cindy Danielson:** Compliance with the approved DMS Plan will be monitored at regular reporting intervals as part of the annual RPPR process. RPPR updates will be made to accommodate this.
Does the 2023 NIH data sharing plan apply for new grants or does it also apply to existing grants?	**Taunton Paine**: The DMS Policy applies to competing applications submitted for the January 25, 2023, receipt date and later dates. So, it will not apply to existing grants until the next competing applications.
How can we share any data that could be patentable or in the process of filing a patent?	**Taunton Paine:** The NIH DMS Policy recognizes that other NIH policies and other Federal laws, regulations, and policies might limit data sharing. Consistent with longstanding guidance, the filing of a patent application to secure intellectual property rights may justify a need to delay disclosure of research findings, as well as any scientific data underlying them, and a delay of 30 to 60 days is generally viewed as a reasonable period to allow for time to file a patent application if needed. However, scientific data that are not the subject of a patent filing or are precompetitive data that are not the subject of a patent application should be shared within the expected timelines. NIH ICOs will review the reasonableness of proposed limitations when they assess DMS Plans or approve updates.

EHLC Workshop Day 1: Questions and Answers	
Environmental researchers often use space-time linked data (e.g., air pollution/temperature/area demographics) that would be "indirect identifiers" for human subjects. With a long-enough time series (or in combination with other data elements), it is easy to figure out precisely where and when participants arise (akin to sequence alignment), even without direct geographic identifiers. Can NIEHS give examples for expectations for such spatiotemporal data that balance data sharing and risk? Environmental epidemiology often uses secondary data for which consent is waived.	**Taunton Paine:** Chris [Duncan] may have additional input, but NIH has provided supplemental information to the DMS Policy regarding protecting privacy that may help to address some of your questions: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-22-213.html Additionally, NIH acknowledges privacy as a potentially justifiable factor for limiting sharing of data, especially when options for mitigating privacy risks have been considered and would be insufficient: https://sharing.nih.gov/faqs#/data-management-and-sharing-policy.htm?anchor=56549
NIEHS had developed several examples plans - can we contribute them to the https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/writing-a-data-management-and-sharing-plan page?	**Christopher Duncan:** Yes - there is a mechanism for this to happen. Let's touch base offline to discuss.
When and where will the recordings be posted?	**Workshop Staff:** Recordings will be posted to the workshop website after the end of the workshop series.

EHLC Workshop Day 1: Questions and Answers	
It is still not clear the differences between the Resource Sharing Plan (still required) and this new DMS Plan. I would think that there is an overlap between them.	**Taunton Paine**: NIH requires all applicants planning to generate scientific data to prepare a DMS Plan that describes how the scientific data will be managed and shared. The DMS Plan is a separate attachment, and not part of the Resource Sharing Plan attachment. Resource Sharing Plans are not required in all cases, but only when other NIH-wide sharing policies apply (e.g., Model Organism Sharing Policy, Research Tools Policy) or when a specific FOA includes other NIH Institute and Center or program-specific requirements. Resource Sharing Plans relate to sharing as expected by the model organism sharing policy, the research tools policy, and other NIH Institute and Center or program-specific requirements as applicable. Resource Sharing Plans are submitted in a different section of the application and are reviewed differently. For more information on the model organisms or research tools policy, see: https://sharing.nih.gov/other-sharing-policies
Getting Started with Data Management **Nicole Contaxis, NYU Health Sciences Library**	
Will PubMed add a Medical Subject Headings (MeSH) tag for the Diagnostic and Statistical Manual of Mental Disorders, DSM, so PubMed users can see and search for papers with a DSM?	**Nicole Contaxis:** Every year the National Library of Medicine adds additional MESH terms, so I would reach out to the NLM with questions related to MESH
How do I find out whether my field has acceptable data standards? Is there a resource that you would recommend?	**Nicole Contaxis:** FAIRsharing.org is a great place to start
I'm confused - the mouse example seems like a failure of carrying out standardized data collection procedures. This is simply bad science. Does a data management plan detail data collection procedure?	**Nicole Contaxis:** An NIH-compliant data management plan includes six elements. While you will not be asked to talk about your data collection procedures explicitly, you will be asked what standards you plan to use, both for data collection and metadata creation

EHLC Workshop Day 1: Questions and Answers	
Excepting what I write in my data management plan, how do I incorporate standards when sharing data? Is this something the repository or journal asks for?	**Nicole Contaxis:** Start with the library and go to repositories that relate to your work and see what is and is not working. Some repositories require use of standards when submitting data, but it does depend on the repository. I would suggest investigating what repositories you will use and then working backwards from there
How can one replicate these teachings in Africa, especially, Nigeria?	**Nicole Contaxis:** All of the recordings will be available on the EHLC website after the workshop here: https://www.niehs.nih.gov/research/programs/ehlc/resources/index.cfm Please feel free to share. Also, you may consider connecting with the NIH Data Science for Health Discovery and Innovation in Africa (DS-I Africa) group https://dsi-africa.org/.
In your example of the Jackson Heart Study (JHS), do we need to repeat everything already publicly available in the DMS if we are using JHS?	**Nicole Contaxis:** As far as I understand, if you are using publicly available data, you should note that fact in your DMS and point towards where that publicly available data exists online.
As far as data sharing and the peer-review deadline, is the requirement to have the data submitted for sharing to the repository of choice, or fully released to the community through the portal? I am thinking about what role the data submit-to-release lifecycle plays in compliance.	**Taunton Paine:** Under the DMS Policy, NIH requires researchers to prospectively plan for how scientific data will be preserved and shared through submission of a DMS Plan. Shared scientific data should be made accessible as soon as possible, and no later than the time of an associated publication, or the end of performance period, whichever comes first. Researchers should indicate proposed timelines for sharing data in the DMS Plan and must comply with the approved DMS Plan including timelines for data sharing. Anticipated repository release timelines should be factored into the proposed data sharing timelines, and any unanticipated delays should be communicated to the NIH funding NIH Institute, Center, or Office.
Is there a type of file extension to store metadata that we should be aware of?	**Taunton Paine:** There is no file extension specifically for metadata, although some metadata standards do employ file types like XML

EHLC Workshop Day 1: Questions and Answers	
FAIRsharing.org Allyson Lister, University of Oxford e-Research Centre	
There are different levels of granularity; flow cytometry data for instance could be raw flow cytometry standard (FCS) files, so others can literally reanalyze data as they see fit, versus counts and calculated frequencies as they might be published; uncertain what needs to be shared.	What needs to be shared depends on your needs; FAIRsharing helps you find the resources that are relevant to your domain. What does your funder require? Or perhaps the journal publisher? We have a policy registry that may include the policy you need to align with, together with any standards or databases they list. You can also simply search the standards that we describe within your domain of interest (e.g., flow cytometry) to find the options that you can use.
For FairSharing.org, are the services available free of charge?	All our services are free. We are an academic project within the University of Oxford.
Introduction to the NIH Common Data Element Repository Robin Taylor, NIH/NLM	
Can this CDE concept be explained for a lay audience? I still don't understand what this is referring to or how it is relevant for me.	It may be more interoperable later. You might have a set of choices or a specific number. These would be difficult to interoperate. So CDEs would have both and can be later combined.
What if there are no NIH endorsed CDEs and the available CDEs conflict?	When I searched the repository, you may see similar CDEs. Sometimes there is some redundancy. But often, the CDEs measure the same thing but in different ways. So, I would say if you were unsure of what CDE to use, go back to your supervisor for how they would want that data collected.
Can more than one member of a team have access to your data form? I am thinking of a case were more than one staff member may need access or a postdoc leaves the lab…	Multiple people can have access to the same things in the CDE. Anyone can log in and see those things.
Are there CDEs for environmental data? I did not see any in your examples.	There are some CDEs for environmental data in the CDE portal. In addition, the NIEHS Disaster Research Response has also been publishing data collection tools and resources including CDEs through the site https://tools.niehs.nih.gov/dr2/ . In general, though, it would be nice to see more contributed CDEs for environmental health research.

EHLC Workshop Day 1: Questions and Answers	
Resources to Get You Started: DMPTool Maria Praetzellis, California Digital Library	
I see specific sharing plans that are available. Is the goal for us to use one of these or to use one of these to create our own?	The goal of the sample plans is to give people a flavor of what researchers have stated. It's to give people some ideas but not to give anyone a script to be used verbatim.
Is there an option to add this on a per project basis (as for a multi-project grant)? Or is this needed?	If I understand correctly, for a multi-grant project, you'd need to create DMPs for each proposal submission.
It doesn't look like the NIH form is available in DMPTool yet. Is there an ETA for when it will be available?	It is available right now. In the tool you will see the policy released on the 25th. It's currently available on the application. Once it goes into play, it will be the only available one.
Does NLM consult with end users or experts in vetting which CDEs are compelling enough to reach "NIH-endorsement" level, and conversely, are there incentives for researchers to volunteer new CDEs that reach that level?	The NIH CDE Governance Committee (GC) does not consult with subject matter experts about the CDEs submitted for consideration for endorsement; it is expected that subject matter experts were informed about the development and refinement of CDEs prior to submission. The GC members, who come from multiple NIH Institutes and Centers (ICs) and have significant experience with CDEs, review CDEs only according to the criteria set by the NIH Scientific Data Council: • Clear definition of the variable as a specified question and a permissible type, set, or range or answers. • Documented evidence of reliability • Human- and machine-readable format • Clear licensing and intellectual property status (prefer Creative Commons or open-source) • Recommended or designated by a recognized NIH body (ICO, NIH research-initiative working group, trans-NIH committee, etc.) NLM consults with end users to inform decision-making about the NIH CDE Repository platform and the CDEs made available there. CDEs are submitted at an institutional or research-initiative level, not at the individual researcher level. Researchers, project officers, and others may be motivated by the FAIR and reusable data that is a result of collecting research data using CDEs.

EHLC Workshop Day 1: Questions and Answers	
Can you please add the link to the NIH form in DMPTool?	The NIH template in the DMPTool can be accessed here: DMPTool_NIHTemplate
Introduction to NHLBI BioData Catalyst® (BDC) **Rebecca Boyles, RTI International**	
If someone is doing a method study, for example, and testing response rates to standard CDEs versus alternately worded (novel) questions, with various (novel) types of response, and they want to publish all these results, what site should they go to create the new CDEs that will be required?	NHLBI BioData Catalyst® can host data with CDEs or novel questions. There are several CDE efforts supported by NIH which each have their own governance structures. One critical resource for CDE collections is the NIH CDE Repository: https://cde.nlm.nih.gov/home.
Monarch Initiative: Fuzzy Phenotype Matching **Kevin Schaper, University of Colorado at Anschutz Medical Campus**	
What is the backbone technology that they are all using?	We are in some transition; we are rebuilding the graph. The big tech shift we have is moving to the Biolink model. The older graph was more purely resource data framework (rdf)-centric. We're moving to a property graph so that we can represent edge properties as defined in the Biolink model. The underlying technology is Neo4j.
KnowWhereGraph in a Nutshell **Krzysztof Janowicz, University of California at Santa Barbara**	
How is the KnowWhereGraph maintained and updated?	Datasets can be updated frequently, and some get snapshots. There are staging versions that guarantee the quality of the data.

EHLC Workshop Day 2: Questions and Answers	
Question	**Response**
The CEDAR Workbench **Mark Musen, Stanford University** ***and*** **10 years of ISA: Lessons Learned and Recent Developments** **Philippe Rocca-Serra, University of Oxford e-Research Centre**	
I am actively using CEDAR to develop a metadata template for work within our HHEAR Data Center. If I think I have found a bug, or have an issue to report with CEDAR, where can I report this? Do you have a public (or authenticated user accessible) issue reporting resource?	**Mark Musen:** I have listed the website where you can find the email list for comments and to report bugs (https://metadatacenter.org). The email address is cedar-support@metadatacenter.org.
We have a number of people who are filling out plans, and metadata is a new thing for them to be capturing and thinking about. I was wondering if you could each take a moment to discuss how researchers can use the tools you talked about	**Mark Musen:** Standalone data-management-planning tools and metadata-authoring tools are not effective unless used with data repositories. Right now, we are collaborating with generalist data repositories such as Dryad and the Open Science Framework (OSF) to integrate CEDAR directly into their data-accessioning software. **Philippe Rocca-Serra:** It depends on the experimental plan, and looking into better ways to capture that and make the most and move upstream would be a major improvement. I think it could bring better tools and make data collection in the lab easier. We need better integration with vendors as well to implement open standards.
The relationship with repositories is critical, and we need to see greater adoption of standards within them and wonder if you have any advice to promote this?	**Mark Musen:** NIH should fund this kind of integration as part of its generalist data repository initiative. NSF is supporting the integration of CEDAR with Dryad; OSF is interested in incorporating CEDAR as well. I think it would be helpful for NIH to set reasonable expectations for data integration and data FAIRness. Investigators need not only to put their data into repositories, but also to annotate their data so that they are FAIR.

EHLC Workshop Day 2: Questions and Answers	
The first presenter showed an example of how to enter metadata for one specific biosample from one individual. We are contemplating sharing longitudinal data collected from over 600 people over 10 years. Are there templates that allow for cataloguing of much larger datasets? Entering data participant by participant would not be feasible.	**Mark Musen:** Absolutely, we are working closely with the Human BioMolecular Atlas Program (HuBMAP), a large consortium supported by the NIH Common Fund. We are working to develop mechanisms to do perform upload of metadata and a validation strategy for those metadata. We view scalability in data annotation as an essential goal.
Because repositories do not capture all data, what do researchers do, what are alternatives?	**Mark Musen:** We work closely with groups who want to submit data. There are a lot of repositories that allow only limited metadata, and the question is what you do with that. My preference is to work with the repository. Most repositories are beginning to recognize that more metadata are better than fewer metadata, and that more metadata are necessary to make datasets FAIR. **Philippe Rocca-Serra:** You can make other packages. Generally, making the data readable and moving towards linked data will help this connectivity.

EHLC Workshop Day 2: Questions and Answers	
I just tried to import a large set of XML templates in CEDAR, and it failed. I am not aware that the tools I use (DistillerSR, for example) can import XMLs from CEDAR or other tools. It seems like there is a lack of compatibility and transferability of templates from one tool to another, regardless of how well they are structured. Is there a standard that can be established for exchanging templates? Building everything from scratch makes it hard to move from one tool to another.	**Mark Musen:** In the case of CEDAR, the templates are represented in standard JSON. We hope that a variety of tools will be able to use that same standard. The idea of rendering metadata in a standard underlying representation will make it possible to create metadata in a variety of ways using a variety of applications. We are interested in creating *ecosystems* for FAIR data, not in one-off tools. Attendee: The European Food Safety Authority (EFSA) is doing some work on data exchange standards, in which companies such as Distiller may be involved, that I think are intended to help solve this problem. Could be worth following up with EFSA on their plans (there is a report forthcoming, but I have not seen it yet). Or with me. Attendee: Thanks Paul. I can follow up with EFSA. If there is a standard way to represent, import, export, or exchange templates themselves, that would be huge in using approaches.
I have sometimes found that sample annotations are completed, but not interpretable in the context of a published report. For example, samples may be labeled 1 to 10, but lack information on how these identifiers relate to experimental groups. Should we require that published metadata be sufficient to reproduce fundamental conclusions?	**Mark Musen:** Yes, I think this is a real challenge. Most metadata are terrible, and we need to encourage investigators to do a good job of annotating their datasets, so metadata and publications can stand on their own. You need somewhere to go for reproducibility, and that should be the rich metadata describing datasets online. **Philippe Rocca-Serra**: We could see several tests formed on the meta data and there are several integration and reusability tests that challenge the FAIR principles.

EHLC Workshop Day 2: Questions and Answers	
With ISA creator no longer being supported, does the ISA team have plans to replace it with other tools that facilitate data collection in the more familiar spreadsheet format?	**Philippe Rocca-Serra:** There is an effort coming from Germany, a group has developed an interface that is integrated into excel. Another one is being developed, but is under wraps. Indeed, there are new components being developed and I also pointed to the nfdi4plant annotated research context (ARC) project. Please see: https://nfdi4plants.org/nfdi4plants.knowledgebase/docs/implementation/Swate.html.
In terms of the Data Management Plan and determining what metadata should be made available, what should be done for "small" data sets for which there is no established electronic repository?	**Mark Musen:** This is where the NIH generalist data repository initiative comes in.
Standard Terminology: Ontology Lookup Services. OBO Foundry. Specific Ontologies **James Overton, Knocean Inc**	
One of the main things is they use the same forms and structures so it's easy to integrate. What about if you want to integrate a format that is not found in the ontology?	The tools that I mentioned such as OBO can give you a head start. In general, I like working with tables and templates. Robot and the ODK contain various tools to take a spreadsheet into ontology terms. I think it shouldn't be too hard to take a small number of terms and fit them in. There's also the larger issue of mapping. There are also tools to support that.
For researchers new to the concept of standardized terminology, how do they decide whether to use a taxonomy, thesaurus, or ontology? Which is better to use for which scenario?	Starting small and building up is the right place to start. My first suggestion is to find a standard that does what you want. The ontology browsers are the correct place to start.

EHLC Workshop Day 2: Questions and Answers	
I have heard a lot about precomposed versus post-composed terms. From the perspective of matching existing ontology terms to a specific use case, the latter avoids a proliferation of highly similar terms. Is there a consensus emerging in the field on this topic?	It's still a hard problem. People have different needs and uses. If you really do mean the general term, you can say that. If you mean something specific, you can compose it. There's no consensus because people have different needs and concerns.
When you are searching for terms and find the same concept represented in different ontologies, what criteria do you use to choose between the ontologies? How do you grade the quality of one ontology over another? For example, many terms in the National Cancer Institute (NCI) Thesaurus are also covered in other ontologies like the Experimental Factors Ontology.	My biased answer is that I like to pick OBO ontologies. I know I can start small and build outwards. I would rather use the OBO ontology that has its own domain and scope.
Using Ontologies: Tutorial on Finding and Requesting Ontology Terms **Nicole Vasilevsky, Critical Path Institute**	
Re-ensuring the ontology terms are open, is it the case that if the ontology is open, then all terms are open? If not, does the license clearly indicate that some terms are not? It would be a shame to invest oneself in each ontology and run into a hitch with licensing.	**James Overton:** I am not a lawyer, but OBO operates on the basis that the (open) license on the ontology file applies to all the contents (i.e., the terms) in that file. **Nicole Vasilevsky:** My understanding is that is has an open license. I would say if its open, it is all open.

EHLC Workshop Day 2: Questions and Answers	
Just clarifying: In OBOFoundry, individual ontologies are meant to be non-overlapping in topic (and/or use case), but they are all interoperable (with each other)?	**James Overton:** Yes, OBO projects should be both orthogonal and interoperable. In practice we sometimes fall short, but that is the goal. OBO ontologies use an open-source development model. The short answer is the people who have permissions to accept changes into the version control repository make the final decision. In practice, we strive for consensus among the developers and the users of the ontology. It works better in practice than you might think! But yes, there can be conflicts. **Nicole Vasilevsky:** The intention is for them to be all interoperable, but this is not always what happens. There are some cases where they are not interoperable.
With so many ontologies available, how does a researcher know if an ontology is the correct one to use?	**James Overton:** Yes, it can be hard to decide which ontology to use. In general, it can be hard to pick one standard among other overlapping standards. Nicole and I gave our advice, but you must make a judgement call. It is not necessarily an "all or nothing" choice with a "wrong" answer, though. There are many options for mapping, translating, or migrating from one ontology to another. Adopting any standard will help you add structure to your data, and structured data is much easier to work with in automated, scalable ways. **Nicole Vasilevsky**: I would start with the OBO ontology. There is never a correct one, but the best approach is to just decide and stick to that ontology. It may be helpful to think about what your collaborators are using.
Who makes final decisions re: changes to any given ontology?	**Nicole Vasilevsky**: It comes down to the ontology owner. Nothing is ever final, terms can be changed or deleted. The person who makes the decisions is the primary curator, but everyone can weigh in on the decision.

EHLC Workshop Day 2: Questions and Answers	
Apologies if I missed this in another talk/session, but if one is annotating data with an ontology, how does one give credit to that ontology (e.g., how do you cite)?	**James Overton**: In general, most scientific ontologies have a "launch" paper that they would like you to cite. For OBO specifically, our metadata registry includes a publication field for this purpose, and you can see that on the project page, e.g., the OBI "Publication" here https://obofoundry.org/ontology/obi.html When using a specific OBO term, it is enough to use the Persistent Uniform Resource Locator (PURL) term, which will lead back to the ontology and its attribution/publication/metadata/etc. **Nicole Vasilevsky:** The foundry webpage should have a field on how to cite it. Usually there is a paper you can cite, and if not just cite the repository itself.
Can you take a tangential step in the process and describe how exactly researchers would incorporate the ontology/terms. Would a researcher include this in the materials and methods section and the dataset?	**Nicole Vasilevsky:** You can use a spreadsheet to keep that for your personal use and submit it as supplemental data. You can use ontology terms in your methods and use the standardized labels. Using the preferred ontology terms helps with text mining.
But Standards Don't Exist for My Domain **Sierra Moxon, Lawrence Berkeley National Laboratory**	
What kind of support can I get for LinkML?	Our slack is on the OBO website. You can also attend our meetings.
Can you tell us what tools can use LinkML?	There's a wide variety of tools. You can use it to make schemas, script your data, among more tools.

EHLC Workshop Day 3: Questions and Answers	
Question	**Response**
Q&A Session after Session 1: Spotlight on Domain-Specific Resources	
PhenX Toolkit **Carol Hamilton, RTI International**	
You [Carol Hamilton] mentioned that you are tracking research or publications that mention PhenX. How are you searching for those uses or mentions? Is that simple searches of titles and abstracts or some other method?	A curation team tracks publications and looks for who cited PhenX and looks for two citations: 1) citing the concept 2) citing the source that it came from or the citation that was used. Funding organizations are also tracked, for example, information on when the FOAs were released and associated publication funding statements. Carol explained that a curation process is in place and that there are efforts underway to try to semi-automate that process. • The participant shared interest in semi-automation of the process and believed the research community would be interested as well in that methodology. The participant asked Carol to share information on this process when it is available. • Carol Hamilton mentioned trying to get permanent identifiers associated with PhenX protocols (e.g., digital object identifiers [DOIs]), so that it is easier for investigators to precisely cite use of PhenX protocols and to be credited for using protocols that promote data sharing. She added Steve Edwards was instrumental in getting the PhenX publications and citation analysis process up and running.

EHLC Workshop Day 3: Questions and Answers	
Chemical Identifiers: Capabilities, Connections and Contradictions **Antony Williams, U.S. EPA**	
Question for Antony Williams: What resource(s) link chemicals and their mass spectral peaks?	There are multiple resources online that will link chemicals to spectral data. I will provide examples in the form of direct links. The sizes of the databases differ in terms of the number of chemicals covered. MassBank: https://massbank.eu/MassBank/RecordDisplay?id=MSBNK-UFZ-UA002901&dsn=UFZ NIST Webbook: https://webbook.nist.gov/cgi/cbook.cgi?ID=C1912249&Mask=200#Mass-Spec PubChem: https://pubchem.ncbi.nlm.nih.gov/compound/Atrazine#section=Spectral-Information Also, the Human Metabolome Database (https://hmdb.ca/), MassBank of North America and many others.
MIATE: Supporting Standardized Collection of Metadata for In Vivo Toxicology Research **Rance Nault, Michigan State University**	
Who can submit data, what's the process of approval?	Tox Data Commons are early in development. The Superfund Research Center is collecting metadata in a standardized way before publication. Rance explained the goal is to demonstrate this works and demonstrate the value before it is opened to the public, taking a stepwise approach. Currently it is a Michigan State University (MSU) resource, but hopefully it will open in the future to the public with formalized approval and depositing processes for the long term.

EHLC Workshop Day 3: Questions and Answers	
Are the Gene Expression Omnibus (GEO) templates available/promoted through GEO? Can you clarify the idea behind Tox Data Commons - will it hold data or only metadata and link to the data? Who can deposit their data there?	The GEO templates are not available or promoted directly through GEO. The template takes advantage of the ability to add custom metadata fields in the upload form. In our planned implementation of the ToxDataCommons, it is possible to hold both data and metadata but there are costs associated with the storage of large datasets such as sequencing data. For these types of data, the goal is to link out to external established repositories while allowing for the storage of smaller data (e.g., body weights) in the commons.
MOLGENIS Catalogues: For Multi-Center Cohort Studies and Beyond **Morris Swertz, University Medical Center Groningen**	
How would one go about piloting, prototyping, or doing some type of integrated analysis with MOLGENIS?	One could use the catalog as a tool to create a standard data dictionary, which ideally might be an existing one (e.g., PhenX). Derived from that is the generated data dictionary that one could use to structure data in DataShield, which does the federation on the level of individual statistical functions. This only works if all the cohorts have the same data structure. The difficult part is getting a DataShield installation into the cohorts as you typically need a local IT person to install the DataShield software. In cohorts, it also important how they define and harmonize the data.
Do you have examples where researchers have harmonized their data and published research findings?	The Life Cycle project (De Moira et al, Eur J Epidemiol. 2021 May;36(5):565–580. doi: 10.1007/s10654-021-00733-9) has information on setting up a catalog. A more specific meta-analysis using DataShield was done using the EU Child Cohort Network (de Moira et al, J Allergy Clin Immunol. 2022 Jul;159(1):82–92. doi:10.1016/j.jaci.2022.01.023).
Did those examples inform the evolution of your tool? If so, how?	Every next conversion has different research questions, needs, and approaches. We have about 400, including a project for the European Agency. We have a rich catalogue, but we still need people to fill in these details and it can be overwhelming. Look at instances of the catalogue, delineate which details are needed, and then reach out to people to fill them in. Morris provided a link to a specific research project: https://pubmed.ncbi.nlm.nih.gov/35150722/

EHLC Workshop Day 3: Questions and Answers	
Work in Progress: The Development of a Semantic Resource Listing for EHS Data: Harmonization Use Case Jeanette Stingone, Columbia University	
Can you share the link to the tool in the chat?	Spreadsheet link is at: https://docs.google.com/spreadsheets/d/1E7xbqV_XQM8Vo1MX1c0dV9P8LEowBfxVnOAMqJ1Vwm4/edit#gid=1846838588
Some resources shared previously have common terminology, and you have been involved with mapping terminology to ontologies. Do you have recommendations or suggestions for making that process easier?	It depends on your purpose. Mapping sometimes is to see how data is harmonized. Sometimes people want to map to go beyond a CDE. Jeanette highlighted there are gaps in the terminologies and ontologies available. My work focuses on taking data and being able to harmonize them and pull data across studies. For best practices, first don't start from the first study. Start from what you think is the standard terminology. Many initiatives like HHEAR have a preferred terminology/ontology. For example, Biolink and translator also have preferred terminologies for specific topics. This is typically where we look for terms. Because sometimes one term can be found in multiple places.
How would I use the spreadsheet to find ontologies related to human health risk assessment?	This use case has been narrowed down to human epidemiology studies, and what you mentioned would hopefully fall under there and fit within one of the domains or subdomains. Not only would the resources be used, but could this process be used for risk assessment? For human epidemiology studies, for different use cases, different tools may be needed or used. There may be areas that are covered and others that haven't been covered. For example, dermal exposure could fall in multiple places. This loop diagram could be used to find semantic resources, does the semantic resource provide quality information for mapping, and move forward from there. • Participant encouraged emailing this group sharing the interest in expanding domains. • Jeanette Stingone stated it would be helpful to hear about expansion around certain domains and gaps. • Participant stated there would also be some exposures and terminology explaining domains that may overlap and would need to be defined.

EHLC Workshop Day 3: Questions and Answers	
A participant suggested adding a field under the source of a chemical (air, food, water, etc.) to distinguish chemicals studied from endogenous exposures/pathways, versus only from exogenous exposures/pathways, versus from both. There are thousands of chemicals of environmental concern such as carbon monoxide (CO), nitric oxide (NO) and hydrogen sulfide (H_2S) that also are produced endogenously 24/7 and more in response to environmental and physiological stimuli	**Jeanette Stingone** agreed with this as a great point and encouraged this be added to the Division of Health for Unaccompanied Children (DHUC) Mural Board.