NIEHS Superfund Research Program White Paper: Enhancing the Integration, Interoperability, and Reuse of SRP-Generated Data Through External Use Cases

Executive Summary

The National Institute of Environmental Health Sciences (NIEHS) Superfund Basic Research and Training Program (SRP) funds diverse research projects spanning biomedical, environmental science, and engineering disciplines. These projects produce a wealth of data. SRP wants to apply this data to answer different questions related to connections between exposures to hazardous substances and health.

In 2019, SRP facilitated science-driven collaborative projects to enhance data integration, interoperability, and reuse. These collaborative projects were based on "use cases" designed to clarify how data management and data sharing could advance research.

Collaborators pursued rigorous research questions and identified current limitations to data management efforts for the SRP. Overall, 19 projects utilized more than 50 datasets, representing multiple science areas. The datasets came from SRP-funded research projects, external collaborators, and state, local, and federal sources. The teams reported their experience and progress during a virtual showcase event in February 2021.

This white paper describes each use case funded by SRP over the two-year period, including innovative approaches to combining disparate datasets and creating user-friendly tools. It also describes the challenges teams encountered, such as inconsistent metadata standards, lack of existing ontologies, and data security, and their recommendations to inform best practices for moving forward.

Introduction

To meet its mandates under the Superfund Amendments Reauthorization Act, the National Institute of Environmental Health Sciences (NIEHS) Superfund Basic Research and Training Program (SRP) funds diverse research projects that span biomedical, environmental science, and engineering disciplines. These projects include developing:

- Advanced techniques for the detection, assessment, and evaluation of the effects on human health of hazardous substances.
- Methods to assess the risks to human health presented by hazardous substances.
- Methods and technologies to detect hazardous substances in the environment.
- Basic biological, chemical, and physical methods to reduce the amount and toxicity of hazardous substances.

Together, these projects produce a wealth of data. SRP seeks to apply this data to gain a more comprehensive understanding of relationships and connections between exposures to hazardous substances and health.

Within any SPR research center and across the entire program, the transdisciplinary nature and coordination of research projects well positions teams to:

- Apply previous findings to new scientific endeavors.
- Accelerate the pace of research by incorporating data across disciplines.

To be most effective in reaching common goals and meeting program mandates, SRP researchers who generate and apply research data should more closely follow the FAIR Data Principles.

The FAIR Data Principles are a set of guidelines for data to be Findable, Accessible, Interoperable, and Reusable.

- Findable data can be easily discovered by both machines and humans and are assigned unique and persistent identifiers.
- Accessible data are available and obtainable.
- Interoperable data use similar formats, language, and vocabularies.
- Reusable data are described sufficiently with robust metadata so experimental conditions can be replicated and understood while being shared with the least restrictive usage license possible.

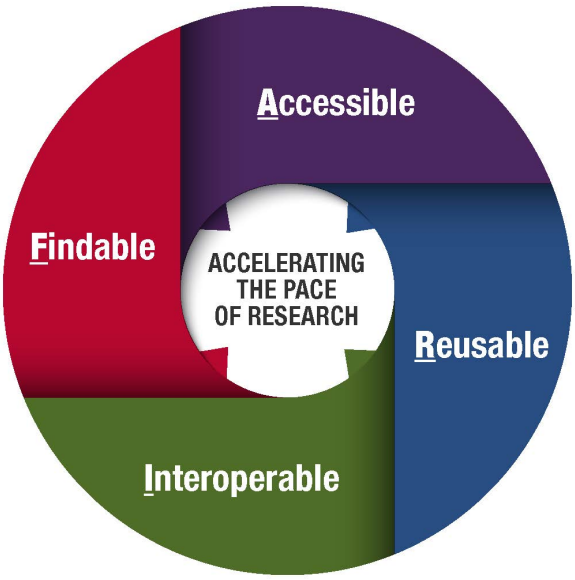

Figure 1. The FAIR Data Principles dictate that data should be Findable, Accessible, Interoperable, and Reusable. By increasing the FAIR-ness of data, SRP grantees can accelerate the pace of research and uncover new insights.

The FAIR Data Principles emphasize robust data management to support scientific discovery (see Figure 1).

However, integrating datasets is challenging, with both known and unknown barriers. Before data generated from individual research projects can be integrated and reused, researchers need to have harmonized workflows, consistent data stewardship, and protocols to make data shareable as it is collected.

In support of FAIR data, the National Institutes of Health (NIH) Data Management and Sharing policy was released in October 2020 to promote the management and sharing of scientific data generated from NIH-funded research. This policy will go into effect in 2023. Implicit in this plan is data stewardship, which is also a key component of the NIEHS strategic plan and related initiatives to enhance data management and sharing, so data generated from NIEHS-funded research grants can be discovered and re-used for downstream investigations (Wilkinson et al. 2016). It will be accomplished through improving the capture, storage, organization, management, integration, presentation, and dissemination of digital biomedical research data.

SRP's data sharing goals are in alignment with those of NIH and NIEHS. SRP has long supported data science activities (see Figure 2). A 2015 Risk E-Learning webinar series set the groundwork for exploring challenges and opportunities for integrating datasets to solve complex environmental health problems. An initial step towards integration was to make SRP data more Findable. For this reason, the SRP began posting data sets on its website, connecting datasets generated from SRP funding with a description of the respective research projects from which the data were generated.

In 2019, SRP facilitated science-driven collaborative projects to foster SRP-generated data integration, interoperability, and reuse. The opportunity allowed all SRP-funded centers to improve their data management, analysis, and sharing activities within their Centers.

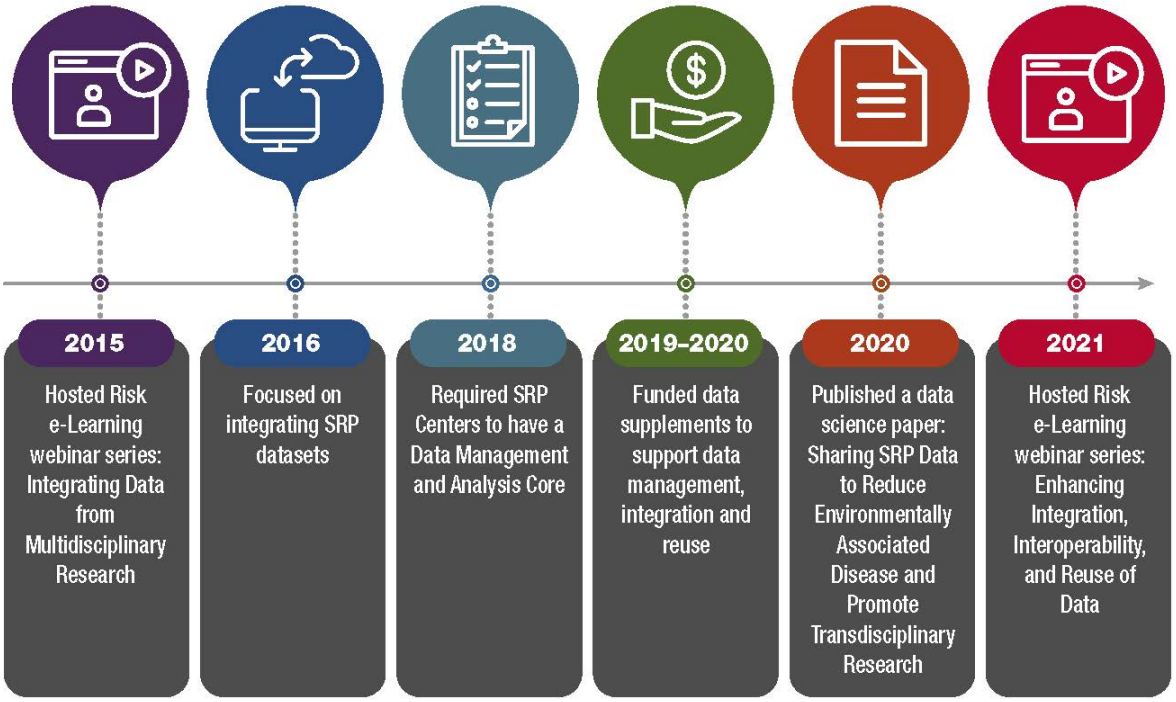

Figure 2. SRP has long supported data science activities. As research evolves in new directions, SRP is committed to pushing the boundaries of SRP science through data sharing and integration, and this is an ongoing and ever-evolving effort.

The opportunity was also designed to better understand the data sharing needs across the program. To accomplish this, SRP Centers could collaborate with another SRP Center, as well as an outside institution, to develop "Use Cases" that utilized data science approaches to advance the interoperability and reuse of diverse and complex SRP data streams (see Figure 3). As part of the Use Case, collaborative teams identified current limitations for data sharing and data interoperability and strategies to address them. Their goal was to inform future data management efforts for the SRP and to inform best practices for moving forward.

These Use Cases brought together researchers exploring similar research questions from different perspectives or disciplines, such as combining biomedical and environmental data or research in animals and humans, to reveal new insights.

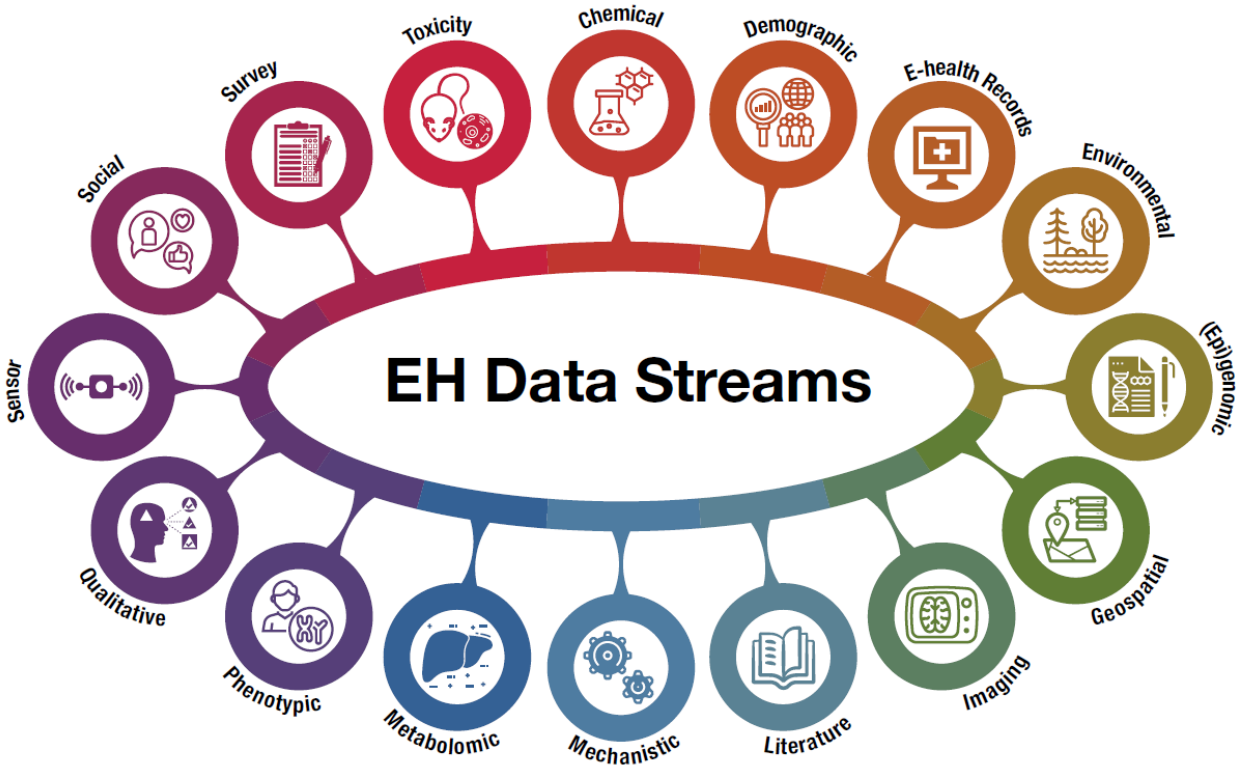

Figure 3. SRP-funded researchers generate a wealth of complex environmental health data spanning diverse data streams, posing a challenge for data integration.

Teams were expected to work closely with experts in data science to identify existing resources to advance FAIR-ness of SRP datasets. This included openly sharing their resulting digital products, including depositing data and metadata in public repositories and sharing of software code in public repositories.

SRP supported 19 Use Cases representing a diverse range of data, research questions, and disciplines. Teams starting at various stages along the spectrum of readiness for data interoperability (see Figure 4) worked together to set the groundwork to answer complex research environmental health questions that individual groups could not tackle alone.

An important goal to inspire these 19 teams by sharing lessons learned and identifying opportunities and challenges for data integration. The teams reported back on their experience and progress during a virtual showcase event, February 18 – 19, 2021.

A description of each Use Case is presented below, along with challenges encountered and recommendations for increasing the FAIR-ness of SRP data.

Use Cases

Leveraging Data Across Human Populations

Arsenic Mass Balance: Integrating Environmental and Biomarker Data across Diverse Populations

Arsenic, naturally found in earth's crust, is known to cause a variety of health problems in humans. Health risk estimates are currently based on drinking water exposure but, depending on the location, other sources are also relevant, including food and potentially dust and air, for example, in regions where inorganic arsenic is a common component of mining waste.

A team of researchers from Columbia University, the University of California (UC) Davis, and the University of New Mexico (UNM) SRP Centers worked together to understand how comparing exposure and excretion across populations can create a more complete picture of potential sources of arsenic.

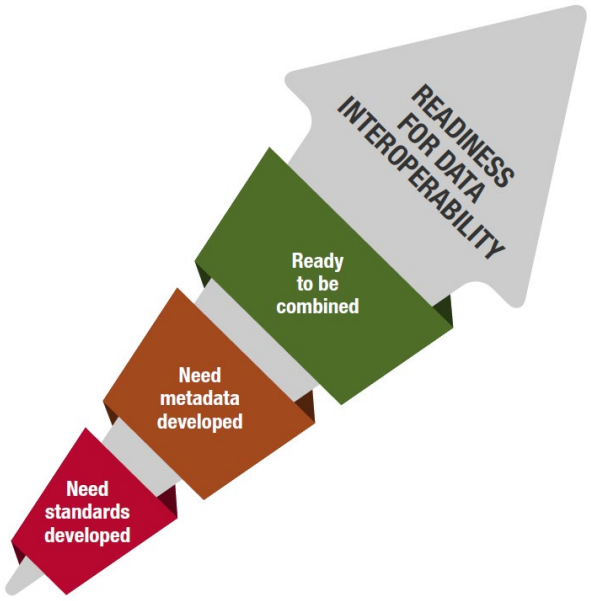

Figure 4. SRP-funded Use Cases began at different points along the spectrum of readiness for data interoperability.

These centers explore the effects of arsenic from different sources in populations in Bangladesh, Chile, and the Navajo Nation in the U.S., respectively (see Table 1). The Use Case sought to collectively analyze arsenic measurements in biological samples, like urine, and environmental samples, like water and dust.

Each center first created a searchable data dictionary, including key names, definitions, and attributes about specific data elements using the same format. The team then integrated the information into one shared data dictionary leveraging the Medical Subject Headings (MeSH) thesaurus to standardize their vocabulary. The team also used the Ecological Metadata Language to standardize their metadata. According to the researchers, the shared data dictionary had to account for different limits of detection for different instruments or different units, depending on the laboratory and analytical instruments that were used by the different groups.

Using their data dictionary, the team successfully harmonized data across the three cohorts allowing them to evaluate the relationship between concentrations of arsenic measured in environmental and biological samples (see Figure 5). Using a mass-balance approach, which dictates that the amount of arsenic entering the body should on average equal the amount that exits the body, the team was able to assess the contribution of different sources of environmental arsenic to levels measured in urine.

Deviations from mass-balance between arsenic intake and excretion led the team to consider sources besides drinking water, some which had not been considered previously. For example, secondary drinking water wells and rice were important factors for the Bangladesh cohort and house dust had to be considered for the Navajo Nation.

The team also learned that utilizing data visualization tools, such as scalable vector graphics (SVG), allowed it to adjust and scale data while maintaining data quality and integrity, without needing to access the protected data.

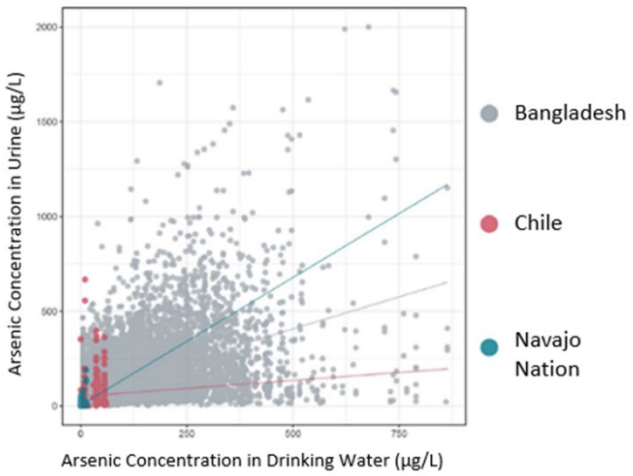

Figure 5. SVG showing a comparison of arsenic in urine as a function of arsenic concentrations in the primary drinking water source in Bangladesh, Chile, and New Mexico.

The Use Case also sought to make their data accessible and reusable by sharing their data, metadata, and analytical code. However, they encountered significant challenges related to data ownership and privacy. For example, data collected by UNM contains private information that is owned by the Navajo Nation. Due to privacy restrictions, raw data is only stored locally at each center. To address this limitation, the researchers shared their analytical code via GitHub. Security and privacy measures that balance public metadata and private health information are under development by the groups.

Through this collaboration, the Use Case team learned the importance of maintaining open lines of communication, utilizing clearly defined ontologies, and converting data to SVG. It allowed the researchers to get a broad picture of how to integrate environmental and biological samples to assess the contribution of different sources of arsenic internal dose. A publication describing results from this analysis is forthcoming.

Table 1. Existing datasets on arsenic across populations utilized by the Columbia, UC Berkeley, and UNM use case.

Institution	Study population	Biological samples (n)	Environmental variables (n)		
		Urine	Water	Dust	Air
Columbia	11,224 adults from Bangladesh	11,226	11,751	-	-
UC Berkeley	630 adults from Northern Chile	610	610	-	585
UNM	629 adults from Navajo Nation	619	619	619	-

Data Harmonization Across SRP Pregnancy and Birth Cohorts

Adverse pregnancy outcomes, like preterm birth and low birth weight, are a significant global public health challenge. Rates of adverse pregnancy outcomes are higher near hazardous waste sites or other sources of environmental pollution. For example, Native American communities in the Southwestern United States have concerns that uranium and arsenic exposures arising from abandoned uranium mine sites have increased the prevalence of metal-associated diseases including immune dysfunction.

Collaborators from the UNM, Dartmouth College, and Northeastern University Puerto Rico Testsite for Exploring Contamination Threats (PROTECT) SRP Centers investigated if their biomonitoring, demographic, and environmental data could be integrated across three populations to better understand the effect of environmental exposures on birth outcomes. By combining data from three different cohorts (see Table 2), the team sought to increase the statistical power of their analysis and enable exploration of a broader set of research questions.

The PROTECT SRP Center investigates the association between exposure to phthalates and preterm birth among pregnant women in Puerto Rico using the PROTECT cohort. They also collect information on other birth outcomes and cytokines, cell signaling molecules that are important regulators of immune response. The Dartmouth SRP Center is exploring the relationship between various prenatal exposures and birth outcomes, such as reduced birth weight, fetal growth restriction and gestational age using the New Hampshire Birth Cohort. They also have data on cytokines in pregnant women. The UNM SRP Center explores the link between prenatal exposures and birth outcomes within the Navajo Birth Cohort Study. They also assessed the potential for a dietary zinc supplement to reduce the toxicity of exposure to metal mixtures, including arsenic, in Thinking Zinc: A Study of Zinc Supplementation to Ameliorate Adverse Effects of Mine Waste Exposure on the Navajo Nation by examining changes in cytokine profiles and other outcomes in study participants.

The team focused on harmonizing key exposure data that would enhance the ability of other human biomonitoring studies to evaluate health endpoints. Specifically, they included biomonitoring results for arsenic across all cohorts and covariates such as fish consumption and socioeconomic status.

The team planned to create a data and methodology infrastructure that could serve as a foundation for current and future studies looking at common toxicants across populations, and at common outcomes of concern across contaminant classes and across populations. They aimed to integrate exposure metadata collected across the three cohorts and to develop a secure web platform to explore associations, such as between exposure biomarkers and outcomes at birth. Their approach was designed to determine the variance introduced in analyses by differences across populations, laboratory methods, classes of toxicants, and collection protocols. These are all critical to interpretating results and developing protocols to standardize the process.

To harmonize and integrate their data, the team first evaluated each cohort's data dictionaries to map and align the common variables. Identifying opportunities for harmonization was challenging because data was collected and stored differently by each center. Their expanded and harmonized data dictionary allowed them to combine data across cohorts. One example of their efforts is the development of innovative methods to enable harmonization, such as adjustments to urinary dilution across cohorts using either specific gravity measurements or creatinine levels. They also explored using novel machine learning methods to impute missing data.

Given data privacy issues associated with data from the Navajo Nation METALS cohort, the team needed to develop a data analysis framework that would be securely hosted by UNM. They leveraged several open-source tools, including Django, a web and Python-based analytical tool, an application gateway called NginX, and Docker as a containerization software. Their web-accessible secure processing platform can facilitate a wealth of new and future scientific discoveries as well as potential cost savings. This work sheds light on new opportunities to enhance data sharing and accelerate the pace of research, even when working with sensitive populations.

The platform has since been used to perform several statistical analyses, graphics, and project hypotheses. For example, the team investigated the association between exposure to arsenic during pregnancy using maternal urinary arsenic concentrations and birth outcomes, including gestational age, birth weight, and head circumference. Analysis can be performed within a single cohort, or across cohorts. Preliminary analyses have revealed that higher arsenic exposure during pregnancy is associated with lower birthweight, and the team is beginning to characterize the variation in this relationship resulting from levels of different types of arsenic.

The team is continuing their analysis of the harmonized data sets. They are working toward a joint paper on the arsenic study and are pursuing new questions on gestational diabetes. The team is also interested in sharing their methodology and tools with other researchers who have significant privacy challenges associated with their cohorts. The team's tools are open-source and hosted on GitHub.

Table 2. Existing datasets on the effect of environmental exposures on birth outcomes utilized by the Northeastern, Dartmouth, and UNM use case.

	Northeastern University	**Dartmouth College**	**University of New Mexico (METALS)**
Cohort	PROTECT Cohort	New Hampshire Birth Cohort Study	Navajo Birth Cohort Study; Thinking Zinc Study
Community	Northern Puerto Rico	Rural New England	Navajo Nation (Indigenous)
Questionnaire Data	Demographics, socioeconomics, behavioral, medical history, diet, maternal stress	Demographics, socioeconomics, lifestyle, medical history, diet, supplement use, occupation, drinking water, and other exposures	Demographics, socioeconomics, diet; home construction; occupational information; activity and resource use; drinking water source; vitamin supplement use
Chemical exposure data	Phthalates	Arsenic and other nutrient and toxic elements	Uranium and mixed metals from mine waste
Biological and environmental samples	Urine, blood	Toenail clippings, urine, drinking water	Urine, blood, drinking water
Health Outcome data	Gestational age, birth weight, birth length, head circumference, birth anomalies/ defects, type of delivery, certain cytokines	Gestational age, birth weight, birth length, birth head circumference, birth anomalies/defects, type of delivery, Apgar scores, maternal/infant infections, labor course; certain cytokines	SRP outcomes on adults in the intervention include cytokines, lymphocyte profiles, DNA damage, antinuclear antibodies (ANA)
No. of Participants	1,450+ pregnant women enrolled with 1,200+ live births to date	2,010 pregnant women with urinary metals assay results collected at ~24-28 weeks of	780 pregnant women in birth cohort test case, cytokines, and other

	Northeastern University	Dartmouth College	University of New Mexico (METALS)
		gestation with 1,877 born as of date of the dataset compilation	outcomes on ~200 to date, Zinc trial target 100
Data Dictionary	PROTECT data dictionary	New Hampshire Birth Cohort Study data dictionary	The database for the zinc study is still in development: birth cohort information is available through data managers directly

Arsenic Epigenetics META: towards a Meta-analysis of Epigenome Data on Arsenic

A team of researchers from the UC Berkeley and Columbia University SRP Centers worked together to enable meta-analyses of multiple Epigenome-Wide Association Studies (EWAS) related to environmental arsenic exposures. Epigenetic changes alter gene expression without directly altering DNA sequences and might serve as biomarkers of environmental exposures. EWAS use genome-wide assays of epigenetic marks, such as DNA methylation, to identify associations between phenotypes or exposures and epigenetic variation across the genome. These studies provide unique insight into the role of the environment on human health, but most studies of arsenic exposure to date have worked with small sample sizes and utilize diverse data processing and analytical methods yielding different results.

The team aimed to pool EWAS studies in two different populations to determine if the influence of arsenic exposure on the epigenome is generalizable across study populations and tissues. Arsenic-related epigenetic dysregulation may provide information about biological pathways linking arsenic to health outcomes and provide a biomarker of previous exposure and disease risk.

Existing datasets included high dimensional DNA methylation data from adults exposed to arsenic in Bangladesh and Northern Chile (see Table 3). Variables included historical arsenic exposure and biomarkers, and demographic characteristics including sex and age. Epigenetic data was generated from distinct population's DNA from different tissues, such as blood and buccal cells. Compatible Illumina platforms were used to generate the epigenetic array data, so the team hoped to develop a harmonized pipeline, pool results from individual EWAS across two different continents for meta-analyses, and preserve data for future use (see Figure 6).

They established consistent classification of exposure across datasets by standardizing genomic annotation of results and implementing quality control and data preprocessing steps to facilitate integration. Pre-processing included data normalization, quality control, and cleaning. They standardized these steps across SRP centers by working collaboratively via GitHub.

Specifically, they created a workflow and protocol beginning with raw image data files obtained from the DNA methylation array technology. Each center performed data processing and conducted EWAS locally, and center-specific code was deposited to GitHub. The team leveraged R packages available through Bioconductor (Gentleman et al. 2004), a free, open-source, and open-development software project to facilitate reproducible research.

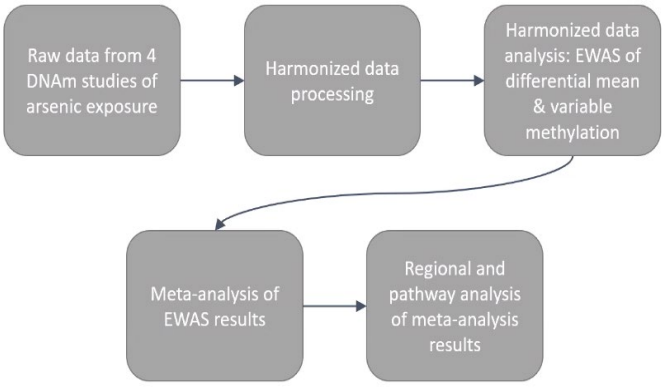

Figure 6. Data processing and analysis pipeline used by the UC Berkeley and Columbia University Use Case team. (Image courtesy of UC Berkeley and Columbia University SRP Centers)

In individual EWAS, the team did not find any common differentially methylated positions. Meta-analyzing their results increased statistical power to identify significant common findings. For example, their meta-analysis identified three differentially methylated positions and nineteen differentially variable positions. In KEGG biological pathway analysis, differentially methylated and variable positions in the genome were related to pathways with potential biological relevance to arsenic exposure like one-carbon pool by folate. One-carbon metabolism is responsible for synthesizing the methyl donor in arsenic metabolism, a process which facilitates urinary excretion and reduces arsenic toxicity.

Their data can only be stored locally at each SRP center due to confidentiality. However, complete results from individual EWAS are available on GitHub and will be uploaded to an Open Science Framework (OSF) repository, to facilitate comparison with other EWAS. The code is available on the project's GitHub, allowing others to utilize it for their own data and the team's written protocol can be used as template for other collaborations.

A publication describing their meta-analyses was published in Environmental Health (Bozack et al. 2021). The team plans to be involved with the development of protocols for epigenetic data sharing and shared practices to ensure transparency and maximize collaboration and data reuse. They hope this can be accomplished in the future by creating an Epigenetics Consortium of Environmental Exposures and working with existing consortia like the successful NIEHS Pregnancy And Childhood Epigenetics (PACE) consortium.

These efforts have led to an ongoing collaboration between UC Berkeley and Columbia SRP Centers around arsenic-induced epigenetic dysregulation. For example, they held a virtual symposium to facilitate and build on their collaboration. They plan to explore additional research questions, such as investigating chronic versus acute arsenic epigenetic signatures and determining the reliability of epigenetic changes as biomarkers of arsenic exposure. They also plan to expand this project to other cohorts.

Table 3. Description of existing DNA methylation datasets from cohorts exposed to arsenic at different life stages utilized by the UC Berkeley and Columbia University use case.

Institution	Dataset Description
Columbia	Data from 80 participants from the Health Effects of Arsenic Longitudinal Study (HEALS) cohort study. Study participants were classified as having low Arsenic (As) exposure and high As exposure based on drinking water As concentrations. Datasets include DNA methylation data (Illumina's 450K array or the HumanMethylationEPIC BeadChip 850k), extensive biomarkers of As exposure (drinking water As, urinary As from multiple time points, blood As, As metabolites in blood and in urine), data on potential co-exposures, demographic information, and nutrition information.
UC Berkeley	Illumina HumanMethylationEPIC BeadChip array data from 40 participants from an adult cohort in Northern Chile where half of the subjects had been exposed to very high levels of naturally occurring arsenic in drinking water as children. Hundreds of arsenic measurements in drinking water are available for over the past 60 years. Data from buccal and blood cell samples were available.

Integrating Omics Data Across Model Organisms

Two Use Cases: Integration and Analysis of SRC-Generated Cardiometabolic Syndrome Data Streams from Animal Models, AND Refining Species-Conserved Adverse Outcome Pathways (AOPs) of AhR-mediated Adverse Effects

Collaborators at the Michigan State University (MSU), University of Louisville (UL), University of Kentucky (UK) SRP Centers sought to combine data from laboratory-controlled animal studies to identify mechanisms by which exposure to Superfund contaminants promote cardiometabolic disease development and progression. Their objective was to combine data to identify better preventative and therapeutic intervention strategies and to improve the development of adverse outcome pathways (AOPs). AOPs are structured ways to represent biological events leading to adverse health effects designed to support greater use of mechanistic data in risk assessment and decision making. In closely related work, a second Use Case for the MSU and the University of Iowa SRP Centers focused on characterizing the molecular mechanisms of AhR-mediated toxicity, a signaling pathway that regulates biological response to chemicals, and refine AOPs through data integration and reuse. Given the common goals and models used, the teams combined forces to tackle the two Use Cases together.

The teams aimed to integrate data on mouse toxicology experiments with different study designs, including transcriptomics (RNAseq), proteomics, metabolomics, and clinical chemistry data (see Table 4). While there were many existing repositories for individual types of data, such as the Gene Expression Omnibus (GEO) and the Metabolomics Workbench, none were designed for all the data types they needed to combine to capture the complexity of animal toxicology experiments.

As a first step, they evaluated the Metabolomics Workbench repository to assess gaps in FAIR-ness to increase data sharing. Here, they implemented the mwtab Python library and package (Smelter and Moseley 2018), available on GitHub, and the Python Package Index (Powell and Moseley 2021) that provides access to all files in the mwTab and JSON formats and access to all of the publicly available datasets. They developed a set of validation methods and found that 5.5% of analyses could not be parsed, meaning that some data could not be converted to the desired uniform format. They also

identified consistency errors with the same data in different formats, such as between mwTab and JSON. Delving in deeper to consistency issues, they evaluated mwTab formatted files more closely and found instances of analyses lacking raw, experimental, or metadata, and inconsistent use of field names across datasets that could prevent comparisons or metanalyses.

One contributing factor the team noted was that data depositions have quadrupled in the last three years, resulting in an increased number of datasets being deposited, processed, and made available. One of the major issues the team identified as a barrier to reusing data or conducting metanalyses is inconsistent use of field names across datasets. For example, they found many mwTab files contain field names that could not be matched to commonly used fields, which would make it difficult for anyone to use the data. Therefore, the team is implementing field harmonization methods directly into the mwTab Python package.

The team found that data in different repositories were often missing essential metadata, such as which species, strain, or sex was involved in the toxicology experiment. They noted many of these repositories were designed with a broader scope, making it difficult to require metadata that are unique to a smaller community. To address these challenges, the Use Case teams took a step back and focused on improving the consistency of metadata collection. The team identified key pieces of metadata that would be needed to standardize and reuse in vivo animal toxicology experiment data, called Minimum Information about Animal Toxicology Experiments (MIATE). A few examples of MIATE include specifying the format, ontology sources, and the units for each field name relating to the experiment, such as animal details, housing conditions, and exposure details. They created infrastructure that expanded upon the Tox Bio Checklist (Fostel et al. 2007) and made their MIATE publicly available.

Focused on data collection, they used the Investigation, Study, Assay (ISA) framework to standardize data and metadata collection using their MIATE checklist and to make their existing datasets more FAIR. They made the framework publicly available on fairsharing.org and GitHub, and datasets collected using the framework have been deposited into the Gene Expression Omnibus [GEO; GSE148339 (Nault et al. 2021), GSE167328 (Jurgelewicz et al. 2021), GSE171941, GSE171942, GSE178168]. They also developed a web-application for finding, accessing, integrating, and reusing datasets.

Building off this important foundation, the team hopes to generate new knowledge by integrating datasets and to continue to improve the FAIR-ness of the Metabolomics Workbench (Powell and Moseley 2021) and other in vivo toxicology datasets. They also hope to collaborate with more SRP partners to further develop and use their framework for MIATE. They plan to publish papers and develop educational materials to improve the how data are collected and reported moving forward.

Through their work, the Use Cases learned that many data repositories appropriate for common data types are broad and do not have metadata standards that are specific to animal toxicology experiments. The team suggested initiatives to reward submission of metadata and tools to make collecting and depositing data and metadata more user-friendly moving forward. They also stressed the importance of training on data standards and activities like problem-focused code-a-thons to help develop data collection tools.

Finally, another important outcome from this team was that encountering similar challenges around integrating related data spurred additional collaboration outside the individual Use Cases. Through meetings and interactions, the centers realized that it made more sense to combine efforts and tackle

the challenges together. Their work provides a strong example of one of the ultimate goals of the initiative -- to build a community of practice and sustainability.

Table 4. Existing datasets on cardiometabolic toxicity utilized by the MSU, UL, and UK use case.

Data Types	MSU Tetrachlorodibenzodioxin	UK Polychlorinated biphenyls (PCBs)	U of L Volatile organic compounds (VOCs), PCBs	U of Iowa PCBs
RNA-Seq/ Transcriptomics	(Nault et al. 2015) (Nault et al. 2016b) (Nault et al. 2016a) (Fader et al. 2017)	PCB126/LDLr-/-mouse/liver RNAseq unpublished	Unpublished liver data from Wahlang et al. (2014)	Gadupudi et al. (2018) Gadupudi et al. (2016a) Gadupudi et al. (2016b) Wu et al. (2016)
(Phospho)Proteomics			Hardesty et al. (2019a) Hardesty et al. (2019b) Unpublished liver proteomics from male & female mice treated according to Lang et al. (2018)	Included above
Metabolomics/ Lipidomics	Nault et al. (2016b) Nault et al. (2017)	Deng et al. (2019) Petriello et al. (2018c) Petriello et al. (2018a)		
Metagenomics (microbiome)	Stedtfeld et al. (2017)	Petriello et al. (2018c)	Unpublished data from Wahlang et al. (2016b) Unpublished data from (Lang et al. 2018)	
Metabolic phenotyping	Included above	Included above Wahlang et al. (2017b)	Included above	

Data Types	MSU Tetrachlorodibenzodioxin	UK Polychlorinated biphenyls (PCBs)	U of L Volatile organic compounds (VOCs), PCBs	U of Iowa PCBs
		Wahlang et al. (2016a) Wahlang et al. (2017a)		
Clinical Chemistry	Included above	Included above Petriello et al. (2018b)	Included above	Included above
Histopathology	Included above	Included above	Included above	
Eicasanoids				Included above

*Available in GEO (transcriptomics), Metabolomics Workbench/MetaboLights (metabolomics), or UCSD MassIVE (proteomics)

Integrating Population Genomic Data to Understand Mechanisms of Chemical Susceptibility and Resistance

Researchers from the Boston University (BU) and Duke University SRP Centers collaborated to better understand the underlying mechanisms controlling susceptibility versus resistance to hazardous chemicals by integrating population genomic data from multiple populations of killifish that differ in their chemical sensitivity. By comparing their data to similar data on genetic variation in rodent models and in humans, they hoped to enhance the use of wildlife as environmental sentinels and models for human health.

The team leveraged two parallel projects exploring the genetic mechanisms underlying the evolved resistance to polycyclic aromatic hydrocarbons (PAHs), based on studies from Duke, and polychlorinated biphenyls (PCBs), from studies out of BU, in multiple populations of Atlantic killifish inhabiting sites contaminated with high levels of these chemicals. Initial analysis of genomic data from both projects identified variation in the aryl hydrocarbon receptor (AHR) signaling pathway as one common feature associated with the differential sensitivity. The AHR signaling pathway regulates the biological response of animals, including humans, to some PAHs and PCBs.

The Use Case project sought to integrate the killifish genomic data to enable a deeper and more extensive analysis of the datasets. Their goal was to enhance the accessibility, interoperability, and reuse of these datasets while allowing them to compare their data with other rodent and human data related to genetic variations and chemical susceptibility and resistance.

The team aimed to integrate existing whole-genome sequencing data for 384 killifish in eight different populations (Reid et al. 2016) of fish from BU with sequenced genome samples from 288 killifish (Osterberg et al. 2018) in nine populations from Duke (see Table 5). These data include restriction site associated DNA sequencing (RAD-seq), whole-genome re-sequencing (WGS), and RNA sequencing (RNA-seq) data. They also used WGS data from a closely related species of Gulf killifish (Oziolor et al. 2019) and a new genome assembly for Atlantic killifish, totaling more than 1,000 killifish genomic datasets.

These datasets were available in several existing repositories, including the National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) and in Dryad. The team took advantage of the fact that these data were already in a standardized form and were quite interoperable, allowing them to focus on collecting and compiling all the data in one place and creating a harmonized bioinformatics analysis pipeline using standard, established methods and tools. Some of the existing tools included FastQC for quality checks, STAR to map RNAseq data, and Samblaster (Faust and Hall 2014) for identifying duplicates in WGS data.

They reanalyzed the data using the new killifish genome assembly, identifying improvements to the existing genome in the process. Then they loaded the harmonized data and associated metadata into the open-source genome browser software JBrowse Buels et al. (2016), which allows the data to be visualized and queried. They performed joint variant calling to better incorporate the different groups of data.

Their platform, freely available at SuperFunBase, allows users to look at a portion of the killifish genome and see human genes that perform the same function. It also allows users to summarize the 'omics data in different ways. For example, they can look at RNA-seq data and see genes that are expressed in particular populations and see where there are variations in gene expression across populations. According to the team, this tool has already been useful in identifying gene variants that may play important functions, such as relating to resistance or susceptibility to harmful exposures. It can also be used to predict specific changes in proteins that would result from these variants.

In addition to making all the underlying data for their platform available in SuperFunBase where they can be downloaded directly by users, the team uploaded their raw read data into NCBI, including GEO and SRA, and the updated killifish genome and gene annotation are available. They also constructed an Open Science Framework (OSF.io) project page to disseminate their work more broadly while promoting collaboration. All code written to produce the genome browser from source data is publicly available in a git repository on BitBucket.

The team was successful in identifying the best reference genome assembly and annotations, integrating data, and deploying their new platform. Data analysis using SuperFunBase to answer their original research questions is in progress. They are also in the process of collecting feedback from the user community to improve the accessibility of their tool and producing publications.

In the long-term, they hope to add population-specific epigenomic data to their analyses and link to complementary databases such as SeqAPASS or the Comparative Toxicogenomics Database. They also aim to develop a strategy to allow researchers to link SNV data in killifish with SNV data in animals and humans to look for similar genes across species, which could help further explain genetic susceptibility to hazardous substances. Additionally, they would like to integrate environmental and biomedical datasets to allow bidirectional, cross-species comparisons that will shed new light on the underlying mechanisms of gene-environment interactions.

Through their Use Case, the team learned that using existing data repositories with established standard data formats was incredibly helpful in moving their project forward. They also found joint variant calling to be particularly useful to improve the accuracy of variant calls by looking at data from all populations simultaneously.

Table 5. Existing datasets on killifish utilized by the BU and Duke use case.

		Killifish Genome Data	RAD-seq Data	RNA-seq Data	SNP Data
BU	**Repository/ Database**	NCBI BioProject Dryad			NCBI SNP Human Genome Variation Society NCBI ClinVar Ensembl
	Dataset Identifier	PRJNA269290 PRJNA323589 dryad.68n87 dryad.t2888			
	Associated Publications	Reid et al. (2017) Reid et al. (2016)		Reid et al. (2016) Oleksiak et al. (2011)	
Duke	**Repository/ Database**	NCBI BioProject			
	Dataset Identifier	PRJNA450424		Unpublished data	
	Associated Publications		Osterberg et al. (2018)		

Integration and Sharing of Xenobiotics-Associated Assays across Species, Phenotypes, and Sites

Researchers at the Boston University (BU) and Oregon State University (OSU) Centers collaborated to understand how to integrate xenobiotic assay data and make it accessible and interactive. Xenobiotics are chemicals, usually man-made, that originate outside of the body. They sought to combine existing mammalian gene expression assay data and chemical annotations related to adverse effects generated at the BU SRP Center with dose-response behavior and morphology data in zebrafish and Superfund site chemical composition data from the OSU SRP Center. Specifically, the collaborators sought to combine data across species to better understand the underlying mechanisms by which exposure to chemicals harm health. Their goal was to establish a data-driven taxonomy of compound classes based on changes to RNA that may represent shared modes of action.

Data inputs from BU included two large publicly available transcriptomic datasets which quantify the expression level of RNAs in a given cell population. Data from the carcinogenome project tested approximately 500 chemicals in human cell lines and data from the adipogenome project included 78 metabolism-disrupting chemicals and controls in mouse cells. They also leveraged publicly available data sets and repositories (e.g., TG-GATEs, DrugMatrix, MSigDB, PubChem). Collectively, this dataset includes information on chemical carcinogenicity, genotoxicity, adipogenicity, connectivity to drugs (CMap), and expression and activity levels of genes and pathways in response to each chemical exposure. OSU data included zebrafish-based morphological and behavioral screens for over 1,200 chemicals and expression profiling data for a subset of these chemicals, in addition to chemical concentration data and Superfund

extract data. OSU data is currently housed internally at the Pacific Northwest National Laboratory in a secure, firewalled data repository, called the Experimental Data Management System (Hobbie et al. 2012) and can be downloaded from http://datahub.pnnl.gov. The teams have over 200 screened chemicals in common.

As a first step, they wanted to develop the infrastructure for storing, retrieving, and querying the data. Leveraging the available data and existing ontologies, such as GeneCards, MSigDB, and Reactome, the team used the R/Shiny interface and commands to develop the front end for two open-source software portals and used application program interfaces (APIs) to allow the portals to talk to each other. They also included a security and privacy system which allows all users read access and required login credentials to make additions and edit.

Their portals, the Xposome Portal, hosted at BU, and the SRP Data Analytics Portal, hosted at OSU, ensure their data is accessible to outside users. For example, the Xposome Portal is an interactive R/Shiny interface that facilitates chemical screening using the compiled high-throughput transcriptomic assay data and the data from the zebrafish assays when available. Users can drill down into information to see what genes are affected by a particular chemical and whether gene expression is increased or decreased, and get more detailed information about effects on particular gene pathways, for example. They can also interact with the data and perform analyses with a built-in tool they developed, called K2Taxonomer, that allows the user to look at changes in gene expression across a group of related chemicals to visualize their similarities or differences from other groups of chemicals in a heatmap.

The Xposome Portal front end includes a Gitter community, a chat and networking platform that allows users to provide feedback for continuous process and usability improvement. Documentation for the Xposome portal is available on GitHub, and additional documentation for both portals will be included in an upcoming paper. The team is finalizing and optimizing the front end for both portals as well as the integration with the back end of the portals, which uses GeneHive, an extensible object storage system for storing and annotating high-throughput data. They are working on dataset manuscripts describing their data to ensure that it is findable through DOI and ISA metadata. In addition to the created portals, all data from OSU can be accessed on their website. They are using APIs and linked interfaces to ensure interoperability, and they are committing their analysis code to GitHub and Docker containers to ensure reproducibility.

The team noted one particular challenge with available ontologies was that chemicals are not clearly and explicitly identified, particularly for mixtures. This is something they hope to continue to address in the future. Looking forward, they hope to integrate across chemical screens to allow for metanalyses within Xposome and incorporate additional datasets and functionalities. They would also like to support additional integration with the CMap portal so users can interact with the data live, and integrate with additional external public repositories (e.g., NURSA, Transcriptomine, CTDbase).

According to the team, one of the largest benefits of their Use Case was training and acquisition of skills related to data management and sharing, including developing user interfaces, designing databases, Dockerization for tool portability, and code development and documentation. They noted that these skills are critical and reusable for future projects. One of the examples of lessons learned during the process was that storing deconstructed expression data as binary data frames is preferable to storing in a database management system. They also identified significant value in incorporating data scientists in

SRP research projects to facilitate interoperability and reuse. The team used GitHub to document and share their lessons learned within their own group and to share with the broader community.

Sharing Environmental Microbiome Data

Deciphering Intra- and Cross-Kingdom Microbial Interactions for Bioremediation of Superfund Pollutants

Studies involving microorganisms that break down pollutants are often conducted with single microbial cultures or simplified bacterial communities in the laboratory. While these experiments are useful to uncover mechanistic insights, they do not capture the complex microbial interactions that exist in nature. Researchers at the Duke University and University of Iowa SRP Centers collaborated to establish a common computational framework and infrastructure, and to standardize sampling approaches to better share, integrate, and analyze large microbiome sequencing datasets collected from natural sediments across SRP centers. Both Duke and the University of Iowa are working to engineer microbes to clean up contaminants in the environment. The team hoped that by sharing data, they could shed light on how complex populations of microbes interact within an environment to provide useful information to improve bioremediation strategies.

The Use Case first sought to establish a reproducible and sharable microbiome bioinformatics pipeline focusing on their existing 16S rRNA high throughput sequencing data from PAH- and PCB-contaminated sediments, respectively, which could then be applied to other data types (see Table 6).

An initial challenge the group faced was the fact that processes and methods across labs were not standardized. Therefore, one of the main goals of this Use Case was to develop a guidance document of standardized practices for environmental sampling, sequencing, and analysis for use in environmental remediation studies. The team closely analyzed their workflows, including sample collection from soil, DNA extraction, and preparing samples for sequencing, all of which can vary significantly and act as a barrier to data sharing.

When the team analyzed their overall workflow, they determined that the lack of standardization in both experimental and computational steps hindered reproducibility of results. To address this problem, the team shared all experimental protocols and analysis code digitally in a GitHub repository. They also included version control to ensure methods were up to date as techniques and tools evolved. These protocols cover methods related to sampling, DNA extraction, and library prepping to get from soil samples to 16S rRNA gene sequences in a reproducible way across centers.

Another component of their collaboration was evaluating their computational procedures looking for ways to maximize reproducibility. For example, differences in software used to conduct analyses and the code used to produce the results. They developed a software container, which is a standalone, transferable, and executable assembly of software, to run their microbiome analyses. Their computing environment is in a container on the Singularity Hub and has all the software packages with source coded needed to run analyses, such as more than 60 R packages pre-installed, which greatly reduces the amount of time required for someone to develop and install the analysis pipeline. This greatly facilitates reproducibility and sharing within and between institutions by allowing anyone to use the exact computing environment to run the same analysis. Their Singularity containers are available, and instructions for running the containers, and protocols are available on GitHub, their data is stored in Sequence Read Archive (SRA).

The team also sought to critically evaluate the FAIR-ness of their data management practices retrospectively to inform their strategy for the future. The team initially thought their microbiome data was reasonably FAIR as it was deposited in SRA with a persistent identifier (e.g., BioProject Number PRJNA382682), making it findable and accessible. However, they quickly learned that the interoperability and reusability of the 16S rRNA data set needed to be improved, particularly in terms of metadata.

The team had not published critical metadata, such as environmental conditions or physical and chemical characteristics of the sediment samples. This Use Case brought to light the fact that SRA and other similar data repositories lack a clear procedure for receiving environmental contaminant data that is necessary for reanalysis and reuse. Until appropriate repositories for environmental data are created and can be linked to sequencing data in SRA, one solution the team proposed is to provide their environmental metadata in a Data in Brief article, which is in progress.

Through their data reanalysis, the team identified the need for robust data management practices rooted in high quality data and metadata, supported by analysis techniques and databases, adequate documentation, and version control, and finally, ensuring ease of access to strategies to make data shareable between institutions.

This Use Case ultimately enhanced the reproducibility of analyses with 16S rRNA environmental microbiome data generated with high-throughput sequencing. These two SRP centers adopted data management best practices like developing standardized protocols and making them publicly available and creating a standardized and accessible microbiome analysis pipeline that allows them to compare results directly, increasing the speed and ease of discovery. Their pipeline provides a variety of statistical tools and techniques that will improve the reproducibility of how they determine relationships between environmental parameters and microbial community composition. Increased documentation and sharing of the metadata will allow others in the SRP community and beyond to re-use valuable microbiome data to understand and engineer environmental microbiomes for bioremediation.

Through their collaboration, the team learned that interoperability and reusability require detailed protocols and guidance documents that accompany data analysis tools. They also found that metadata should be as complete as possible to allow outside researchers to reuse data for more diverse analyses.

Table 6. Existing datasets on microbiome sequences utilized by the Duke and University of Iowa use case.

Collaborator	Dataset Description	PMCID
University of Iowa	16S high-throughput sequencing data from several sediment sampling locations within a PCB-contaminated wastewater lagoon. The raw, unprocessed sequences (16S) are deposited into the Genbank Sequence Read Archive under Bioproject number PRJNA382682	Mattes et al. (2018)
Duke	16S/18S/fungal ITS Ion Torrent and lllumina MiSeq amplicon metagenomic data sets from several sediment sampling locations along the Elizabeth River as well as a former wood treatment facility in Yadkinville, NC and bench	Ikuma and Gunsch (2012) Czaplicki and Gunsch (2016) Chang et al. (2019)

Collaborator	Dataset Description	PMCID
	scale reactor work in addition to a single 16S amplicon metagenomic/ metabolomic dataset from Fundulus (fish gut and gills). Although not all data have been published to date, raw and processed data are available from all these experiments.	

Two Use Cases: Integrating and Creating Broad Access for Transcriptomic, Microbiome and Physicochemical Datasets of Phytoremediator and Phytostabilizer Plants; AND Data Interoperability for Investigating Biogeochemical Controls on Metal Mixture Toxicity Using Stable Isotopes and Gene Expression

Soil and water with high levels of toxic metals, including cadmium, lead, mercury, and arsenic, can be harmful to human and environmental health. Traditional approaches to decontaminate heavy metals include excavating and removing soils, which can be costly and impractical. Using plants to take up metals and stabilize them is a cost-effective alternative, but there are many complex interactions and genes and pathways involved that are not well characterized. The University of Arizona (UA) SRP Center team worked with researchers from the University of California (UC) San Diego SRP Center to understand interactions between metals, microbes, and plants that help some plants tolerate contaminants and stabilize metals in contaminated soils. Specifically, they sought to investigate genomic, transcriptomic, microbiome, and physicochemical properties to identifying genes and pathways that enable plants to grow and stabilize metals in semi-arid environments.

In a closely related Use Case, the UA SRP Center team collaborated with researchers from an SRP-funded individual research project at the Colorado School of Mines (CSM) to explore how remediation of mining waste affects diversity in terrestrial and aquatic systems. Researchers at CSM examined stream impacts of mining and the processes involved in recovery following clean-up, however complex interactions among metal mixtures are not well captured by current predictive toxicity models used by regulatory agencies. The group conducted toxicology experiments with aquatic organisms to uncover the molecular mechanisms involved in interactions between metals to improve model predictions. The UA team has tackled the issue of mining waste on the terrestrial side, looking at plants that can take up and stabilizing metals. They have tested a phytoremediation strategy called compost-assisted phytostabilization that promotes plant and root growth that locks metals underground.

In the first Use Case, both groups have data on chemical concentrations, bioavailability to plants, genomics, metagenomics, ionomics, and bulk soil and rhizosphere microbial diversity. In the second Use Case, both groups have data on ionomics, contaminant concentrations, physicochemical characteristics, and toxicity. The UA group also has data on microbial diversity over time after remediation, while the CSM group has data on water quality, diversity and abundance of benthic organisms, and algal biomass before and after remediation. See Table 7 for more information about the existing datasets.

The teams aimed to integrate their data and create a data analysis portal to enable robust analyses that would shed light on the complex relationships between environmental factors, microbial communities, and remediation success. However, they first needed to take preliminary steps to make their data more standardized and interoperable.

They developed standard operating procedures for gathering, cleaning, and storing data. They also invested a significant amount of time mapping data to existing ontologies. For example, they used Environment Ontology (ENVO) for metal concentrations, Plant Ontology for plant structures, BioCollections Ontology for observations and measurements, and Population and Community Ontology for diversity metrics. Through this process, they found that existing ontologies lacked the level of detail required to describe their data, so they contributed to these ontologies by adding new terms as needed.

For example, they developed a new branch of the ENVO ontology to allow users to search for data on calcium in different systems (e.g., fresh water, topsoil, garden soil), and added diversity metric terms to the Population and Community Ontology. Another ontology-related challenge they encountered was finding social science ontology terms for resiliencies related to remediation. To overcome this challenge, they temporarily created terms in an internal ontology, but hope to see this develop further in a publicly maintained ontology.

They are in the process of analyzing and annotating transcriptomic datasets, and are working to combine all aquatic, terrestrial, and greenhouse data collections into an open-source format database, which will be hosted in the UCSD Superfund Portal. They hope this tool will help users explore commonalities between terrestrial and aquatic impacts of metals and to explore how diversity changes in response to remediation in both systems.

Sequence data for both Use Cases is being deposited in NCBI's SRA using minimum information standards (i.e., Minimum Information about any (x) sequence -MixS), while environmental data will be deposited in the Knowledge Network for Biocomplexity (KNB) data repository. All their code is available on GitHub. They are preparing data for final publication, including association with ontologies, standard metadata (e.g., MixS, DataCite), open formats, and a data dictionary. A paper on how they used ontologies for integrating environmental science data from terrestrial and aquatic systems for hypothesis generation and explorations is in process, as well as a collaborative publication on soil microbiome and plant gene expression in metal contaminated mine tailings.

Another important outcome of these Use Cases was the ability to train the team on data science (e.g., GitHub, CyVerse, ontologies, makefiles, visualizations). They also trained several students in large scale data analyses using multiple platforms and new programming.

The Use Case supplement helped to bring together scientists in both institutions in analyzing, interpreting, and preparing data for FAIR public access. By combining transcriptome, microbiome, and other large data sets, the UA and UC San Diego team is deriving new models on how microbiomes enable phytostabilization at Superfund sites. In the future, they plan to collaborate to characterize the effects of beneficial microbes and use that data to inform recommendations and solutions. Their approach has also enabled them to consider mining impacts at more of an ecosystem level. While they are just beginning to integrate data, the team thinks that being able to compare impacts of metal contamination on diversity across systems will provide a better understanding of the mechanisms that relate contamination and diversity.

Table 7. Existing datasets microbiome sequencing and other characteristics utilized by the UA, UC San Diego, and Colorado School of Mines use case.

Use Case	Data Type	Notes
UA and UC San Diego	Transcriptomic	Data sets from 24 phytostabilizer plant samples from shoots and roots grown in compost-amended mine tailings. Transcriptome data are the sum of an organism's RNA transcripts and provide a picture of gene expression in specific cells or tissue under specific conditions.
	Microbiome	Generated from the rhizosphere-influenced and bulk soil samples collected from the pots grown with quailbush exposed to compost-amended mine tailings and potting soil. These data provide microbiome community diversity metrics.
	Ionomic	Metal and elemental content (As, Cd, Cu, Fe, K, Mn, Na, Pb, Zn) on leaf, shoot, and root samples taken from the same plants as for the transcriptome samples above. These data sets provide information on the accumulation of toxicants and elemental nutrients in each sample.
	Physicochemical	Physicochemical characteristics of mine tailings, compost, and potting soil, including pH, electrical conductivity, total organic carbon, total nitrogen, and metal/elemental content of the mine tailings, compost, and potting soil samples. These data indicate the state of the growth medium prior to planting and after plant growth and establishment.
UA and CSM	Water Chemistry	Data from North Fork Clear Creek, Colorado before and after acid mine drainage treatment, including total and dissolved concentrations of major and trace elements, as well as water pH and conductivity. Discharge data allows for computation of metal loads. These data identify spatiotemporal trends in metal loading and remediation effectiveness.
	Biological	Data from North Fork Clear Creek obtained from benthic sampling performed over the past 3 years including the total abundance of benthic organisms, taxonomic benthic diversity, and algal biomass. These data indicate the biological response of stream communities to improved water quality associated with acid mine drainage remediation (Kotalik et al. 2021)
	Modeling Toxicity	Model-computed toxicity for field data using the measured water composition, including dissolved organic carbon and water hardness. Toxicity of copper and zinc over the study period was calculated to determine the effectiveness of the acid mine drainage treatment to aquatic health.
	Mixture Toxicity Assays	Mortality for exposures to Cu, Cd, Ni, and Zn in mixtures-based toxicity studies in Daphnia Magna RNA seq data from Daphnia magna from metal mixture toxicity studies. Quantitative reverse transcription polymerase chain reaction (RT-qPCR) data for selected biomarkers from metal mixture toxicity studies.

Use Case	Data Type	Notes
	Chemical Analyses	Samples from the Iron King Mine and Humboldt Smelter Superfund site (IKMHSS) mine tailings in Arizona that were collected prior to and during phytostabilization were measured for pH, moisture content, color, texture, and to identify rocks and minerals present in the sample. (Root et al. 2015). Total elemental composition and speciation (As, Pb, and Zn) was used to determine the impact of phytostabilization on metal mobility and bioavailability (Hammond et al. 2020).
	XAS and X-ray Diffraction	XAS and X-ray diffraction of biogeochemical processes affecting metal(loid) molecular stabilization and mobility in the root zone of plants during phytoremediation (Hammond et al. 2018).
	Microbial analysis	Analysis of bulk and rhizosphere samples from IKMHSS during compost-assisted phytoremediation (Honeker et al. 2019; Hottenstein et al. 2019; Valentin-Vargas et al. 2018).
	rRNA	16S rRNA and gene abundance and activity during IKMHSS phytoremediation quantified using quantitative PCR and quantitative reverse transcription PCR (Honeker et al. 2019; Nelson et al. 2015).
	Metagenomics	Metagenomic sequencing analysis of bulk and rhizosphere samples from IKMHSS during compost-assisted phytoremediation (unpublished).

Geospatial Platforms and Visualization of Environmental Data

Linking Data from Laboratory and Field Investigations of Mercury Transformation, Bioaccumulation, and Remediation

The Agency for Toxic Substances and Disease Registry has identified mercury as one of the top chemicals of concern to human health. Exposure increases the risk of diabetes, respiratory disease, and reproductive and developmental disorders. For fetuses, infants, and children, exposure has severe, adverse effects on the developing nervous system and interferes with cognitive thinking and memory.

Mercury contamination can persist in aquatic ecosystems worldwide and is the most frequent cause of fish consumption advisories across the U.S. Methyl mercury, the form that biomagnifies in the aquatic food web, is controlled by a range of geochemical, microbiological, transport, and ecological processes. Collaborators from the Dartmouth College SRP Center and SRP-funded individual research projects at Duke University and University of Maryland-Baltimore County (UMBC) set out to better understand the range of factors controlling mercury movement and transport in aquatic environments to evaluate the effectiveness of remediation strategies.

By compiling data from both controlled laboratory and microcosm experiments and larger scale mesocosm and field observations, the team also hoped to improve the context and scalability of lab data. Their aim was to create a centralized data platform to compare controlled experiments and field observations to improve insights into the environmental relevance of experimental findings. They also sought to categorize environmental systems based on which factors predominantly influenced methyl mercury production and bioaccumulation, and to identify data gaps that could inform future studies.

The team leveraged diverse field datasets (see Table 7) collected by Dartmouth, the University of Connecticut, and SERC that included multiple estuaries, rivers, and marshes on the U.S. East Coast. These datasets included information about the water (e.g., pH, dissolved oxygen, temperature), sediment (e.g., organic carbon), and mercury concentration data from water, sediment, and biota. The UMBC/SERC also contributed data from field experiments in a Maine salt marsh on the potential impact of sorbents to reduce mercury bioavailability. Duke researchers collected lab and experimental data on mercury methylation from microcosms and field mesocosms, including information about water, sediment, and biota. Data used in the compilation included information collected through SRP and other NIEHS funding from all the teams, as well as from other projects.

Their initial plan was to integrate their data using an unstructured database, but they encountered challenges in combining data from individual labs. Even though all collaborators were mercury scientists, each lab named and stored measurements differently that best suited the specific research goal. To address this challenge, they pivoted to an approach involving loading their data into a structured database (PostgreSQL). Their first step was to create a consistent naming convention across labs, leveraging existing ontologies, including some nomenclature from ENVO, CHEBI, and UNITS. Their terms included parameter names, definitions, and unit conversions (i.e., molar to weight) that would make their data talk to each other.

Next, the team developed a sharable data analysis pipeline for data cleanup and transformation in OpenRefine. OpenRefine is an open-source software which visualizes and manipulates large quantities of data all at once. They wrote code to tell the application how to translate data values from each collaborator's separate Excel files into one normalized database using their standard terms. This process generated a new file that reproduces the same automated cleaning process so other researchers from separate organizations can replicate this Use Case's pipeline.

In total, the Use Case has compiled and harmonized five lab and field datasets, produced a consistent naming convention between labs, and completed a data dictionary describing the contents, format, and structure of the database and its elements. They are in the process of mapping from individual datasets to a form a unified public data repository, and plan to store their code and infrastructure documentation on GitHub.

Following its completion, the team discussed strategies to maintain their repository after the conclusion of the Use Case project and how it might be funded and maintained for longevity. They also considered the challenge of data ownership rights for publicly available data. They discussed the challenge of storage timeframes being longer than funding periods of grants, and how research efforts should focus on creating analysis pipelines that ingest source data hosted in a central repository, rather than attempting to develop a general-purpose platform.

They hope to expand data sharing with other mercury scientists as the challenges faced during this project would not be isolated to just their institutions. They also plan to increase communication about their data sharing activities to other stakeholders in the U.S. Environmental Protection Agency (EPA) and U.S. Geological Survey (USGS) and cross-train scientists with data experts with an emphasis on involving SRP trainees. The creation of this data repository and the challenges and future directions will be described in a paper to be submitted for publication.

Table 7. Existing data used by the Dartmouth, Duke, and UMBC/SERC Use Case team

Collaborator	Data Type	Data Source
Dartmouth and collaborators at the University of Connecticut	Field sampling: data collected across several years at sites in Maine, New Hampshire, Connecticut, New Jersey, New York; included dissolved and particulate Hg and MeHg, temperature, salinity, pH, dissolved oxygen, dissolved organic carbon, etc. in water, sediment, and biota, as appropriate.	Buckman et al. (2021); Taylor et al. (2019)
UMBC and collaborators at the Smithsonian Environmental Research Center	Field survey of marsh and adjacent mud flats across the Chesapeake Bay (Virginia and Maryland), and Maine. Field trial of the impact of activated carbon on Hg bioavailability in a marsh in Maine. Both data sets include detailed sediment and porewater geochemistry, and Hg methylation rates.	Gilmour et al. (2018)
Duke	Outdoor field mesocosm experiments: Hg methylation and MeHg bioaccumulation factors in water, sediment, biofilms, and animals. Also compared passive sampling technologies for predicting Hg methylation and MeHg bioaccumulation potentials and evaluated activated carbon amendments in altering bioavailability and methylation.	(Neal-Walthall, in press)

Making Fish Contaminant Data FAIR to Improve Fish Consumption Advisories

Fish consumption advisories are meant to help people make informed choices about consuming fish caught from local waters. However, guidelines vary significantly from location to location, or may be inconsistent within the same water body if recommendations are made by more than one jurisdiction (see Figure 7). Researchers at the Boston University (BU) and Dartmouth College SRP Centers worked to create a searchable data platform containing fish tissue and environmental contaminant data from their centers as well as publicly available data.

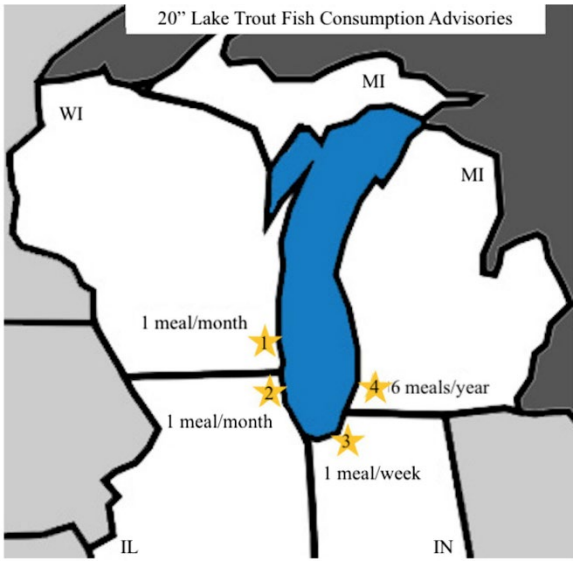

Figure 7. Differences in advisories issued by different states bordering Lake Michigan (Cleary et al. 2021)

By comparing similar types of environmental science data, they sought to compare fish consumption advisories to ultimately determine if they are protective of sensitive populations. Since there are no comprehensive U.S. databases of chemical contaminants in fish tissue, they also wanted to create a searchable database that allows determination of temporal and spatial evaluations.

Both centers focus on fish consumption as pathways of exposure to mercury, polychlorinated biphenyls (PCBs), and other contaminants. Before their collaboration, data existed in center data files, government reports, ecological risk assessments, and government websites. According to the team, utilizing this data required significant effort to make them accessible, and they were not interoperable or reusable. By combining and integrating their data, they saw potential to address contaminated fish consumption, particularly by vulnerable populations.

The team gathered EPA fish contaminant datasets including the National Rivers and Streams Assessment (2008-2009, 2013-2014), the National Coastal Condition Assessment (2000-2006, 2010), the National Lake Fish Tissue Study (1999-2003), and the Great Lakes Environmental Database (1999-2018). The composition and coverage of these datasets were compared with center data collected at Superfund and contaminated sites, which include diverse marine and freshwater fish species, sample types (e.g., fillet, whole body), and concentrations of contaminants, including PCBs and mercury. The U.S. EPA datasets helped to illuminate the potential and highlight the gaps that are seen across Superfund-generated data.

These data were not standardized in terms of how contaminants were described, how samples were characterized, when and how they were taken. Additionally, a uniform ontology was not available, creating challenges for data integration and normalization, particularly for contaminant nomenclature. To integrate data into a single centralized repository, the team mapped metadata between data sources and normalized inputs using a customized ontology they developed. Their ontology aggregated and extended existing ontologies, including ecological, physiological, and environmental terms from EnvO and contaminant terms from ChEBI.

The team is building a relational database to combine their SRP center and external data so users can query all datasets at once (see Figure 8). They began with a defined organization for column mapping and data types. The team recorded this schema and stored the utility code, which facilitates migration, duplication, and sharing, on GitHub.

Figure 8. The team began by sanitizing each center's disparate data, which involves fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data, to form their centralized repository. (Image courtesy of Boston University and Dartmouth College SRP Centers)

The team plans to explore opportunities for funding to expand their work in order for external partners to submit data, which will require investigating quality control and security considerations. They will also need to consult with other researchers aggregating data to establish QA/QC protocols. Ultimately, they hope to expand data sharing with other fish contaminant and public health scientists, state and federal risk assessors, and other SRP centers.

The repository, currently internal, will underpin an interactive map visualization to provide a broad view of contamination nationwide including PCBs, mercury, and other pollutants, including perfluorooctane sulfonic acid (PFOS). This tool should help make data FAIR while creating opportunities for scientific collaboration among SRP environmental health researchers, local researchers, citizen science and community groups, Native American tribes, and federal and state government.

In a preliminary mixtures analysis using data from the Great Lakes Fish Monitoring and Surveillance Program collected from 1999-2018, the team reported that fish contaminant levels tended to increase with fish size. They also showed trends, such as PCB levels decreasing over time, mercury levels as well as pesticides, have remained relatively stable over the past two decades and where data exist, levels of PFOS are increasing. These data may be useful supporting information for the success of cleanup projects or regulatory decisions. Some contaminants were reported to continue to exceed the levels for safe fish consumption. Using data from the National Rivers and Streams Assessment, they were also able to conduct a preliminary analysis revealing higher correlations of PCB chemicals in urban sites compared to non-urban sites. According to the team, by increasing the FAIR-ness of data and facilitating these kinds of analyses, communities and decisionmakers will have stronger tools available to evaluate the protectiveness of fish consumption advisories.

By increasing communication through workshops and other events, the team also plans to boost engagement of stakeholders in the U.S. EPA, state departments of environment and health, and regional and state stakeholders.

Validate and Develop Visualization and Reproducibility Documentation for Source-Receptor Relationships for Toxicants

Collaborators from the University of Rhode Island and Massachusetts Institute of Technology (MIT) SRP Centers set out to understand the link between sources of chemical emissions and their concentrations in the environment. More than 600 sites across the U.S. are contaminated by per- and polyfluoroalkyl substances (PFAS), but the extent of transport away from these sites to potential human exposure pathways, such as inhalation, is virtually unknown. These source-receptor relationships link pollution emissions to their migration and deposition, as well as to human exposure and finally to resulting health effects. Such information is critical for accurate risk assessment and the development of effective remediation policies. By quantifying and visualizing source-receptor relationships and potential health and environmental impacts, the team sought to provide more detailed data to inform decision making.

They planned to investigate the commonalities in atmospheric deposition pathways of two classes of chemicals by integrating modeling data and measurement data from both centers. By combining and comparing their data, they hoped to develop robust source-receptor relationships for PFAS and PAHs,

information that is essential for attributing exposures and health effects to Superfund sites or other sources of pollutants (see Figure 9).

Their existing datasets included PFAS and PAH modeling data generated using the GEOS-Chem atmospheric transport model, a global 3-D chemical transport model for atmospheric composition driven by meteorological input from the Goddard Earth Observing System (GEOS) of the NASA Global Modeling and Assimilation Office. GEOS-Chem is open access and available through GitHub. GEOS-Chem models link pollutant emissions, chemical transformations, and fluxes back to surface ecosystems. They also used statistical and semi-empirical modeling data layers and atmospheric modeling data for a subset of PFAS.

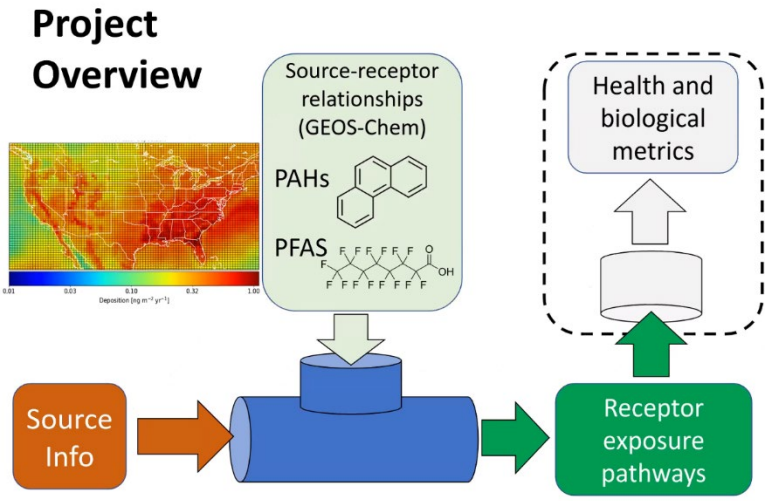

Figure 9. Graphical Overview of MIT and URI Use Case project. (Image courtesy of the MIT SRP Center)

An initial challenge was the lack of existing ontologies, so they developed an ontology for metadata specific to atmospheric toxicant modeling that can be extended to help integrate information from additional models. The team developed and implemented a consistent procedure for harmonizing the spatial and temporal resolution of the data to enable geographic comparison. This involved setting up a common data formatting, or metadata, hierarchy that worked for these two datasets, but that could also be generally applicable to other chemical transport model outputs generated with GEOS-Chem. Using the SWEET ontologies, their hierarchy started with source type, such as specific emitter or class of emitters. Each of these sources could emit a suite of chemical species which then potentially reach human receptors through different exposure pathways.

The team used a NetCDF (Network Common Data Form), which is a file format for storing multidimensional variables such as temperature, humidity, pressure, wind speed, and wind direction. Each of these variables can be displayed through a dimension, such as time, in geographic modeling software by making a layer or table view from the NetCDF file. The group used Python tools to transform and visualize the NetCDF data into their PFAS- and PAH-specific GEOS-Chem source-receptor model.

The researchers wrote an application programming interface (API) wrapper into this process, making their actions reproducible and accessible, and broadly applicable to different toxicants. An API wrapper is a language-specific (e.g., Python) package or kit that encapsulates multiple processes to make complicated functions easy to use.

Together they created visualizations and interactive maps of source-receptor relationships for the Northeast U.S. region. All metadata, and visualization tools and model code are archived on GitHub and have been deposited to the Open Science Framework as part of this project.

A manuscript exploring contrasting source-receptor relationships for PAHs and PFAS is still in progress, and they hope to continue work to link exposure pathways to health endpoints. In the future, they are

considering containerizing the code they created to allow non-data scientists to host and apply the process on a local server to manipulate the data. This functionality would be of interest to those at state and local levels who want to know what sources and contaminant levels mean.

Developing a Spatial Approach for Toxic Transferal from Industrial and Vacant Land Uses to Green Infrastructure

Disaster events such as flooding can spread harmful contaminants from current and former industrial sites into neighboring communities. Implementing new green infrastructure, such as rain gardens or food forests, has been linked to improved public health outcomes, decreased flood damage, and decreased concentrations of toxics in stormwater runoff. However, little is known about whether these systems are vulnerable to toxics transfer during extreme weather and other disasters. Researchers at the Texas A&M University (TAMU), Brown University, and UC San Diego SRP Centers worked together to understand how land use, such as vacant or industrial land and green space, affect people's exposure to harmful chemicals and impact community resilience and health.

They aimed to integrate diverse city, local, and federal data, with SRP center datasets on spatial land use, including green infrastructure, flood plains, vacant lot uses, public health outcomes, industrial land uses, and sociodemographic conditions (see Table 8) to create interactive maps that could show how different factors contribute to an area's vulnerability to toxicant transfer or flooding, for example.

The main challenge to combining these datasets was the fact that they are not processed, documented, and indexed in widely used data discovery systems. They are also not formatted consistently in terms of nomenclatures, metadata, and access protocols. Finally, the team commented that it is time consuming to prepare data to be input into urban development and environmental exposure models due to metadata deficiencies and data type misalignment. The team noted that insufficient metadata descriptions and lack of standards-based tools for data processing, management, exploration, and visualization are a barrier.

Working closely with the research data management librarian at Brown University, the team created a metadata application profile (MAP) to capture the minimum reporting information to enable data discovery and machine-actionable potential of their project data. Their MAP defined core elements for the uniform description of data using international standards, persistent identifiers, vocabularies, and ontologies. Their MAP allowed data to be incorporated into a data discovery platform developed by UC San Diego with National Science Foundation funding called the Data Discovery Studio (DDStudio), as well as a data analysis and visualization platform called SuAVe.

DDStudio provides, ontology-based metadata enhancement, data discovery, and integration with data science software modules. It uses machine learning to recognize and extract historical land use data, parse, and geocode the directories, and crosswalk among metadata. The team combined six geospatial datasets at the U.S. census tract scale representing three different geographical areas around the country and did a community-level analysis. Once the data were derived, cleaned, and spatially assigned using GIS software, they were integrated and visualized using the Toxicological Prioritization Index (ToxPi) software (Marvel et al. 2018). ToxPi normalized and equally weighted the six data sets on a scale of zero to 1. Two of the datasets, social vulnerability and health outcomes, contained subcategories. The resulting pie charts were used to visualize threats to a community by assigning a vulnerability score for each neighborhood.

The team created an online interactive dashboard, called the Toxics Mobility Vulnerability Index, to show how factors, such as impervious surfaces, floodplain area, and social factors, affect the community's vulnerability to toxicant transfer, flooding, or health outcomes, like high blood pressure (see Figure 10). This tool was designed for community members and decision makers to visualize factors that affect human health and make informed decisions. It will also be very useful in the context of climate change and remediation strategies, since it identifies specific factors that contribute to flood risk, such as impervious surfaces, so decision makers are better equipped to identify solutions to mitigate, for example redistribution of hazardous contaminants that may result from sea level rise or extreme weather.

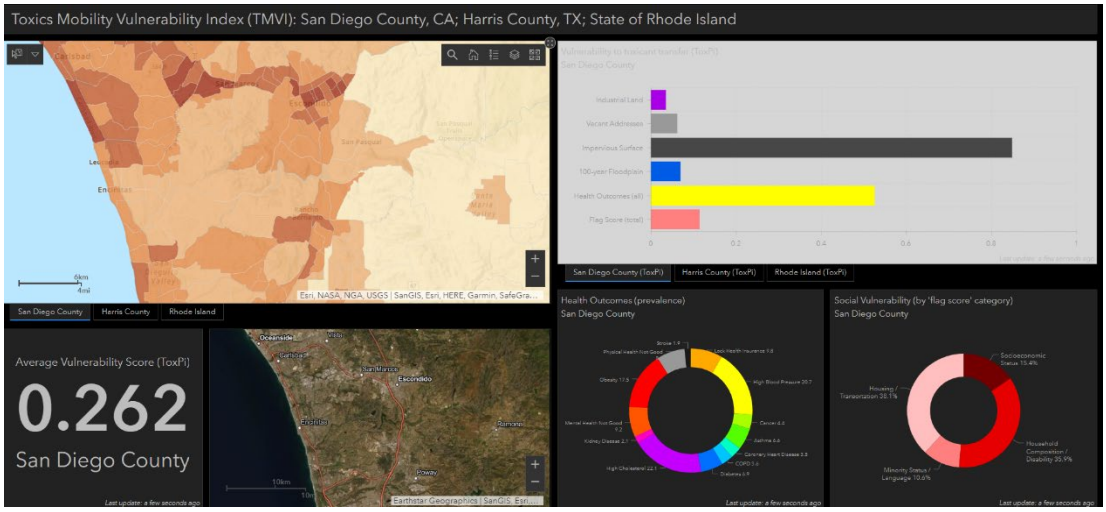

Figure 10. The Toxics Mobility Vulnerability Index

The team leveraged their partnership to test the reproducibility of their approach and results for three different areas: San Diego County, California; Harris County, Texas; and Rhode Island. They were also able to do a comparative analysis across the three areas to look at differences in green land cover in relation to other vulnerability factors. The data are shared in repositories including Dataverse, ArcGIS Online, Brown Digital Repository, and DDStudio, and the demonstrated access, interoperability, and reuse with Jupyter notebooks and the publicly available interactive dashboards. They are also using GitHub for managing and sharing code for extracting data.

The team learned that following established data formats was critical to creating a consistent description of the datasets so they could be used by their data discovery platform. They also learned that it is important to quickly demonstrate the benefits of FAIR data sharing, such as developing the interactive map that showed how integrating data provides unique insight.

Sharing their data has encouraged its use beyond their intended aims, such as being leveraged for analyzing data related to COVID-19. They also identified a future project with the city of San Diego to optimize their climate action planning using existing data. Their Use Case efforts have made them well positioned to help address this need and connect data on climate and health and look at toxicant flows. Additionally, they are incorporating a citizen science component to this project leveraging an existing youth advocacy campaign.

This Use Case combined several datasets originally generated independently by TAMU, Brown University, and UC San Diego into three larger, comparable datasets for coastal Texas (Harris County, Texas), the state of Rhode Island, and San Diego County, California. All registered datasets can be found through ArcGIS Online based on their metadata, via online user interface or via Python code in a Jupyter Notebook.

Table 8. Existing datasets on geospatial and sociological characteristics utilized by the TAMU, UC San Diego, and Brown University use case.

Dataset	Year	Source	Reference	Scale	Registered in
Green Infrastructure Quantity and Quality	2016	Texas Natural Resources Information System DataHub	USGS	US Census Tract	AGOL, DDS; with respective GUIDs
Vacant Addresses	2016	U.S. Postal Service	HUD	US Census Tract	AGOL, DDS; with respective GUIDs
Vacant Lands		SANDAG			AGOL, DDS; with respective GUIDs
Public Health Outcomes (14 factors)	2016	https://www.cdc.gov/500cities/index.htm	CDC	US Census Tract	AGOL, DDS; with respective GUIDs
Social Vulnerability/Sociodemographic Conditions	2016	https://svi.cdc.gov/data-and-tools-download.html	CDC	US Census Tract	AGOL, DDS; with respective GUIDs
Industrial Land Uses	2016	Multiple (created from land use data); includes Brown's Historical Industrial Site database that contains information on 6655 manufacturing sites in Rhode Island.	Local	US Census Tract	
Flood Damage (Flood Plain)	2016	https://www.fema.gov/faq-details/GIS-Data Spatial Hazard Events and Losses Database for the US (SHELDUS)	FEMA	US Census Tract	
San Diego historical business locations		City of San Diego	City of San Diego		AGOL, DDS, Brown Digital Repository, Dataverse; DOIs, GUIDs

Metadata standards used: ISO 19115/19139, schema.org; AGOL = ArcGIS Online; DDS = Data Discovery Studio; vocabularies included CINERGI ontology for automated metadata enhancement and domain vocabularies (e.g., NAICS).

Data Sharing Tools, Workflows, and Platforms

Improving the Robustness and Toxicological Significance of Nontarget Chemical Identification in High Resolution Mass Spectrometric Data

Superfund sites contain complex chemical mixtures, where mixture components may be known or unknown. Samples of these mixtures sometimes lack data on methods of detection, environmental occurrence, or toxicity. High-resolution mass spectrometry (HRMS) allows researchers to look for and measure multiple unknown chemicals, an approach called non-targeted analysis.

Researchers at the Duke University and the UC Davis SRP Centers worked together with an external collaborator from the U.S. Environmental Protection Agency (EPA) to explore how to harmonize and combine sources of HRMS data to improve non-targeted analysis of complex mixtures of environmental contaminants. They aimed to develop methods with open-source data analysis software by performing an intercomparison study, in which both centers shared sample spectra between each lab and compared results of non-targeted analysis. They also sought to develop new approaches to link toxicity data with non-targeted screening results.

This type of analysis is vital for assessing environmental health risks after contamination events but making this data interoperable and reusable has many challenges. For example, different HRMS laboratory instruments used to collect the spectra in many cases have vendor-specific or proprietary data formats, software, and lab protocols, which hinders data sharing and reusability. Additionally, no online repositories or fully developed ontologies currently exist to make this type of data accessible and archivable.

The team observed that metadata within HRMS datasets are not always interoperable with different software due to differences in the format of the data. Data formats in this dataset include SQLite databases of spectral libraries and mzML spectral data structures, which differ depending on laboratory tools and need to be standardized.

The team decided to focus on two primary datasets (see Table 9). The first dataset is a water quality analysis assessment Duke SRP grantees performed in 2018 on North Carolina surface waters after Hurricane Florence. The researchers looked at 36 water samples and performed 72 different mass spectrometry analyses, resulting in almost 50 gigabytes of mass spectral data. Together, they established and tested data sharing protocols and algorithms. These included specifics for compound identification/annotation and specific performance metrics like confidence scale. A second dataset was produced by UC Davis SRP grantees in collaboration with the Yurok Tribal Environmental Program, which was interested in the environmental impacts associated with contaminated sediments around abandoned lumber processing sites.

The analysis workflow to process storm water data used by UC Davis required that raw data was converted into an open-source format, called mzML, before conversion into the appropriate data format for analysis in the choice open-source software, MS-Dial. Among various choices that affect the analytical and data analysis process, the cascade of data conversions has the potential to result in inaccurate conversions. For example, UC Davis analysis of the Duke dataset generated roughly 8,500 molecular features after converting Hurricane Florence stormwater data, and 98 compounds had a match in the spectral library. By comparison, the data analysis workflow from Duke generated only about 2,400 compounds with 641 matches to compounds in a spectral library.

According to the researchers, discrepancies between the number of generated and identified spectra can occur depending on data processing, compound annotation, and compound library matches. Current libraries used to store and annotate non-targeted datasets usually include a number of repositories using different architectures. To overcome this challenge, they are improving format conversion parameters and library harmonization to improve the FAIR-ness of spectral libraries.

The team built software to combine and translate several of the most important open-source libraries of mass spectrometry data, such as MassBank of North America and National Institute of Standards and Technology (NIST) libraries. They are translating these spectral libraries, which are in NIST's spatial data format, into an open and fully accessible SQLite database that is easily accessible and interoperable. By combining different libraries together, the researchers have access to a higher resolution of analytical data to match against unknown compounds in environmental samples. To facilitate interoperability and reusability, the team will make the code used to harmonize libraries publicly available via GitHub.

After harmonization of data analysis software packages, the Use Case plans to connect their data to the EPA CompTox Chemicals Dashboard, a public database containing 875,000 chemicals and their associated toxicological information. Typical results of non-targeted analysis will generate thousands of molecular features. To generate a priority list, compounds are labeled as contaminants of emerging concern by using EPA and other collaborative toxicology data. Afterwards, integrating results compiled from this project into the EPA dashboard will also make HRMS data collected at these SRP Centers available to a broad range of researchers and interested stakeholders.

Through their work, the researchers learned that open data formats are critical for analyzing scientific data and advancing the pace of research. According to the team, the knowledge gained through this Use Case collaboration allowed them to make considerable progress towards increasing HRMS data harmonization, accessibility, and interoperability, such as developing an open-source mass spectral library. The collaborators noted there is still work to be done to streamline data sharing, conversion, and cross-platform analysis to make non-targeted compound identification more FAIR.

Table 9. Existing HRMS datasets utilized by the Duke and UC Davis use case.

Institution	Dataset	Spatiotemporal details	Data metrics	Format	Analysis Instrument
Duke	Non-targeted analysis results from ESI (+) HRAM MS/MS analysis of North Carolina river water	Twelve locations, in two river basins in North Carolina were sampled before and after Hurricane Florence made landfall in 2018	Data metrics: 36 water samples measured (72 LC MS/MS analyses, ESI+/ ESI+/--) representing 47.7 GB of mass spectral data)	.RAW	Thermo Orbitrap instrument platforms

Institution	Dataset	Spatiotemporal details	Data metrics	Format	Analysis Instrument
UC Davis	Non-targeted analysis results from ESI (-) HRAM MS/MS analysis of Yurok Tribe Sediment	Six locations near abandoned lumber processing mill sites on Yurok Tribal Land near Klamath, CA	Data metrics: 38 sediment samples measured (76 LC MS/MS analyses, ESI+/ ESI+/--) representing 215 GB of mass spectral data)	.D	Agilent instrument platforms

Improvement of Small Molecule Biosensor Probe Development and Biomedical Applications through the Integration and Reuse of SRC Data Sets

Researchers at the UC Davis and UC San Diego SRP Centers explored how to re-purpose and integrate their existing datasets (see Table 10) to improve biosensor probes used to detect and quantify pollutants in the environment or humans. Through this collaboration, the researchers sought to improve the specificity of their probes and improve their ability to detect new target chemicals in the environment.

Most analytical methods used to detect toxicants and biomarkers of exposure can be expensive and complex to utilize. To address this challenge, UC San Diego researchers develop protein biosensors, small proteins which emit light upon contact with a small molecular weight toxicant, for rapid on-site detection of environmental contaminants such as arsenic, cadmium, and organochlorides. UC Davis researchers develop immunoassays, which rely on antibodies from living organisms to bind to chemicals, to design probes to measure specific chemicals and biomarkers of exposure. These tools can be useful for detecting contaminants on-site and for studying human exposure and disease.

The Use Case initially aimed to establish a data management structure and share protocols to link their experimental data, such as probe sequences and binding affinities, to other types of data from metabolomics, gene expression, plant remediation, and community outreach projects from their SRP Centers. Their goal was to provide a centralized and easily searchable resource to identify a wider range of environmental contaminants and design better detection tools. According to the team, this ended up being an ambitious aim because of the many differences in their data, which initially prevented their data sets from being reusable and interoperable. Instead, they focused on exploring ways to make their data reusable for future data mining and machine learning and providing context for this data to be interpreted by other investigators.

The main challenge to sharing their data was different formats, quality, and collection methods across groups. The researchers engaged in continuous communication via email and virtual meetings to clarify parameters and standards to reformat their data and metadata so it could be uploaded shared, reused, and understood by the other organization. Most data from biosensor probes were stored in notebooks at UC Davis and Google drives at UC San Diego and the researchers had to establish a framework to make their data digital and organized.

To make their data more accessible, the team tried to establish specific rules for data entry and sharing, and a common format and vocabulary for datasets. For example, molecular probe sequences (Nbs and SEBs) and small molecule selection data (concentrations reported in a uniform way although it varies depending on the mode of collection and differences between the labs) were in text format where possible.

To some extent, the team was successful in reformatting their existing datasets to be interoperable with other probe data and is currently entering their data in the UC San Diego SRC Data Management Portal, which integrates data from UC San Diego SRP Center projects in a consistent format and facilitates a more standardized data analysis process. They established this platform as the standard method for data storage for their future data collections. An important lesson is to format the data prior to collection if possible.

The team noted that public repositories to access probe data were not available, which interfered with their efforts to improve accessibility. Another challenge to make probe data FAIR is the absence of existing ontologies. The researchers engaged in conversations to begin developing ontologies for probe data to allow other scientists to reuse and understand it, but this is still in early stages. In the long-term, the team aims to make their data publicly available. They are developing public repositories within their Center's websites, but these are in early stages and not ready to be deployed.

Through their work, the team learned that developing methodologies, protocols, architecture, and standards prior to data collection are a necessity. They observed that improving data architecture and standardizing data after collection is difficult and time consuming. They also stressed the importance of collecting data in one specific format, such as text-based, since converting raw data that is instrument-specific can be challenging.

According to the team, the Use Case collaboration allowed them to establish a framework to collect new biosensor data that will improve data interoperability and re-use. The collaborators plan to use this framework to collect larger datasets that could be used to improve probe sequencing.

Table 10. Existing biosensor probe datasets utilized by the UC San Diego and UC Davis use case.

Institution	Compound	Probe Type	Sample Type	Associated Publications
UC Davis	TCC	Nanobody	N	
	3-PBA	Nanobody	human urine, environmental and food samples	Kim et al. (2012)
	BDE-47	Nanobody	furniture samples	Bever et al. (2014)
	TBBPA	Nanobody	spiked soil and serum	Wang et al. (2014)
	sEH	Nanobody	human samples in process	Cui et al. (2015)

Institution	Compound	Probe Type	Sample Type	Associated Publications
	Ochratoxin A	Nanobody	cereal	Liu et al. (2014)
	Carbaryl	Nanobody	spiked cereal	Liu et al. (2019)
	Trizophos	Nanobody	spiked water, soil, apple	Wang et al. (2019a)
	Fipronil	Nanobody	serum from exposed animals	Wang et al. (2019b)
	CIF	Nanobody	human samples in process	Vasylieva et al. (2019)
	2,4-D	Nanobody	environmental samples	In Progress
	CNAP	Nanobody	environmental samples	In Progress
	TETS	Nanobody	exposed animals	In Progress
UC San Diego	PBDE-100	Synthetically Evolved Receptors and Biosensors	environmental samples	In Progress
	Triclosan	Synthetically Evolved Receptors and Biosensors	N	In Progress
	Phananthrene	Synthetically Evolved Receptors and Biosensors	N	In Progress
	Polychlorinated biphenyl	Synthetically Evolved Receptors and Biosensors	N	In Progress
	TCDD	Synthetically Evolved Receptors and Biosensors	N	In Progress
	PCB-Aroclor	Synthetically Evolved Receptors and Biosensors	N	In Progress

Improvement, Harmonization, and Merging of Data Streams related to DNA Damage

About one-third of Superfund sites are contaminated with known DNA damaging contaminants. DNA damage can lead to mutations that result in disease. Researchers at the MIT SRP Center developed a high-throughput technology called CometChip, which measures the movement of DNA under electric current to quantify the level of DNA damage. This technology has been widely adopted by researchers, including researchers at the UNM SRP Center, who aim to use the tool to assess DNA damage from metal exposures and explore whether zinc supplementation reduces DNA damage.

The MIT and UNM SRP Centers are working together to harmonize existing data analysis approaches to share data generated from CometChip (see Figure 11) to better understand how environmental contaminants, such as metals, can modulate DNA damage and repair. The team aimed to merge existing datasets (see Table 11) from human and animal models to understand how levels of DNA damage in humans compare to mouse models. They also wanted to understand how CometChip data relates to absolute levels of DNA lesions in human tissues, which could reveal new insight into environmentally induced DNA damage in humans. Results from this project will inform future efforts to leverage animal data and facilitate further understanding of the mechanisms involved in DNA damage.

Interpreting the resulting cell cluster images from the CometChip assay can be challenging, and there are no standard methods for data analysis. For each experiment, thousands of comet images must be collected and analyzed. Different researchers use different analysis software and imaging conditions, resulting in variability between groups. The team also noted that there are currently no public repositories or ontologies for these data, which hinders data sharing and interoperability.

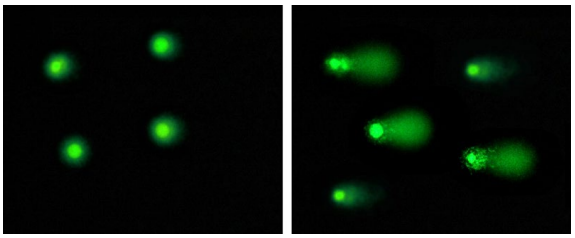

Figure 11. Image of DNA damage captured from CometChip analysis. A normal cell is on the left, whereas a cell exposed to a DNA-damaging chemical is on the right. (Image courtesy of Bevin Engelward)

The researchers set out to establish parameters and standards for Comet images so that they can be compatible with basic microscopes, establish an open-source image processing and analysis platform, and develop statistical approaches to optimize image analysis.

The team combined datasets from one million cell cluster images obtained from MIT CometChip assays with over 10,000 cell cluster images from UNM CometChip assays to analyze DNA damage in randomly selected samples from 253 individuals from the Navajo Nation (see Table 11).

Through this exercise, the MIT team observed several weaknesses in their analysis software, such as the need to adjust analysis parameters by hand, as well as issues when comparing results from different laboratories due to a lack of standards in algorithm and parameter selection.

To help users calibrate their data and facilitate data interpretation, the team used MIT's image data to generate a standard curve that compares cell irradiation to DNA strand breaks. This tool enables researchers to covert image data into quantified DNA strand breaks, which helps to estimate absolute levels of DNA damage.

The Use Case team is customizing an existing open-source, web-based platform used for data sharing and management, called SEEK, to create MIT SEEK, a repository for sharing Comet metadata between the two centers. This platform includes custom Excel spreadsheets to easily transfer metadata into the

SEEK platform, and custom unique identifiers to link samples with their corresponding protocols (Word documents or PDFs).

To increase the interoperability of metadata uploaded to this repository, the researchers defined their metadata variables. They also adapted their existing analysis software for automated data formatting, and to automatically export data analytics, including metadata, into Excel. Once the results from this analysis have been published, they plan to make their metadata publicly available by transferring it into FAIRDOM Hub, an open-access platform, built upon the SEEK software, used to share and publish data, models, and protocols.

The collaborators also created ontologies for Comet data and are working with researchers outside of their Use Case, who are currently using the CometChip, to finalize these ontologies to ensure that they meet the needs of other communities and allow for data interoperability across different institutions.

Additionally, they identified existing repositories that would be useful to increase the FAIR-ness of Comet data. For example, they are exploring the use of GitHub to make their software code publicly available. They also plan to publish their protocol(s) through their SRP websites and on NextGen Protocols, an open-access website developed by the MIT researchers to share protocols and standard operating procedures. They plan to make their metadata publicly available by transferring it into FAIRDOM Hub, an open-access platform, built upon the SEEK platform, used to share and publish data, models, and protocols. They will also share the raw data using an open-access hub hosted by MIT.

Through their work, the team learned that it is critical to have standards and protocols to make data FAIR before experiments are performed, since doing this retroactively can be time consuming. They also observed that training new graduate students and postdocs is key to advancing data science. The researchers noted that creating metadata in real time as experiments are being done is extremely important to ensure metadata quality and completeness, and to allow programmers to move metadata readily and easily into the SEEK platform.

According to the Use Case, this project changed their outlook on data management and organization. Prior to this project, there was no formal Data Management Plan for CometChip data, and this collaboration allowed the team to develop an overall plan to manage and analyze these data, which will not only benefit MIT and UNM, but will also be useful for other investigators looking at DNA damage.

Table 11. Existing CometChip datasets utilized by the MIT and UNM use case.

Institution	Dataset	Description	Variables
UNM	Navajo Birth Cohort Study: 253 individuals (202 pregnant women and 51 men)	> 10,000 cell cluster images from CometChip DNA damage analysis > Controlled exposure of human cells to DNA oxidation damage using hydrogen peroxide	DNA damage levels
MIT	CometChip data collected under the auspices of the MIT SRP; *in vitro* studies using mammalian cells	> 1 million cell cluster images from CometChip DNA damage analysis > standard curves for gamma irradiation to estimate percent comet tail for specific levels of DNA damage	DNA damage levels

Development of Interoperable Data Platforms to Define Best Practices and Data Sharing for Flow Cytometry

Flow cytometry, a technique used to detect and measure characteristics of cells, allows unprecedented detail in studies of the immune system and other areas of cell biology. Tens of thousands of cells can be quickly examined, and the data gathered are processed by a computer. However, this type of data can be complex to analyze and harmonize.

Collaborators at the UNM and University of Louisville (UL) SRP Centers set out to develop a platform to store and share diverse datasets obtained by flow cytometry. By integrating existing datasets, this Use Case aimed to better understand the effects of chemical exposures on circulating blood cells that cause immune injury or cardiovascular disease. Integrating existing datasets will also allow the researchers to apply flow data more broadly to understand if animal models can predict human immune responses to environmental exposures.

While the UNM and UL teams both utilize flow cytometry methods, their research has a different focus on exposures and cell endpoints. UNM focuses on immune-mediated injury induced by metals and metals mixtures, while UL focuses on cardiovascular diseases induced by volatile organic compounds (VOCs). Both centers complement their human studies with studies in animals to identify the molecular and cellular mechanisms that contribute to toxicity (see Table 12).

The researchers noted that flow cytometry data does not typically follow the FAIR principles since experimental approaches vary across labs and its interpretation can be hindered if not linked to sufficient metadata. They focused on creating and implementing usable protocols and platforms for storing, handling, and accessing flow cytometry data. The team worked closely with a bioinformatics consultant who provided data science training and advice on best practices to achieve their goals.

As a first step, the team developed a template for sharing flow cytometry metadata, which described parameters such as the type of instruments used to collect data, detector voltages, and how raw data were analyzed, which are specific to individual experiments. Their goal was to produce a structured form that can easily be incorporated into the data collection process and facilitate data integration. This template was based on a MiflowCyt (Lee et al. 2008) template, which was developed by experts in data science to establish the minimum information required to record and report data from flow cytometry experiments.

Using this template, the team documented experimental details and data, such as nomenclature and file names, for the mouse studies conducted at UNM and the human studies conducted at UL. Completed forms, metadata, and templates were uploaded into a test portal on the Environment Data Initiative Staging Environment.

The researchers also aimed to develop flow cytometry ontology tools that identify specific populations of cells of environmental health interest, such as peripheral blood mononuclear cells studied at UNM to explore the effects of arsenic on immune system suppression. They used the Open Biological and Biomedical Ontology Foundry website to identify a potential ontology for flow data and established the cell ontology, a structured controlled vocabulary for cell types, as a starting point. In addition, they identified that FlowCL (Courtot et al. 2015), a software package that performs semantic labelling of cell populations, would expand the ontology and would help reference it to certain cell populations. This

strategy allowed the team to successfully convert some of their datasets into a format that can be analyzed by others.

Through this data sharing effort, the collaborators hope to increase the breadth of toxicological outcomes from flow cytometry experiments performed at different institutions and advance the understanding of cell types different researchers are using. To increase access to their metadata template and data, and inform other research, the UNM team is creating a website that will allow easy access to their data portal. The UL team plans to establish their own portal and website as well.

From this collaboration, the researchers learned that analysis and data sharing is not possible without a common vocabulary and standard data reporting format. Each institution had a specific way of describing and collecting data, which hindered initial data integration and interoperability. They also learned that in future experiments, it is necessary to record much more detail about instrument settings, voltages and thresholds, and other data analysis steps.

Table 12. Existing flow cytometry datasets utilized by the UMN and UL use case.

Institution	Organism	Dataset
UNM	Mouse	> 100 mice for evaluation of cell surface marker expression and subset analysis of T, B, natural killer, and erythroid markers from bone marrow, spleen, and thymus tissues.
UL	Human	316 subjects assessed for 15 types of circulating angiogenic cells and platelet aggregates.

Improving Synchrotron-Based Data Access, Analysis and Workflows: Measuring the Concentration, Speciation and Distribution of Contaminants in Environmental and Biomedical Matrices

Researchers at the Columbia University, University of Arizona (UA), Dartmouth College, and UNM SRP Centers collaborated to explore how Synchrotron-based spectroscopic data from their Centers can be combined to better understand chemical speciation and the environmental and biochemical factors that control chemical form, retention, transport, and distribution of contaminants in diverse samples.

Synchrotrons use electrons accelerated to near light speed and steered by magnets to create beams of light that cause the chemical elements within a sample to fluoresce. Synchrotron data provides elemental abundance, distribution, and speciation data and is a highly sought-after technique. Detailed chemical speciation information (X-ray absorption spectroscopy) is obtained by comparing light absorbance patterns in the sample with known chemical references. The source of the reference materials is key in this analysis. Elemental imaging via X-ray microprobe maps shows where elements are within an intact sample, with a host of environment, biological and biomedical applications. While researchers are usually only interested in data for a few specific elements in elemental mapping, the synchrotron collects a full spectrum, providing information on a broad range of elements. Data for elements outside of the investigator's initial hypothesis are never re-used, despite their potential value.

Archiving this wealth of data would allow more researchers to leverage valuable data that is challenging to obtain.

By developing a series of spectral databases and verifiable and traceable reference materials with appropriate metadata, the team aimed to automate, integrate, and improve synchrotron data analysis to better quantify the distribution of contaminant species in environmental samples and within biological tissues. Specifically, they focused on how to combine their existing spectra from different environmental samples, images and spectra for specific elements and reference materials, spectra from biological samples, and metadata, including markers related to space and time (see Table 13). By creating and sharing a library for reference spectra, researchers could save time in future analyses because they would not be starting from scratch with every analysis. They also wanted to create automated user-friendly, web-based workflows to ensure uniformity in data analysis and robust QA/QC and make data collection formats more consistent and compatible moving forward.

The team worked through many challenges related to existing synchrotron data not being FAIR. Most important, different synchrotron beamlines are customized in different ways for each analysis and have their own evolving software, resulting in data with many different formats that are difficult to reproduce. They also noted the lack of organized data archives and the fact that legacy data is lost if it is not published. Another challenge was the lack of standard reference materials and libraries, data processing approaches, and inconsistent metadata. The team also noted a lack of ontologies for synchrotron data, and the fact that existing datasets are highly variable (e.g., ranging from field sites to individual particle scale) with inconsistent organization and storage by individual researchers. Finally, the team explained how establishing metadata standards for these samples are difficult because samples are usually selected for analysis because of their uniqueness.

An important component of their collaboration was focused on developing statistical approaches to determine reference integrity and data quality, and to create unsupervised spectral analysis tools. They wanted to enable advanced statistical analyses of synchrotron data from specific locations in three-dimensional space and time series, which included organizing their data and linking with appropriate standardized metadata. They also worked to develop a more standardized ontology for their data.

They began with partial processing to link metadata for different types of experiments to the data, environmental samples, and reference materials. They also needed to write small pieces of code that could be used to convert data and allow a single software package to process all the data. The team implemented a universal format for data independent of the collection platform and developed standard operating procedures for data collection and processing, and approaches to standardize how results are reported. Finally, they used the SMAK program (Sam's MicroAnalysis toolKit), developed by a beamline scientist, Sam Webb, at the Stanford synchrotron, to work with all the imaging data and facilitate additional analyses. The team's work helped standardize data collection so that future data can be more easily assimilated.

They created user-friendly web-based workflows within a storage system, called the Biological Elemental Imaging Database (BEID), for microprobe analysis and laser ablation ICP-MS data, and a related website and interoperability widget. While the database is not currently publicly available, BEID will ensure uniformity in data analysis, allow users to directly upload to the database, and conduct automated quality checks on submitted data. Work is ongoing regarding terms and conditions of data use and access and their data conversion widget, which slowed down the database. A short-term work

around involved taking the preview and compare capabilities of the database offline, and reverting to flat image files for data, but the team noted this was limiting as the data are meant to be interactive. They plan to continue working to address the software compatibility issues that will allow BEID to be optimally functional through internal testing and revisions and will publicly launch the database when it is final.

While the final database is not fully functional, the team made significant progress in leveraging and combining their data to answer new questions. Focusing on arsenic as an example, the team was interested in tools to predict whether arsenic is present or could be present in water based on specific properties of soil, such as pH, presence of iron and organic carbon, and redox state. They integrated data on reference standards (Nghiem et al. 2020) for iron and arsenic and worked to create unsupervised spectral analysis tools that integrate statistical clustering with important environmental outputs like water quality. They also wanted to leverage their data to determine if observed patterns were generalizable so that dense data collected at one site could be used to make predictions for other locations with less robust data. Leveraging existing data across the Use Case, they compared data from Vietnam with data from Minnesota using unsupervised approaches and identified similarities in iron minerology. They are in the process of preparing a publication related to this work and other collaborative data analyses that shed light on factors that control arsenic levels and toxicity in rice.

Through their work, the team learned that setting data up to be FAIR at the beginning of a project is much easier than working backward. They also learned that it is important to work closely with those who build the synchrotron machines to collaborate and build dataflows where the required metadata is automatically collected. The researchers noted that the most direct and clear way to share data with enough information to be useful is within publications as well as within a centralized database. According to the team, the Use Case collaboration directly increased their understanding of the systems they are investigating and their ability to link data types that previously could not be combined. Their efforts made considerable progress towards enabling interdisciplinary research and integrating data across domains.

Table 13. Existing synchrotron datasets utilized by the Columbia, Dartmouth, UA, and UNM use case

Collaborator	Project (dates of data collection)	# of Samples	# of References	Example variables
Columbia	As remediation of NPL sites (2007-present)	>200	>200	As, Fe, Mn references characteristic of neutral pH environments and aquifers, NPL-site characterization
	As mitigation in Bangladesh (2006-present)	>1000	>200	As, Fe, and Mn speciation in natural environmental sediments/soils
Arizona	Phytoremediation of mine tailings (2005-present)	>1000	>200	As, Pb, U, Zn, Fe, Mn, and S reference compounds characteristic of

Collaborator	Project (dates of data collection)	# of Samples	# of References	Example variables
				sulfide ore and mine-impacted soils/sediments/plant tissues
Dartmouth	Trace Core (2006-2017)	>500	>100	Synchrotron and elemental imaging data spanning, animal and human tissue specimens, related laser ablation ICP-MS
UNM	As/U immobilization (2011-present)	>100	>100	U, As, V, Fe and Mn reference spectra for natural and synthetic minerals and mineral mixtures typical of mining impacted areas

Combining and Harmonizing Disparate Datasets: Challenges and Opportunities

The Use Case Showcase represented a unique opportunity for teams to discuss the common challenges of sharing and integrating data and strategies to keep the momentum of their efforts moving forward.

For example, this exercise highlighted the need for more standardized data collection, research planning, and rigorous standard operating procedures to maximize data quality and reduce the amount of effort required to combine data in the future. Additional overarching challenges and strategies for addressing them are described below.

Metadata

A common challenge among Use Cases was missing or inconsistently reported metadata. Several groups struggled with integrating data because key information on how the data were collected, such as environmental sampling, lab protocols, and data formats, were not reported. Participants agreed that understanding what metadata is needed and having consistent metadata standards before setting up an experiment would help avoid having to work backwards to re-convert or re-annotate the data later, which can be time and resource consuming.

Since some repositories do not require metadata that is specific to types of studies, such as toxicology experiments or environmental studies, participants discussed the need for standardizing and improving metadata to facilitate data integration and reuse. The group agreed that a key aspect of standardizing metadata is standardizing data collection processes as well.

They also discussed the utility of minimum information checklists compared to robust quality metadata, and how to strike the right balance to identify what information is essential for understanding

experiment conditions and to make the data useful for reanalysis. They discussed how minimum requirements are useful to ensure that at least some metadata is included, but in some cases, such as MIxS or MIATE, it may be too minimal to be of use to researchers seeking to reuse the data. While robust metadata is the goal, participants also acknowledged an additional burden might be placed on researchers if too much metadata is required.

More standardized and robust metadata allows for more efficient data sharing and reuse. Since machine-readable metadata are essential for automatic discovery of datasets and services, standardized and descriptive metadata is an essential component of increasing data FAIR-ness. Participants discussed the best ways to incentivize researchers to include more robust metadata over time.

They also discussed where metadata should be stored and how relationships with raw data can be maintained if they exist in separate repositories.

Ontologies

Ontologies are unique and stable identifiers for entities ranging from chemicals to social factors recommended for use by a community. They can also characterize properties of and relationships between entities. Ontologies can facilitate data integration by standardizing vocabulary for describing different entities and relationships between them. At the beginning of their projects, some teams had ontologies available that worked well for their data while others only had partially applicable ontologies available. A common challenge was having to work backwards to fit previously collected data into an ontology after the fact, particularly when existing ontologies were not as well suited to the team's data.

Other teams did not have existing ontologies available to use as a starting point. They noted that it was difficult to reach consensus on ontologies even among a small group of collaborators, raising questions about how to scale up standard ontologies to larger, potentially international, networks.

They discussed the need for additional expertise in ontology development to help others understand how different disciplines and researchers interpret or use the same terms, and the utility of creating data dictionaries to define all the ways people describe a concept to improve interoperability (see Figure 12).

Participants agreed that standardizing ontologies for each domain is important for continuing efforts to improve data sharing across multiple laboratories. They discussed that some existing ontologies may work well as a launching point, but that each team should work closely with the ontology developers to add new terms and other enhancements that will make them more representative of the diversity and complexity of their data. They suggested that researchers work together to standardize terminology and improve existing ontologies continually as research evolves. There was general consensus that using an ontology should be coupled with an expectation that researchers contribute to its continual improvement.

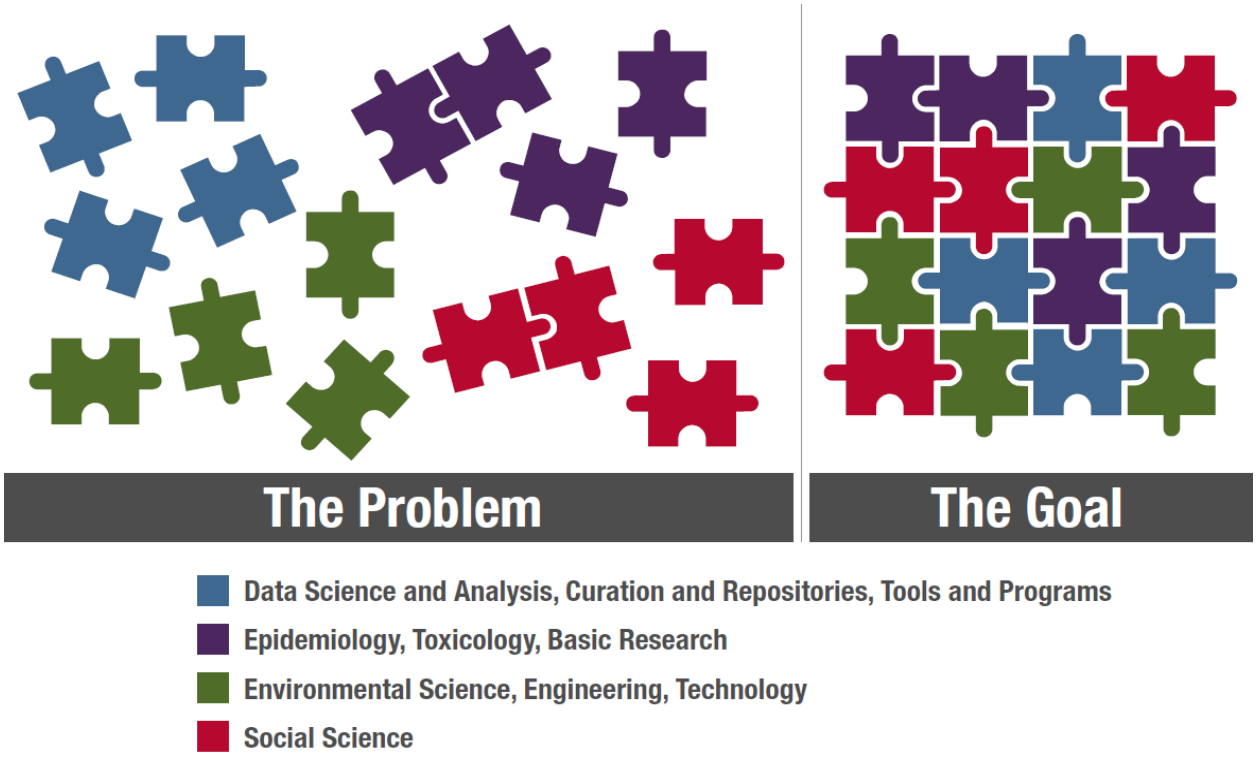

Figure 12. Data from different domains and data streams, and even within similar domains and streams, do not always talk to each other. The goal is to design research studies with interoperability in mind.

Data Repositories

Use Case teams noted that some data types lacked existing repositories to store data, or that existing repositories were too broad or too narrow in scope to be useful. They discussed the utility of having more general repositories built around a common scientific theme, such as contamination occurrence, and using unsupervised methods to compile raw data and widgets to toggle between domain-specific repositories.

They also explained how some existing repositories do not require enough metadata to make the data useful, and again touched on the idea that separate repositories for metadata and raw data would be helpful while maintaining links between the two.

While GitHub was noted to be a useful tool for storing code, participants agreed it is not well suited for long-term storage. Instead, the groups discussed Open Science Framework and containerization dockers, which allow users to deploy portals on their own servers, as favorable alternatives.

Another common topic of discussion for human health data was centered on data privacy, security, and personally identifiable information. Participants shared specific challenges, such as IRB approvals and data restrictions, particularly for working with tribal communities.

Analyzing Data

There was general agreement that reproducible analysis pipelines would be incredibly beneficial for moving the field forward. Several Use Case teams relied heavily on version controlled and sharable data analysis pipelines to ensure their process was reproducible and repeatable. One successful example is the use of containers to package the entire computing environment. This allows users to quickly run the same analyses with the same parameters and compare results.

Participants discussed the need for robust reference libraries, and the challenge of balancing large common libraries with many specific libraries. They agreed that interoperable and accessible reference libraries specific to their data types are necessary to make data FAIR and to avoid duplicating efforts. Some participants proposed the idea of giving users tools, such as an interoperability widget, to pull data from different repositories.

Another challenge some Use Case teams encountered was vendor-specific data formats (e.g., mass spectrometry data and synchrotron data), many of which are proprietary. They explained how this creates a barrier to pulling data into workflows and data sharing. They discussed how open data formats are needed from the onset, because often processing vendor-specific formats into open formats can result in lost data, links, or intrinsic information. The group agreed that it is important to convene relevant stakeholders, those responsible for defining raw data formats, and end users.

Success Stories

While all groups made significant strides towards improving the FAIR-ness of SRP data, some data was farther along in the readiness for interoperability spectrum at the beginning of the project. These groups were able to create sophisticated tools to help decision makers and communities better understand potential risks to their health.

For example, collaborators at Texas A&M University (Vasylieva et al.), Brown University, and UC San Diego SRP Centers created an online interactive map combining their individual datasets with city and local data. The map shows how factors like social vulnerability, impervious surfaces, and green space contribute to health risks.

Researchers from the Boston University and Dartmouth University SRP Centers created a searchable platform of publicly available data on contaminants in fish, environmental factors, and SRP data, that will underpin an interactive mapping tool showing contaminants in fish and health risks by region.

Other teams demonstrated the importance of close collaboration among researchers, data scientists, and data librarians. For example, researchers from the University of Iowa SRP Center worked closely with a data librarian and the Center's data management and analysis core to gain more insight on their metadata. The data services librarian brings unique insight by having conversations with each project to understand their specific needs and how to facilitate better data management and sharing practices across the broader University of Iowa SRP Center. They also offer a one-credit data management course, which has been a valuable resource for trainees. Similarly, the Research Data Management librarian at Brown University worked closely with researchers from Brown, TAMU, and UC San Diego on increasing the FAIR-ness of their Use Case datasets.

Recommendations

Despite the challenges encountered, there was clear agreement across Use Case teams that combining datasets can provide incredible insight and value to local, state, and federal decision makers for improving public health. These datasets can also be useful tools for community groups and other stakeholders. The teams discussed several recommendations to continue their momentum and increasing the FAIR-ness of data.

To harmonize data across disciplines, participants agreed that researchers need to practice good data stewardship and embrace data sharing best practices from the beginning by collecting, processing, and storing data in a standardized way. By working closely with data scientists, informaticists, and data management experts, SRP researchers will be able to develop more robust data documentation, data management protocols, data modeling, and more reproducible data that can be shared and maintained long-term.

One recommendation was for teams to develop a coordinated communication platform at the onset, such as Zoom or Slack, to encourage communication between researchers and data scientists and across research groups. They suggested these tools could also be used as platforms for training and education – a critical need agreed upon by all participants.

Another key theme that emerged was the question of sustainability since local portals or repositories developed would require consistent funding to maintain. The group suggested federal support, and that NIEHS, could consider taking on this role and maintaining environmental health data repositories.

The participants shared how federating disparate repositories to link between them would be beneficial. Since the ultimate goal of FAIR data is to make it machine readable, established metadata repositories linked through persistent identifiers to data repositories could be dispersed but connected with rich metadata. Dataverse is a successful example of this idea.

To overcome the challenge of private data, participants suggested making metadata records available in relevant repositories to make it compliant with FAIR Data Principles. Even if the underlying data cannot be shared publicly, users can search and find data of interest and request it following established protocols.

They also discussed incentivizing making data FAIR. One idea was to certify data at different levels based on the metadata included beyond the minimum requirements. Similarly, they discussed how data and software attribution, similar to the credit researchers get when their work is cited in other papers, could help incentivize making data available and FAIR.

Additionally, participants had specific recommendations for SRP researchers, including encouraging scientists to get involved in relevant stakeholder groups, such as the Genomic Standards Consortium, to advance needs related to environmental science for metadata standards. They also recommended working closely with data scientists during research planning and discussed how SRP researchers could include grant numbers when depositing data or creating repositories or ontologies so funding agencies can track the legacy of their investments over time.

Future Directions

The outcomes and lessons learned from the 19 Use Cases represent a culture change within the SRP research community and shed light on the need for improving how data is collected, annotated, and disseminated in the future. Key findings and novel approaches developed will be shared through several publications from Use Case teams, but an important goal of this exercise is to keep the momentum moving forward, building from these advances.

One of the major themes that emerged from the discussions is the incredible value of and need for additional training and capacity building throughout the research and data lifecycle. Specifically, the participants identified a need for training in what metadata to collect, how to implement standard vocabularies, and data management as important next steps. They also discussed the utility of "Code-A-Thons," potentially across the broader NIH network.

There was consensus among the participants that data science training for both principal investigators and trainees is critical. Training researchers in data science best practices gives them to the ability to plan projects to collect data with the FAIR principles in mind from the outset and to push their research in new and innovative directions. Investing in opportunities for trainees is also important to give them the skills in data management that they can apply to their own research and future careers. With these tools, SRP trainees will be well positioned to become leaders in the next generation of data-savvy researchers.

Another common theme was the need for early coordination and broader communication among researchers, data scientists, data librarians, and other stakeholders. In particular, data scientists and librarians should be involved in designing research studies to ensure data is collected with FAIR principles in mind. Data scientists could enhance research and analysis efforts by creating standard analysis pipelines and simple tools for researchers to use, such as containerizing workflows to make them more user-friendly.

The participants also discussed the need to encourage additional partnerships as a way of bringing diverse stakeholders together, including public health and community stakeholders, to continue discussions and identify solutions. The ultimate goal is to build a sustainable community of practice to meet the challenges of this rapidly evolving field.

Making data FAIR requires a range of interrelated activities throughout the research data lifecycle to facilitate data discoverability, access, and analysis. It requires a robust data management infrastructure, data-knowledgeable researchers, and skilled data scientists. Collectively, these components will form a data management and sharing ecosystem that can continue to address new and emerging challenges (see Figure 13).

Figure 13. SRP multiproject centers integrate diverse biomedical and environmental science and engineering research with community engagement, research translation, training, and data science to shed new light on a central research question. By expanding SRP data science initiatives beyond individual centers to the broader SRP, NIEHS, and scientific communities, SRP-funded grantees are helping to build and refine a data management and sharing community of practice to support a broader ecosystem of ecosystems.

References

Bever CR, Majkova Z, Radhakrishnan R, Suni I, McCoy M, Wang Y, et al. 2014. Development and utilization of camelid vhh antibodies from alpaca for 2,2',4,4'-tetrabrominated diphenyl ether detection. Anal Chem 86:7875-7882.

Bozack AK, Boileau P, Wei L, Hubbard AE, Sille FCM, Ferreccio C, et al. 2021. Exposure to arsenic at different life-stages and DNA methylation meta-analysis in buccal cells and leukocytes. Environ Health 20:79.

Buckman KL, Mason RP, Seelen E, Taylor VF, Balcom PH, Chipman J, et al. 2021. Patterns in forage fish mercury concentrations across northeast us estuaries. Environ Res 194:110629.

Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. 2016. Jbrowse: A dynamic web platform for genome visualization and analysis. Genome Biol 17:66.

Chang R, Yang B, Zhu QJ. 2019. Theoretical studies on the electronic structure parameters and reactive activity of neu5gc and neu5ac under food processing solvent environment. Molecules 24.

Cleary BM, Romano ME, Chen CY, Heiger-Bernays W, Crawford KA. 2021. Comparison of recreational fish consumption advisories across the USA. Curr Environ Health Rep 8:71-88.

Courtot M, Meskas J, Diehl AD, Droumeva R, Gottardo R, Jalali A, et al. 2015. Flowcl: Ontology-based cell population labelling in flow cytometry. Bioinformatics 31:1337-1339.

Cui Y, Li D, Morisseau C, Dong JX, Yang J, Wan D, et al. 2015. Heavy chain single-domain antibodies to detect native human soluble epoxide hydrolase. Anal Bioanal Chem 407:7275-7283.

Czaplicki LM, Gunsch CK. 2016. Reflection on molecular approaches influencing state-of-the-art bioremediation design: Culturing to microbial community fingerprinting to omics. J Environ Eng (New York) 142.

Deng P, Barney J, Petriello MC, Morris AJ, Wahlang B, Hennig B. 2019. Hepatic metabolomics reveals that liver injury increases PCB 126-induced oxidative stress and metabolic dysfunction. Chemosphere 217:140-149.

Fader KA, Nault R, Zhang C, Kumagai K, Harkema JR, Zacharewski TR. 2017. 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD)-elicited effects on bile acid homeostasis: Alterations in biosynthesis, enterohepatic circulation, and microbial metabolism. Sci Rep 7:5921.

Faust GG, Hall IM. 2014. Samblaster: Fast duplicate marking and structural variant read extraction. Bioinformatics 30:2503-2505.

Fostel JM, Burgoon L, Zwickl C, Lord P, Corton JC, Bushel PR, et al. 2007. Toward a checklist for exchange and interpretation of data from a toxicology study. Toxicol Sci 99:26-34.

Gadupudi GS, Klaren WD, Olivier AK, Klingelhutz AJ, Robertson LW. 2016a. PCB126-induced disruption in gluconeogenesis and fatty acid oxidation precedes fatty liver in male rats. Toxicol Sci 149:98-110.

Gadupudi GS, Klingelhutz AJ, Robertson LW. 2016b. Diminished phosphorylation of creb is a key event in the dysregulation of gluconeogenesis and glycogenolysis in pcb126 hepatotoxicity. Chem Res Toxicol 29:1504-1509.

Gadupudi GS, Elser BA, Sandgruber FA, Li X, Gibson-Corley KN, Robertson LW. 2018. Pcb126 inhibits the activation of ampk-creb signal transduction required for energy sensing in liver. Toxicol Sci 163:440-453.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol 5:R80.

Gilmour C, Bell T, Soren A, Riedel G, Riedel G, Kopec D, et al. 2018. Activated carbon thin-layer placement as an in situ mercury remediation tool in a penobscot river salt marsh. Sci Total Environ 621:839-848.

Hammond CM, Root RA, Maier RM, Chorover J. 2018. Mechanisms of arsenic sequestration by Prosopis juliflora during the phytostabilization of metalliferous mine tailings. Environ Sci Technol 52:1156-1164.

Hammond CM, Root RA, Maier RM, Chorover J. 2020. Arsenic and iron speciation and mobilization during phytostabilization of pyritic mine tailings. Geochim Cosmochim Acta 286:306-323.

Hardesty JE, Wahlang B, Falkner KC, Shi H, Jin J, Wilkey D, et al. 2019a. Hepatic signalling disruption by pollutant polychlorinated biphenyls in steatohepatitis. Cell Signal 53:132-139.
Hardesty JE, Wahlang B, Falkner KC, Shi H, Jin J, Zhou Y, et al. 2019b. Proteomic analysis reveals novel mechanisms by which polychlorinated biphenyls compromise the liver promoting diet-induced steatohepatitis. J Proteome Res 18:1582-1594.
Hobbie KA, Peterson ES, Barton ML, Waters KM, Anderson KA. 2012. Integration of data systems and technology improves research and collaboration for a superfund research center. J Lab Autom 17:275-283.
Honeker LK, Gullo CF, Neilson JW, Chorover J, Maier RM. 2019. Effect of re-acidification on buffalo grass rhizosphere and bulk microbial communities during phytostabilization of metalliferous mine tailings. Front Microbiol 10:1209.
Hottenstein JD, Neilson JW, Gil-Loaiza J, Root RA, White SA, Chorover J, et al. 2019. Soil microbiome dynamics during pyritic mine tailing phytostabilization: Understanding microbial bioindicators of soil acidification. Front Microbiol 10:1211.
Ikuma K, Gunsch CK. 2012. Genetic bioaugmentation as an effective method for in situ bioremediation: Functionality of catabolic plasmids following conjugal transfers. Bioengineered 3:236-241.
Jurgelewicz A, Dornbos P, Warren M, Nault R, Arkatkar A, Lin H, et al. 2021. Genetics-based approach to identify novel genes regulated by the aryl hydrocarbon receptor in mouse liver. Toxicol Sci 181:285-294.
Kim HJ, McCoy MR, Majkova Z, Dechant JE, Gee SJ, Tabares-da Rosa S, et al. 2012. Correction to isolation of alpaca antihapten heavy chain single domain antibodies for development of sensitive immunoassay. Anal Chem 84:6919.
Kotalik CJ, Cadmus P, Clements WH. 2021. Before-after control-impact field surveys and novel experimental approaches provide valuable insights for characterizing stream recovery from acid mine drainage. Sci Total Environ 771:145419.
Lang AL, Chen L, Poff GD, Ding WX, Barnett RA, Arteel GE, et al. 2018. Vinyl chloride dysregulates metabolic homeostasis and enhances diet-induced liver injury in mice. Hepatol Commun 2:270-284.
Lee JA, Spidlen J, Boyce K, Cai J, Crosbie N, Dalphin M, et al. 2008. Miflowcyt: The minimum information about a flow cytometry experiment. Cytometry A 73:926-930.
Liu X, Xu Y, Xiong YH, Tu Z, Li YP, He ZY, et al. 2014. Vhh phage-based competitive real-time immuno-polymerase chain reaction for ultrasensitive detection of Ochratoxin A in cereal. Anal Chem 86:7471-7477.
Liu Z, Wang K, Wu S, Wang Z, Ding G, Hao X, et al. 2019. Development of an immunoassay for the detection of carbaryl in cereals based on a camelid variable heavy-chain antibody domain. J Sci Food Agric 99:4383-4390.
Marvel SW, To K, Grimm FA, Wright FA, Rusyn I, Reif DM. 2018. Toxpi graphical user interface 2.0: Dynamic exploration, visualization, and sharing of integrated data models. BMC Bioinformatics 19:80.
Mattes TE, Ewald JM, Liang Y, Martinez A, Awad A, Richards P, et al. 2018. PCB dechlorination hotspots and reductive dehalogenase genes in sediments from a contaminated wastewater lagoon. Environ Sci Pollut Res Int 25:16376-16388.
Nault R, Fader KA, Zacharewski T. 2015. RNA-seq versus oligonucleotide array assessment of dose-dependent tcdd-elicited hepatic gene expression in mice. BMC Genomics 16:373.
Nault R, Fader KA, Ammendolia DA, Dornbos P, Potter D, Sharratt B, et al. 2016a. Dose-dependent metabolic reprogramming and differential gene expression in TCDD-elicited hepatic fibrosis. Toxicol Sci 154:253-266.
Nault R, Fader KA, Kirby MP, Ahmed S, Matthews J, Jones AD, et al. 2016b. Pyruvate kinase isoform switching and hepatic metabolic reprogramming by the environmental contaminant 2,3,7,8-tetrachlorodibenzo-p-dioxin. Toxicol Sci 149:358-371.

Nault R, Fader KA, Lydic TA, Zacharewski TR. 2017. Lipidomic evaluation of aryl hydrocarbon receptor-mediated hepatic steatosis in male and female mice elicited by 2,3,7,8-tetrachlorodibenzo-p-dioxin. Chem Res Toxicol 30:1060-1075.
Nault R, Fader KA, Bhattacharya S, Zacharewski TR. 2021. Single-nuclei RNA sequencing assessment of the hepatic effects of 2,3,7,8-tetrachlorodibenzo-p-dioxin. Cell Mol Gastroenterol Hepatol 11:147-159.
Neal-Walthall N NU, Hsu-Kim H. Utility of diffusive gradient in thin-film (dgt) passive samplers for predicting mercury methlylation potential and bioaccumulation in freshwater wetlands. Manuscript in Prep.
Nelson KN, Neilson JW, Root RA, Chorover J, Maier RM. 2015. Abundance and activity of 16s rRNA, amoa and nifh bacterial genes during assisted phytostabilization of mine tailings. Int J Phytoremediation 17:493-502.
Nghiem AA, Shen Y, Stahl M, Sun J, Haque E, DeYoung B, et al. 2020. Aquifer-scale observations of iron redox transformations in arsenic-impacted environments to predict future contamination. Environ Sci Technol Lett 7:916-922.
Oleksiak MF, Karchner SI, Jenny MJ, Franks DG, Welch DB, Hahn ME. 2011. Transcriptomic assessment of resistance to effects of an aryl hydrocarbon receptor (AHR) agonist in embryos of atlantic killifish (fundulus heteroclitus) from a marine superfund site. BMC Genomics 12:263.
Osterberg JS, Cammen KM, Schultz TF, Clark BW, Di Giulio RT. 2018. Genome-wide scan reveals signatures of selection related to pollution adaptation in non-model estuarine atlantic killifish (fundulus heteroclitus). Aquat Toxicol 200:73-82.
Oziolor EM, Reid NM, Yair S, Lee KM, Guberman VerPloeg S, Bruns PC, et al. 2019. Adaptive introgression enables evolutionary rescue from extreme environmental pollution. Science 364:455-457.
Petriello MC, Brandon JA, Hoffman J, Wang C, Tripathi H, Abdel-Latif A, et al. 2018a. Dioxin-like PCB 126 increases systemic inflammation and accelerates atherosclerosis in lean ldl receptor-deficient mice. Toxicol Sci 162:548-558.
Petriello MC, Charnigo R, Sunkara M, Soman S, Pavuk M, Birnbaum L, et al. 2018b. Relationship between serum trimethylamine n-oxide and exposure to dioxin-like pollutants. Environ Res 162:211-218.
Petriello MC, Hoffman JB, Vsevolozhskaya O, Morris AJ, Hennig B. 2018c. Dioxin-like PCB 126 increases intestinal inflammation and disrupts gut microbiota and metabolic homeostasis. Environ Pollut 242:1022-1032.
Powell CD, Moseley HNB. 2021. The mwtab Python library for RESTful access and enhanced quality control, deposition, and curation of the metabolomics workbench data repository. Metabolites 11.
Reid NM, Proestou DA, Clark BW, Warren WC, Colbourne JK, Shaw JR, et al. 2016. The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. Science 354:1305-1308.
Reid NM, Jackson CE, Gilbert D, Minx P, Montague MJ, Hampton TH, et al. 2017. The landscape of extreme genomic variation in the highly adaptable atlantic killifish. Genome Biol Evol 9:659-676.
Root RA, Hayes SM, Hammond CM, Maier RM, Chorover J. 2015. Toxic metal(loid) speciation during weathering of iron sulfide mine tailings under semi-arid climate. Appl Geochem 62:131-149.
Smelter A, Moseley HNB. 2018. A Python library for FAIRer access and deposition to the metabolomics workbench data repository. Metabolomics 14:64.
Stedtfeld RD, Stedtfeld TM, Fader KA, Williams MR, Bhaduri P, Quensen J, et al. 2017. TCDD influences reservoir of antibiotic resistance genes in murine gut microbiome. FEMS Microbiol Ecol 93.
Taylor VF, Buckman KL, Seelen EA, Mazrui NM, Balcom PH, Mason RP, et al. 2019. Organic carbon content drives methylmercury levels in the water column and in estuarine food webs across latitudes in the northeast united states. Environ Pollut 246:639-649.
Valentin-Vargas A, Neilson JW, Root RA, Chorover J, Maier RM. 2018. Treatment impacts on temporal microbial community dynamics during phytostabilization of acid-generating mine tailings in semiarid regions. Sci Total Environ 618:357-368.

Vasylieva N, Kitamura S, Dong J, Barnych B, Hvorecny KL, Madden DR, et al. 2019. Nanobody-based binding assay for the discovery of potent inhibitors of cftr inhibitory factor (cif). Anal Chim Acta 1057:106-113.
Wahlang B, Song M, Beier JI, Cameron Falkner K, Al-Eryani L, Clair HB, et al. 2014. Evaluation of aroclor 1260 exposure in a mouse model of diet-induced obesity and non-alcoholic fatty liver disease. Toxicol Appl Pharmacol 279:380-390.
Wahlang B, Petriello MC, Perkins JT, Shen S, Hennig B. 2016a. Polychlorinated biphenyl exposure alters the expression profile of micrornas associated with vascular diseases. Toxicol In Vitro 35:180-187.
Wahlang B, Prough RA, Falkner KC, Hardesty JE, Song M, Clair HB, et al. 2016b. Polychlorinated biphenyl-xenobiotic nuclear receptor interactions regulate energy metabolism, behavior, and inflammation in non-alcoholic-steatohepatitis. Toxicol Sci 149:396-410.
Wahlang B, Barney J, Thompson B, Wang C, Hamad OM, Hoffman JB, et al. 2017a. Editor's highlight: PCB126 exposure increases risk for peripheral vascular diseases in a liver injury mouse model. Toxicol Sci 160:256-267.
Wahlang B, Perkins JT, Petriello MC, Hoffman JB, Stromberg AJ, Hennig B. 2017b. A compromised liver alters polychlorinated biphenyl-mediated toxicity. Toxicology 380:11-22.
Wang J, Bever CR, Majkova Z, Dechant JE, Yang J, Gee SJ, et al. 2014. Heterologous antigen selection of camelid heavy chain single domain antibodies against tetrabromobisphenol A. Anal Chem 86:8296-8302.
Wang K, Liu Z, Ding G, Li J, Vasylieva N, Li QX, et al. 2019a. Development of a one-step immunoassay for triazophos using camel single-domain antibody-alkaline phosphatase fusion protein. Anal Bioanal Chem 411:1287-1295.
Wang K, Vasylieva N, Wan D, Eads DA, Yang J, Tretten T, et al. 2019b. Quantitative detection of fipronil and fipronil-sulfone in sera of black-tailed prairie dogs and rats after oral exposure to fipronil by camel single-domain antibody-based immunoassays. Anal Chem 91:1532-1540.
Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018.
Wu X, Yang J, Morisseau C, Robertson LW, Hammock B, Lehmler HJ. 2016. 3,3',4,4',5-pentachlorobiphenyl (pcb 126) decreases hepatic and systemic ratios of epoxide to diol metabolites of unsaturated fatty acids in male rats. Toxicol Sci 152:309-322.