Review

Michelle L. Heacock*, Sara M. Amolegbe, Lesley A. Skalla, Brittany A. Trottier, Danielle J. Carlin, Heather F. Henry, Adeline R. Lopez, Christopher G. Duncan, Cindy P. Lawler, David M. Balshaw and William A. Suk*

# Sharing SRP data to reduce environmentally associated disease and promote transdisciplinary research

https://doi.org/10.1515/reveh-2019-0089
Received November 6, 2019; accepted January 6, 2020; previously published online March 3, 2020

**Abstract:** The National Institute of Environmental Health Sciences (NIEHS) Superfund Basic Research and Training Program (SRP) funds a wide range of projects that span biomedical, environmental sciences, and engineering research and generate a wealth of data resulting from hypothesis-driven research projects. Combining or integrating these diverse data offers an opportunity to uncover new scientific connections that can be used to gain a more comprehensive understanding of the interplay between exposures and health. Integrating and reusing data generated from individual research projects within the program requires harmonization of data workflows, ensuring consistent and robust practices in data stewardship, and embracing data sharing from the onset of data collection and analysis. We describe opportunities to leverage data within the SRP and current SRP efforts to advance data sharing and reuse, including by developing an SRP dataset library and fostering data integration through Data Management and Analysis Cores. We also discuss opportunities to improve public health by identifying parallels in the data captured from health and engineering research, layering data streams for a more comprehensive picture of exposures and disease, and using existing SRP research infrastructure to facilitate and foster data sharing. Importantly, we point out that while the SRP is in a unique position to exploit these opportunities, they can be employed across environmental health research. SRP research teams, which comprise cross-disciplinary scientists focused on similar research questions, are well positioned to use data to leverage previous findings and accelerate the pace of research. Incorporating data streams from different disciplines addressing similar questions can provide a broader understanding and uncover the answers to complex and discrete research questions.

**Keywords:** data integration; data sharing; environmental health data; exposures and health; transdisciplinary research.

## Introduction

Large quantities of data are generated in scientific research each day and large-scale datasets are becoming the norm in many domains of science (1). This, together with advances in technology, provides a unique and timely opportuntiy to leverage data assets across all scientific domains. With more efficient data collection methods and rapid innovations in computer processing and data analytics, researchers are tapping into previously inaccessible data and extracting new information. Innovations that make data more accessible, such as improved data storage and curation, open possibilities for researchers to use data in ways not previously possible, but also require us to think about how data are generated, managed, and leveraged to improve scientific understanding.

***Corresponding authors: Michelle L. Heacock and William A. Suk,** Superfund Research Program, National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), Department of Health and Human Services (DHHS), Research Triangle Park, NC, USA, E-mail: heacockm@niehs.nih.gov (M.L. Heacock); suk@niehs.nih.gov (W.A. Suk)
**Sara M. Amolegbe, Lesley A. Skalla and Adeline R. Lopez:** MDB, Inc., Durham, NC, USA
**Brittany A. Trottier, Danielle J. Carlin and Heather F. Henry:** Superfund Research Program, National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), Department of Health and Human Services (DHHS), Research Triangle Park, NC, USA
**Christopher G. Duncan, Cindy P. Lawler and David M. Balshaw:** National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), Department of Health and Human Services (DHHS), Research Triangle Park, NC, USA

Open Access. © 2020 Michelle L. Heacock, William A. Suk et al., published by De Gruyter. CC BY This work is licensed under the Creative Commons Attribution 4.0 Public License.

Encouraging the sharing of research data by making it available in public repositories is important because it enables new discoveries, stimulates new collaborations, increases the utility of research results, and strengthens scientific transparency and rigor. When combined with other data, high-quality, well-described data may become more useful, thus maximizing research dollars.

Advances in analytical technologies such as machine learning, a method of data analysis that automates analytical model building based on the idea that systems can infer from data and identify patterns without explicit programming by humans, are providing new opportunities to integrate data. Because one single data type cannot capture the complexity of disease, integrative methods combining data from multiple domains are emerging to provide a more comprehensive view of the different dimensions involved in health and the environment (2).

Researchers in biology-related fields have used data science tools to understand gene-environment interactions and biological mechanisms of disease. The National Institutes of Health (NIH) defines data science as "the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data" (3). In chemical science and engineering fields, researchers have used data science to monitor chemical processes, gain an understanding of chemical activity, and predict activity of a chemical based on its properties, but there is still a lot of potential growth in this field (4). As data science rapidly evolves, it creates new opportunities to address complex environmental health science (EHS) questions by integrating these types of data, as well as many other data types from a variety of sources.

Data integration presents many challenges, which include being able to access and use data from different fields, requiring sound practices in data stewardship that preserve and improve content, accessibility, and usability of data and metadata (5). Different research disciplines, and even discrete communities within them, often format and describe data differently, pointing to the need to standardize these processes to build upon and integrate data streams (6). The FAIR Data Principles, which are guidelines to improve the reusability of data, were developed to facilitate these goals and have been widely accepted across the various areas of science (7). The FAIR Principles – Findable, Accessible, Interoperable, and Reusable – emphasize the importance of good data management to support discovery. The principles include:

- Developing more descriptive metadata, which provides contextual information about data, that is indexed in searchable resources with unique identifiers so it is findable;
- Making it retrievable by personal identifier with clear permissions so it is accessible to both humans and machines;
- Using shared vocabularies and data standards so it is interoperable; and
- Publishing data in sources with rich enough metadata, including versions of data and software used, to enable proper reuse.

An important aspect of these guidelines centers on interoperability, or the harmonization of data from different resources with regard to structure, formatting, and annotation, which will facilitate the ability of researchers to open their content to integrative analysis (8). This requires standard formats and ontologies, which encompass formal naming and definition of categories and properties of data, as well as annotation of the data with rich and descriptive metadata that facilitates integration of data streams to reuse data (7).

An important step in identifying and combining available and relevant data is assessing potential bias associated with each data type (9). Potential sources of bias may be identified based on how the data were produced or used, requiring descriptive metadata so that scientists can assess the risks of bias associated with each data type so that the risk can be incorporated into data analysis.

In order for other researchers to properly reproduce published data, descriptive metadata should include versions of software and hardware used throughout the study, software code, and versions of data produced or used (10, 11). Tracking workflow and using a version control system (e.g. GitHub, Bitbucket, Jupyter Notebooks, Code Ocean, etc.) during research and analysis helps simplify making code available at the time of publication and ensures that the correct version of a data source is reported (11). Following FAIR principles enables the use of powerful analytics tools to access data for machine learning and prediction (12). It also provides an opportunity for researchers to extract maximum benefit from investments in research.

## Initiatives in data sharing

To accelerate the translation of research into knowledge and maximize the utility of federally funded data, federal agencies have set forth policies and guidelines to expand the scope of data stewardship. In February 2013, a memorandum from the Executive Office of Science and

Technology Policy was released, intended to ensure that, "the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community" (13). Following this directive, federal agencies have issued their own policies and guidelines for data sharing and management (5).

For example, the NIH issued the Genomic Data Sharing Policy in 2014 to promote robust sharing of genomic data and to provide appropriate protections for research involving human data (14). More recently, NIH released its first Strategic Plan for Data Science in June 2018, which outlines a roadmap to modernize how data are stored, managed, standardized, and published. The strategic plan highlights the need to make data resources accessible, well organized, secure, and efficiently operated (15). The U.S. Geological Survey (USGS) requires every research project funded or managed by the USGS to include a data management plan within their project work plan prior to initiation of the project (16). The National Science Foundation (NSF) also has a data sharing policy, which requires investigators to share with other researchers the primary data, samples, physical collections, and other supporting materials created or gathered in the course of work under NSF grants (17). As technology continues to advance, federal agencies aim to maximize the value of data generated through federally funded efforts and accelerate the pace of scientific discoveries.

The NIH National Institute of Environmental Health Sciences (NIEHS) emphasizes in its strategic plans that development of innovative data science and data-driven approaches is integral to EHSs. The new plan affirms the importance of developing data sharing platforms and integrating and synthesizing data and research findings in a way that will make a meaningful impact on public health (15). As part of the NIEHS, the Superfund Basic Research and Training Program (SRP) is well positioned to utilize data approaches and maximize the impact of its data to answer complex environmental health questions. The SRP was established with a broad set of mandates focused on understanding the effects of hazardous substances on human health and developing methods and technologies to detect and reduce the amount and toxicity of hazardous substances in the environment (18).

To address these mandates, the NIEHS SRP has funded transdisciplinary research in order to address broad, complex issues related to exposures to hazardous substances. The SRP multiproject center grants are the mainstay of the program, where SRP centers within universities build teams of scientists and engineers working in different fields to tackle complex but targeted problems in environmental health. The centers consist of several projects and cores, deliberately designed to address discrete research questions that contribute to a critical piece of the center's research focus. Researchers are also asked to integrate among projects and cores, forming close collaborations toward a broader understanding of the center's common theme. Usually that integration is at a level of sharing samples and perspectives, but there is an important opportunity to integrate data meaningfully. As described in the SRP 2015–2020 strategic plan, the SRP aims to support integration of multidisciplinary environmental health research data to achieve problem-solving, solution-oriented research. The strategic plan outlines SRP goals to facilitate the development of coordination and data management infrastructure to enhance grantee capacity to integrate diverse data from multidisciplinary research across multiple platforms (19).

## Utilizing data from the Superfund Research Program

SRP grantees have generated a wide range of data types from research on exposures, health outcomes and underlying biological mechanisms, fate and transport of chemicals through the environment, and technologies to reduce contaminants in the environment through an understanding of chemical properties and biogeochemical interactions. Figure 1 provides a summary of the types of science conducted within the SRP, based on a map of publications by SRP grantees, which span the fields of math, physics, chemistry, engineering, computer science, biotechnology, earth sciences, biology, the health professions, infectious disease, medical specialties, brain research, and social science. This diversity of research sheds light on the potential opportunities for data sharing across disciplines and research projects. By applying research findings of one discipline into the context of another, we may be able to uncover new, and unforeseen, connections.

SRP publications from 1995 to 2018 (n = 8458) mapped to the University of California, San Diego Map of Science (Sci2) tool (20) which visualizes 13 main disciplines of science within a network of 554 subdisciplines. SRP publication records from Clarivate Analytics' Web of Science were loaded into Sci2 and overlaid on the Map of Science based on journal title, using the Science Map via Journals visualization option. Disciplines are labeled by color, and circles are proportional to the number of publications. One thousand one hundred and twenty-five out of 1310 journals were mapped to 297 subdisciplines and 12 disciplines. One hundred and eighty-five journals are

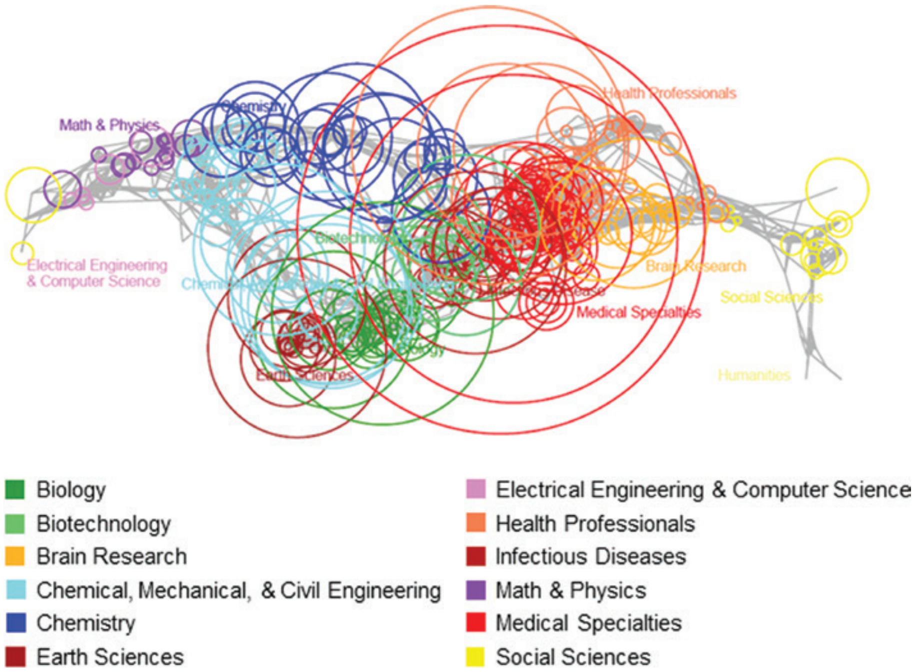

**Figure 1:** Map of SRP publications by discipline.

considered unclassified by Sci2 map and therefore do not appear in visualization.

Centers generate varied data types and formats across research projects that represent the different fields within a given center and across the program as a whole. The opportunity to maximize data generated lies with the shared research question and goals that are a part of each center and the shared mandates across the SRP. Additionally, for studies that capture non-targeted data (e.g. mass spectrometry analyses), only a portion of a rich dataset collected as part of a project may be useful to their specific research objectives, but could be useful to another research project, particularly when brought together with other similar data sets or data streams to fill in data gaps or reinforce research results. Sharing and integrating datasets, regardless of the size, could help answer challenging questions about the role of environmental exposures, their interplay with biological response in human disease (e.g. resilience and susceptibility to exposures), and how to reduce those exposures to improve human health.

## Where we are now: SRP efforts to advance data sharing

In order to encourage and support sharing of data, SRP initiated efforts to increase dialogue surrounding the need for data sharing and integration, facilitated access to publicly available datasets, and is now adding a requirement for a center core focused on data management and analysis.

## Programmatic efforts to facilitate data sharing dialogue

To address the increased need to share data in response to strategic plans across many agencies, the SRP and its grantees have been proactive in recognizing the importance of leveraging, integrating, and reusing data. To encourage data sharing, integration, and reuse between SRP grantees, the SRP has made use of time at annual meetings, convening sessions and keynote speakers, to highlight successful data sharing, facilitate conversations, and receive feedback from its grantees on how research findings from one field might be used in other analyses. SRP researchers have also been involved in workshops to promote data integration, including the National Academy of Sciences, Engineering, and Medicine workshop on informing environmental health decisions through data integration (21).

In 2015, the SRP hosted a webinar series, Integrating Data from Multidisciplinary Research, to explore the challenges and opportunities in integrating datasets to solve complex environmental health problems (22). The series, which included presenters from the NSF, the USGS, and the NIEHS Children's Health Exposure Analysis Resource (CHEAR), introduced the importance of data sharing and highlighted projects that enable data discovery and facilitate sharing of research and expertise. The webinar series provided a way for SRP staff and grantees to see how other EHS researchers are tackling data sharing. The webinars also provided the opportunity to interact with data sharing experts in the field, ask questions, and share experiences. Below we describe strategies that could be employed across the broader environmental health fields using specific examples from the SRP. These data sharing goals are not unique to the SRP and may be shared with the wider research community.

SRP program staff also encourage collaborations between researchers with shared interests through the K.C. Donnelly Externship. The externship provides a vehicle for SRP trainees to collaborate with researchers outside their home institutions while broadening their expertise and skills. The opportunity also facilitates research collaborations and data sharing among grantees (23).

## Developing an SRP dataset library

Datasets are important products of scientific research. Since the inception of the program in 1987, SRP grantees have produced more than 12,000 publications as part of their SRP-funded research projects, but only a small percentage of these publications have associated datasets available in public repositories. It is likely that much of the data generated as part of the research described in these publications are so-called dark data, where the data are either not shared at all or lost in supplementary information (24). According to an International Data Corporation study, more than 90% of digital data are dark data (25).

Datasets archived in journal websites as supporting files are not indexed, and therefore are effectively hidden from discovery. Another study found that only about 20% of data availability statements in publications indicate that data are deposited in a repository (26).

Supplemental data that may no longer have value in one field may be useful to another research project and should therefore, in many cases, also be publicly available, meaning that they are available online for anyone to use. There is likely to be additional data beyond what is captured in underlying primary publications, including study data reflecting null findings that are often not published, which could be relevant and repurposed by another researcher to address a different research question. To begin to make data more findable and accessible, and to provide the research context from which the data were generated, the SRP created and maintains a searchable catalog of publicly available datasets from SRP-supported publications (27).

The dataset listing is housed on the SRP website, which is generally used to communicate major research findings and events from across the program. The datasets are linked to their associated SRP projects and users can browse or search dataset tables by dataset title, data type, data repository, organism, dataset accession number, and SRP project(s), making the data more findable and accessible. Each dataset record contains metadata, including data type, experiment type, organism, and a link to the actual dataset and information (e.g. abstract and project updates) on related SRP publications and projects. This added context about how the data were generated within an SRP center provides users with additional information that is not normally available in a dataset repository. Users can also search the SRP publications associated with the datasets for specific terms, such as a target chemical, to find datasets associated with research-specific topics. This is not a data repository, but a listing of SRP publication-associated datasets, which encourages access to shared SRP data and provides one way to increase findability of SRP data. Specifically, these datasets are connected to a description of the research project from which the data were generated, providing critical additional information on the context of the dataset that may not be clear in a large data repository of diverse datasets.

To populate this dataset listing, PubMed and Clarivate Analytics' Data Citation Index on Web of Science are regularly searched to identify publications and public datasets associated with SRP funding. In addition, dataset information is provided by SRP grantees to program staff. Datasets are added to the existing SRP database and displayed on the SRP website (3415 datasets included as of January 21, 2020). Examples of data types include DNA and protein sequences, gene expression, genotype and phenotype, proteomics, exposure, environmental science, and clinical trial data. Table 1 provides a summary of the number of datasets by general data type in the SRP database. The database is currently populated primarily by nucleotide sequence and gene expression data. Because SRP research being conducted across the program is much broader in scope, as shown in Figure 1, there is potential to broaden the publicly available dataset data to include much more diverse data types.

Providing researchers and the public a way to find and access SRP-funded data increases the impact and visibility of SRP research and promotes data sharing and collaboration. However, identifying SRP-funded datasets and

**Table 1:** Number of datasets by general data type in the SRP database (as of January 21, 2020).

| General data type | Number of datasets in the SRP database |
|---|---:|
| 3D structural | 7 |
| Carcinogenicity | 2 |
| Chemistry and Chemical Biology | 9 |
| Clinical trials | 5 |
| Environmental science data | 3 |
| Exposure data | 3 |
| Gene expression | 191 |
| Genotype/phenotype | 11 |
| Geospatial data | 1 |
| Metabolomics | 1 |
| Nucleotide sequence | 3103 |
| Protein sequence | 6 |
| Protein structure | 66 |
| Proteomics | 3 |
| Software applications | 1 |
| Source code | 2 |

tracking data sharing can be difficult. While many scientific journals and funding agencies mandate making data freely available to the public, data sharing and the practice of data citation are still in early development. Data mentions and citations vary greatly across research fields and formal citation of datasets using digital object identifiers (DOIs) and data citations is limited, with the majority of datasets archived in the journal's supplementary materials (28). This is a problem because when data are provided in the supplemental information of a publication, they can be difficult to search, find, and reuse. Datasets that are archived in data repositories and assigned DOIs or accession numbers are much more likely to be cited and therefore tracked.

### Enabling data integration through data management cores

In its most recent multiproject center request for applications, SRP added a requirement for a Data Management and Analysis Core (DMAC) (RFA-ES-18-002, https://grants.nih.gov/grants/guide/rfa-files/rfa-es-18-002.html). The primary purpose of the DMAC is to support the management and integration of data across the center, which includes establishing, coordinating, and monitoring steps for collecting, processing, and analyzing data. As it is critical for data scientists to work closely with the subject matter/research domain experts, the SRP intentionally designed the DMACs so this interaction is central to this research support core. For this reason, the DMAC will work closely with project and core leaders to ensure high data quality throughout the data life cycle. The DMAC will also work with project and core leaders to identify opportunities for integrating project and core-generated data with other existing datasets. DMAC leads will also be interacting across the program so that challenges and opportunities can be discussed along the way. The goal of the DMAC is to create shared data standards and systems between biomedical and environmental science and engineering projects, improving the extent to which different projects can exchange and interpret data. By facilitating data integration, the DMAC aims to accelerate the impact of the center's research (29). Including data management plans that consider data collection for a given research project in the context of the larger center research questions can help to maximize the potential of future data integration among research projects.

## Opportunities to connect disciplines and improve public health

Incorporating a variety of expertise in research can strengthen study design, data collection and analysis, data structure, and data use. Multidisciplinary research programs, like the SRP centers, can make use of their existing data and research infrastructures to answer new questions.

To leverage resources and maximize productivity, it is important to develop ways to inventory existing data as well as integrate and store new data that is continuously created. Creating repositories to store and collect data, as well as interfaces to facilitate data integration and analysis, will be important to maximize the benefits of research by making data FAIR.

For example, in the field of geosciences, a data repository is being developed to catalog geoscience data, known as the Community Inventory of EarthCube Resources for Geosciences Interoperability (CINERGI) data portal (30). The goal of CINERGI is to compile an inventory of geoscience data while developing mechanisms to ensure that different resources have consistent and easy-to-interpret descriptions, traceable origins, and documentation that is as complete as possible. To be useful, such portals must have sufficient metadata and targeted keywords so that users can easily find the data across varying domains. Data must also be annotated to indicate its version or date (31).

Another opportunity to integrate data may come from the use of a knowledge base, a kind of repository that accumulates, organizes, and links growing bodies of information related to core datasets, and typically requires more human curation of information beyond annotation needed for databases (3). Knowledge bases can enable data sharing by providing additional information and context about different types of data and how they are organized for researchers from other disciplines. They may also help bridge the gap between how different researchers consider, describe, and approach data (32).

SciCrunch is an example of a knowledge base. It is a data sharing and display platform designed to help communities create their own portals to provide access to research resources, data, literature, and tools to both their own communities and across communities. SciCrunch currently hosts community portals for Research Resource Identifiers (RRIDs), the Neuroscience Information Framework, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Framework, and the Drug Design Data Resources (33). CyVerse is another example, which provides life scientists with powerful computational infrastructure to handle huge datasets and complex analyses to enable data-driven discovery (34).

NIH also supports the creation of knowledge bases that make data accessible for reuse (35). For example, the Library of Integrated Network-Based Cellular Signatures (LINCS) Program catalogs changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents. By generating this type of data and making it public, the LINCS project aims to improve our understanding of how cells respond to various genetic and environmental stressors and provide detailed information of cell pathways and networks that may be used to develop therapies to restore perturbed pathways and networks to normal states (36). Efforts are also underway to support the development of databases and informatics resources to combine human and model organism databases and enable crosstalk across mechanistic pathways (35).

By putting seemingly disparate and distinct datasets into perspective, researchers can start to see how the pieces fit within the larger program research objectives. Over time, this may encourage the translation of data to knowledge and application through the convergence of knowledge management and information architecture, which requires the organization, labeling, and navigation of information in an effective way (32).

## Research infrastructure facilitates data sharing and integration

SRP has research infrastructure in place that can facilitate data sharing. Over the years, large cohort studies funded by the SRP have amassed a variety of biospecimens and data on exposures and health effects. These large, ongoing cohorts are a rich source of data that may be leveraged when combined with other studies. SRP researchers also have a wide variety of mechanistic data based on animal studies. Moving forward, SRP researchers have the opportunity to combine epidemiology and exposure data to improve the assessment of risk from environmentally relevant exposures. For example, SRP grantees and colleagues are developing software focused on integrating multiple sources of evidence, including epidemiology and mechanistic data, toward a better understanding of health outcomes (37). SRP researchers are also developing tools to standardize dose-response data modeling and creating computational approaches to predict toxicity values for use in health assessments on a large number of chemicals (38, 39).

Spatial data collected over time on the fate and transport of contaminants may also help to inform what we know about the potential for exposure and health effects. Ongoing studies combine exposure data with health outcome information so we can start to visualize potential links between exposure and disease (40–42). Other studies are using spatial modeling to characterize the fate and transport of contaminants into the environment (43, 44). It is important to harmonize these different types of spatial data to provide researchers with a comparable view of data from different studies. For example, an NIEHS-funded study linked birth record data to the CalEnviroScreen dataset, which contains spatial data on exposures and environmental effects, to evaluate cumulative exposures related to preterm birth using environmental and social information (45). In another study, NIEHS-funded researchers are working to operationalize built environment measures at varying spatial scales so that environmental data can be linked more appropriately to health outcome data (46).

SRP research infrastructure has also provided an avenue for researchers to move quickly to respond to disasters. In the aftermath of Hurricane Katrina, SRP provided emergency supplemental funding to study health effects. As part of this funding, SRP grantees developed the NIEHS Environmental Health Sciences Data Resource Portal that integrated advances in geographic information systems and data visualization technologies to serve as a national resource to track environmental hazards after disasters. The portal provided a way for NIEHS researchers and partners to work together more effectively and obtain shared datasets and applications. It also functioned as a way to test new technologies to help advance environmental health research (47). Using the portal, SRP investigators provided Gulf Coast leaders with tools and data to monitor health impacts, assess and reduce human exposures to contaminants, and develop remediation strategies (48).

### Parallels between health and remediation research

The diversity of SRP research may create opportunities that link findings across a variety of fields, such as health and remediation research. For example, both fields include research related to microbial communities. As with humans, the microbiomes of plants, animals, and the environment play a role in growth, development, and vulnerability to contaminants. Microbiomes are influenced by different environments and this data may inform microbial research in other fields.

In order to integrate microbiome data across fields, there is a need for standardized methods and metadata collection to make robust interstudy comparisons about microbial communities. Efforts are underway to develop policies and priorities for organizing microbiome studies and creating a catalog of microbiomes (49). NIH also created the NIH Interagency Strategic Plan for Microbiome Research FY 2018–2022 to support collaborative research, develop robust and consistent standards, support open and transparent data, and further develop analytical technologies (50).

Many of these efforts focus primarily on the animal and human microbiomes. Given the inclusion of both biomedical and environmental science research within a center, SRP data provides the unique opportunity to take this one step further, with data focused on microbe-microbe interactions in the environment (51) as well as biogeochemical interactions that help specific bacteria degrade pollutants (52). Exploring how microbes break down contaminants in the environment is important to the field of environmental remediation, but could also translate to other fields of research, such as how microbes respond to contaminants in the body. SRP grantees have data on specific contaminants and specific microbes, and researchers may be able to use that data to accelerate research on the human gut microbiome. The gut microbial ecosystem may also depend on where we live (53). Integrating spatial and temporal data with information about the microbes may further our understanding of the human gut microbiome, an area of science that is not well understood.

## Challenges ahead

As scientists begin embracing the use of big data, grantees are identifying many challenges involved in finding and using datasets and collaborating with computational scientists. Some challenges include identifying useful datasets, handling data that are available in disparate formats, and the need to ensure privacy of research participants.

### Privacy

Many environmental health researchers have expressed concerns about privacy of research participants when, for example, cohort datasets are shared (54). It is possible for data to be de-identified before sharing or storing in a public database. However, for many datasets, its usefulness hinges on knowing potentially identifying characteristics. Examples include location data in a pollution study or specific race in studies aiming to identify health disparities. When several attributes are associated with individuals, even de-identified data can be re-identified. A recent study found that 99.98% of Americans could be identified in an anonymized dataset that includes 15 demographic attributes (55). Concerns about privacy are complex and real, and such risks may discourage people from participating in research studies in the future.

Various approaches have been identified to address these concerns, which include adopting privacy-enhancing technologies for data analysis, ensuring data security and improving procedures for controlled access, and making privacy requirements more consistent (56). For example, the NIH All of Us Research Program, an effort to accelerate research and improve health by gathering data from 1 million or more people living in the US, commits to ensuring confidentiality and integrity of all specimens and data. As part of the Precision Medicine Initiative, they are building security practices into development and identifying strong safeguards to ensure privacy. They also reevaluate their methods regularly to keep up with advancing technology (57). Improving procedures that automatically reduce potential identifiers would be critical to make the process of data sharing more efficient and attractive to the research community.

## Training

Others have pointed out the large investment needed to train the next generation of scientists to more routinely work with large datasets (58). As biomedical research is accelerating its reliance on computational, mathematical, and statistical methods, supporting training requires new emphases on analytical skills (59). The overarching needs are twofold: bringing more trainees from computational biology into EHS fields to perform data analysis and mining, as well as providing more scientists in the EHSs the knowledge they need to work with both data and data science experts.

For example, trainees with computational skills could develop scripts to automate processes, which would otherwise be executed by laborious and time-consuming step-by-step approaches, making data analysis more efficient. It is also important to address the disparity between scientists without data science expertise and those who have been recently trained with these much-needed analytical skills. Training in data science is not just important for early-career scientists but should extend to mid- and late-career scientists who have a deep expertise in a given area of science and could benefit from a better understanding of data science and how it can increase the robustness of their data. Scientists at all levels from all disciplines must know enough about databases, data storage, and data mining to ask the right questions of their computational biologist colleagues, inform data scientists on whether an approach is good for the field, and work with them to evaluate the data, including recognizing limitations (60). While this training may not provide the skillset needed to independently take on data science approaches, it can serve as a foundation for the subject matter expert to more efficiently partner with the data scientist. In addition to training subject matter experts to understand basic data science, data science experts in key disciplines related to data analysis and integration are also important to ensure data are used thoughtfully and effectively.

To gain a better understanding of how existing data science and environmental health resources (e.g. trainee pipelines, mentors, research) can be used, to identify what is needed, and to build capacity in data science for environmental health researchers, the NIEHS convened a workshop in 2018. The workshop, Developing a Data Science Competent EHS Workforce, brought together data experts from relevant research disciplines to examine data science resources and provide recommendations on how these resources can address EHS-specific training goals in data science training (61).

## Ontologies/semantic reasoning

Enabling data sharing, integration, and searching requires semantic standards, such as ontologies, a need that has been recognized by many scientists and scientific organizations (62). The use of ontologies can facilitate data integration by standardizing vocabulary for describing different entities and relationships between them. Ontologies can improve metadata, automatic data verification, and support for queries of data (63). In order to improve data sharing across multiple laboratories, researchers will need to work together to standardize terminology and improve existing ontologies to keep up with the evolution of research. Representation of information using standardized ontologies not only improves data aggregation and integration, but also opens up the use of automatic processing to derive facts that are not expressed in the ontology specifically, known as semantic reasoning (62, 64). Utilization of standardized ontologies also facilitates the development and application of machine learning methods that can integrate data from different datasets to generate new knowledge. Due to the transdisciplinary nature of EHS research, developing a common language may be especially challenging, but it will yield valuable knowledge for many fields beyond EHSs. The effort that will be required to develop ontologies and common languages for these data has benefits reaching a variety of fields.

These benefits are exemplified in the existing Comparative Toxicogenomics Database (CTD), which gathers and analyzes data about chemical-gene interactions and chemical-disease and gene-disease relationships, to construct networks of chemical-gene-disease relationships. The CTD integrates data from real-world exposure studies into the database and includes a standardized vocabulary for diverse study designs and types of data. It also provides much-needed real-world context to exposure information that will enable its use in developing hypotheses about connections between real-world exposures and diseases (65).

# Conclusion

Integrating diverse data, such as data from the SRP, can provide insights into links between exposure and disease. Developing ways for information from different projects to be organized and linked could facilitate data sharing and serve to maximize data integration across discrete research disciplines. However, this can only be done by first taking steps to encourage and facilitate greater data sharing and focus on a more standardized way of describing and storing data. In order to harmonize data across disciplines, researchers need to practice good data stewardship and embrace data sharing from the project onset by collecting, processing, and storing data in a standardized way. If data from different projects are collected and analyzed using standard terminology, scientists can look to data from other research projects to inform both their own research and new research questions.

Despite challenges, increased sharing, integration, and reuse of data from SRP and other EHS programs will yield rewards for environmental health as well as many other research fields. SRP centers have research that span the scientific disciplines, with researchers who are used to working in a research-diverse team environment. Extending this collaboration to the level of data sharing, such as by enabling data sharing and integration through the DMACs, can advance centers' research and greatly enhance their impact. Centers will then be well positioned to combine a variety of data streams to accelerate the pace of research, leverage previous research findings, and answer challenging environmental health questions. This strategy of embedding a required partnership between a data scientist and the research domain expert is one that is transferrable to other areas of science to help increase the potential of research findings through maximizing research investments and turn research into knowledge.

**Research funding:** Funding by the National Institute of Environmental Health Sciences.

**Conflict of interest:** None declared.
**Informed consent:** Not applicable.
**Ethical approval:** Not applicable.

## References

1. Bui AAT, Van Horn JD, NIH BD2K Centers Consortium. Envisioning the future of 'big data' biomedicine. J Biomed Inform 2017;69:115–7.
2. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. Inf Fusion 2019; 50:71–91.
3. National Institutes of Health. NIH strategic plan for data science. 2018 [cited 16 Nov 2018]. Available at: https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf.
4. National Academies of Sciences, Engineering, and Medicine. Data science: opportunities to transform chemical sciences and engineering: proceedings of a workshop – in brief. Washington, DC: The National Academies Press; 2018. doi: 10.17226/25191.
5. Peng G. The state of assessing data stewardship maturity – an overview. Data Sci J 2018;17:7.
6. Hendler J. Data integration for heterogenous datasets. Big Data 2014;2(4):205–15.
7. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.
8. Sansone SA, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, et al. Toward interoperable bioscience data. Nat Genet 2012;44(2):121–6.
9. Elliott JH, Grimshaw J, Altman R, Bero L, Goodman SN, Henry D, et al. Informatics: make sense of health data. Nature 2015;527(7576):31–2.
10. McIntosh LD, Juehne A, Vitale CRH, Liu X, Alcoser R, Lukas JC, et al. Repeat: a framework to assess empirical reproducibility in biomedical research. BMC Med Res Methodol 2017;17(1):143.
11. Stodden V, Miguez S. Best practices for computational science: software infrastructure and environments for reproducible and extensible research. J Open Res Softw 2014;2(1):e21.
12. Wise J, de Barron AG, Splendiani A, Balali-Mood B, Vasant D, Little E, et al. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discov Today 2019;24(4):933–8.
13. Holdren JP. Increasing access to the results of federally funded scientific research. 2013 [cited 16 Nov 2018]. Available at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.
14. National Institutes of Health. National Institutes of Health genomic data sharing policy. 2014 [cited 16 Nov 2018]. Available at: https://osp.od.nih.gov/wp-content/uploads/NIH_GDS_Policy.pdf.
15. National Institute of Environmental Health Sciences. 2018–2023 strategic plan: advancing environmental health sciences, improving health. 2018 [cited 16 Nov 2018]. Available at: https://www.niehs.nih.gov/about/strategicplan/strategicplan20182023_508.pdf.
16. U.S. Geological Survey. Fundamental science practices: scientific data management. 2017 [cited 16 Nov 2018]. Available at: https://www.usgs.gov/about/organization/science-support/survey-manual/5026-fundamental-science-practices-scientific-data.
17. National Science Foundation. Dissemination and sharing of research results. 2018 [cited 16 Nov 2018]. Available at: https://www.nsf.gov/bfa/dias/policy/dmp.jsp.
18. National Institute of Environmental Health Sciences. Program mandates: Superfund Research Program. 2015 [cited 3 April 2019]. Available at: https://www.niehs.nih.gov/research/supported/centers/srp/about/program/index.cfm.
19. Superfund Research Program. Strategies to attain SRP objectives and goals: 2015–2020. 2015 [cited 20 Nov 2018]. Available at: https://www.niehs.nih.gov/research/supported/centers/srp/assets/docs/srp_strategies_to_attain_objectives_and_goals_2015_508.pdf.
20. Börner K, Klavans R, Patek M, Zoss AM, Biberstine JR, Light RP, et al. Design and update of a classification system: the UCSD map of science. PLoS One 2012;7(7):e39464.
21. National Academies of Sciences, Engineering, and Medicine. Informing environmental health decisions through data integration: proceedings of a workshop – in brief. Washington, DC: The National Academies Press; 2018. doi: 10.17226/25139.

22. National Institute of Environmental Health Sciences. Integrating data from multidisciplinary research. 2015 [cited 16 Nov 2018]. Available at: https://www.niehs.nih.gov/research/supported/centers/srp/events/riskelearning/integrating_data/index.cfm.
23. National Institute of Environmental Health Sciences. KC Donnelly externship award supplement. 2018 [cited 25 Sept 2019]. Available at: https://www.niehs.nih.gov/research/supported/centers/srp/training/donnelly/index.cfm.
24. Patil C, Siegel V. Shining a light on dark data. Dis Model Mech 2009;2(11–12):521–5.
25. Gantz J, Reinsel D. Extracting value from chaos. 2011 [cited 3 April 2019]. Available at: https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.
26. Federer LM, Belter CW, Joubert DJ, Livinski A, Lu YL, Snyders LN, et al. Data sharing in PLoS One: an analysis of data availability statements. PLoS One 2018;13(5):e0194768.
27. National Institute of Environmental Health Sciences. Datasets listing: Superfund Research Program. 2018 [cited 16 Nov 2018]. Available at: https://tools.niehs.nih.gov/srp/data/datasets.cfm.
28. Zhao M, Yan E, Li K. Data set mentions and citations: a content analysis of full-text publications. J Assoc Inf Sci Technol 2018; 69:32–46.
29. Department of Health and Human Services. Superfund Hazardous Substance Research and Training Program (P42 clinical trial optional). 2018 [cited 16 Nov 2018]. Available at: https://grants.nih.gov/grants/guide/rfa-files/RFA-ES-18-002.html.
30. Zaslavsky I, Bermudez L, Grethe J, Gupta A, Hsu L, Lehnert K, et al. CINERGI: Community Inventory of EarthCube Resources for Geoscience Interoperability. 2015 [cited 4 April 2019]. Available at: https://www.earthcube.org/group/cinergi.
31. Griffin PC, Khadake J, LeMay KS, Lewis SE, Orchard S, Pask A, et al. Best practice data life cycle approaches for the life sciences. F1000Res 2017;6:1618.
32. Pugatch J, Grenen E, Surla S, Schwarz M, Cole-Lewis H. Information architecture of web-based interventions to improve health outcomes: systematic review. J Med Internet Res 2018;20(3):e97.
33. SciCrunch. About SciCrunch. 2019 [cited 4 June 2019]. Available at: https://scicrunch.org/page/scicrunch.
34. CyVerse. CyVerse: The Project. 2019 [cited 5 June 2019]. Available at: http://www.cyverse.org/about.
35. National Institutes of Health. Genomic Community Resources (U24). 2018 [cited 3 April 2019]. Available at: https://grants.nih.gov/grants/guide/pa-files/par-17-273.html.
36. National Institutes of Health. The LINCS consortium. 2018 [cited 25 Nov 2018]. Available at: http://www.lincsproject.org/.
37. Marvel SW, To K, Grimm FA, Wright FA, Rusyn I, Reif DM. ToxPi graphical user interface 2.0: dynamic exploration, visualization, and sharing of integrated data models. BMC Bioinformatics 2018;19(1):80.
38. Wignall JA, Muratov E, Sedykh A, Guyton KZ, Tropsha A, Rusyn I, et al. Conditional Toxicity Value (CTV) predictor: an in silico approach for generating quantitative risk estimates for chemicals. Environ Health Perspect 2018;126(5):057008.
39. Wignall JA, Shapiro AJ, Wright FA, Woodruff TJ, Chiu WA, Guyton KZ, et al. Standardizing benchmark dose calculations to improve science-based decisions in human health assessments. Environ Health Perspect 2014;122(5):499–505.
40. Aschengrau A, Gallagher LG, Winter M, Butler L, Fabian MP, Vieira VM. Modeled exposure to tetrachloroethylene-contaminated drinking water and the occurrence of birth defects: a case-control study from Massachusetts and Rhode Island. Environ Health 2018;17(1):75.
41. Edwards SE, Maxson P, Miranda ML, Fry RC. Cadmium levels in a North Carolina cohort: identifying risk factors for elevated levels during pregnancy. J Expo Sci Environ 2015; Epidemiol 25(4):427–32.
42. Pezzoli K, Leiter RA. Creating healthy and just bioregions. Rev Environ Health 2016;31(1):103–9.
43. Mainhagu J, Morrison C, Truex M, Oostrom M, Brusseau ML. Measuring spatial variability of vapor flux to characterize vadose-zone VOC sources: flow-cell experiments. J Contam Hydrol 2014;167:32–43.
44. Yang Q, Smitherman P, Hess CT, Culbertson CW, Marvinney RG, Smith AE, et al. Uranium and radon in private bedrock well water in Maine: geospatial analysis at two scales. Environ Sci Technol 2014;48(8):4298–306.
45. Huang H, Woodruff TJ, Baer RJ, Bangia K, August LM, Jellife-Palowski LL, et al. Investigation of association between environmental and socioeconomic factors and preterm birth in California. Environ Int 2018;121(Pt 2):1066–78.
46. Strominger J, Anthopolos R, Miranda ML. Implications of construction method and spatial scale on measures of the built environment. Int J Health Geogr 2016;15:15.
47. Pezzoli K, Tukey R, Sarabia H, Zaslavsky I, Miranda ML, Suk WA, et al. The NIEHS Environmental Health Sciences Data Resource Portal: placing advanced technologies in service to vulnerable communities. Environ Health Perspect 2007;115(4):564–71.
48. Landrigan PJ, Wright RO, Cordero JF, Eaton DL, Goldstein BD, Hennig B, et al. The NIEHS Superfund Research Program: 25 years of translational research for public health. Environ Health Perspect 2015;123(10):909–18.
49. Pylro VS, Mui TS, Rodrigues JLM, Andreote FD, Roesch LFW, the Working Group Supporting the INCT Microbiome. A step forward to empower global microbiome research through local leadership. Trends Microbiol 2016;24(10):767–71.
50. National Institutes of Health. Interagency strategic plan for microbiome research FY 2018–2022. 2018 [cited 16 Nov 2018]. Available at: https://commonfund.nih.gov/sites/default/files/Interagency_Microbiome%20Strategic_Plan_Final_041918_508.pdf.
51. Mao X, Polasko A, Alvarez-Cohen L. Effects of sulfate reduction on trichloroethene dechlorination by dehalococcoides-containing microbial communities. Appl Environ Microbiol. 2017;83(8):e03384–16.
52. Yan J, Bi M, Bourdon AK. Purinyl-cobamide is a native prosthetic group of reductive dehalogenases. Nat Chem Biol 2018; 14(1):8–14.
53. Tasnim N, Abulizi N, Pither J, Hart MM, Gibson DL. Linking the gut microbial ecosystem with the environment: does gut health depend on where we live? Front Microbiol 2017;8:1935.
54. Yewell J. Big data presents big challenges, big opportunities in environmental health. 2015 [cited 4 April 2019]. Available at: https://factor.niehs.nih.gov/2015/8/science-bigdata/index.htm.
55. Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 2019;10(1):3069.
56. Joly Y, Dyke SOM, Knoppers BM, Pastinen T. Are data sharing and privacy protection mutually exclusive? Cell 2016;167(5):1150–4.
57. National Institutes of Health. Precision Medicine Initiative: Privacy and Trust Principles. 2019 [cited 25 Sept 2019]. Available

at: https://allofus.nih.gov/about/program-overview/precision-medicine-initiative-privacy-and-trust-principles.

58. Brownson RC, Samet JM, Bensyl DM. Applied epidemiology and public health: are we training the future generations appropriately? Ann Epidemiol 2016;27(2):77–82.
59. Dunn MC, Bourne PE. Building the biomedical data science workforce. PLoS Biol 2017;15(7):e2003082.
60. Garmire LX, Gliske S, Nguyen QC, Chen JH, Nemati S, Van Horn JD, et al. The training of next generation data scientists in biomedicine. Pac Symp Biocomput 2016;22:640–5.
61. National Institute of Environmental Health Sciences. Workshop on developing a data science competent EHS workforce. 2018 [cited 17 May 2019]. Available at: https://www.niehs.nih.gov/news/events/pastmtg/2018/data-science/index.cfm.
62. Mattingly CJ, Boyles R, Lawler CP, Haugen AC, Dearry A, Haendel M. Laying a community-based foundation for data-driven semantic standards in environmental health sciences. Environ Health Perspect 2016;124(8):1136–40.
63. Zhang H, Guo Y, Li Q, George TJ, Shenkman E, Modave F, et al. An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. BMC Med Inform Decis Mak 2018;18(Suppl 2):41.
64. Mattingly CJ, McKone TE, Callahan MA, Blake JA, Hubal EA. Providing the missing link: the exposure science ontology ExO. Environ Sci Technol 2012;46(6):3046–53.
65. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, et al. The Comparative Toxicogenomics Database: update 2019. Nucleic Acids Res 2019;47(D1):D948–54.