Risk e-Learning Webinar Series: Enhancing Integration, Interoperability, and Reuse of Data

Opening Remarks

William Suk, Ph.D., M.P.H.

May 17, 2021

Director, Superfund Research Program
National Institute of Environmental Health Sciences

NIH Data Science Goals and Objectives (released 2018)

Data Infrastructure	Modernized Data Ecosystem	Data Management, Analytics, and Tools	Workforce Development	Stewardship and Sustainability
•Optimize data storage and security •Connect NIH data systems	•Modernize data repository ecosystem •Support storage and sharing of individual datasets •Better integrate clinical and observational data into biomedical data science	•Support useful, generalizable, and accessible tools and workflows •Broaden utility of and access to specialized tools •Improve discovery and cataloging resources	•Enhance the NIH data-science workforce •Expand the national research workforce •Engage a broader community	•Develop policies for a FAIR data ecosystem •Enhance stewardship

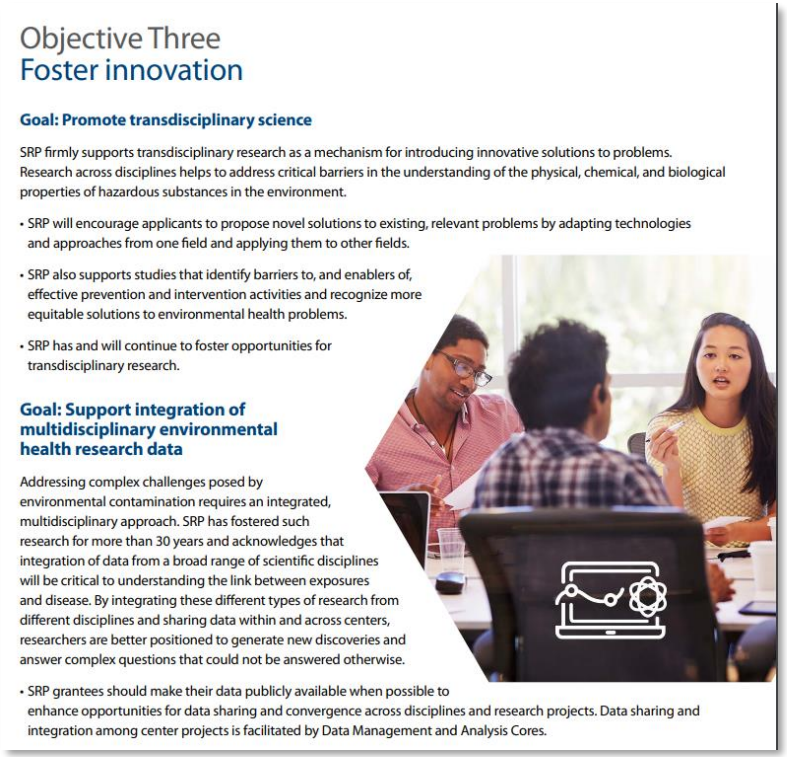
https://datascience.nih.gov/strategicplan

NIEHS Strategic Plan 2018-2023

https://www.niehs.nih.gov/about/strategicplan/strategicplan20182023_508.pdf

SRP Strategic Plan 2020-2025

https://www.niehs.nih.gov/research/supported/centers/srp/assets/docs/srp_strategic_plan_202025_508.pdf



Objective Three
Foster innovation

Goal: Promote transdisciplinary science

SRP firmly supports transdisciplinary research as a mechanism for introducing innovative solutions to problems. Research across disciplines helps to address critical barriers in the understanding of the physical, chemical, and biological properties of hazardous substances in the environment.

- SRP will encourage applicants to propose novel solutions to existing, relevant problems by adapting technologies and approaches from one field and applying them to other fields.
- SRP also supports studies that identify barriers to, and enablers of, effective prevention and intervention activities and recognize more equitable solutions to environmental health problems.
- SRP has and will continue to foster opportunities for transdisciplinary research.

Goal: Support integration of multidisciplinary environmental health research data

Addressing complex challenges posed by environmental contamination requires an integrated, multidisciplinary approach. SRP has fostered such research for more than 30 years and acknowledges that integration of data from a broad range of scientific disciplines will be critical to understanding the link between exposures and disease. By integrating these different types of research from different disciplines and sharing data within and across centers, researchers are better positioned to generate new discoveries and answer complex questions that could not be answered otherwise.

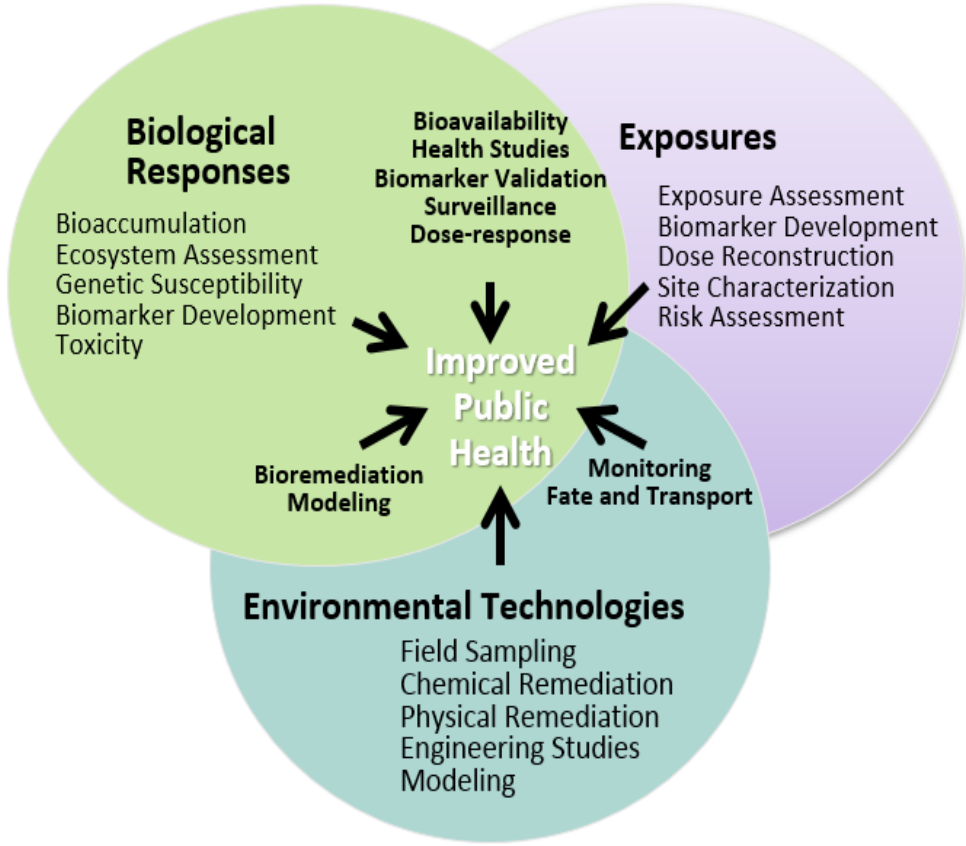
- SRP grantees should make their data publicly available when possible to enhance opportunities for data sharing and convergence across disciplines and research projects. Data sharing and integration among center projects is facilitated by Data Management and Analysis Cores.

SRP Mandates under SARA

University-based basic research program established in 1986 under Superfund Amendments Reauthorization Act (SARA)

Development of:

- Advanced techniques for the detection, assessment, and evaluation of the human health effects 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



SRP Mandates: https://www.niehs.nih.gov/research/supported/centers/srp/about/program/index.cfm

Better Positioned for the Data Management and Sharing Policy

Final NIH Policy for Data Management and Sharing

Notice Number:
NOT-OD-21-013

Key Dates

Release Date:	October 29, 2020
Effective Date:	January 25, 2023

Related Announcements

NOT-OD-21-014 – Supplemental Information to the NIH Policy for Data Management and Sharing: Elements of an NIH Data Management and Sharing Plan

NOT-OD-21-015 – Supplemental Information to the NIH Policy for Data Management and Sharing: Allowable Costs for Data Management and Sharing

NOT-OD-21-016 – Supplemental Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research

NOT-OD-20-013 - Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Issued by

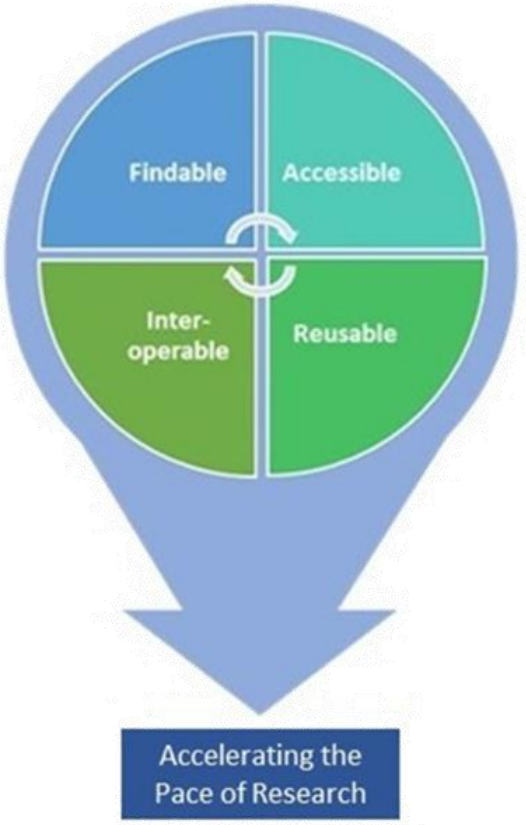
Office of The Director, National Institutes of Health (OD)

Purpose

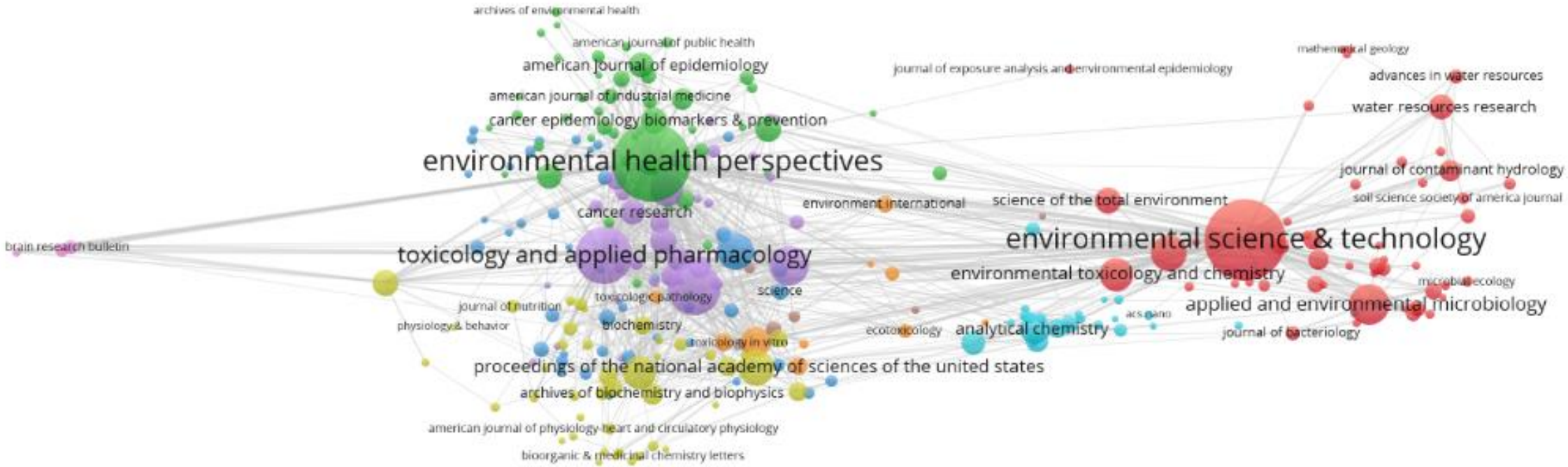
Summary

The National Institutes of Health (NIH) is issuing this final NIH Policy for Data Management and Sharing (DMS Policy) to promote the management and sharing of scientific data generated from NIH-funded or conducted research. This Policy establishes the requirements of submission of Data Management and Sharing Plans (hereinafter Plans) and compliance with NIH Institute, Center, or Office (ICO)-approved Plans. It also emphasizes the importance of good data management practices and establishes the expectation for maximizing the appropriate sharing of scientific data generated from NIH-funded or conducted research, with justified limitations or exceptions. This Policy applies to research funded or conducted by NIH that results in the generation of scientific data.

Background



Transdisciplinary SRP Research



Integrating SRP and SRP-related Data

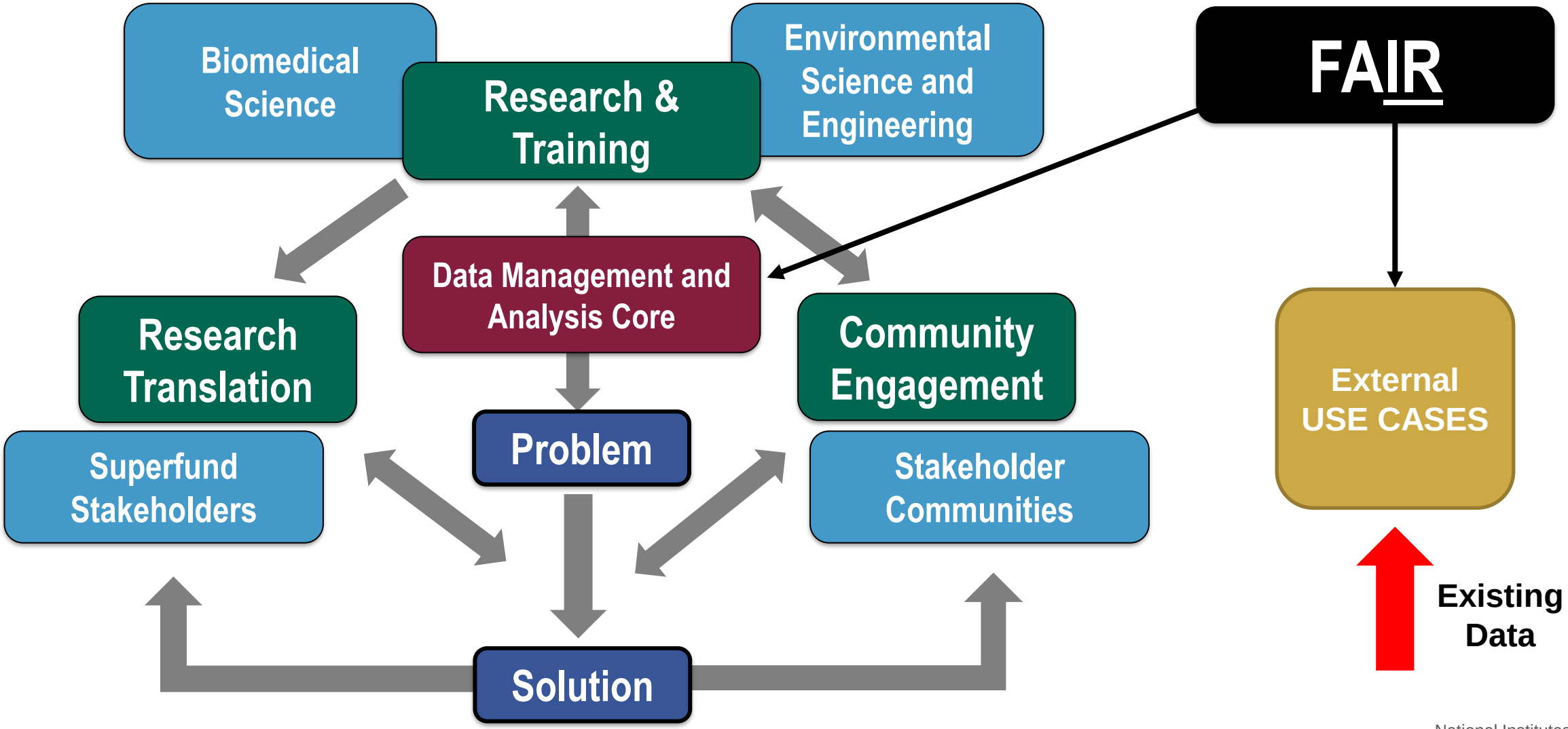
- SRP generates a wealth of data through biomedical, environmental science, and engineering research.
- New scientific connections can be revealed by integrating diverse data.
- To combine and reuse data, researchers must harmonize data workflows, have consistent and robust data stewardship, and embrace the idea of data sharing.

Creation of data management and analysis cores (DMACs)

Administrative Supplement Funding: External and Internal Use Cases (EUCs and IUCs)

Source: Heacock ML, Amolegbe SM, Skalla LA, Trottier BA, Carlin DJ, Henry HF, Lopez AR, Duncan CG, Lawler CP, Balshaw DM, Suk WA. Sharing SRP Data to Reduce Environmentally Associated Disease and Promote Transdisciplinary Research.

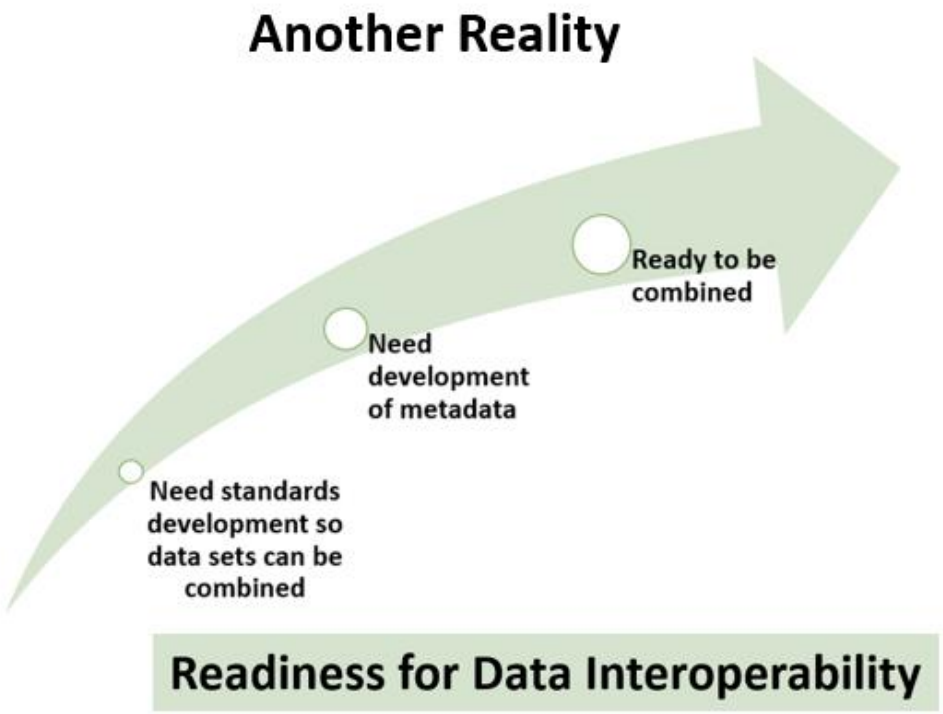
Enhancing Collaborative Science at the Level of the Data



Data Sharing Integration and Reuse Supplements

External Use-Cases (EUCs):

- Address real-world, science-driven research question,
- Address integration of two or more data streams (e.g., biomedical with environmental science and engineering),
- Address barriers to **Interoperability** and **Reuse**
- Assess effectiveness of strategies to improve the FAIRness of SRP data.



Use Cases will be applied to share best practices and help identify needs in data integration and re-use

Existing Data: Data Sharing Integration and Reuse Supplements (External Use Cases)

The Reality

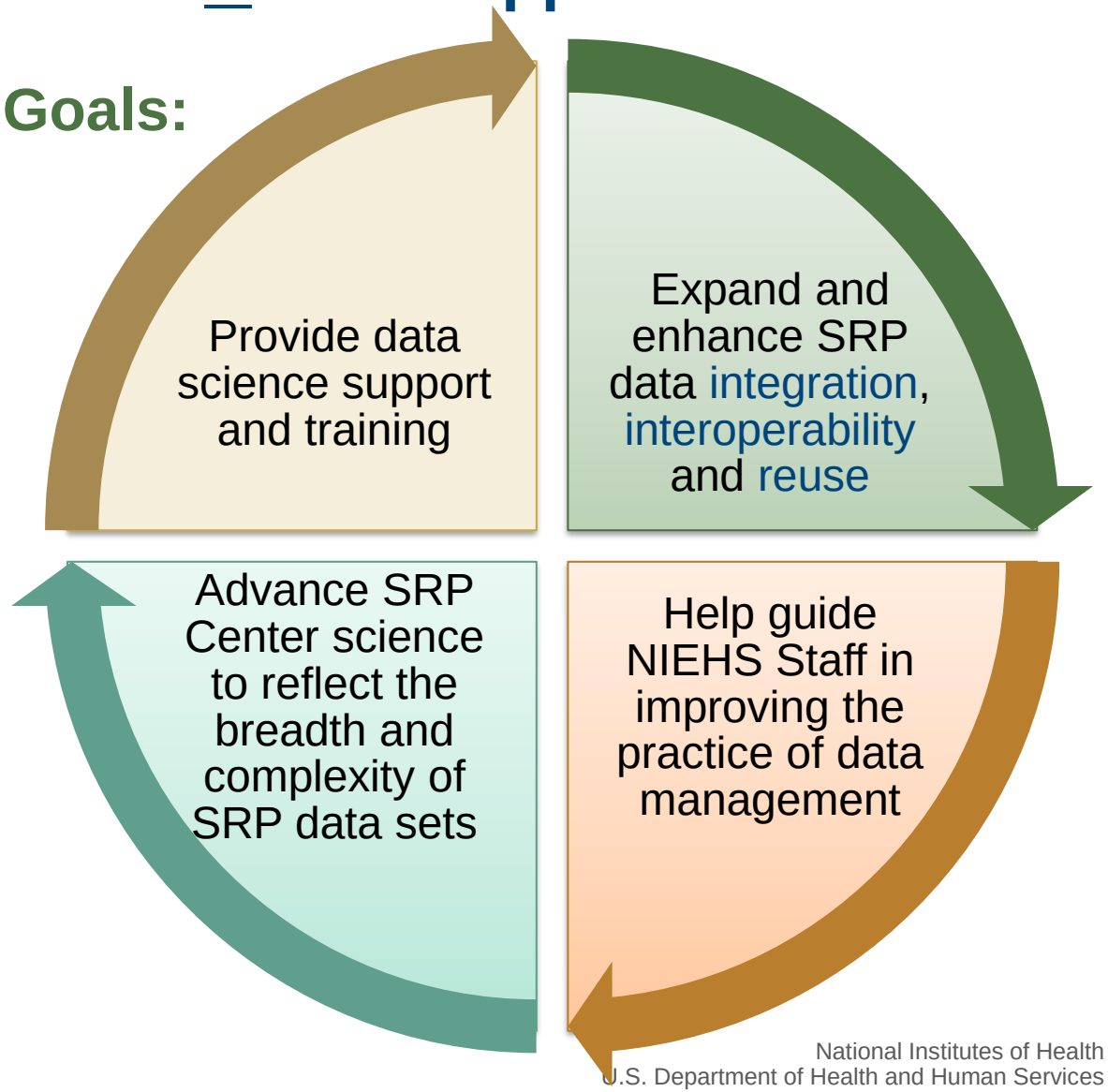
Challenges

- Center research is directed to common research goals
- Research approaches are collaborative and trans-disciplinary
- Centers are structurally and scientifically integrated

Advantages

- Data science expertise
- Ontologies/data dictionaries/standard vocabularies
- Availability of repositories
- Costs

Goals:



Listing of the External Use Cases

Title	Centers (and R01s)
Arsenic Epigenetics META: towards a Meta-analysis of Epigenome Data on Arsenic	Berkeley and Columbia
Arsenic Mass Balance: Integrating Environmental and Biomarker Data across Diverse Populations	Berkeley, Columbia, UNM
Validate and Develop new Visualization and Reproducibility Documentation for Atmospheric Deposition of Toxicants	MIT and URI
Cross-kingdom microbial interactions for bioremediation of Superfund pollutants	Duke and Iowa
Data Assembly, Management and Creating Broad Access for Combining Transcriptomic, Proteomic, Microbiome and Biogeochemical Data Sets of Phytoremediator/Phytostabilizer Plants	Arizona and SD
Data Harmonization Across 3 SRP Pregnancy and Birth Cohorts	Dartmouth, NE, and UNM
Data Interoperability for Integrating Aquatic and Terrestrial Remediation and Recovery Data from Mine-Contaminated Sites	Arizona and CSM (R01)
Development of Interoperable Data Platforms to Define Best Practices and Data Sharing for Flow Cytometry	Louisville and UNM
Improvement of small molecule biosensor probe development and biomedical applications through the integration and reuse of SRC data sets	Davis and SD
Improvement, Harmonization, and Merging of Data Streams related to DNA Damage	MIT and UNM
The Xposome: Integration and Sharing of Xenobiotics-Associated Assays across Species, Phenotypes, and Site	Boston and OSU
Improving Synchrotron-Based Data Access, Analysis and Workflows Measuring the Concentration, Speciation and Distribution of Contaminants in Environmental and Biomedical Matrices	Arizona, Dartmouth, and UNM
Improving the Robustness and Toxicological Significance of Nontarget Chemical Identification in High Resolution Mass Spectrometric Data	Davis and Duke
Integrating individual and population genomic data to understand mechanisms of chemical susceptibility and resistance	Boston and Duke
Integrating longitudinal spatial data on industrial land use and green infrastructure conditions in Rhode Island and Houston, Texas	Brown, SD, and TAMU
Integration and Analysis of SRC-Generated Cardiometabolic Syndrome Data Streams from Animal Models	Kentucky, Louisville, and MSU
Linking Data from Laboratory and Field Investigations of Mercury Transformation, Bioaccumulation, and Remediation	Dartmouth, Duke (R01), and UMBC (R01)
Making Fish Contaminant Data FAIR to Improve Fish Consumption Advisories	Boston and Dartmouth
Refining species-conserved adverse outcome pathways for AhR-mediated toxicities through the integration and re-use of Superfund Research datasets	Iowa and MSU

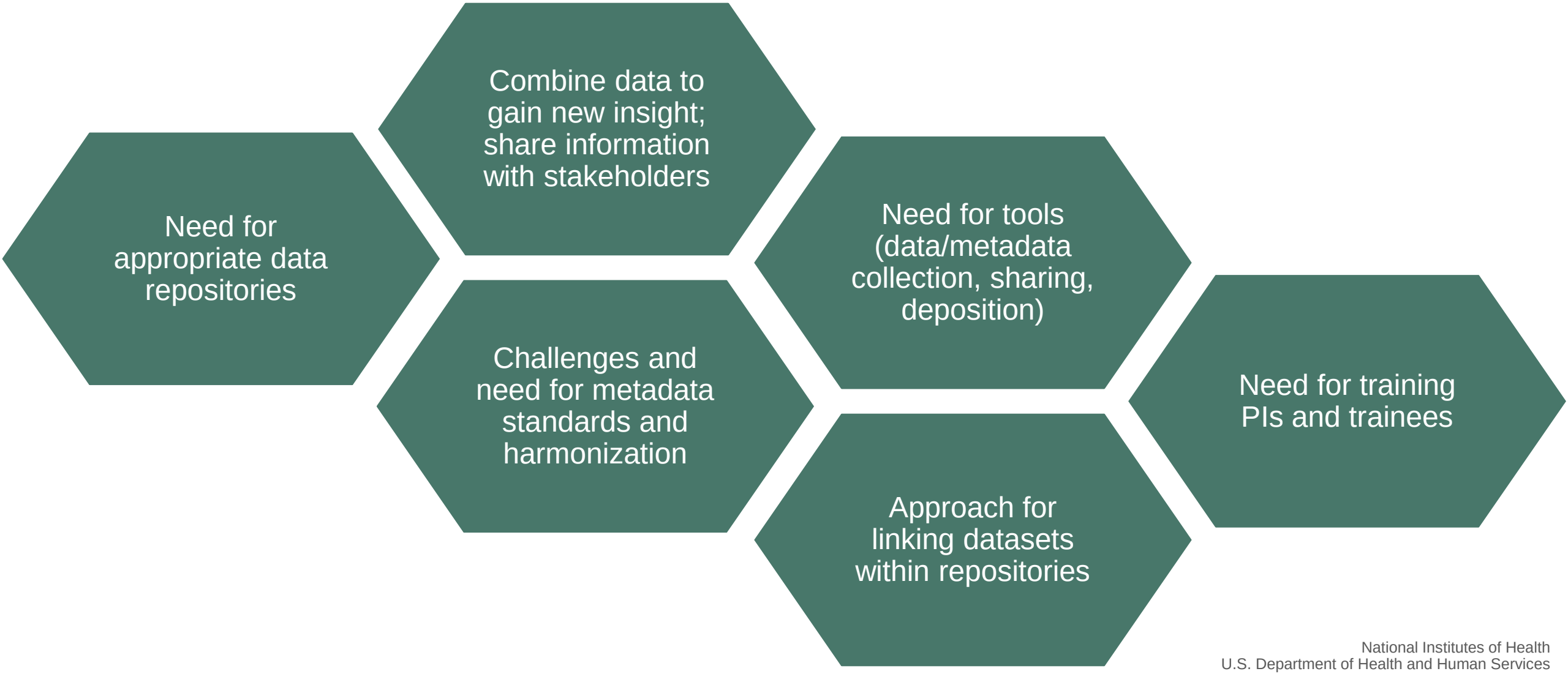
Type of Collaboration	# EUCs
P42 + P42	11
P42+P42+P42	5
P42+R01+R01	1
P42+R01	1
# with 4 collaborators	1

Data Sharing Integration and Reuse Supplements

EUC Research Question Topics And Examples

Topic	Example
Plant-microbiome-metal interactions	Can we create broad access to gain a mechanistic understanding of plant-microbiome-metal interactions that mediate plant survival and phytostabilization of toxic metals in metal-contaminated mine soils
Spatial Approach for exploring associations	How do vacant and industrial land uses impact green infrastructure conditions often presumed to enhance community resilience and public health?
Chemical metabolism across organisms	Can we support exploratory data analysis, hypothesis generation, and elucidation of chemicals' modes of action and adverse effects by integrating and sharing chemical-associated assays across species?
Genome integrity and the epigenome	How does genetic variation influence sensitivity and resistance to the Superfund chemicals, polychlorinated biphenyls (PCBs) and polycyclic aromatic hydrocarbons (PCBs)?

Common Themes Across the External Use Cases

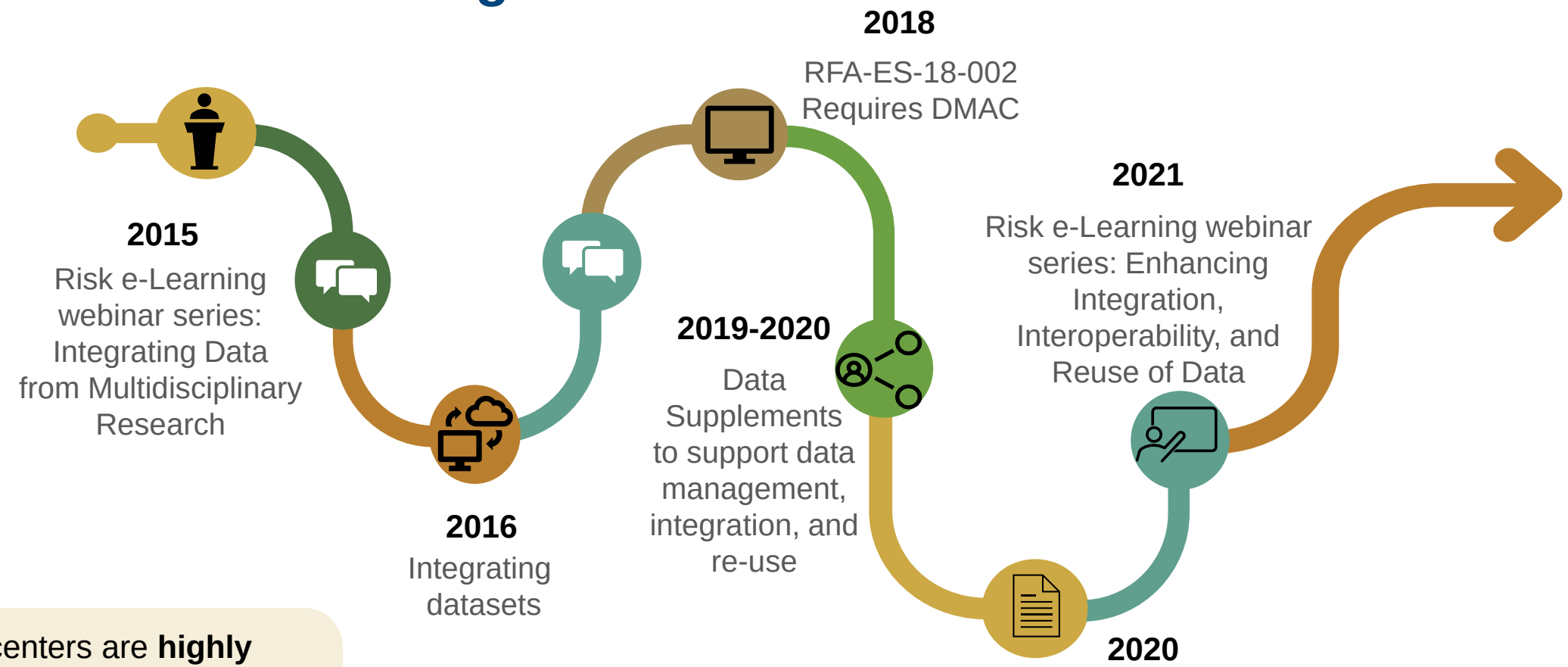


Key takeaways

- Data integration and reuse across a defined research program
- Data scientists collaborating with researchers
  - How researchers use the data
  - Removing barriers by facilitating access
  - Maximizing research potential
- Training is critical
  - Training sessions for everyone involved (trainees, researchers)
  - Recommended training topics and materials (guides, videos, etc.)
  - Need for a Data Science term glossary

Perspective from researchers and program staff

SRP Data Science / Sharing Activities



https://www.degruyter.com/document/doi/10.1515/reveh-2019-0089/html

SRP P42 centers are **highly integrated** - all research projects work to answer questions from a collective discipline standpoint versus a domain specific.

Snapshot of Webinar Series

Session I – Data Sharing Tools, Workflows, and Platforms
Monday, May 17, 2021, 1:00 PM-3:00 PM EDT
To register, visit the EPA's CLU-IN Training & Events webpage
The first session will introduce tools, strategies, workflows, and platforms developed by SRP researchers to organize existing data obtained from measuring environmental contaminants to facilitate interoperability. We will also hear about the U.S. EPA's CompTox Chemicals Dashboard, a compilation of information from many sites and databases developed to organize chemical data and address data gaps.

Session II – Geospatial Platforms for Analysis and Visualization Across Environmental Data
To register, visit the EPA's CLU-IN Training & Events webpage
Thursday, June 3, 2021, 2:00 PM-4:00 PM EDT
In the second session, presenters will describe efforts to combine and analyze data sets from SRP Centers and other sources using geospatial platforms. This session will also feature a speaker supported by NSF who will provide discuss Hydroshare, an online system to share hydrologic data and models.

Session III – Integrating Omics Data Across Model Organisms and Populations
Friday, June 18, 2021, 1:00 PM-3:00 PM EDT
To register, visit the EPA's CLU-IN Training & Events webpage
The third and final session will feature SRP-funded researchers collaborating to combine omics (e.g., genomics, proteomics) data within and across model organisms as well as studies in human populations. We will also hear from The Global Alliance for Genomics and Health about their work to incorporate semantic data models for sharing of genomic data.

Thank you!