National Institutes of Health
Office of Data Science Strategy

The Importance of Ontologies to Connect Data, Enhance Software, and Create a Data and Digital Health Ecosystem

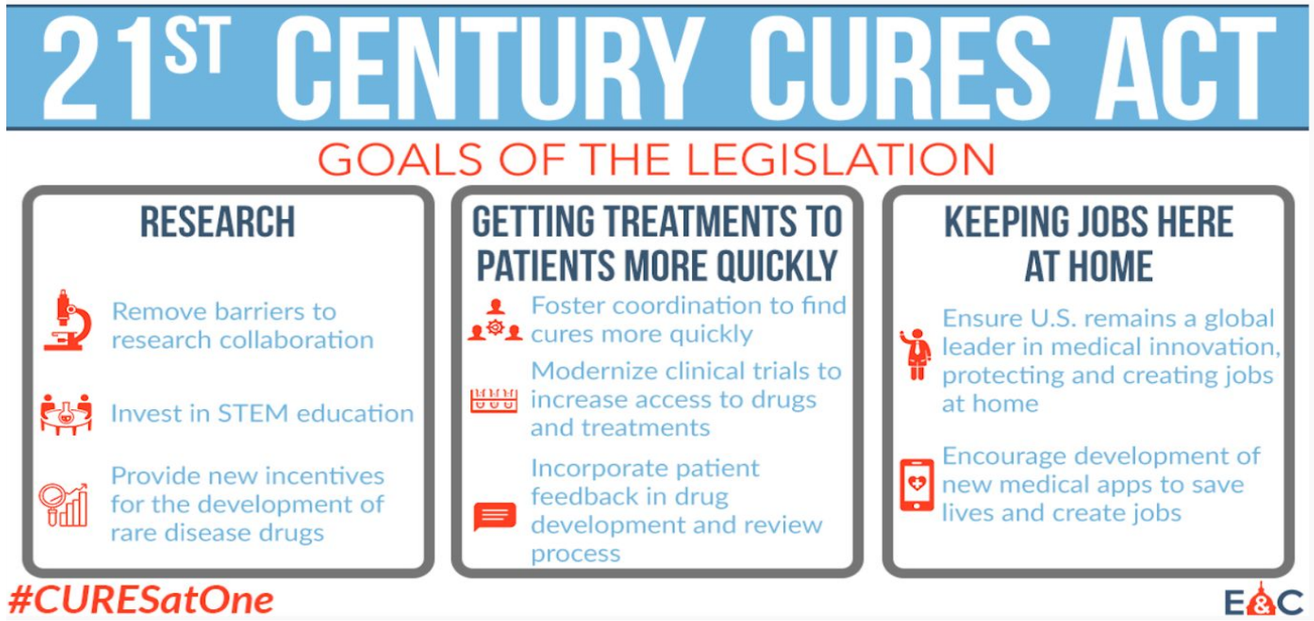
Susan Gregurick, Ph.D.
Associate Director for Data Science
and
Director, Office of Data Science Strategy

2022

Data sharing is expected

"Increasing the pool of researchers who can access data and decreasing the time it takes for them to review and find new patterns in that data ***is critical to speeding up development of lifesaving treatments for patients."***

- Joe Biden



"[The NIH Director] ***may require recipients of NIH awards to share scientific data****, to the extent feasible, generated from such NIH awards …"*

- 21st Century Cures Act

"Only now that the new Cures Act privacy protections are in place, are ***we moving forward on the exciting new authority to require data sharing****."*

- Francis Collins

BREAKING NEWS: Open Access

"The White House announcement today is an astronomical win for innovation and scientific progress."

-Ron Wyden, U.S. Senator from Oregon

"We are enthusiastic to move forward on these important efforts to make research results more accessible.."

- Lawrence Tabak, Performing the Duties of the NIH Director

"AAAS, the nonprofit publisher of the Science family of journals, supports the objectives of the White House OSTP and has a long history of advocating for equitable access to scientific research and data..."

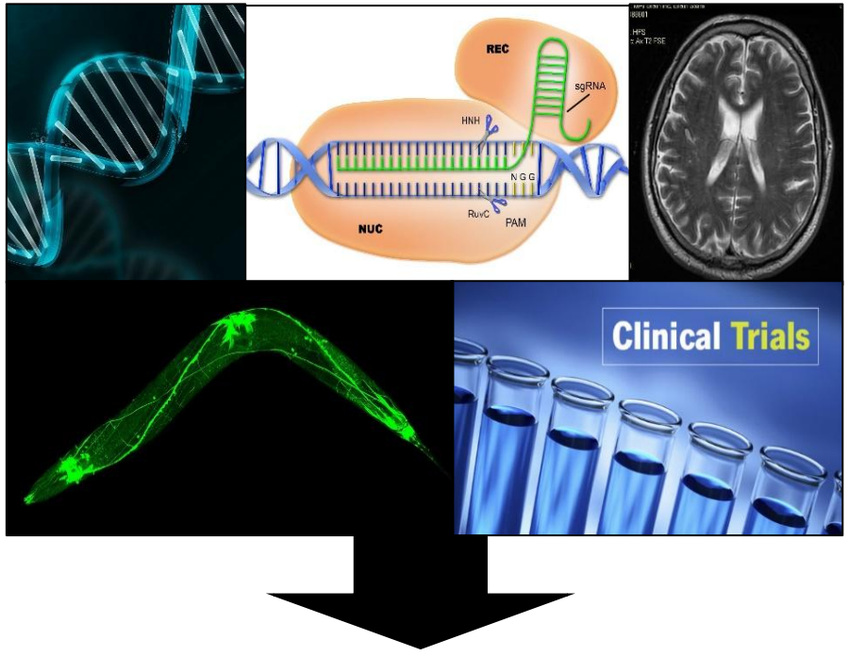
- Sudip Parikh, Chief Executive Officer, AAAS

Data sharing advances the NIH mission

Sparks New Research Collaborations

Enhances Rigorous Study Design

Makes High-Value Datasets Available



Enables Unique Data Combinations

Facilitates Study Validation

Stimulates New Research Inquiries

Maximizes Data Collection
(reduces redundancy/maximizes participant contributions)

Fosters Stewardship
(provides transparency/accountability for taxpayer funds)

Accelerates the Research Enterprise
(for all the reasons stated above!)

NIH Policy for Data Management and Sharing

- **Two basic requirements:**
  - Submission of a Data Management & Sharing “Plan” for all NIH-funded research
  - Compliance with the ICO-approved Plan
- **Effective January 25, 2023** ***(replaces 2003 Data Sharing Policy)***

Policy Expectations

SHARING SHOULD BE …

- **The default practice**
  - Data sharing should be maximized
  - Justifiable limits for technical/ethical/legal factors

- **Responsibly implemented**
  - Outline protection of privacy, rights, and confidentiality
  - Abide by existing laws, regulations, and policies

- **Prospectively planned for at all stages of the research process**

COMMUNITY Resources: What's a good plan?

Recommended elements of a plan:

- **Data type -** Data to be preserved and shared
- **Related tools, software, code -** Tools and software needed to access/manipulate data
- **Standards -** Standards to be applied to scientific data/metadata
- **Data preservation, access, timelines -** Repository to be used, persistent unique identifier, and when/how long data will be available
- **Access, distribution, reuse considerations -** Factors for data access, distribution, or reuse
- **Oversight of data management -** How plan compliance will be monitored/ managed and by whom

NOT-OD-21-014 – Suppl. Info.: Elements of an NIH Data Management and Sharing Plan

COMMUNITY RESOURCES: Where should the data go?

Encourages use of established repositories

Helps investigators identify appropriate data repositories

- E.g., use of persistent unique identifiers, attached metadata, facilitates quality assurance

- Refers to list of NIH-supported Data Repositories

NIH ICs may designate specific data repository(ies)

NOT-OD-21-016 – Suppl. Info.: Selecting a Repository for Data Resulting from NIH-Supported Research

FAIR Data, TRUST Repositories, Data Management and Sharing

Data is the new oil!

Genetic Expression and Variation Analysis

Microbiome Analysis

Cellular Structure and Functional Analysis

Neuroscience Analysis

Genomic and Phenotypic Analysis

Neuronal Image Analysis

NIH makes available almost 200pb of data on 3 clouds

Metabolomics Analysis

Whole Genome Sequence Analysis

Single-Cell ‘Omics Analysis

Microscopy Image Analysis

Cryo-Electron Microscopy Analysis

Clinical Analytics, new applications of FHIR

Collaborations to Make Data FAIR and AI/ML Ready

NIH supported collaboration, bringing together expertise in biomedicine, data management, and artificial intelligence and machine learning (AI/ML) to make NIH-supported data AI-ready for AI/ML analytics.

FY21-FY22: 73 Awards, $21M

Most common biomedical focus areas:

Alzheimer's and Parkinson's disease, cardiovascular disease, cancer, and aging

Most common data types:

imaging, EHRs, -omics, microbes/pathogens, speech

NHGRI | NIA | NIBIB | NIDA | NIDCD | NIDCR | NIEHS | NIGMS | NIMH | NINDS | NCI | NLM | NIMHD | NIDDK | NICHD | NIAID | NIAMS | NHLBI

Improving the AI/ML-Readiness of Data

Phil Brown, Northeastern University

T32-ES023769

Goal: Prepare researchers for successful careers as data analysts, ready to exploit the power of available AI/ML frameworks.

Research: Provide modules for a rich foundation in AI/ML for training to prepare data for AI and ML applications in a rigorous and reproducible way, understand the ethical issues around AI and ML, as well as receive hands-on training around FAIR principles for storing and accessing such data.

John Gilmore, University of North Carolina Chapel Hill

R01-MH123747

Goal: Study imaging and image analysis methodologies to identify children at high risk for schizophrenia.

Research: Bridge missing timepoint imaging data (data imputation) using Out-of-Distribution Detection (ML) from existing data at different timepoints.

Carl Kesselman, University of Southern California

U01-DE028729

Goal: The FaceBase consortium is a distributed network of researchers investigating craniofacial development and dysmorphology. FaceBase Hub infrastructure stores, represents, and serves relevant data to the research community

Research: Streamline curation using ML approaches to improve metadata descriptive elements while maintaining required restrictions on data handling.

Implementing FAIR Data Sharing

NIH strongly encourages
open access data sharing repositories
as a first choice
https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

Scaled implementation options for sharing datasets

Datasets up to **2 gigabytes**

PubMed Central

- Stores publication-related supplemental materials and datasets directly associated publications.

Datasets up to **20*gigabytes**

Generalist Repositories

- Datasets associated with publications or otherwise and links to PubMed.

High priority datasets **petabytes**

Cloud Partners
(STRIDES Program)

- Store and manage large scale, high priority NIH datasets.

Support for NIH Data Repositories

NIH supports a variety of data repositories and knowledgebases of **differing sizes** and **complexities** and at **different levels of maturity**

- Each has the **potential** to bring **value** to a given research area, but tend to be at **different stages** of maturity demonstrating that they have the appropriate practices in place to reliably manage the data they ingest and make available

- **Spectrum** of **ability** and **readiness** to adhere to the characteristics that are desirable for a data repository that are aligned with **FAIR** (**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable) and **TRUST** (**T**ransparency, **R**esponsibility, **U**ser focus, **S**ustainability, and **T**echnology) principles

- **Developing metrics** for evaluating the **usage**, **utility**, and **impact** of a given repository is **evolving** and likely a function of several aspects

NEW: The Generalist Repository Ecosystem Initiative

Solicit applications from generalist repositories working together to:

- Implement consistent capabilities (NOT-OD-21-016)
- Create better access to & discovery of NIH funded data
- Conduct outreach & train on FAIR data practices
- Engage the research community

Expected Outcomes

- Make data sharing easier
- Improve discoverability
- Increase reproducibility of research
- Encourage secondary use of data

HARVARD DATAVERSE · DRYAD · figshare · MENDELEY DATA · OSF · Vivli CENTER FOR GLOBAL CLINICAL RESEARCH DATA

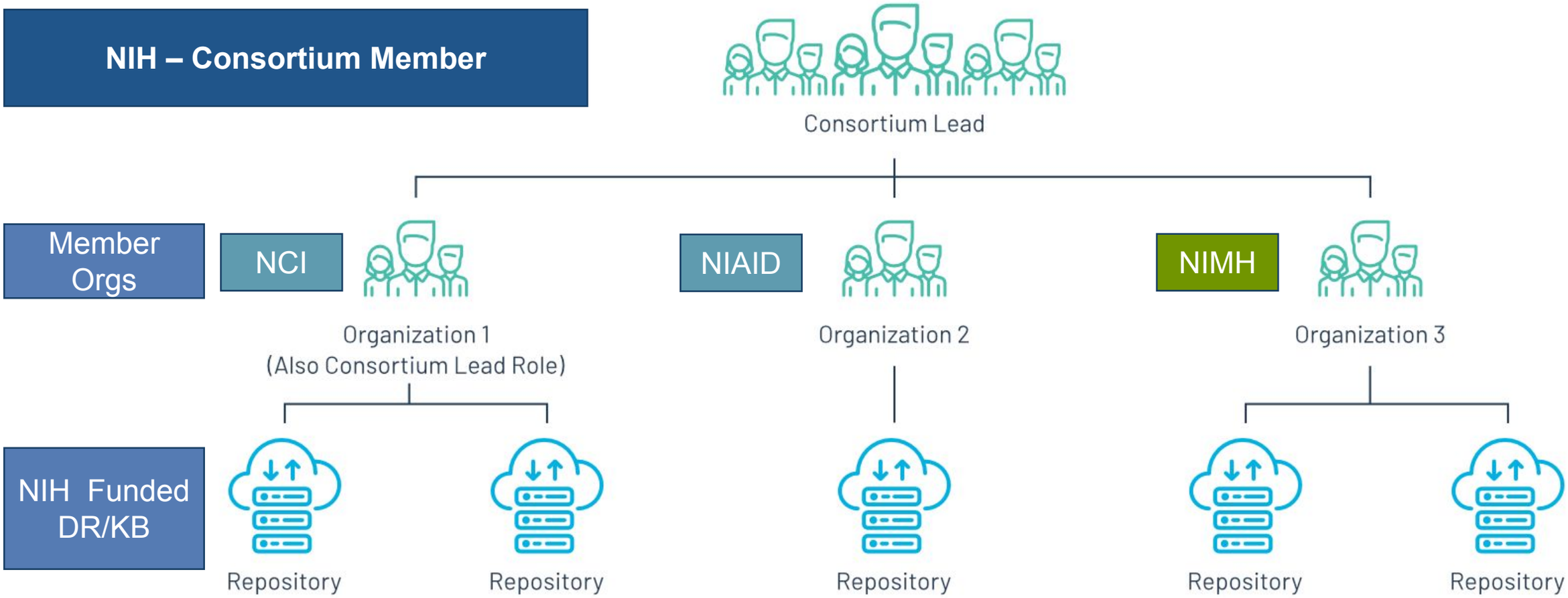
GREI Objectives

- Align with Desirable Characteristics for Data Repositories
- Implement Browse & Search for NIH Funded Data
- Develop Consistent Metadata Models
- Conduct Limited Q/AC of the NIH Funded Data
- Enable Connectivity of Digital Objects
- Use Case Support Including (X-Repository Use Cases)
- Implement Open Metrics
- Develop Educational Materials
- Conduct Broad Outreach (Workshops)
- Commit to “Co-opetition”

Software & work products developed under award will be openly shared

DataCite

NIH became a DataCite consortium member to meet a critical need to mint digital object identifiers, thereby supporting the implementation of FAIR principles for data generated from NIH funded and conducted research.



Positioning repositories for sharing

monarch INITIATIVE

Mouse Phenome Database at The Jackson Laboratory

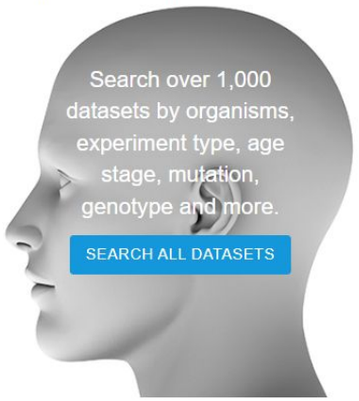
ALLIANCE of GENOME RESOURCES

UniProt

DBAASP v3.0

Database of antimicrobial activity and structure of peptides

Explore our repository:



National Toxicology Program
U.S. Department of Health and Human Services

Chemical Effects in Biological Systems

FY21-FY22: 21 Awards, $4.4M

Biomedical focus areas: Alzheimer's, traumatic brain injuries, obesity nutrition, mental health, immune response, environmental data, vision, ontology

Data types: imaging, behavioral measures, clinical, claims data, EHRs, - omics, environmental health data, brain data, speech and language

IMMPORT
BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY

WORLDWIDE PDB
PROTEIN DATA BANK

Findable Accessible Interoperable Reusable

TRUST
Transparency Responsibility User focus Sustainability Technology

NHGRI | NIA | NIAID | NICHD | NIDA | NIDCD | NIDCR | NIDDK | NIEHS | NIMH | NINDS | NIGMS

Enhance FAIR data sharing

- **Challenge** Data is heterogeneous in formats, identifiers, schemas, etc., and is challenging to interpret and integrate

- **Supported Activity**: Provide improved method for graph representations, and community **Quality Control Dashboard** and a **schema diagram**.

- **Collaboration with C-PAM** – an NIH-funded program that generates precision disease animal models with patient-specific variants in cells and organisms such as worms, zebrafish, frogs, mice, and rats, – to develop a **data capture schema** to enable the resulting data to be computable by resources such as Monarch, NCATS Biomedical Data Translator, model organism databases, and diagnostic tools.

- **Outcome:** Sophisticated comparative information systems to collect, search, and compare the diverse and often mutually incompatible model organism data being generated.

Data Repository (DR) & Knowledgebase (KB) Program

An NIH program to support investigator-initiated, sustainable data resource development driven by critical research needs

- Fill a scientific need or gap
- Encourage adoption of good data management practices
- Engage the research community to contribute and use data
- Govern data life-cycle and preservation

In 2020-2022: 64 applications reviewed & 16 awarded

ICTV

BioPortal

UniProt

RED FLY

ctd

DISEASE ONTOLOGY

reactome

SHARC
SOUTHERN HIV & ALCOHOL RESEARCH CONSORTIUM

The PhenX Toolkit

GENEONTOLOGY
Unifying Biology

HemOnc.org

BindingDB

Resource Watch

Pan-Neurotrauma Data Commons
U24NS122732-01

Principal Investigator(s):
ADAM R FERGUSON (contact), PHD
Karim Fouad, PHD
Jeffrey S. Grethe, PHD
Vance P Lemmon, PHD

Co-Investigators
John Bixby, PhD
Ubbo Visser, PhD
Michael Beattie, PhD
Jacqueline Bresnahan, PhD
J Russell Huie, PhD
Abel Torres-Espin PhD

Consultants
Maryann Martone, PhD
Alison Callahan, PhD

Federal Agency Information
9. Awarding Agency Contact Information
ERNA Petrich
NATIONAL INSTITUTE OF NEUROLOGICAL DISORDERS AND STROKE
erna.petrich@nih.gov
301-496-9248
10. Program Official Contact Inform
LINDA LOUISE Bambrick
NATIONAL INSTITU
DISORDERS

PAR-20-089 and PAR-20-097

The Importance of Ontology Efforts in our Data Ecosystem

- PhenX Toolkit Provide investigators with enhanced standard measurement protocols (e.g., questionnaires, bioassays, physical measurements) will improve the quality and consistency of data collection

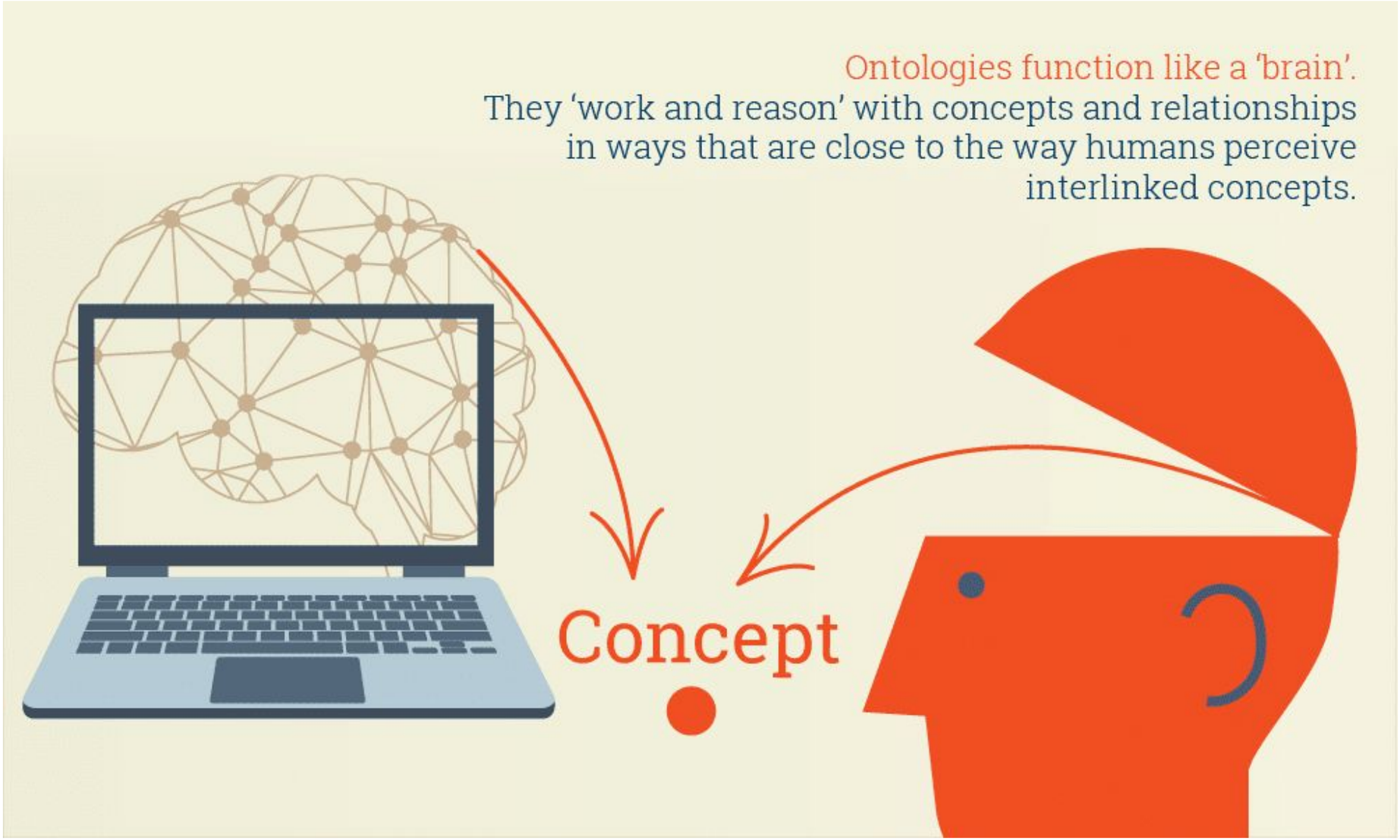
- VIOLIN 2.0: Vaccine Information and Ontology LInked kNowledgebase: Implement a pipeline for automatic knowledge harvesting, standardization, and integration using advanced informatics technologies by building a minimal information standard and its ontology representation

- The Human Disease Ontology: Provide a sustainable approach for linking the growing bodies of information related to core datasets across genomic and proteomic resources adding disease data from new biomedical research and clinical domains and modeling how disease information should be clinically understood and organized within the DO's disease classification.

- Gene Ontology Consortium and Knowledgebase: Develop and refine the Gene Ontology to reflect current biological knowledge by continued development and QC of the ontology, focusing on key biological areas of importance, and working closely with the consortium and the broader expert community

- HemOnc Knowledgebase: Grow and refine the HemOnc ontology. Expand the current base of concepts and relationships and model complex relationships not easily represented by binary relationships.

Let's Focus on Ontologies



Motivations

Describe biological entities

Enabling computer reasoning with biomedical data

Provide reference encyclopedic knowledge

Specify information models

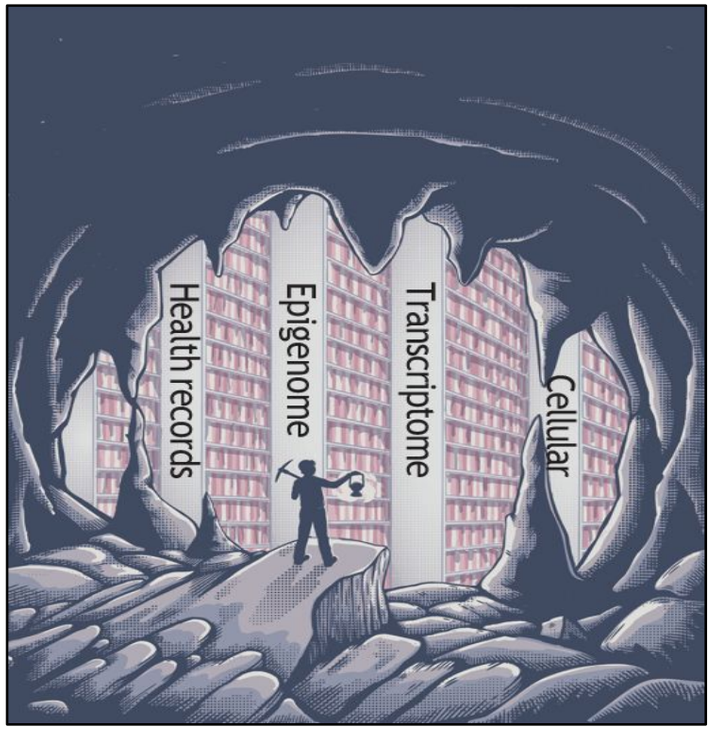
Provide semantics for data & information integration

Enable NLP/AI/Deep Learning

Specify data exchange formats

Where most of us are

A data stack view



A data web view

Modified from P. Khatri, ODSS Data Sharing and Reuse Seminar, Sept 2021

Overview of NIH Efforts in Harmonization

Program defined Common Data Elements (CDEs) and Metadata collected at onset of awards: HEAL, RADx, RECOVER, Accelerating Medicines Partnership Common Fund Data Ecosystem

Mapping of data models, through FHIR, and normalizing data values: N3C, AllofUs, and NCI Data Aggregator

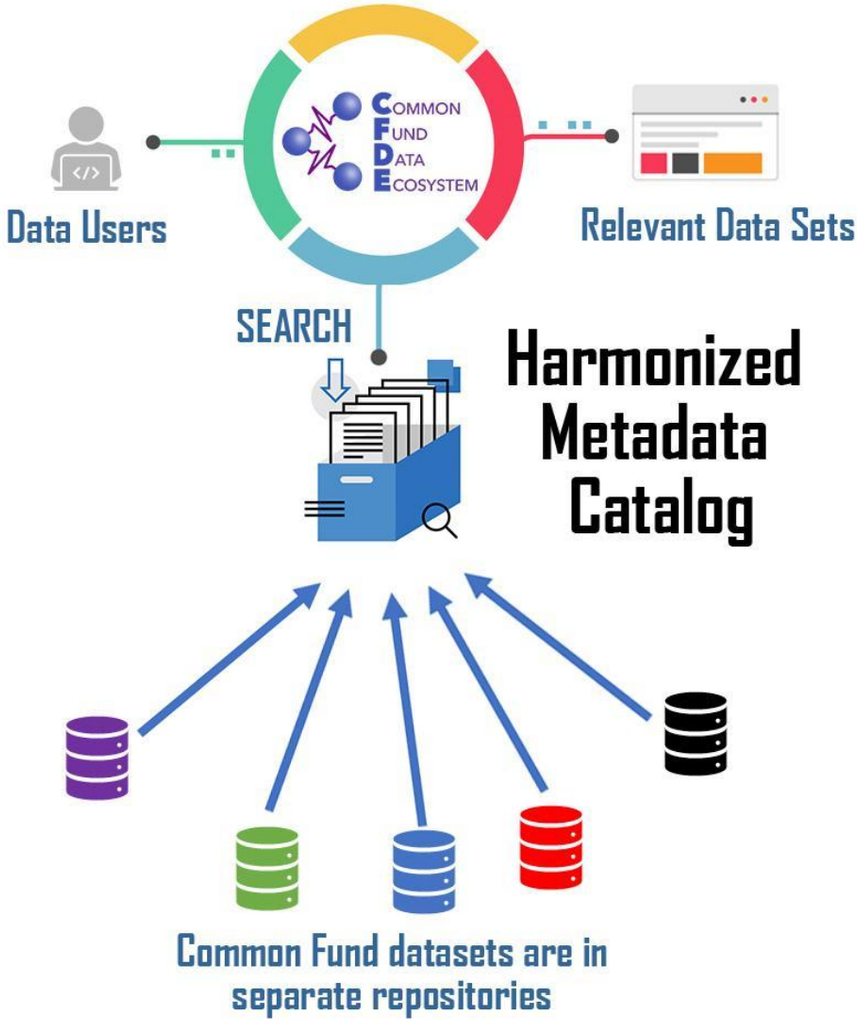
Harmonizing biologic concept across model organisms: Alliance for Genomic Research

Community engaged, iterative development of metadata/CDEs/data models: AMP-AIM, NIDCR FaceBase

Enabling search across Common Fund data sets

Common Fund programs

exRNA RESEARCH PORTAL
GTEx
HMP
HuBMAP
IDG
NIH LINCS PROGRAM
Kids First
METABOLOMICS
GlyGen
4DN
SPARC

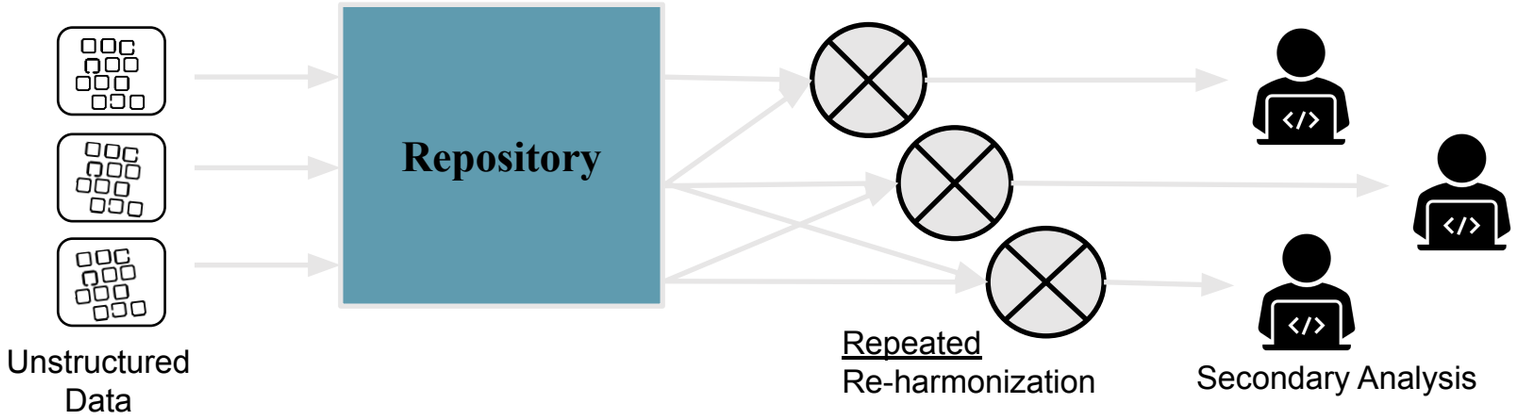


Common Fund Data Ecosystem accelerates and democratizes discovery by harmonizing data descriptors through **a common metadata model**

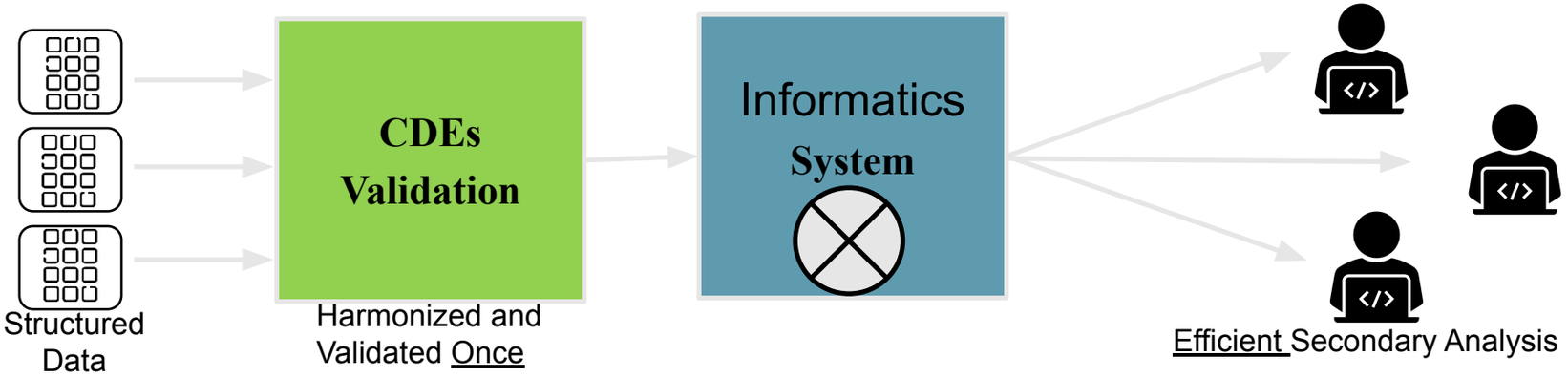
Slide curtesy of Dr. Chris Kisinger

CDEs and Harmonization=Interoperable

Option: Collect unstructured data from each study independently and harmonize data on the backend – highly inefficient – e.g. 70% researchers/postdoc time is spent on data wrangling (QAQC, validation, harmonization). B. Mons



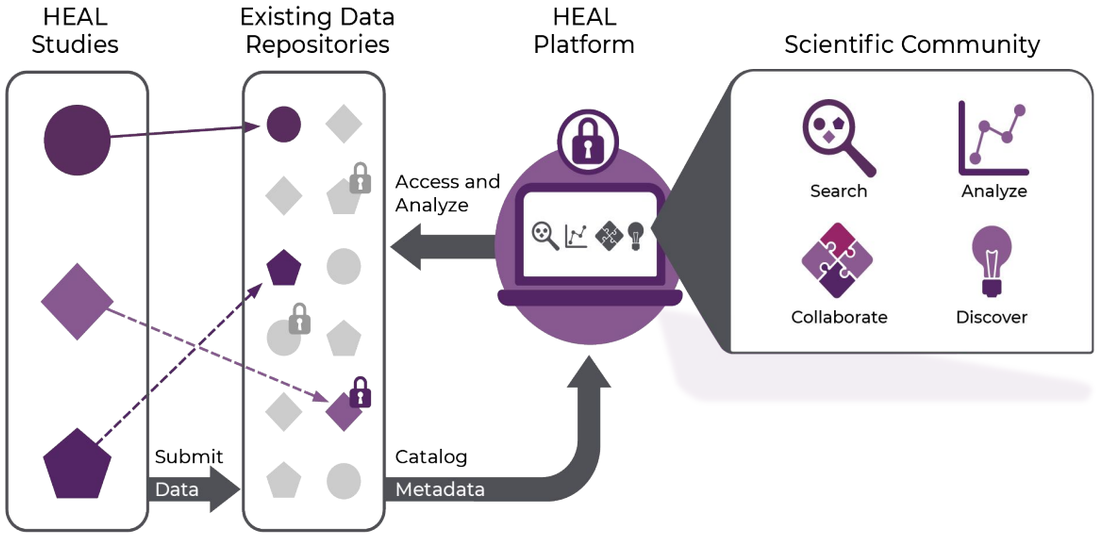
Many larger NIH programs collect structured data using CDEs and validate data on submission – supports FAIR



Slide curtesy of Dr Matt McAuliffe

HEAL Initiative: Metadata models & CDEs

- The HEAL data ecosystem is distributed – allows investigators to use domain-appropriate repositories while centralizing some standards (metadata, CDEs)
- Metadata powers the search tools (Gen3-based Platform and NCATS translator-like knowledge graph)



<u>Study-level metadata model</u>:

- Process: Developed by 2 technical teams with NIH input over 1 year; input from HEAL research teams solicited and integrated
- Challenges: breadth/diversity of HEAL data, powering search for various user groups, clearly indicating to PIs what fields mean

HEAL pain clinical core CDEs:

- <u>Process:</u> internal NIH working group, with feedback from investigators and researchers to identify core measures; supplemental measures are tracked and made available but not required
- <u>Challenges</u>: consensus across a large, diverse program; licensing; adherence and use; tracking of use and licensing
- Currently translating for NLM repository

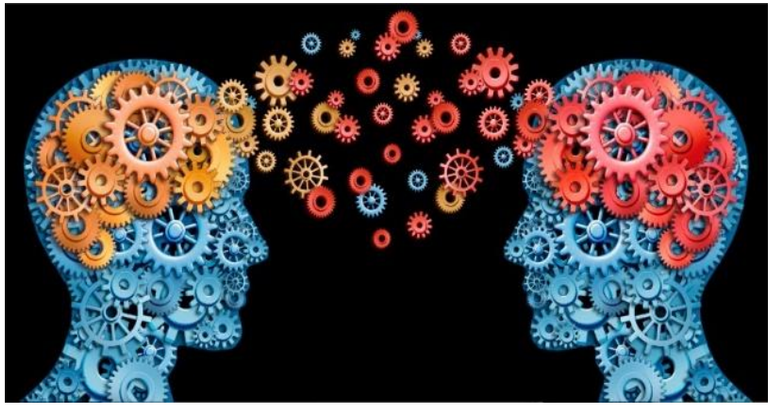
Slide curtesy of Dr. Jess Mazerik

Data and System Interoperability

- Use standards whenever possible (e.g. GA4GH, DICOM, FHIR, CDEs, …)
- Support Findable, Accessible, Interoperable, and Reusable - (FAIR) *data principles*

- **Data interoperability**
  - Structured data (via CDEs) – supports curation and aggregation of data across studies (datasets).
  - UMLS concept coding (NLM CDE governance guidance) supports searching across instances for specific concepts.
  - Data and file formats
  - BRICS PPRL process is called a GUID

- **System Interoperability**
  - Research Authentication Services (RAS)
  - Programmatic access to data and metadata (DATS2.2)
  - Common software APIs to data and metadata



Slide curtesy of Dr Matt McAuliffe

RADx Data Hub and Data Harmonization

The Data Hub is a robust data analytics platform compliant with FAIR and CARE Principles

Data Coordinating Centers

- RADx-UP (Duke University; The University of North Carolina at Chapel Hill)
- RADx Tech (UMASS)
- RADx-rad (UCSD; UTHealth The University of Texas Health Science Center at Houston)
- Digital Health Technologies

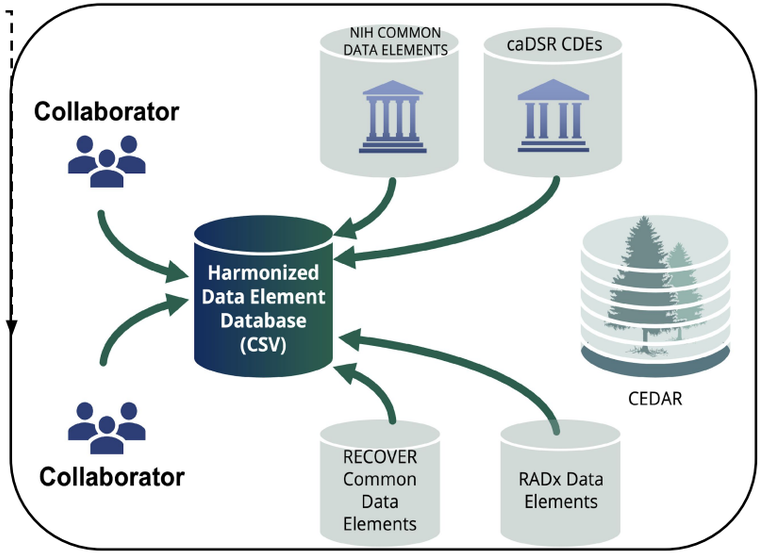
RADx Data Hub Repository

Studies are discoverable in RADx Data Hub and dbGaP catalogue listings

- Supports Common Data Elements
- Supports Standard Metadata
- Data Management
- Data Curation and Harmonization
- Researcher Auth Service (RAS)

Data Harmonization

- TIER 1: Required Common Data Elements
- TIER 2: Program Specific Data Elements
- Integrated Webservice/User Interface
- User-accessible workflow to accommodate new data elements



* Rapid Acceleration of Diagnostics (RADx) https://www.nih.gov/research-training/medical-research-initiatives/radx

Advanced Harmonization Plan

1 Apply Knowledge to Enhance Methods

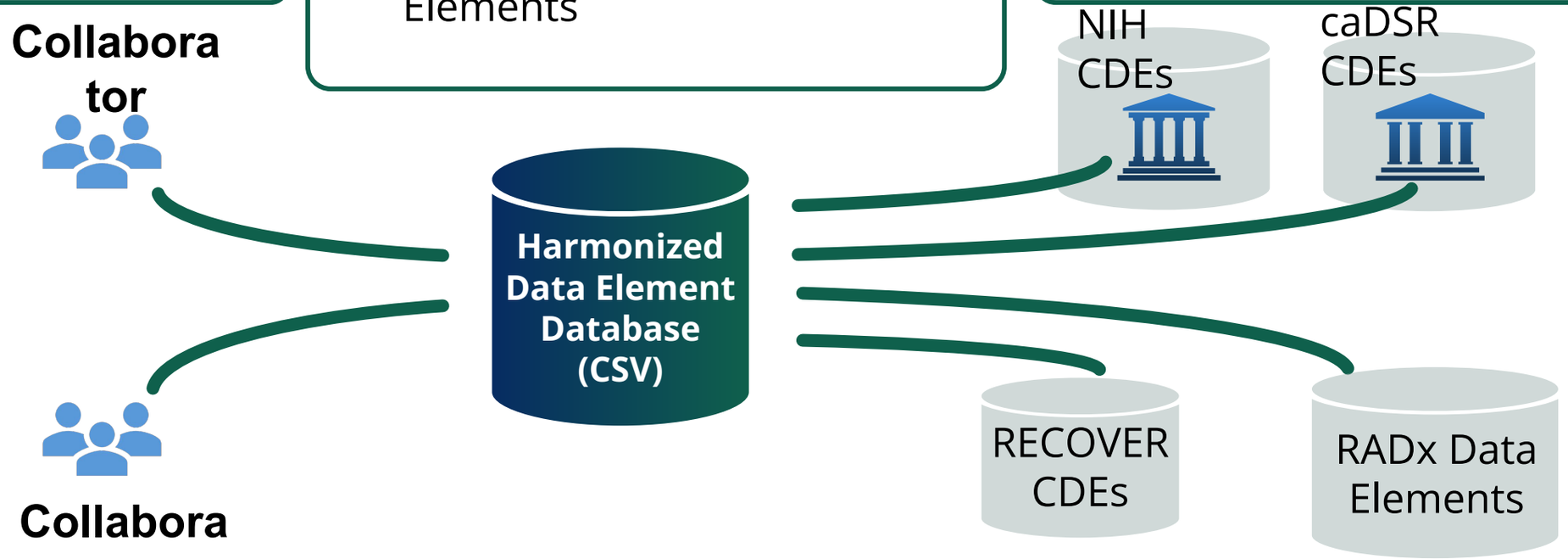
- Apply previous experiences and algorithms
- Reuse other community/ collaborator project results

2 Collect additional data elements

- Incorporate NIH Common Data Element Repository*and RADx program data elements
- Establish set of 'gold standard' Common Data Elements

3 Make Harmonized Resources Easier to Use

- Integrated web service and user interface
- Reusable, user-accessible workflow to accommodate new data elements



Addressing Gaps in the Data Sharing Ecosystem

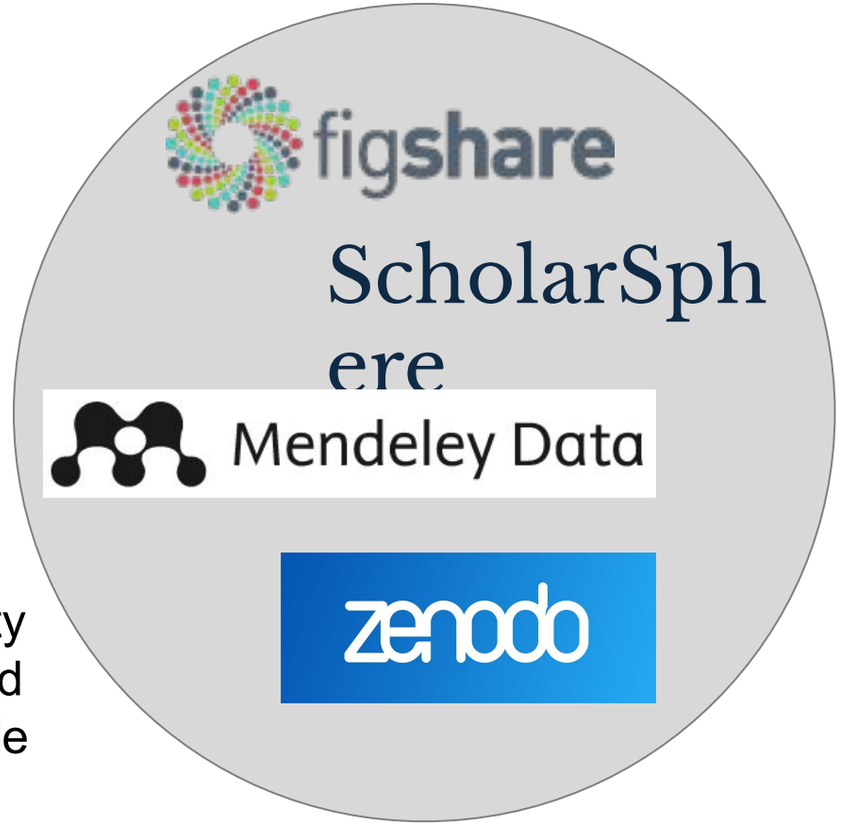
“Domain” or “Specialized” Repositories

PRO: Highly detailed descriptive information; High quality data
CON: Narrow focus; High cost of biocuration

“Hybrid” Repositories

Addresses Gaps: Flexibility to adapt to new species and assays, with minimum viable information for reusability; “Self-serve” style of data curation with structure to guide scientists to produce quality (meta)data.

“Generalist” Repositories



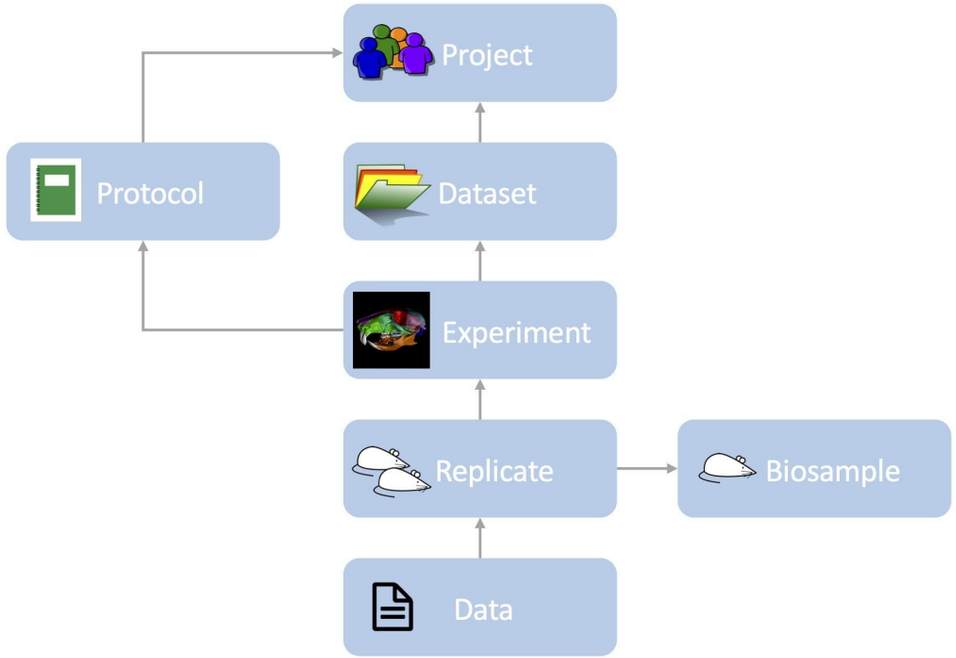
PRO: All types of data and science; Highly scalable
CON: Minimal structure for detail on data; Quality concerns

Slide curtesy of Drs. Schuler and Kesselman

Structuring Metadata and Data Models

Creating consistent, minimal and shareable, metadata or data models

- **Balance between *simplicity* and *detail*** – a highly detailed data model is advantageous for reproducibility but presents a major barrier to entry and thus reduces participation
- **User in-the-loop** – Begin with established standards, published data structures, and then involve your user community in the design and evaluation of the repository's data model
- **Standardize** – adopt vocabulary for all relevant metadata attributes (species, anatomy, gene…)



FaceBase Example

Slide curtesy of Drs. Schuler and Kesselman

Challenges of Serving a Diverse Community

- Repositories must evolve at the “speed of science” and avoid painting themselves into a corner by becoming overly specialized and rigid
- Repositories must provide enough structure and guidance to empower scientists to create truly “FAIR” datasets at every point of the data lifecycle, without becoming overly prescriptive
- Population studies for example, entail numerous unique measures, such as questionnaires and clinical reports, that cannot be standardized away
- One-size-fits-all approach will not satisfy the diverse and broad scientific inquiries needed by those we hope to serve

PROMOTING INTEROPERABILITY OF DATA AND DATA RESOURCES

Impact: Promote FAIRness not just across data resources, but also other digital research objects (such as code) to create a fully interoperable digital research ecosystem

IDEAS:

- Develop and implement **open and computable schemas, metadata, data dictionaries/models, and Common Data Elements (CDEs)** to enable:
  - Interoperability across data and systems
  - Federated search across repositories
- Develop **tools** that support annotation of data to facilitate creation of interoperable FAIR data
- Develop **minimum standards/schemas for APIs** to promote computational interoperability across resources

Thought Question

How /Should we create effective harmonization capability across (at least) larger NIH projects (RADx, RECOVER, AMP, HEAL, BRAIN…)?

Outcome: Minimal set of consistent and computable, findable data elements with consistent data model

NIH Data and Technology Advancement (DATA) National Service Scholar Program

- One- or two-year **national service sabbatical** in high-impact NIH programs
- Experts in **Data science and technology** to advance NIH mission
- About 5 fellows each year

https://datascience.nih.gov/data-scholars-2022

Example Projects include:

- Eye Health Data Interoperability
- AI-Ready Data for Pandemic Preparedness
- Automating Consumer Health Information
- Wearables Predicting Clinical Outcomes
- Multi Modal Cancer Data
- AI/ML for Genomics
- AI/ML for Medical Image and Clinical Data
- Accelerating Medicines Partnership

2021 Data Scholars and Projects

Dr. Anne Deslattes Mays
Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD)
Cross-project use cases for Kids First Data Resource & INCLUDE Data Hub

Dr. Priyanka Ghosh
National Center for Biotechnology Information in the National Library of Medicine (NLM)
Scalable, advanced search methods in Sequence Read Archive (SRA).

Dr. Lara Clark
National Institute of Environmental Health Sciences (NIEHS)
Software/code, documentation, tutorials, manuscripts, and outreach for environmental health.

Dr. Jaleal Sanjak
National Center for Advancement of Translational Science (NCATS)
Data integration applied across 7,000 rare diseases

Dr. John Gachago
Office of Data Science Strategy (OD)
Electronic health records for health disparities research and ethical use of machine learning/artificial intelligence (ML/AI) techniques.

Dr. Ansu Chatterjee
All of Us Research Program (OD)
Multimodal data integration and record linkages, and ethical data science and machine learning for research on large biomedical databases.

ODSS DATA SHARING & REUSE SEMINAR SERIES

Highlighting exemplars of data sharing/reuse monthly on 2nd Friday

Past Speakers:

Karen E. Adolph, PhD
Databrary: Secure and Ethical Sharing of Research Video as Data and Documentation

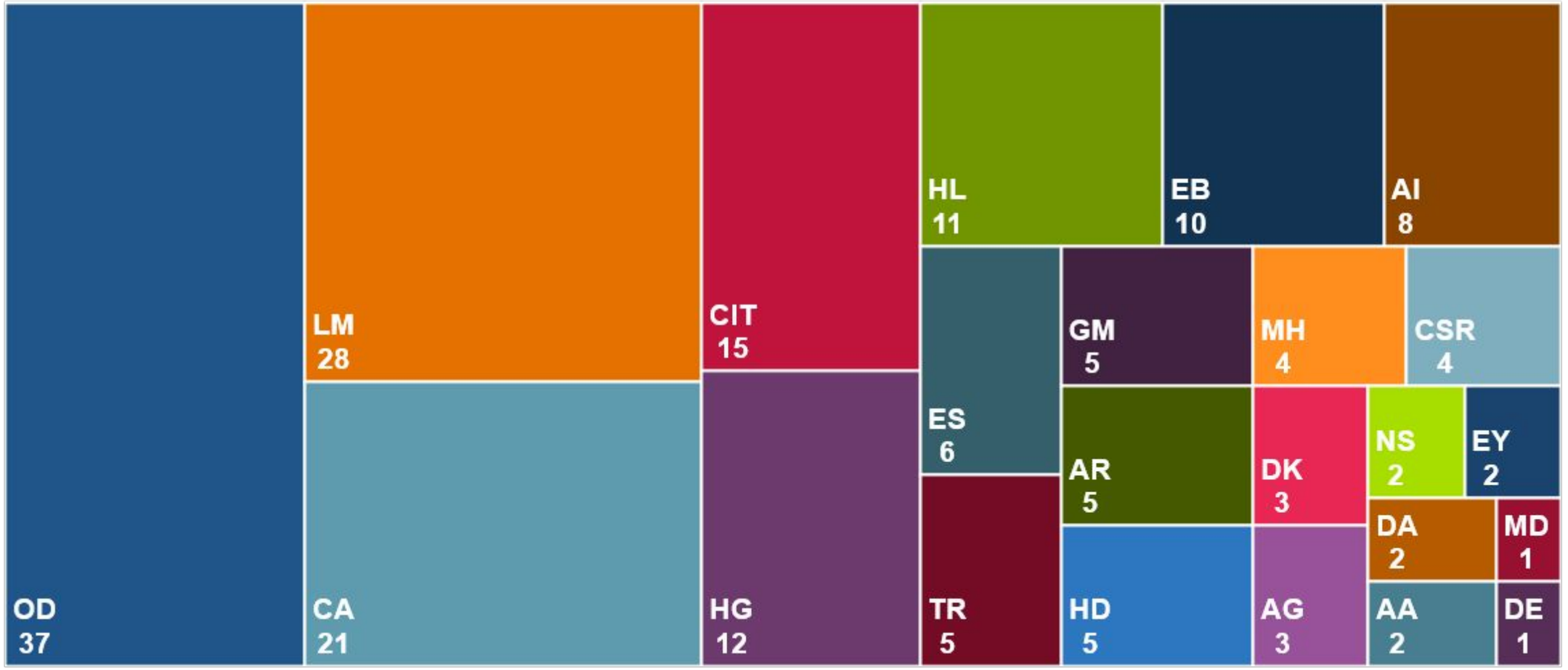
Purvesh Khatri, PhD
Adventures of a Data Parasite: Accelerating Clinical Translation Using Heterogeneity in Public Data

Alexander Ropelewski
The Brain Image Library: A Resource for Sharing Microscopy Data

https://datascience.nih.gov/nih-data-sharing-and-reuse-seminar-series

Catalyzing Data Science Across NIH

More than 200 NIH staff from 23 ICOs contributed to these activities



Office of Data Science Strategy

www.datascience.nih.gov

A modernized, integrated, FAIR biomedical data ecosystem