June 22-23, 2022

Welcome to Day 1

NIH Cloud Platform Interoperability Spring 2022 Virtual Workshop

NIH National Heart, Lung, and Blood Institute | BioData CATALYST

AnVIL

NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center

NIH National Library of Medicine National Center for Biotechnology Information

Welcome

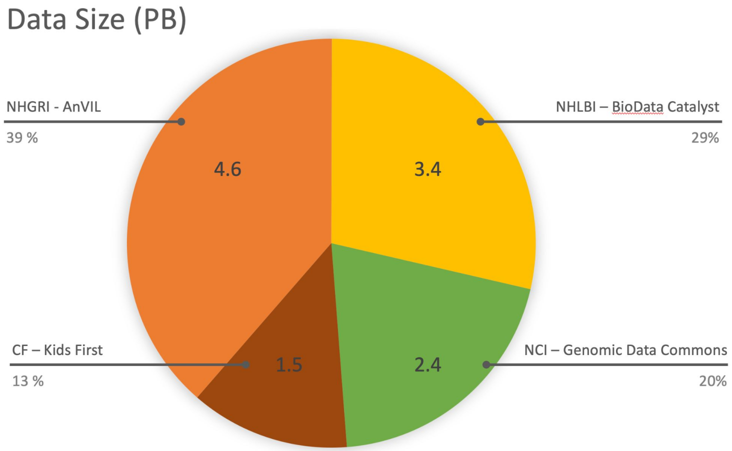
Michael Schatz (Johns Hopkins University)
Anthony Phillipakis (Broad Institute)

Introductions

Michael Schatz
Johns Hopkins University
Computer Science and Biology

Anthony Philippakis
Broad Institute
Chief Data Officer & Institute Scientist

NCPI & The NCPI Dataset Catalog



Researcher Auth Service

Data Repository Service

Fast Healthcare Interoperability Resources

12Pb / 828k participants and growing!

Cross-platform accessibility through several key technologies

Today’s Agenda

Agenda

Day 1: Wednesday, June 22, 2022

11:00 AM - 11:10 AM – Welcome and goals of meeting

Anthony Phillipakis (Broad Institute) and Michael Schatz (Johns Hopkins University)

FAIR Data, Computing, Cataloging Resources Across the NIH and Global Communities

The vision shared in this session will provide foundation and direction to chart the collaborative future of NCPI.

11:10 AM - 11:30 – The NIH Strategic Vision for Data Science, Successes and Opportunities for the Next 5 Years

Laura Biven (ODSS)

11:30 AM - 11:50 AM – ODSS All-Hands Meeting Report Out: Data and Compute Infrastructure

Tanja Davidsen (NCI)

11:50 - 12:05 PM – Interoperability in Data Mesh

Samia Rahman (Seagen)

12:05 PM - 12:20 PM – Integrating Data and Knowledge Across Multiple Species: The Importance of Biological Concept Harmonization

Carol Bult (The Jackson Laboratory)

12:20 PM - 12:35 PM – Playing telephone with data access: success with GA4GH DRS

Titus Brown (UC Davis)

12:35 PM - 1:05 PM – Break

1:05 PM - 1:50 PM Panel Discussion with Commercial Cloud Vendors

Patrick Combes - Amazon Web Services
Jer-Ming Chia - Microsoft Azure
Adrish Sannyasi - Google Cloud Platform
Moderator: Michael Schatz

1:50 PM - 2:00 PM – Break

2:00 PM - 4:00 PM – Parallel Breakout Sessions

20 min - ***Data Mesh***
20 min - ***Reproducibility***
20 min - ***Resource and service readiness for AI/ML***
20 min - ***Engaging partnerships (i.e., GA4GH, Elixir, CFDE, Alliance of Genomic Resources)***
5 min - Moderators prepare report back
25 min - Report back

Day 1 Breakout Moderators

Parallel Session 1	Allison Heath	Brian O'Connor
Parallel Session 2	Valentina Di Francesco	Mike Feolo
Parallel Session 3	Chris Wellington	Stan Ahalt
Parallel Session 4	Kathy Reinold	Adam Resnick
Parallel Session 5	Michael Schatz	Rachel Liao

4:00 PM – Conclusion of Workshop Day 1

Tomorrows's Agenda

Day 2: Thursday, June 23, 2022

11:00 AM - 11:05 AM – Welcome and start of Day 2

Stephen Mosher (Johns Hopkins University)

Interoperability Driven Science

Cloud platform interoperability enables scientific discovery. Here we will learn of the latest advances in NCPI demonstration projects and related cloud platforms.

11:05 AM - 11:20 AM – The ELIXIR Cloud for European Life Sciences

Jonathan Tedds (ELIXIR)

11:20 AM - 11:35 AM – Sex chromosome complement aware alignments

Melissa Wilson (ASU)

11:35 AM - 11:50 AM – Genome-wide Sequencing Analysis to Identify the Genes Responsible for Enchondromatoses and Related Malignant Tumors.

Nara Sobreira (JHU)

11:50 AM - 1:05 PM – Working Group Updates

15 min - Community/Governance WG

Bob Grossman (University of Chicago)

Stanley Ahalt (University of North Carolina at Chapel Hill)

15 min - Systems Interoperation WG

Jack DiGiovanna (SevenBridges)

15 min - FHIR WG

Robert Carroll (Vanderbilt University Medical Center)

15 min - NCPI Outreach WG

Stephen Mosher (Johns Hopkins University)

15 min - Search WG

Dave Rogers (Clever Canary)

Kathy Reinold (Broad Institute)

1:05 PM - 1:35 PM – Break

Technical Aspects of Interoperability

Technologies that enable interoperability are important to develop with stakeholders involved to promote the usability of the technical standards and products. In this session, we will hear about technologies enabling interoperability and their successful implementations in research.

1:35 PM - 1:50 PM – The Texas Advanced Computing Center (TACC) as an Interoperable Cloud Resource for Biomedical Research

Dan Stanzione (TACC)

1:50 PM - 2:05 PM – FHIR for Genomics: The Path Forward

Mullai Murugan (Baylor College of Medicine)

2:05 PM - 2:20 PM – Supporting Genomic Data Sharing through the Global Alliance for Genomics and Health

Heidi Rehm (Broad Institute)

2:20 PM - 2:35 PM – Interoperability Opportunities & Challenges with the Cloud and STRIDES

Nick Weber (NIH STRIDES)

2:35 PM - 3:10 PM – Concurrent Breakouts

Topic 1: Bringing researchers to cloud computing

Topic 2: Reproducibility and Interoperability of batch and ad hoc analyses

Topic 3: What technologies and data types are missing across platforms?

Topic 4: Diversifying genomic data science

Topic 5: Flagship use cases for interoperability

Day 2 Breakout Moderators

Topic 1: Bringing researchers to cloud computing	Tiffany Miller
Topic 2: Reproducibility and Interoperability of batch and ad hoc analyses	Jack DiGiovanna
Topic 3: What technologies and data types are missing across platforms?	Ken Wiley
Topic 4: Diversifying genomic data science	Asiyah Lin
Topic 5: Flagship use cases for interoperability	Michael Schatz

3:10 PM - 3:50 PM – Report Back

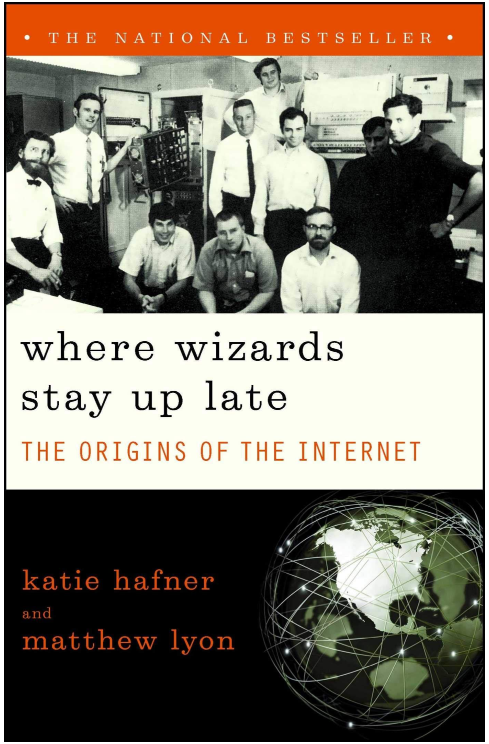
5 minutes for report prep; 5 minute report per group; 10 minutes open discussion

3:50 PM - 4:00 PM – Summary, Future Directions, & Meeting close

Michael Schatz (Johns Hopkins University)

4:00 PM – Meeting close

Driving thoughts on interoperability



We should not underestimate the importance of interoperability...

- If we are successful, we will catalyze the creation of an open and federated data ecosystem.
  - Others have done it before (SWIFT, the internet, the web).
- If we fail, we will degenerate into a collection of monolithic data silos
  - Others have done this before too (medical records in US hospitals)...

FAIR Data, Computing, Cataloging Resources Across the NIH and Global Communities

11:10 AM - 12:35 AM EDT

The NIH Strategic Vision for Data Science, Successes and Opportunities for the Next 5 Years

Laura Biven (ODSS)

National Institutes of Health
Office of Data Science Strategy

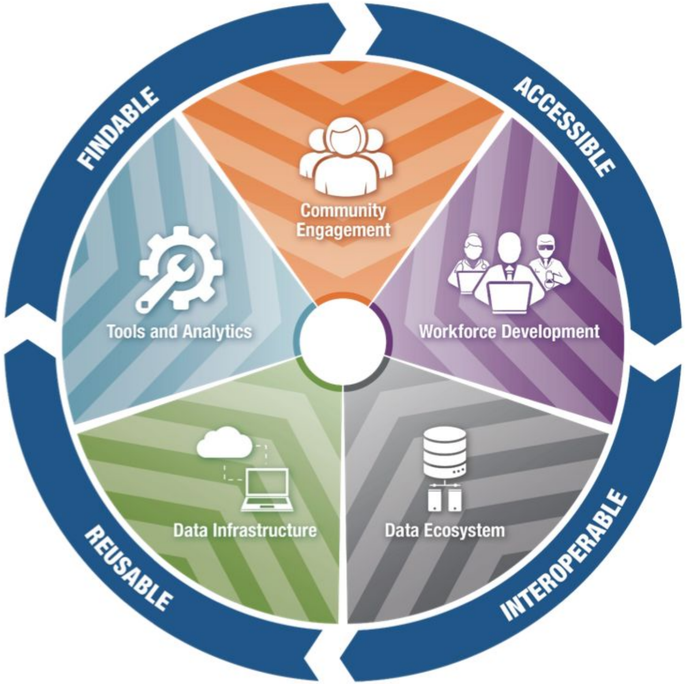
The NIH Strategic Vision for Data Science, Successes and Opportunities for the Next 5 Years

Laura Biven, Ph.D.
Lead, Integrated Infrastructure and Emerging Technologies
Office of Data Science Strategy

6/22/2022

Our vision has been built on the Strategic Plan for Data Science

- Support common infrastructure and architecture on which more specialized platforms can be built and interconnected.
- Leverage commercial tools, technologies, services, and expertise; and adopt and adapt tools and technologies from other fields for use in biomedical research.
- Enhance the nation's biomedical data-science research workforce through improved training programs and novel partnerships.
- Enhance data sharing, access, and interoperability such that NIH-supported data resources are FAIR.
- Ensure the security and confidentiality of patient and participant data in accordance with NIH requirements and applicable law.
- Improve the ability to capture, curate, validate, store, and analyze clinical data for biomedical research.
- With community input, develop, promote—and refine as needed—data standards, including standardized data vocabularies and ontologies, applicable to a broad range of fields.



NIH Strategic Plan for Data Science – Goals & Objectives

Data Infrastructure
- Optimize data storage and security
- Connect NIH data systems

Modernized Data Ecosystem
- Modernize data repository ecosystems
- Support storage and sharing of individual datasets
- Better integrate clinical and observational data into biomedical data science

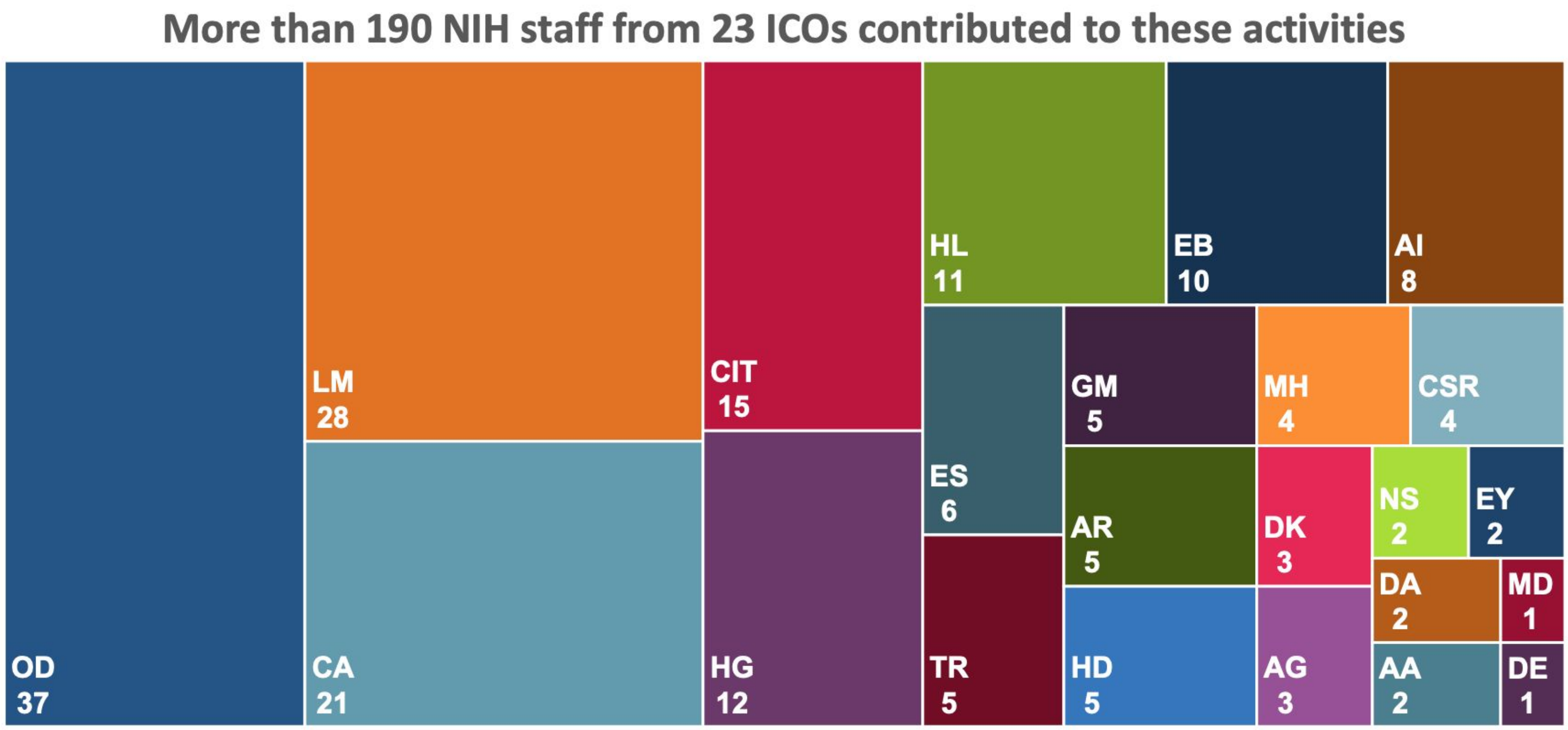
Data Management, Analytics, and Tools
- Support useful, generalizable, and accessible tools
- Broaden utility of, and access to, specialized tools
- Improve discovery and cataloging resources

Workforce Development
- Enhance the NIH data science workforce
- Expand the national research workforce
- Engage a broader community

Stewardship and Sustainability
- Develop policies for a FAIR data ecosystem
- Enhance stewardship

https://datascience.nih.gov/strategicplan

Catalyzing Data Science Across NIH



NIH Strategic Plan for Data Science – Infrastructure

Data Infrastructure

Optimize data storage and security

Connect NIH data systems

STRIDES – Science Technology Research Infrastructure for Discovery, Experimentation and Sustainability

- Initiative aims to modernize biomedical research by reducing economic and process barriers in utilizing commercial cloud services
- *https://cloud.nih.gov/*

RAS – Research Authentication Service

- *Providing easy single-on authentication ("passport access) across platforms and "visa" authorizations for data within the platforms*
- *https://datascience.nih.gov/researcher-auth-service-initiative*

NCPI – NIH Cloud Platform Interoperability

- *Creating a research data mesh across NIH data platforms*
- *https://datascience.nih.gov/nih-cloud-platform-interoperability-effort*

https://datascience.nih.gov/data-infrastructure

NIH Strategic Plan for Data Science – Data Ecosystem

Modernized Data Ecosystem

- Modernize data repository ecosystems
- Support storage and sharing of individual datasets
- Better integrate clinical and observational data into biomedical data science

Enhance FAIRness through Administrative Supplements

- *Implement desirable characteristics for data repositories - NOT-OD-22-069*
- *Improve the AI/ML-Readiness of NIH-Supported Data NOT-OD-22-065*
- Training and research to utilize FHIR for clinical research

Data Repositories & Knowledgebases

- *Active funding opportunities for early-stage & established data repositories and knowledgebases, PAR20-089 & PAR20-097*
- *Enhance us of Common Data Elements*

Generalist Repository Ecosystem Initiative

- *Enhance discoverability and use of NIH funded and generated research data*
  - *Establish consistent capabilities*
  - *Encourage sharing/reusing data*

https://datascience.nih.gov/data-ecosystem

NIH Strategic Plan for Data Science – Software & Tools

Data Management, Analytics, and Tools

Support useful, generalizable, and accessible tools

Broaden utility of, and access to, specialized tools

Improve discovery and cataloging resources

Smart & Connected Health Program

- Accelerate innovations in computer science and engineering to support the transformation of health and medicine
- *Funding Opportunity Announcement - NOT-OD-21-011 (NSF partnership)*

Enhancing Sustainability & Reuse of Research Software

- FHIR implementation
- *Enhancement of Software Tools for Open Science - NOT-OD-22-068*
- *Develop digital technologies to for data from wearable & related devices*

Foster Software Communities of Practice

- *Developed Best Practices for Sharing Software*
- https://bit.ly/3t2SJ71

https://datascience.nih.gov/tools-and-analytics

NIH Strategic Plan for Data Science – Workforce & Training

Workforce Development

Enhance the NIH data science workforce

Expand the national research workforce

Engage a broader community

DATA National Service Scholars Program

- *Recruit* experienced data and computer scientists and engineers to advance high-impact NIH programs
- https://datascience.nih.gov/data-scholars-2022

Coding-it-Forward Summer Fellowship

Other Teaching & Learning Programs

- *Graduate Data Science Summer Program*
- NIH Codathon Program
- Science Teachers Boot Camps

https://datascience.nih.gov/workforce-development

NIH Strategic Plan for Data Science – Policy & Stewardship

Stewardship and Sustainability

Develop policies for a FAIR data ecosystem

Enhance stewardship

New Data Management & Sharing Policy

- ***NOT-OD-21-013* – Final Policy**
- *NOT-OD-21-014 (DMS Plan); NOT-OD-21-015 (Cost); NOT-OD-21-016 (choosing a repository)*
- *https://sharing.nih.gov/*

Data Stewardship Programs

- NNLM Data Service
- Data Curation Network (DCN) Workshop Series

Incentives

- DataWorks! Prize – FASEB & NIH as partners
- *herox.com/dataworks*

https://osp.od.nih.gov/


We recognize the challenges ahead

A Few Challenges

- **Federated Data Infrastructure**; creating a facile data/compute 'mesh' across multiple data platforms/systems. Enabling hybrid computing is another challenge
- **Efficient Access to Controlled Access Data**; enabling automation and maintaining data governance and policy adherence
- **Integration of Clinical, Real World Data, Administrative Data, Digital Health Data;** linking data across platforms and systems is another issue
- **Searching for Data, Building Cohorts, Developing Knowledge**


What's Next

Supporting a Federated Biomedical Research Data Infrastructure

1) Harmonizing data infrastructure and services

IMPACT: To improve FAIRness of **data, computing, and modeling** resources across the NIH

IDEAS:

A NIH (Research) Data Mesh, A de-centralized and distributed system that harmonizes services across the NIH. These may include (not limited to):

- A searchable **NIH-wide resource** of available resources cross NIH
- **Common APIs**, metadata models, digital object identifiers and indexing approaches within existing NIH IC Cloud Ecosystems (e.g., CRDC, AnVIL, BDC, KidsFirst, All of Us, dbGaP/SRA, RECOVER) and new ones
- **User single sign-on** system that uses smart tokens to aide communications within mesh, data access control, auditing and accounting – building on the current RAS work
- **Enable greater data availability** for ICs utilizing STRIDES

Supporting a Federated Biomedical Research Data Infrastructure

2) ENHANCING RESOURCES FOR AI/ML

IMPACT: To improve **data, computation, and modeling** infrastructure for AI/ML analyses across the NIH

IDEAS:

- Increase efforts for data and metadata standardization and indexing
- Develop tools for **ethical AI** and reduce biases in training sets for AI/ML models
- Develop **synthetic datasets** used to train AIs when real data is too scarce or sensitive to use
- Enable **iterative model training** as more data becomes available
- Incorporate **SDOH** data in model training
- Support IC and research communities to more effectively adopt AI with adequate **infrastructure and tools**
- Leverage data collected **from passive sensing devices** to make localized AI models, collect de-identified EHR information

3) 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 standardized schemas, metadata, data formats, and Common Data Elements (CDEs)** to enable:
  - Interoperability across data and systems
  - Federated search across repositories
- Develop **tools** that support structured and standardized metadata and annotation of data to facilitate creation of interoperable FAIR data
- Develop **minimum standards/schemas for APIs** to promote computational interoperability across resources
- Encourage the use of open data and metadata formats

Research Inspired Clinical Data Science

4) Foster Broad Research Use of Healthcare Data

Impact: Researchers able to easily find, access, and use/re-use clinical datasets to make discoveries that improve health and validate published findings

IDEAS :

- Adoption of data and metadata standards to facilitate large-scale analysis across studies, ethical use, and reproducibility.
- Develop data and metadata standards in new and emerging communities such as RWD and SDOH
- Enable linkages of SDoH data with other datatypes, including clinical, healthcare, administrative and RWD
- Develop governance and policy frameworks to guide when/how data linking is appropriate, and how linked data can be shared and used in the context of identifiability risk, consent (even if de-identified), and regulatory parameters – *the value of linkage needs to be balanced with ethical risks of linking and using linked data.*

Research Inspired Clinical Data Science

5) Adopt Health IT Standards for Research – use what exists

Impact: Researchers benefiting from well-established and/or required health IT and related standards based on driving use cases

IDEAS:

- Implement AGILE programs that convene researchers and developers in working groups where they test, implement, adopt, and provide feedback on health IT technologies and standards based on scientific use cases — further driving improvements and adoption
- Enhance Fast HealthCare Interoperability Resources (FHIR) to bridge the data gap between the clinical settings and clinical research
- Enhance use of FHIR for Cohort Discover

Create a Data and Software Ecosystem

6) CULTURE OF SOFTWARE DEVELOPMENT

Impact: Bridging communities and building capacity to transform the development and use of cutting-edge software technologies

IDEAS:

- Develop a community of sustainability
- Establish NIH-wide **metrics and best practices** for data, computational models, and software **sustainability.**
- Provide incentives or programs for software engineers to work with biomedical and behavioral researchers

Create a Data and Software Ecosystem

7) CUTTING-EDGE SOFTWARE TECHNOLOGIES

Impact: Leverage innovative, new technologies to provide data scientists, researchers, and clinicians the powerful tools they need

Ideas:

- Develop tools/workflows to automate the mapping of data elements to common language standards and ontologies
- Leverage novel passive data collection technologies to enable high impact data science applications, downstream
- Facilitate FAIR computational models that incorporate clinical, behavioral, population, and policy data (aka, "above the skin")

Strengthening a Broader Community in Data Science

8) Strengthen data science expertise and diversity

Impact: A multidisciplinary and diverse workforce can accelerate data science research, increase collaboration, and result in more innovative thinking

IDEAS:

- Provided support to use MOOCs to get hands-on practice in ML/AI, etc.
- Pairing a variety of technical trainings (e.g., using cloud, managing data) with domain-specific training
- Leverage NIH IC repositories/analysis platforms as a training resource
- Enhance training in data management and FAIR data, including training in the ethical collection and use of data
- Enhance training in under-represented and under resourced communities
- Developing a community of practice to bridge investigators across disciplines

NEXT STEPS: UPDATING THE STRATEGIC PLAN FOR DATA SCIENCE

- Refine Key Ideas and Concrete Steps
- Develop Evaluation Metrics
- Draft Updated Strategic Plan for Data Science
- Community Engagement and Feedback
- NIH Leadership, Fall 2022
- Finalize Updated Strategic Plan Document by 2023

Recognize that cultural change, not just technological advancement, is necessary for advancing NIH’s data science

ODSS All-Hands Meeting Report Out: Data and Compute Infrastructure

Tanja Davidsen (NCI)

Report out:
ODSS All-Hands Meeting

Data and Compute Infrastructure

Tanja Davidsen, June 2022

Goals for the All-Hands Meeting held on December 13 and 14th, 2021

- *Share the findings and recommendations from the progress that we have made on the NIH Strategic Plan for Data Science.*
- *Hearing from our NIH colleagues, who contribute to data science efforts, new opportunities and challenges in data science.*
- *Network with colleagues across the NIH who are implementing the current strategic plan for data science and learn about successes and opportunities from different tactic teams.*
- *Lay the foundation as a starting point to* ***update the NIH Strategic Plan for Data Science***

Structure of the meeting

- <u>**Parallel breakout sessions (Day 1 and Day 2):**</u>
  - **Breakout 1 Data and Computing Infrastructure**
    - **Co-Chairs:** Valentina Di Francesco (NHGRI), Tanja Davidsen (NCI)
  - **Breakout 2 FAIR Data, Repositories, and Data Sharing**
    - **Co-Chairs:** Lisa Federer (NLM), Michelle Heacock (NIEHS)
  - **Breakout 3 Clinical Data Science**
    - **Co-Chairs:** Susan Wright (NIDA), Valerie Cotton (NICHD)
  - **Breakout 4 Software, Tools, and Methods**
    - **Co-Chairs:** Heidi Sofia (NHGRI), Dana Wolff-Hughes (NCI)
  - **Breakout 5 Intramural Data Science Challenges**
    - **Co-Chairs**: Kim Pruitt (NLM), Matt McAuliffe (CIT)

- <u>**Active participation**</u> from 124 NIH staff across all NIH Institutes and Centers

DAY 1 BREAKOUTs Focused on Bold Ideas to move NIH forward

Charge Questions - Gaps and Opportunities in Data Science: Bold Ideas, move the field forward

- New cutting-edge technologies and enhancing interoperable platforms, repositories, and data
- Engage communities, develop new partnerships, enhance data science
- Create or enhance programs that include diverse perspectives
- Gaps and opportunities in training, workforce development
- How to foster data stewardship & sustainability

DAY 2 BREAKOUT Focused on concrete steps to implementation key ideas from day 1

Charge questions for all break outs: - IMPLEMENTATION - Making it all happen

- Top 2-3 activities or ideas from day
- Concrete steps to implement activities or ideas
- New capabilities, needs and opportunities in training and/or workforce development
- Partnerships and collaborations (internal, external)
- Evaluate the impact and measure success

All Hands Breakout Group: Data & Compute Infrastructure

Valentina Di Francesco (NHGRI), Tanja Davidsen (NCI)

NIH National Institutes of Health
Turning Discovery Into Health

Data and Compute Infrastructure

OVERVIEW OF FOUR KEY IDEAS

1) NIH Wide Data, Computing, Modeling Resources
2) Enhancing Resources for New AI/ML
3) Enhancing Technical Capabilities & Assistance Across NIH
4) Impact and Evaluation

Data and Compute Infrastructure

1) KEY IDEA: NIH WIDE RESOURCES

IMPACT: To improve FAIRness of **data, computing, and modeling** resources across the NIH

CONCRETE STEP: INFRASTRUCTURE

- **Initial step:** Assessment of current NIH data repositories, computing platforms and modeling resources: Find holes and overlap
- **Future step: A NIH Data Mesh,** A de-centralized and distributed system that harmonizes services across the NIH. These may include (not limited to):
    - A searchable **NIH-wide catalog** of available resources cross NIH
    - **Common APIs**, metadata models, digital object identifiers and indexing approaches within existing NIH IC Cloud Ecosystems (e.g., CRDC, AnVIL, BDC, KidsFirst, All of Us, dbGaP/SRA, RECOVER) and new ones
    - **User single sign-on** system that uses smart tokens to aide communications within mesh, data access control, auditing and accounting – building on the current RAS work
    - **Enable greater data availability** for ICs utilizing STRIDES

Data and Compute Infrastructure

1) KEY IDEA: NIH WIDE RESOURCES

CONCRETE STEPS: PARTNERSHIPS AND COLLABORATIONS

- **Encourage Data Science Points of Contact (POCs) at each IC**
  - POC already established for ICs deep in data science
  - Identify a POC at NIH ICs where data science is not an established program
- Build on and grow **existing partnerships** with other agencies and centers
  - DOE, VA Million Vets, DoD, UK Biobank, NSF CloudBank to apply best practices
  - Ensure consistent support to awardees funded by NSF and NIH
- Engage third parties to conduct ML Data Analytic Boot Camps

Data and Compute Infrastructure

2) KEY IDEA: ENHANCING RESOURCES FOR AI/ML

IMPACT: To improve **data, computation, and modeling** infrastructure for AI/ML analyses across the NIH

CONCRETE STEPS:

- Increase efforts for data and metadata standardization and indexing
- Develop tools for **ethical AI** and reduce biases in training sets for AI/ML models
- Develop **synthetic datasets** used to train AIs when real data is too scarce or sensitive to use
- Enable **iterative model training** as more data becomes available
- Incorporate **SDOH** data in model training
- Support IC and research communities to more effectively adopt AI with adequate **infrastructure and tools**
- Leverage data collected **from passive sensing devices** to make localized AI models, collect de-identified EHR information

Data and Compute Infrastructure

3) KEY IDEA: ENHANCING TECHNICAL CAPABILITIES & ASSISTANCE ACROSS NIH

IMPACT: To **improve the resources available to NIH staff** who provide stewardship of the data and compute infrastructure across NIH ICOs

CONCRETE STEPS

- **Sustainability**
  - Establish NIH-wide **metrics and best practices** for data, computational models, and software **sustainability**
  - **Develop data retention metrics** to help determine what data should be retained and at what level of availability
  - Socialize **existing tools (e.g., Dockstore) and new computing advances** to be used across NIH
- **Guidance/Best Practices**
  - Guidance on **interoperability** so all ICs' data and models can be shared more easily
  - Establish a **data steward service** to guide data from generation/curation to a long-term repository
- **Funding**
  - Support ICs with limited data science expertise

Data and Compute Infrastructure

4) EVALUATION AND IMPACT

What does impact/success look like?

- NIH data resources and data services are less siloed, and appear more like a well-integrated and robust ecosystem
- Data, models and computing infrastructure are more widely and easily accessible
  - A novice user without computational experience (e.g., bench biologist, student) easily finds data/models to export to a workspace of their choice to conduct analysis or further explore
- Adoption of RAS more broadly across NIH-supported infrastructure

Cross-Cutting Themes

National Institutes of Health
Turning Discovery Into Health

Cross-Cutting Themes

OVERVIEW OF THREE KEY IDEAS

1) Enhancing Technical Capabilities & Assistance Across NIH
2) Integrate Social Determinants of Health Data into the NIH Data Ecosystem
3) Strengthening NIH Engagements

Cross-Cutting Themes

1) <u>KEY IDEA</u>: ENHANCING TECHNICAL CAPABILITIES & ASSISTANCE ACROSS NIH

<u>Impact:</u> NIH ICs will be able to share and contribute to data science training resources

<u>Concrete Steps:</u>

- Readily available **data science training resources (MOOCs, curricula, videos, tutorials) in a central location**
  - Provided support to use MOOCs to get hands-on practice in ML/AI, etc.
  - Pairing a variety of **technical trainings** (e.g., using cloud, managing data) **with domain-specific training**
  - Provide **support** for individuals who want to take advantage of these resources
- Highlight **successful NIH IC repositories/analysis platforms, interoperability use cases,** leveraging these platforms as a training resource
- **Training in data management and FAIR data**, including **training in the ethical collection and use of data**
- Enhance training in under-represented and under resourced communities
- Developing a community of practice to bridge investigators across disciplines
- Develop a **mentorship program** for NIH staff

Cross-Cutting Themes

2) <u>KEY IDEA</u>: Integrate Social Determinants of Health Data into the NIH Data Ecosystem

<u>Impact:</u> SDOH data are the economic and social conditions that influence an individual and group differences in health outcomes. Increasing the collection and use of SDOH data will enable a greater holistic understandings.

<u>Concrete Steps:</u>

- Partner with HL7 (Gravity) and ONC and communities to promote SDOH standards development with attention to harmonization
- Enable linkages of SDoH data with other datatypes, including clinical, healthcare, administrative and RWD
- Support activities with under-represented groups to expand use of SDOH data models and data collection

Cross-Cutting Themes

3) KEY IDEA: Strengthening NIH Engagements

Impact: Increase the ability to share information on activities across NIH will reduce redundant efforts and increase effective outreach to researchers and scientific communities

Concrete Steps:

- Coordination of NIH-wide efforts with organizations (GA4GH, RDA, etc) and agencies (FDA, DOE, etc)
- Develop mechanisms for continuous input from within and outside of NIH on what is working/not working
- **Consider how to balance consistent DMSP guidance across NIH with the need for discipline- or IC-specific guidance**
- Bring in new expertise and communities, including those that traditionally have not been engaged by the data science community (e.g., institutions, geographic areas, historically underrepresented groups, career levels)

NEXT STEPS: UPDATING THE STRATEGIC PLAN FOR DATA SCIENCE

- Refine Key Ideas and Concrete Steps
- Develop Evaluation Metrics
- Draft Updated Strategic Plan for Data Science
- Community Engagement and Feedback
- Present to SDC, Fall 2022
- Finalize Updated Strategic Plan Document by 2023

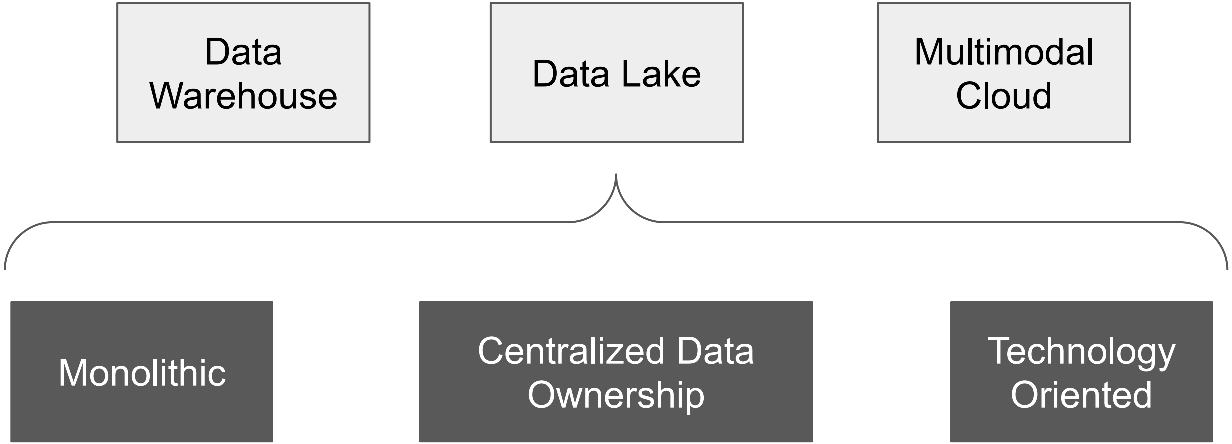
Interoperability in Data Mesh

Samia Rahman (Seagen)

Agenda

- Evolution of Analytical Data Architectures
- What is Data Mesh?
- Achieve Interoperability in Data Mesh

Evolution of Analytical Data Architectures



https://www.oreilly.com/library/view/data-mesh/9781492092384/

What is Data Mesh?

“Data Mesh is a sociotechnical approach to share, access and manage analytical data in complex and large scale environments - within or across organizations”

Domain Oriented Ownership	Data as a product	Self-serve data platform	Federated Computational Governance

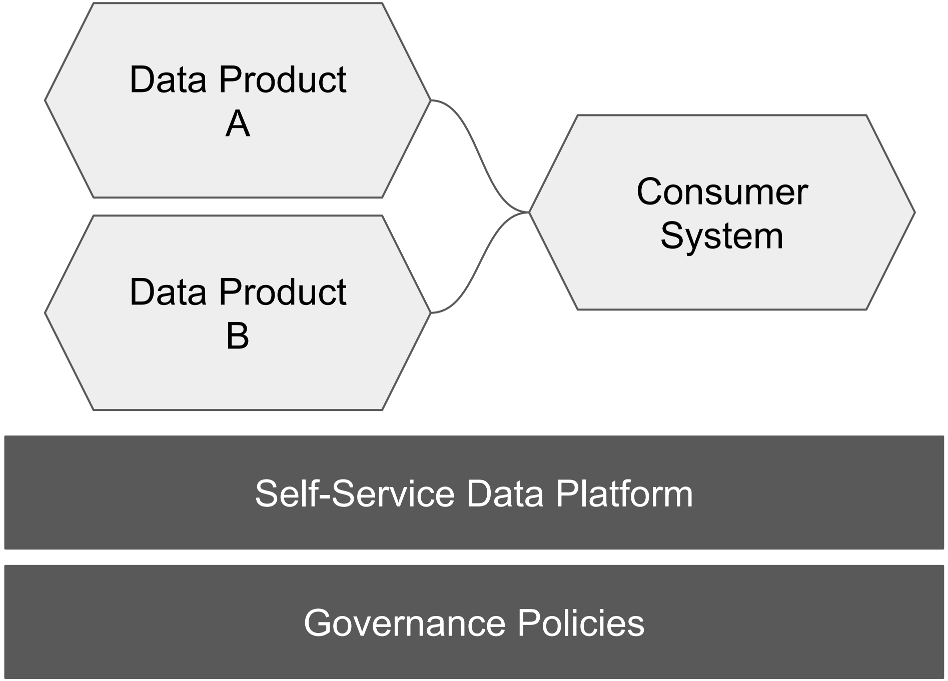
https://www.oreilly.com/library/view/data-mesh/9781492092384/

Data as a Product

Data Product

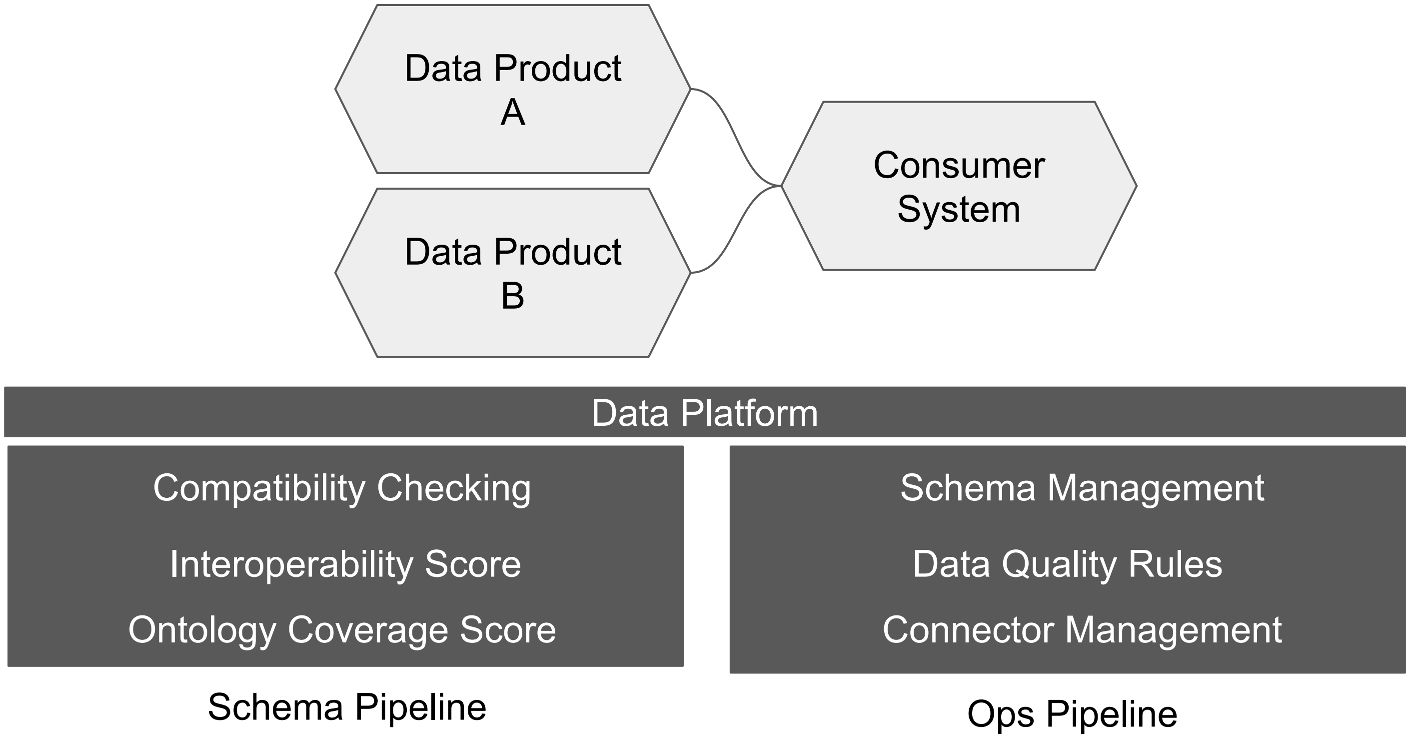
- Findable
- Accessible
- Interoperable
- Reusable
- Valuable
- Trustworthy
- Secure
- Natively Accessible

Achieving Interoperability



https://www.confluent.io/blog/distributed-domain-driven-architecture-data-mesh-best-practices/

Achieving Interoperability



https://www.confluent.io/blog/distributed-domain-driven-architecture-data-mesh-best-practices/

Integrating Data and Knowledge Across Multiple Species: The Importance of Biological Concept Harmonization

Carol Bult (The Jackson Laboratory)

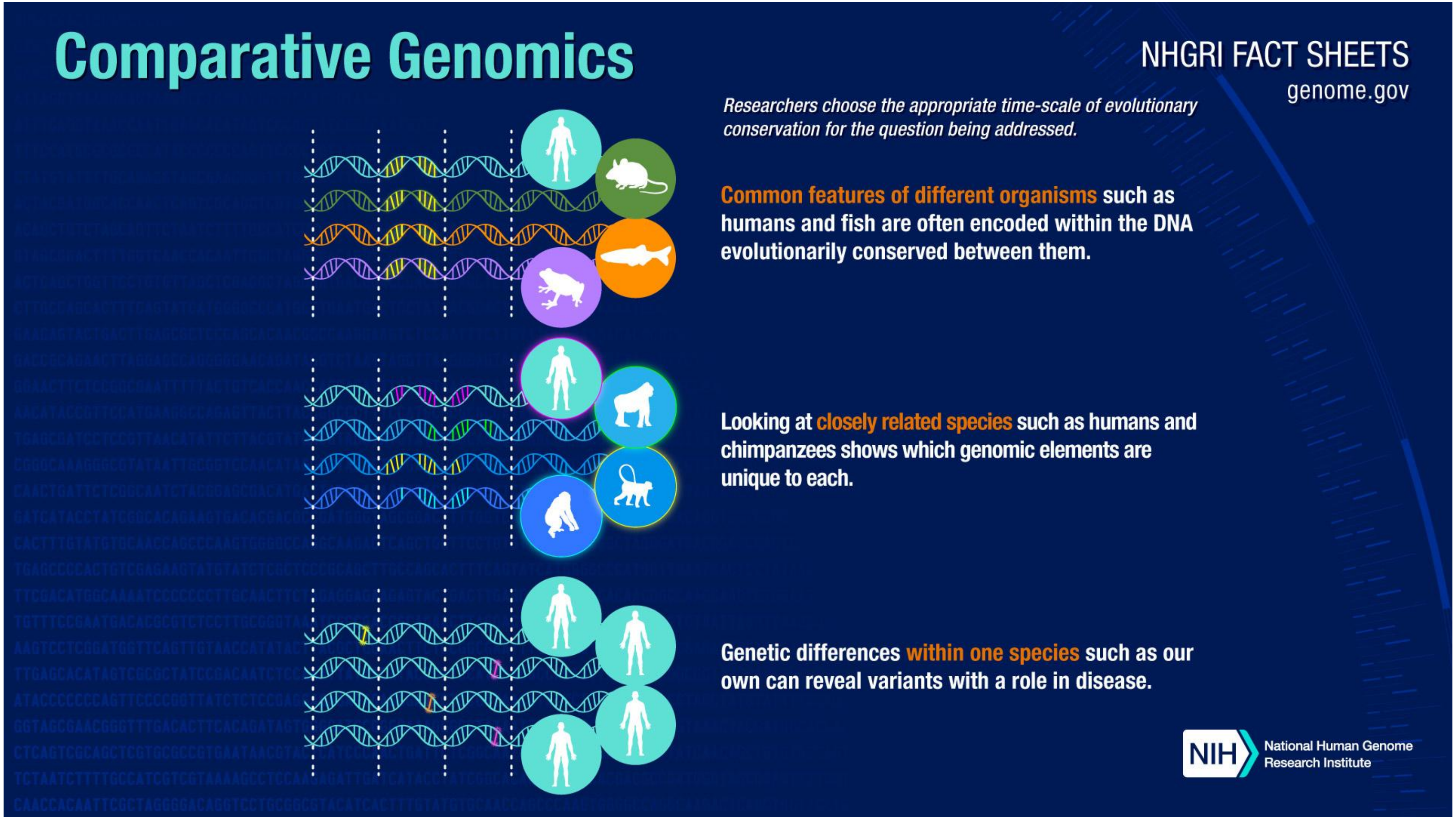
Integrating Data and Knowledge Across Multiple Species: The Importance of Biological Concept Harmonization

Carol Bult, Ph.D.

The Jackson Laboratory

NCPI Spring 2022 Virtual Workshop

June 22-23, 2022



https://www.genome.gov/about-genomics/fact-sheets/Comparative-Genomics-Fact-Sheet

Beyond genomic elements:
Model Organism Databases and the Gene Ontology Consortium

Expression, Function, Phenotype, Disease

Individual MODs represent similar types of data entities/knowledge

Data Entities
- Genomes
- Genome Features
- Alleles & Variants
- Models
- Reagents

Standard nomenclatures

Annotations/Associations
- Function
- Disease
- Phenotype
- Expression
- Interaction
- Regulation

Bio-ontologies

Data
- Movies
- Figures/Images/Pictures
- High Throughput datasets

Standard data formats

But…

Each MOD has unique user interfaces and APIs for data access

The Alliance of Genome Resources: Building a "knowledge commons" for comparative genomics

- Common mechanisms for accessing expertly curated annotations from MODs and GOC
  - Enhanced support for comparative genome biology

- Sustainable genome resource development
  - Shared modular infrastructure to reduce costs of resource development and maintenance

https://alliancegenome.org

Alliance of Genome Resources (2022) Genetics 220(4)
Alliance of Genome Resources (2020) NAR 48(D1):D650-658
Alliance of Genome Resources (2019) Genetics 213(4):118901196

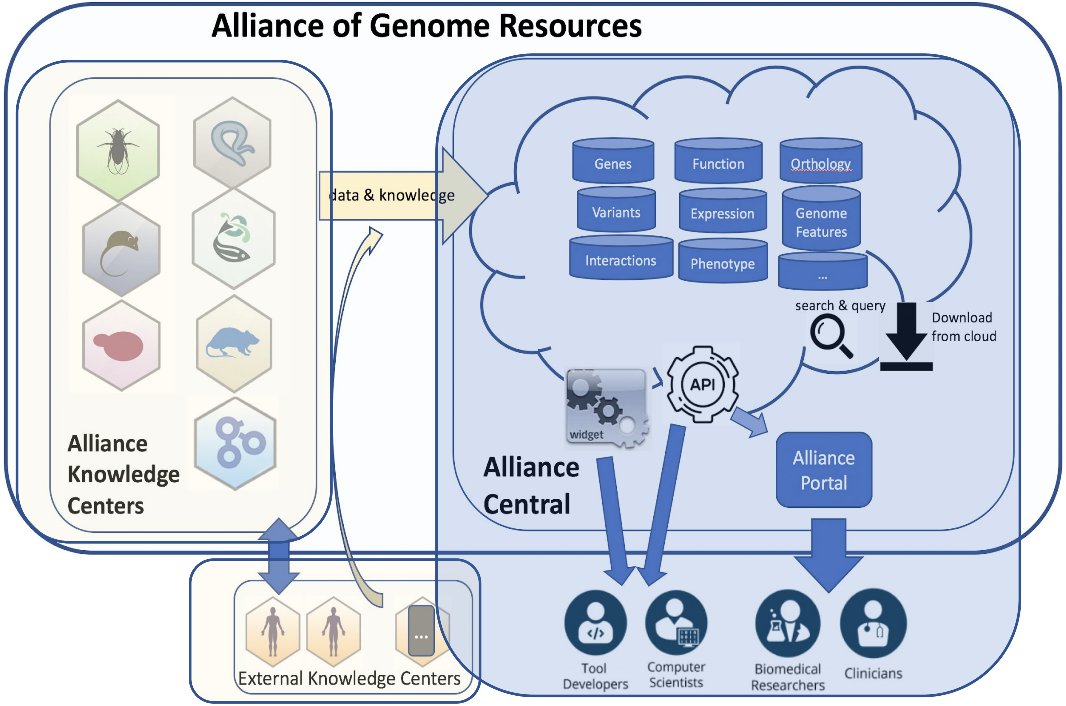
The Alliance “knowledge commons” has two components

Alliance Knowledge Centers: Knowledgebases

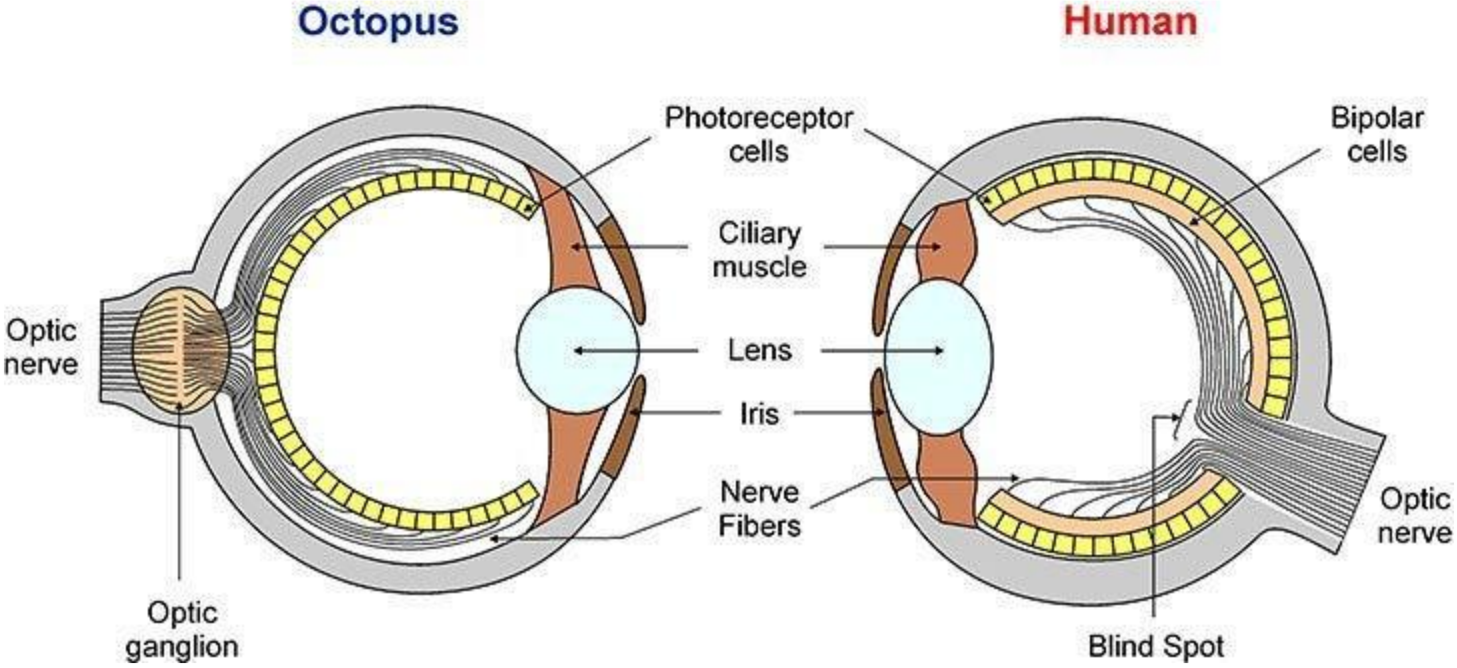
Data Acquisition and Expert Curation
Nomenclature and knowledge representation standards
Organism- specific resources and reagents
Organism-specific community engagement

Alliance Central: Data and infrastructure

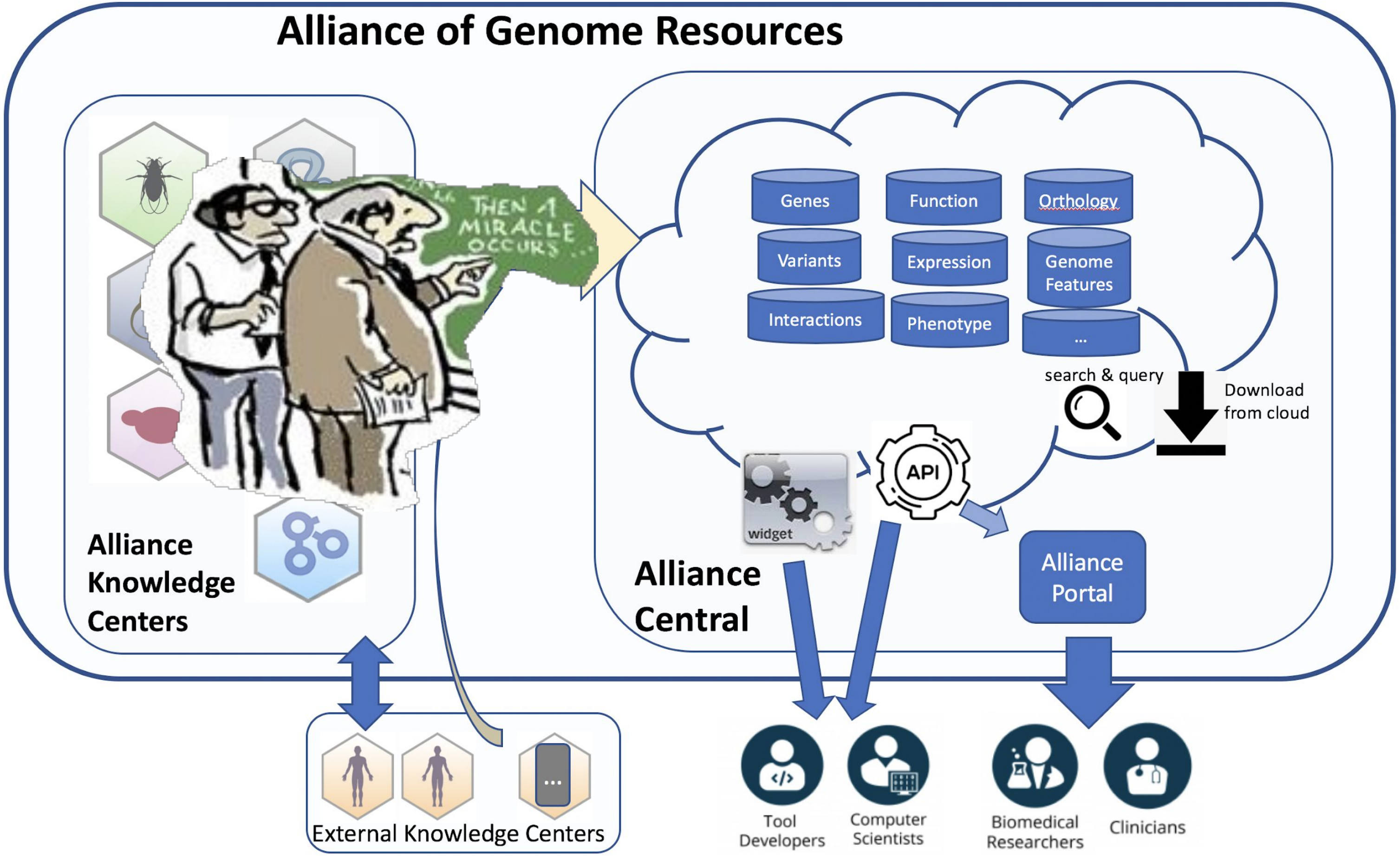
Data management
Programmatic and web data access
Shared user interface development
Platform for tool development



Common data types does not mean common curation processes or biological concept representation



https://i.imgur.com/g936SSE.jpg



Apologies to Sydney Harris…

Harmonization of disease annotations

Case 1: A gene in a model organism genome that is an *ortholog* to a human gene which is associated with (or causal for) a disease

Case 2: A genotype *on a specific genetic background* with expression of phenotype(s) that models the human disease phenotype(s).

The context of an annotation matters for interpretation and computation/prediction

	Gene	Species	Association	Disease	Evidence	Based On
Case 2	Zic3 Annotation details	*Mus musculus*	is implicated in	visceral heterotaxy	TAS	
	zic3 Annotation details	*Danio rerio*	is implicated in	visceral heterotaxy	TAS	
Case 1	Zic3	*Mus musculus*	implicated via orthology	visceral heterotaxy	IEA	ZIC3 (Hsa) zic3 (Dre)
	zic3	*Danio rerio*	implicated via orthology	visceral heterotaxy	IEA	ZIC3 (Hsa) Zic3 (Mmu)

Zic3 Annotation details	*Mus musculus*
zic3 Annotation details	*Danio rerio*

Name	Type	Experimental Condition	Modifier	References
$Zic3^{Bn}/Zic3^{+}$ [background:] BNT/LeJ	Genotype			PMID:10942421 PMID:10861288 Show All 5
$Zic3^{Bn}/Zic3^{Bn}$ [background:] BNT/LeJ	Genotype			PMID:10942421 PMID:10861288 Show All 5
$Zic3^{tm1Bca}/?$ [background:] involves: 129S7/SvEvBrd	Genotype			PMID:11959836
$Zic3^{tm1Bca}/?$ [background:] involves: 129S7/SvEvBrd * C57BL/6	Genotype			PMID:11959836
$Zic3^{tm1Bca}/Zic3^{+}$ [background:] involves: 129S7/SvEvBrd * C57BL/6	Genotype			PMID:11959836
$Zic3^{tm1Bca}/Zic3^{tm1Bca}$ [background:] involves: 129S7/SvEvBrd * C57BL/6	Genotype			PMID:11959836

Name	Type	Experimental Condition	Modifier	References
AB + MO1-zic3	Fish	**has condition:** standard conditions		PMID:22285814

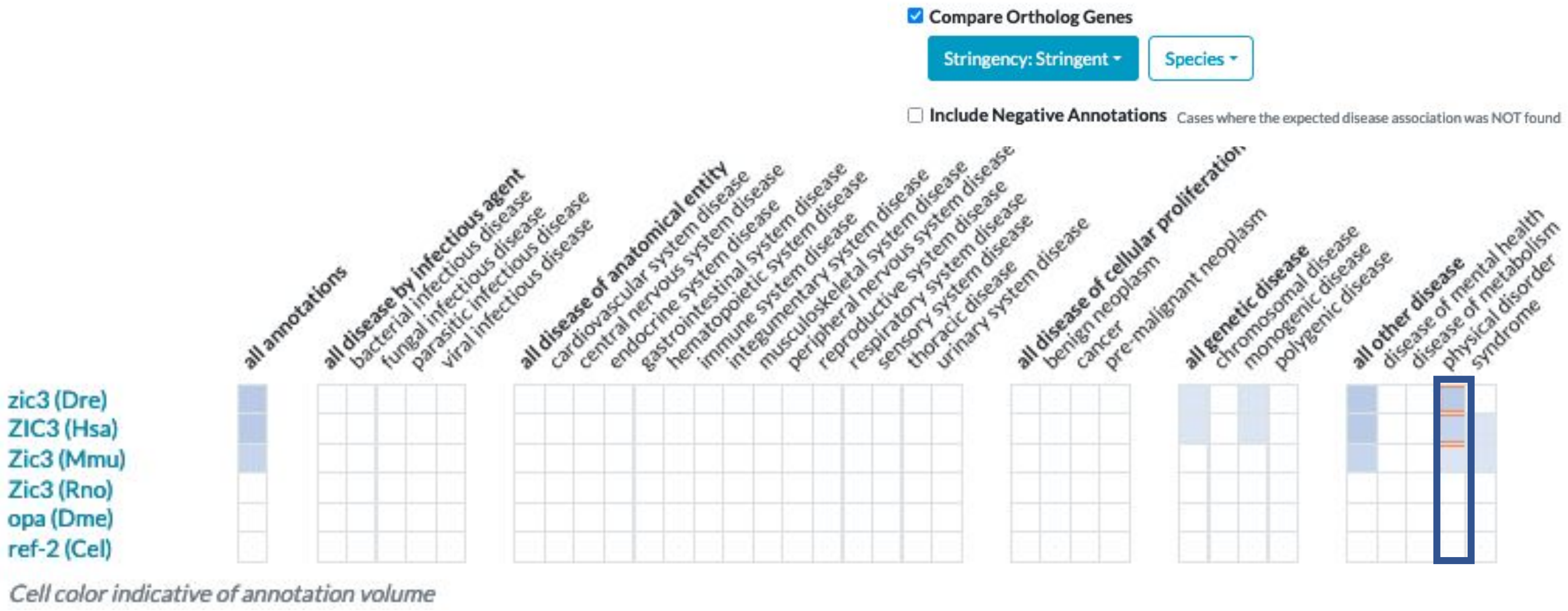
Common orthology

Orthologs for human
ZIC3

Species	Gene Symbol	Count	Best	Best Reverse	Ensembl Compara	HGNC	Hieranoid	InParanoid	OMA	OrthoFinder	OrthoInspector	PANTHER	PhylomeDB	SonicParanoid	ZFIN
Mus musculus	Zic3	10 of 10	Yes	Yes	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	
Rattus norvegicus	Zic3	10 of 10	Yes	Yes	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	
Danio rerio	zic3	10 of 10	Yes	Yes	☑		☑	☑	☑	☑	☑	☑	☑	☑	☑
Drosophila melanogaster	opa	4 of 9	Yes	No	☐		☑	☑	☐	☐	☑	☑	☐	☐	
Caenorhabditis elegans	ref-2	6 of 9	Yes	Yes	☐		☐	☑	☐	☑	☑	☑	☑	☑	

Summary of orthology algorithms

Comparative Disease Annotation Using Ribbon Annotation Summaries



Orthologs for zebrafish zic3 with disease annotations

https://www.alliancegenome.org/gene/ZFIN:ZDB-GENE-030708-2

Species	Gene	Association	Disease	Evidence	Source	Based On	References
Homo sapiens	ZIC3	is implicated in	situs inversus	IAGP	RGD		PMID:9354794
Homo sapiens	ZIC3	is implicated in	visceral heterotaxy	IAGP	OMIM *via* RGD		RGD:7240710
Mus musculus	Zic3 Annotation details	is implicated in	visceral heterotaxy	TAS	MGI		MGI:63130 PMID:1018005 Show All 6
Danio rerio	zic3 Annotation details	is implicated in	right atrial isomerism	TAS	ZFIN		PMID:30120289
Danio rerio	zic3 Annotation details	is implicated in	spina bifida	TAS	ZFIN		PMID:22285814
Danio rerio	zic3 Annotation details	is implicated in	visceral heterotaxy	TAS	ZFIN		PMID:22285814

The goals of the Alliance align with principles in the NIH Data Science Strategic Plan

- **Modernizing the Data Ecosystem**
  - **Separate data-centric and knowledge-centric activities**
  - **Develop shared modular infrastructure**
    - Efficiency (reduction in duplication of effort)
    - Knowledge commons platform
    - Cloud based
  - **Adherence to FAIR principles**
    - Data standards
    - Data integration
    - Harmonized annotation context

High quality, "computation-ready" data for comparative genomics

- Findable
  - Uniquely and persistently identifiable
- Accessible
  - Retrievable by machine or human
- Interoperable
  - Open, well-defined vocabulary
- Reusable
  - Machine process-able

Wilkinson et al. 2016, Sci Data 3:160018. doi: 10.1038/sdata.2016.18

What’s next?

- Extension of the Alliance knowledge commons to other model organisms
  - *Xenopus* sp. (Xenbase) integration underway
- Interoperation with human-centric data commons and disease specific genomics resources

Acknowledgements

Alliance Executive Steering Committee

- Carol J. Bult
- Brian Calvi
- J. Michael Cherry
- Anne Kwitek
- Chris Mungall
- Norbert Perrimon
- Paul Sternberg
- Paul Thomas
- Monte Westerfield

Alliance Scientific Advisory Board

Helen Berman, Brian Oliver, Gary Bader, Shawn Burgess, Andrew Chisholm, Phil Hieter, Calum MacRae, Alex Bateman, Titus Brown, Michelle Southard-Smith

Alliance of Genome Resources All Hands meeting (Stanford University December 2018)

NIH ***U24HG010859***

Playing telephone with data access: success with GA4GH DRS

Titus Brown (UC Davis)

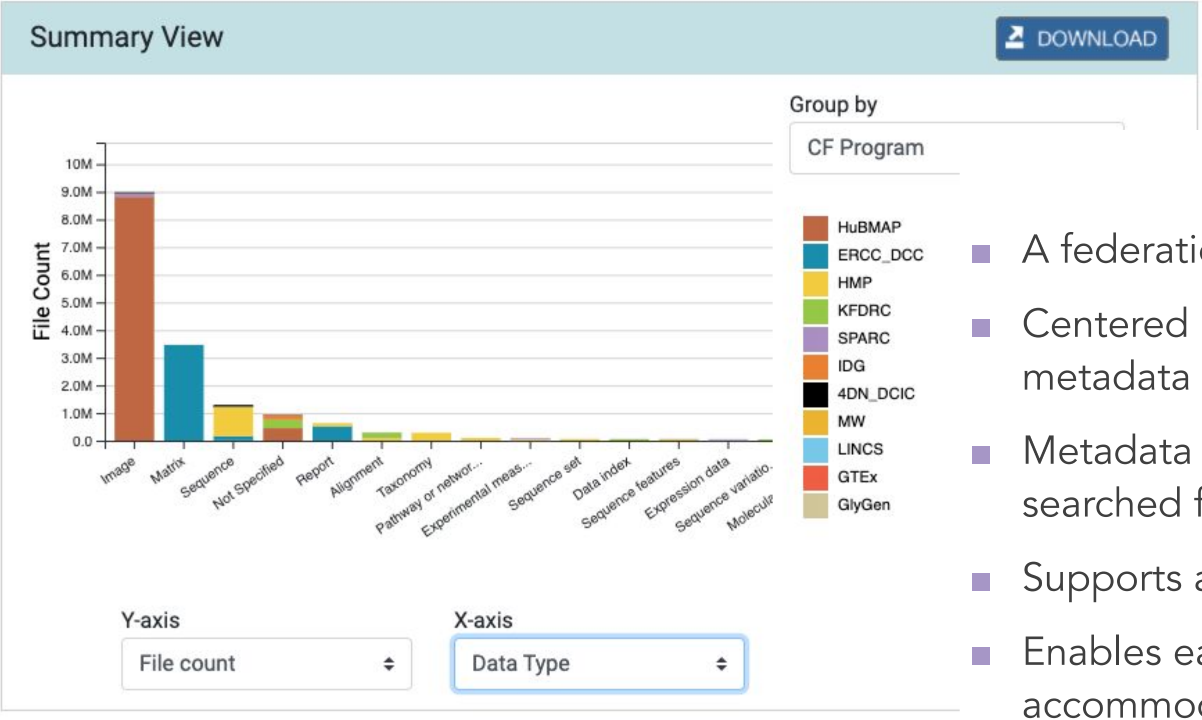
PLAYING TELEPHONE WITH DATA ACCESS: SUCCESS WITH GA4GH DRS

C. TITUS BROWN

JUNE 22, 2022

nih-cfde.org

Common Fund Data Ecosystem



- A federation system
- Centered on a catalog that ingests metadata from 10 Common Fund DCCs
- Metadata model is indexed and searched from a centralized portal
- Supports a variety of data types
- Enables easy expansion to accommodate new data types

THE CROSSCUT METADATA MODEL (C2M2)

Goal: DCCs to share structured, detailed metadata about their experimental resources across the ecosystem.

Not a warehouse

No data replication

Users directed to DCCs as primary resource

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center

GTEx Portal

NIH Human Microbiome Project

4D Nucleome

RNAseq, Variant files

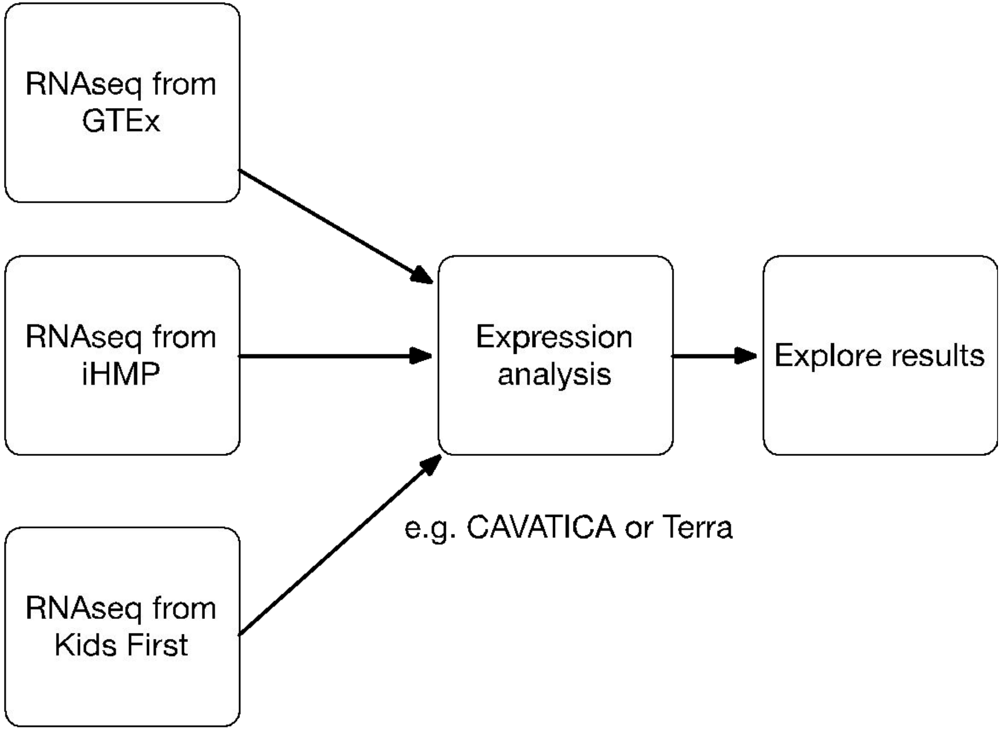
Metagenomic, Electrophysiology

Metadata ONLY

- File type
- Organism
- Assay
- Patient information

CFDE Catalog

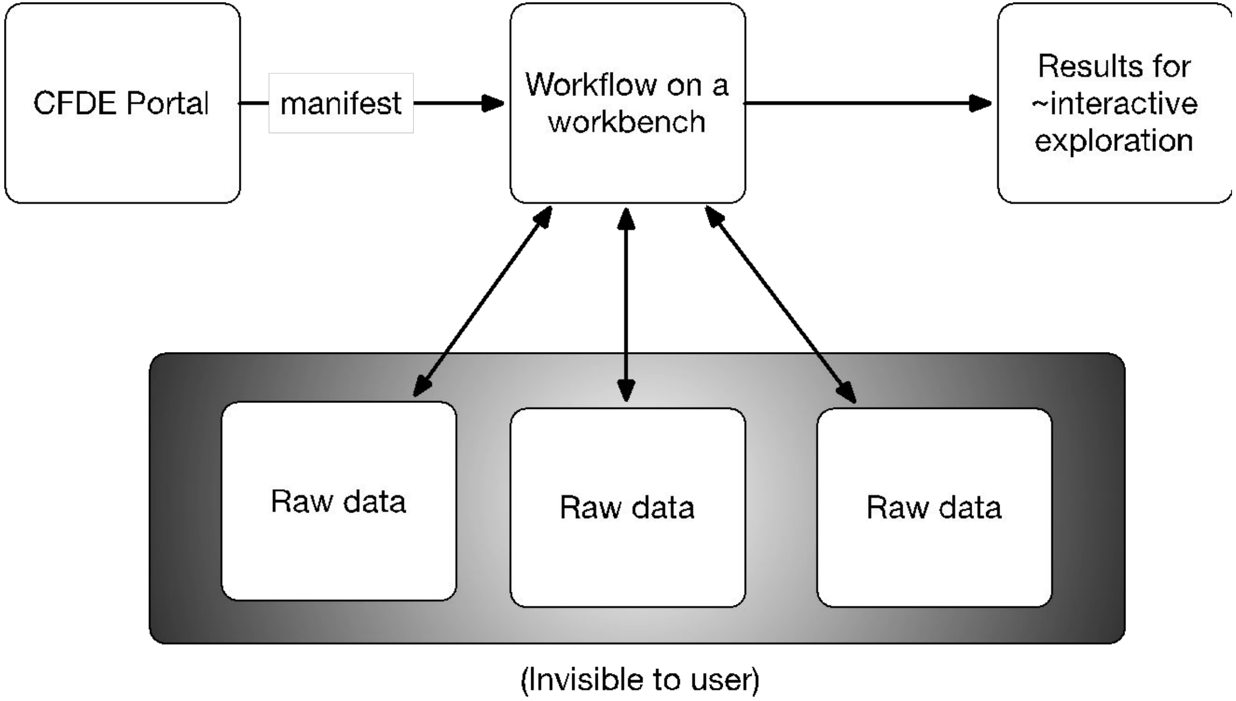
USE CASE AND PHILOSOPHY



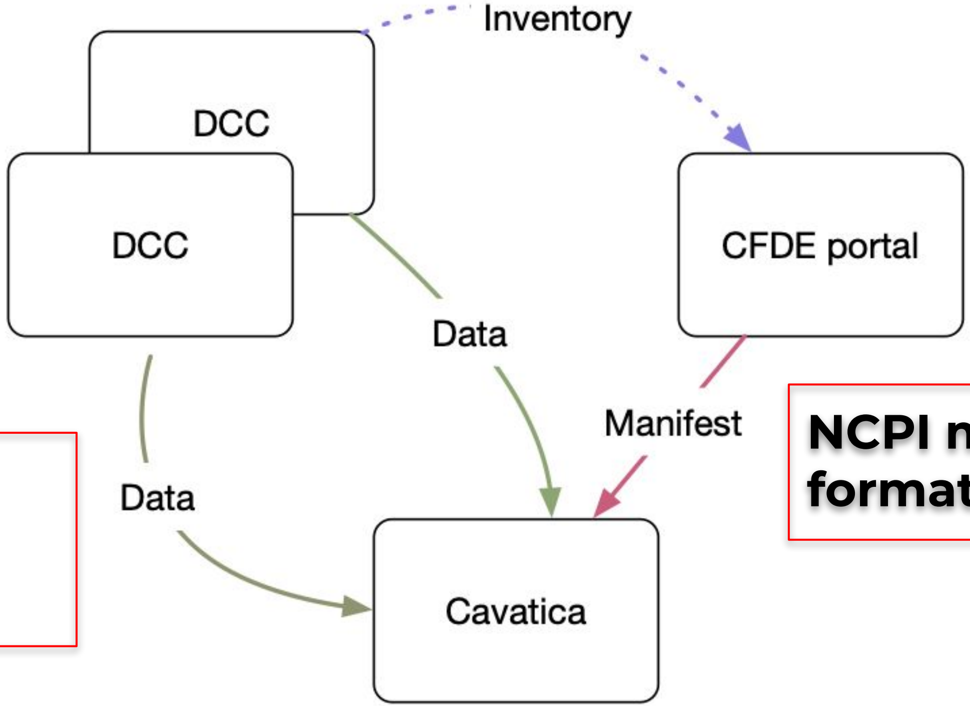
This needs to be achieved in a *standards-compliant* way so that new CF programs, new DCCs, new workflows, new workbenches, and new analyses techniques can be employed seamlessly.

If a biomedical data scientist (shell + R/Python) cannot do this, effectively *no one* can. So we start there.

WORKFLOW FROM USER PERSPECTIVE:



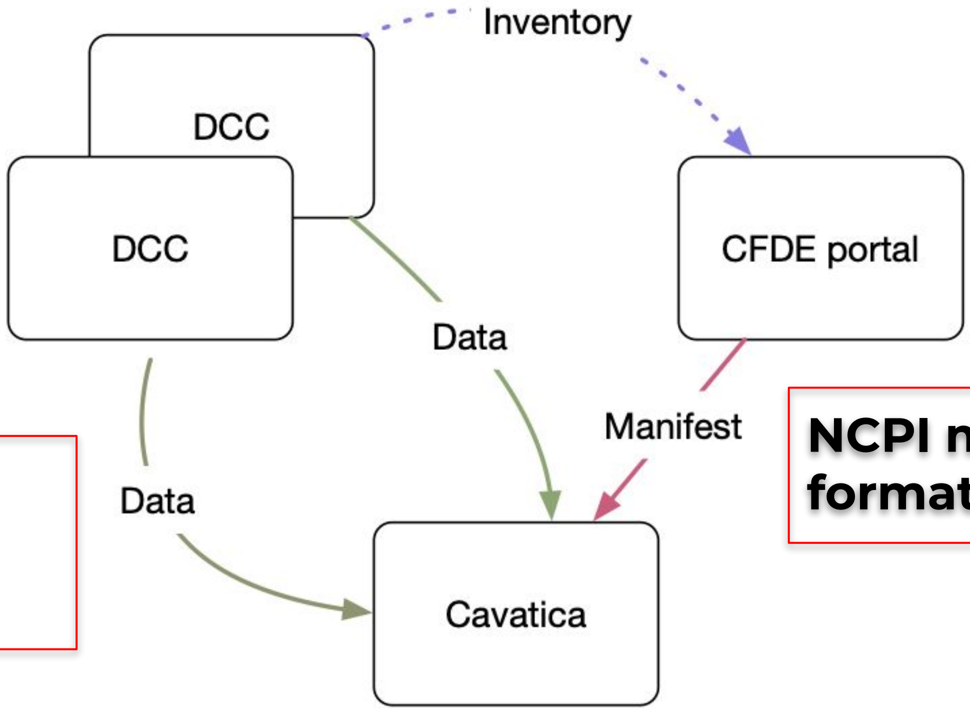
BUT WHAT'S GOING ON BEHIND THE SCENES??



DRS services (+ RAS)

NCPI manifest format

BUT WHAT'S GOING ON BEHIND THE SCENES??



DRS services (+ RAS)

NCPI manifest format

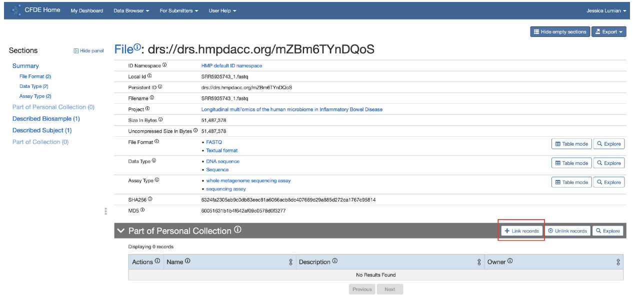
For three data sets, this involves coordination across (at least) 5 entities.

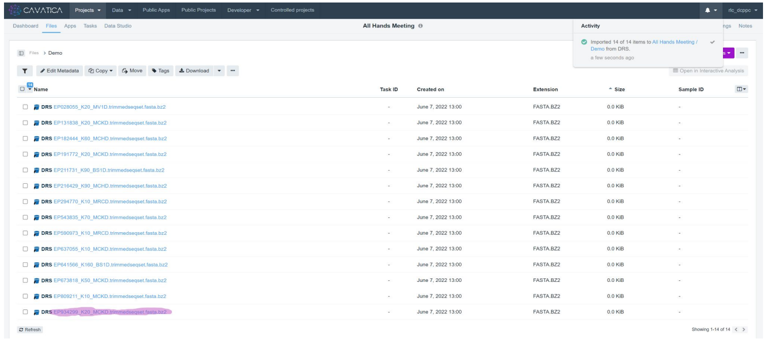
IMPORTANTLY – THIS WORKS!

Video at https://bit.ly/2022-drs-1

name	drs_uri	study_registration	study_id	participant_id	specimen_id	experimental_s
SRR5935743_1.fastq	drs://drs.hmpdacc.org/mZBm6TYnDQoS	tag:hmpdacc.org,2022-04-04:	IBDMDB	PRJNA395569_H4039	SAMN07509920	whole metagenc
SRR5950647_1.fastq	drs://drs.hmpdacc.org/DGLpSDR29kHp	tag:hmpdacc.org,2022-04-04:	IBDMDB	PRJNA395569_H4042	SAMN07532775	whole metagenc

CFDE portal → NCPI manifest format → CAVATICA workbench

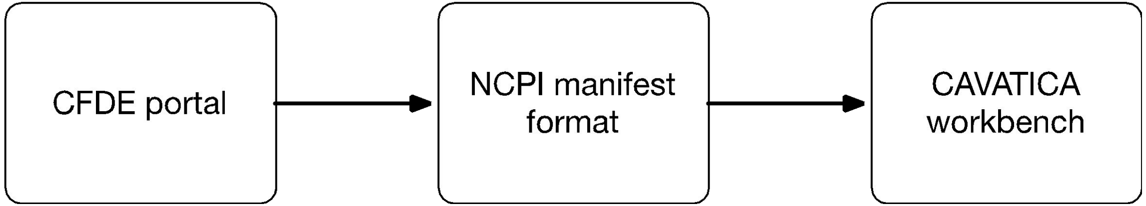


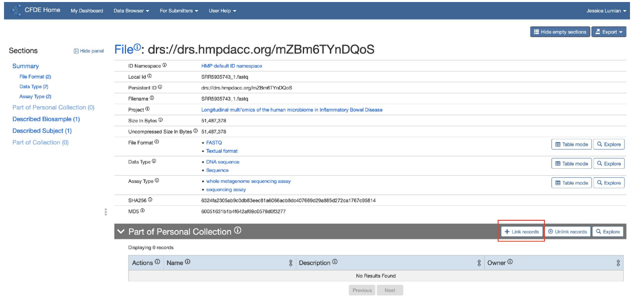


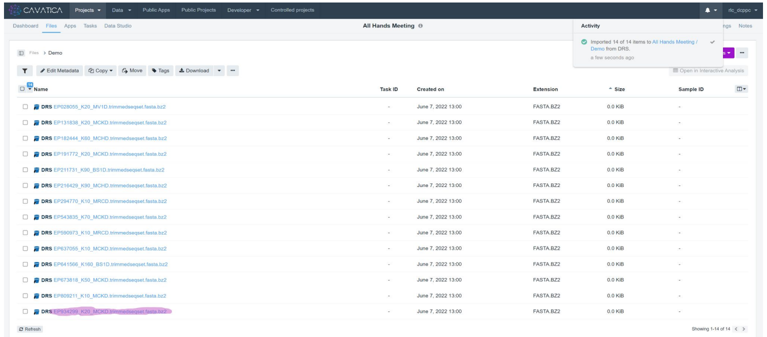
IMPORTANTLY – THIS WORKS!

Video at https://bit.ly/2022-drs-1

name	drs_uri	study_registration	study_id	participant_id	specimen_id	experimental_s
SRR5935743_1.fastq	drs://drs.hmpdacc.org/mZBm6TYnDQoS	tag:hmpdacc.org,2022-04-04:	IBDMDB	PRJNA395569_H4039	SAMN07509920	whole metagenc
SRR5950647_1.fastq	drs://drs.hmpdacc.org/DGLpSDR29kHp	tag:hmpdacc.org,2022-04-04:	IBDMDB	PRJNA395569_H4042	SAMN07532775	whole metagenc







For three data sets, this involves coordination across (at least) 5 entities.

TOWARDS A TRULY FEDERATED FUTURE… AND BEYOND?

- The GA4GH DRS standard offers a truly universal vision for dealing with many "annoying" technical details of data access – including:
  - Access to restricted data.
  - Multi-cloud hosting.
  - Multiple access methods.
  - Changing hosting locations over time.
  - Support for long-term access to sunsetted data sets (e.g. requester-pays).

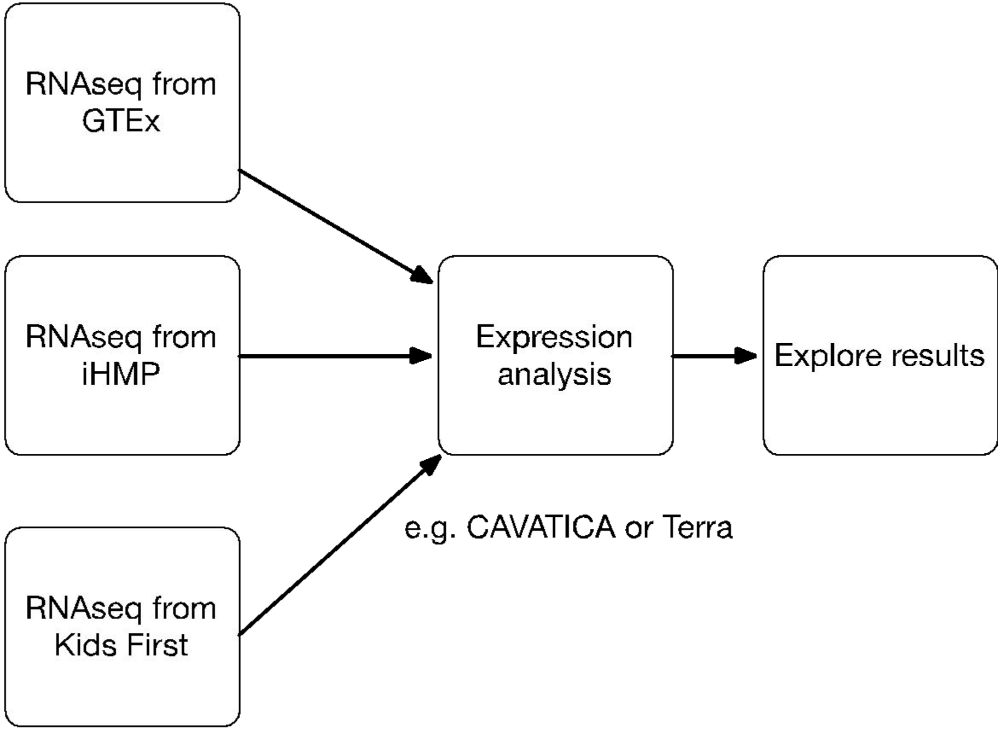
DRS and associated GA4GH protocols will support important aspects of *federating* data sets, including storage, hosting, access, and ownership.

CHALLENGES WITH DRS

Perspective: DRS is not in wide practical use, so many points of practical friction!

- Rapidly evolving standard; different platforms support different versions and specific "corners" of the standard are unsupported on various platforms.
- Very hard to test - no simple interoperability tests, no compliant command-line APIs.
- Challenges remain with requester-pays, which is important for sunsetting programs.
- Sunsetting programs must also figure out who mints DRS IDs, and who provides/maintains access to the data.
- In practice, it is very important to have "user proxies" testing all of this out!

USE CASE AND PHILOSOPHY



This needs to be achieved in a *standards-compliant* way so that new CF programs, new DCCs, new workflows, new workbenches, and new analyses techniques can be employed seamlessly.

If a biomedical data scientist (shell + R/Python) cannot do this, effectively *no one* can. So we start there.

GA4GH DRS is a fundamental building block for connecting biomedical data repositories to analysis workbenches!

Much work remains to iron out the wrinkles, but we're getting close!

THANK YOU!

You can reach me at Titus Brown, ctbrown@ucdavis.edu.

Many thanks to Amanda Charbonneau, Bob Carter, Victor Felix, Owen White, and others!

Break

Resuming at 1:15 PM EDT

Panel Discussion with Commercial Cloud Vendors

Moderator: Michael Schatz

Jer-Ming Chia - Microsoft Azure
Adrish Sannyasi - Google Cloud Platform

Break

Resuming at 2:10 PM EDT

Breakout Session

2:00 PM - 4:00 PM EDT

Breakout Session

Discussion Topics

1. Data Mesh
2. Reproducibility
3. Resource and service readiness for AI/ML
4. Engaging partnerships (i.e., GA4GH, Elixir, CFDE, Alliance of Genomic Resources)

Group 1 Report Back (Allison & Brian)

Data Mesh	Reproducibility	Resource and Service Readiness for AI/ML	Engaging Partnerships

• Definition of data mesh vs. lake
  ◦ More of a social framework
 	
• Key tech/aspects
  ◦ Mission
  ◦ Use cases and serendipitous findings
  ◦ Metrics
  ◦ Specific technical standards and API choices
	• Definition
	  ◦ FAIR
	  ◦ Data and algorithm reproducibility
	  ◦ Technical reproducibility vs. reuse
	• Incentives
	• Who wants technical reproducibility vs. reuse of algorithms
	• Researchers wanting
		• Much metadata needed
		• Training models
		• How to reduce bias
		• What is the security model for AI?
		• How are AI models shared
		  ◦ Testing
		  ◦ Checker tools
		• ML as a service
			• Current group participating in NCPI - what are the next groups to include in NIH?
			  ◦ Examples AoU a different way of looking at things that could bring diversity to the interop of NCPI
			  ◦ Other NIH projects that could join adding new data types
			• Groups outside of NIH
			  ◦ Already very closely aligned with GA4GH
			    ▪ Help with GA4GH clients
			    ▪ Help with GA4GH validation and test frameworks
			  ◦ HL7… already projects using and a working group expanding this model
			  ◦ Elixir Cloud, H3ABionet, and other organizations?

See EasyRetro and our notes document

Group 2 Report Back

Data Mesh	Reproducibility	Resource and Service Readiness for AI/ML	Engaging Partnerships
Data Mesh seems to be what NCPI is working towards Discussion of CDFE taking in metadata and using RAS/DRS to create central data catalog. How do we do plan to do this across NCPI? What is returned when a user searches a "central catalog" How can we define this? The ability to connect data in different clouds (AWS, GCS, on prem, ect.)	Do we need need to share interim data products? What services/technologies are suitable for a more processed layer? Some historical solutions discussed in dbGaP data. It seems like we need to develop a central location for metadata for searching. Is there an elastic indexing that can be used to avoid this. What is reproducibility mean at it's core? Workflow from raw data to result. Can we expect small differences in numerical values at the end ?	What does AI Ready mean? What are the computational requirements? Microsoft Research at Cambridge put out a data readiness framework ML/AI live in data science where there in intersection with computational methods and domain Do we have examples of use of SRA or other large data sets for ML/AI project Can we make use of ideas from Data Sheets and Model Cards to structure metadata	Did not discuss.

Group 3 Report Back

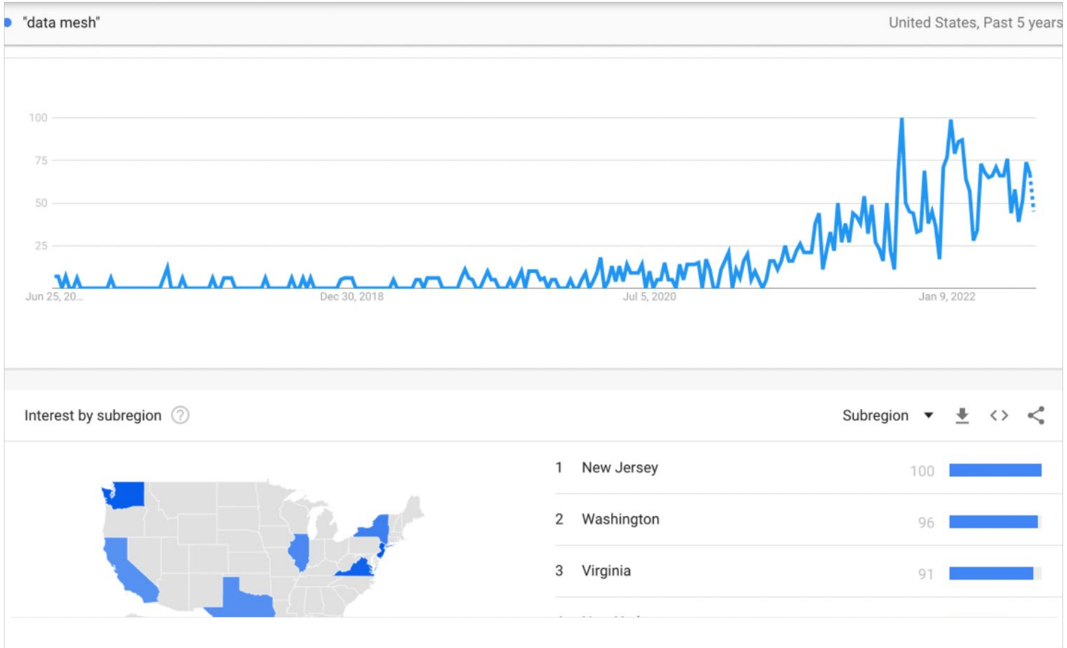
Data Mesh	Reproducibility	Resource and Service Readiness for AI/ML	Engaging Partnerships
• Technical issues are often the easy part relative to the policy, DUAs, and training • What is the value of the data, the motive force or amount of science in it? This isn't always obvious, especially with AI/ML tools, but it is important to spend our efforts reasonably. • Cloud providers provide only one level of authorization (permission to access data), not permission for resource use - passports moving more toward solving this • Important to define scientific use cases, but often these emerge from the community	• Reframe to how to improve provenance, both "downstream" (how has this dataset been used) and "upstream" (what datasets went into this). Interoperability makes it easier to propagate faulty data and results. • Development of standards for provenance (possibly with GA4GH) • What is the appropriate level of reproducibility? "Perfect" versus "good enough". • How do we address reproducibility of the human interpretation of the data?	• Validated gold standard datasets • Community rating and ranking of models and datasets • Training for the community on how to do ML / validation / ethics around ML • Important to recognize that we often do not know the most valuable data or use cases, which makes provenance and standards even more important	• Cloud providers (cost controls, billing). Can we join together to explain our collective requirements to the cloud providers so that they can make their offerings more reasonable for this community to use with reasonable effort. • Documenting requirements can also engage resellers and others who can build solutions • Model organism communities are a rich source of partnerships and data that we could be using to better understand the science of biology, and this is increasingly important in validating some of the results that are being discovered in human contexts.

Group 4 Report Back

Data Mesh	Reproducibility	Resource and Service Readiness for AI/ML	Engaging Partnerships

• Is there a common shared understanding?
• Prioritization of domain-centered ownership/tooling.
  ◦ **Domain-driven design** (DDD)
  ◦ How to implement?
  ◦ Need to define "domains"
• Do we need yet another term? Is it applicable to biomedicine in the same way it is for the enterprise setting?
• Feasibility of mandate? - Are data providers incentivised to participate in the mesh?
• Data product creation and value creation - who's value proposition are we following?
• Importance of being kind - socio-technical implementations and their incentives
• Domain-based engagement - need to develop tools for data product creation.
• Favorite new expression "extreme interoperability"
	• https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778115/ -- Association for Computing Machinery
	• Need all 3 - repeatability, replicability, reproducibility
	• What can NCPI do to help? How far do we go? Python 2.7?
	• Capturing provenance – tools and mechanism for enabling a common framework of provenance.
	• The role for documentation standards that are computable per a domain
	• Reproducible Infrastructure – serverless infrastructure?
	• Dynamic nature of reproducibility -
	• Code versions
	• Standards versions
	• Relationship to harmonization
	• Data products as containing all required information
		Requirements:
		• Data needs to be "all the same"
		• Domain knowledge intersection with the algorithmic output
		• Biology←→ Computational
		• Validation process
		• AI as commodity/product – validated models for reuse
		• NCPI may be uniquely positioned to test the requirement of broad-based data to inform model development
		• AI readiness requires upfront planning for outcomes-based research
		
		Other context of AI implementations:
		Data QC
		Algorithm selection
		Meta-data annotation
			Expand beyond US-centric view:
			• GA4GH - more on the standards side
			• Elixir - technical
			
			RWD - emerging sources:
			• Need to connect with clinical data owners – getting closer to domain experts
			• Can the healthcare enterprise itself be a partner?
			• Can tools and resources be developed to support data product generation closer to "source"
			
			Assay platform developers:
			• Illumina, PacBio, Oxford
			• Imaging
			• EHRs

Group 4- Data Mesh (additional materials)



Domain-oriented decentralized data ownership and architecture

Data as a product

Self-service data infrastructure as a platform

Federated computational governance

Group 4 - Reproducibility (additional materials)

Are there shared definitions:
http://repscience2016.research-infrastructures.eu/img/CaroleGoble-ReproScience2016v2.pdf

Repeatability *(Same team, same experimental setup): The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.*

Replicability *(Different team, same experimental setup): The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.*

Reproducibility *(Different team, different experimental setup): The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.*

Add *Reliability* . . .

Group 5 Report Back

Data Mesh	Reproducibility	Resource and Service Readiness for AI/ML	Engaging Partnerships
• Socio-technical approach • Socio: team composition, use-cases, incentives, use-cases, transparency on costs • Technical: exposing more of the hidden layer of data, simplifying and self-service tooling • Consent: DUO/DUOS as a leading example to harmonize data	• Levels of reproducibility from capturing workflows to ensuring reuse with other data sets • Difficult to be 100% byte-for-byte reproductible: external databases, random behavior in software (by design), fully capturing the software and hardware stack • Notebooks are a useful model for capturing code with all parameters and tools involved • Integration tests are most valuable to ensure systems can talk to each other	• Hardware, software, frameworks, tools, data, model zoo • Usually starts w/ sensors on sequencers (nanopore) is raw current, can collect data on a platform. Raw data, index in data, then look for patterns. Expression data to molecular mechanisms. NLP is great for EHR. ML is being injected almost everywhere. • Transformer models being built by highly resourced orgs like private companies, should we tap into those rather than the dev of our own solutions? • Can we use ML for clinical medical data - incentives and costs - if we could use AI/ML to clean up datasets would that save time and money?	• Tiers of partnership, some partners may contribute standards/tools or data or analysis; Need to lay out expectations. Defining mutual benefits is where the challenge exists • Many standards in development by GA4GH that could support NCPI efforts (e.g. variant spec) • Opportunities to leverage existing tools/workflows from different data centers that are not part of NCPI (e.g. international datasets, consent-limited cohorts) • National resources are being built around the world – crucial to ask, how to partner? Consider tiers within the data mesh that allow for variable engagement • Vanderbilt biobank, bringing that to the cloud? In the process of a cloud migration, working w/ Terra, migrating ETL from epic to omop, currently underway. Even in the cloud, right now expect only Vanderbilt investigators to have access.