

# Update: The Intersection Between Policy-Based Data Management and Emerging Health Science Data Standards

---

Deep Patel – [deep.patel@nih.gov](mailto:deep.patel@nih.gov)

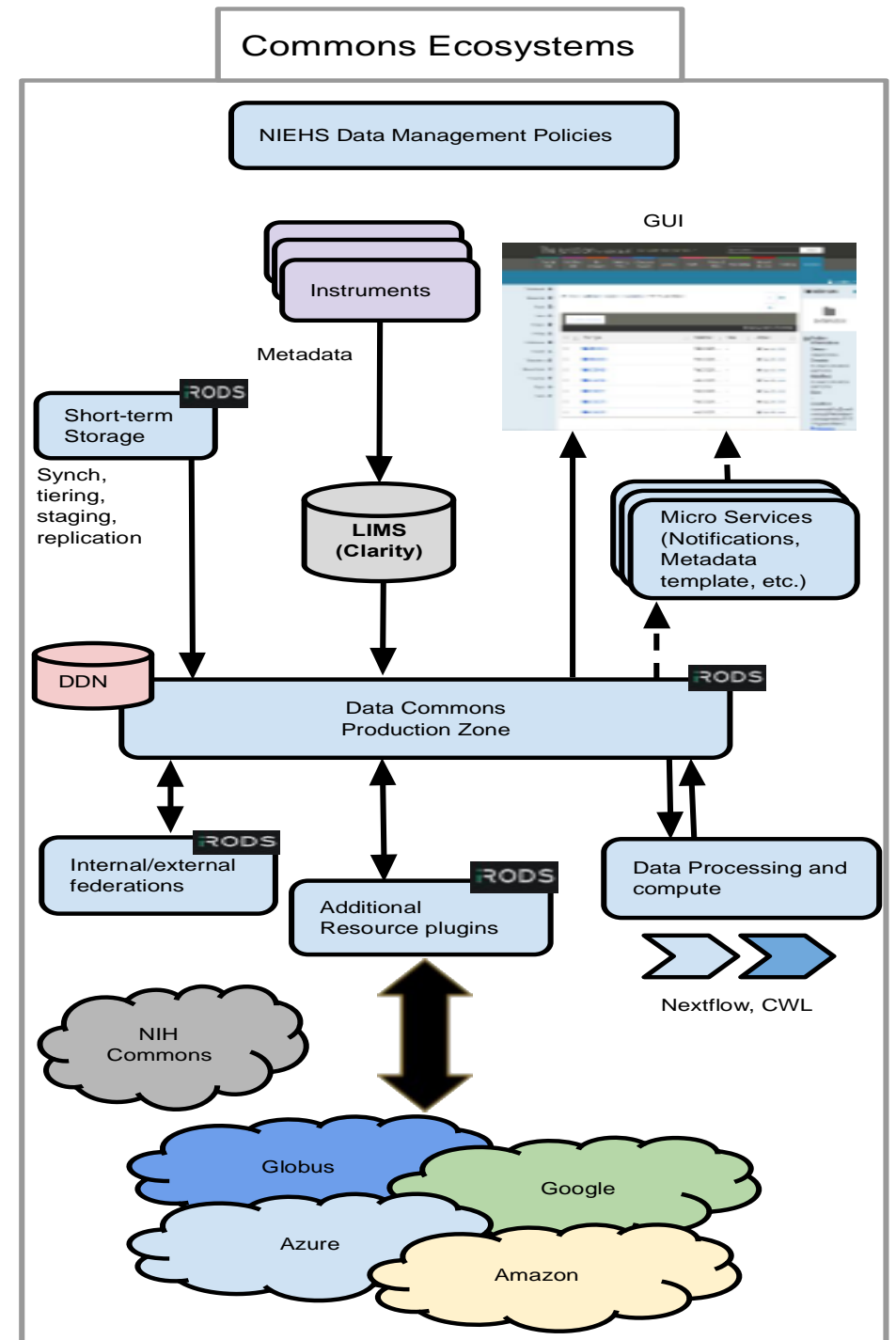
Mike Conway – [mike.conway@nih.gov](mailto:mike.conway@nih.gov)



National Institute of  
Environmental Health Sciences

# NIEHS and iRODS

- NIEHS Data Commons
  - Manage Epigenomics Sequencing Core Data
  - Workflows for pre-processing and ingest using Nextflow, integrate with Clarity LIMS
  - Data Commons as delivery mechanism gathering metadata and pipeline results



# NIEHS and iRODS

- Develop indexing and search tools/plugins for Projects and Samples
- Currently in planning to add Spatial Transcriptomics data to Commons



MiSeq: 431, NextSeq: 883 & NovaSeq: 572  
Run Total: 1,886

The screenshot shows the metaInx web interface. The top navigation bar includes the metaInx logo and a search bar with the text "Epigenomics Projects" and "Enter a search". A "Search" button is located to the right of the search bar. Below the search bar, there is a "Valid options:" section with a close button (X). This section lists two options: "1. Text Search Example => FSHD2" and "2. Include search fields for advanced search. Example => ProjectID: RNASeq\* Hypothesis: sci-RNAseq". Below these options, there is a note: "Wrap text search with double-quotes for exact search. Default operator is OR, use AND explicitly if needed." A table of available search fields is displayed below the note.

Available Fields	Description	Example
ASPNumber	Animal Study Protocol approval number	2011-0016
Background	Project background	
Hypothesis	Project hypothesis	
IRBNumber	Institutional Review Board protocol number	12-N-0095
Organism	Organism used for research study	Homo Sapiens
PIName	Principal investigator's NIEHS name	
ProjectID	Abbreviated project title used for data delivery	
ProjectNumber	Funding approval number issued by the Epigenomics core	FY18-Pilot-Granny-Smith-001
ProjectTitle	Project title	
Relevance	Project relevance	
RequesterEmail	Project submitter's email	
RequesterName	Project submitter's name	
SequencingFacility	Sequencing facility	
url	Data commons location path	
Validation	Project validation	

# Increasing Interoperability and Cloud-Readiness

- CHORDS as an example of embracing Trans-NIH platforms and use of Cloud
  - Using Gen3 Platform on AWS
  - Patient-Centered Outcome Research (PCOR) project looking at the intersection of climate change and public health
  - Collaboration with University of Chicago and the Biomedical Research Hub

The screenshot displays the CHORDS Data Explorer interface. At the top, there is a navigation bar with links for 'Submit Data', 'Dictionary', 'Catalog Query', 'Documentation', 'About', 'test', and 'Logout'. Below this, the 'CHORDS Catalog' logo is visible, along with 'Data Explorer' and 'Profile' links. The main content area is divided into three tabs: 'Geospatial Data Resources' (selected), 'Population Data Resources', and 'Geospatial Tool Resources'. A 'Filters' sidebar on the left shows search facets for 'Project Sponsor' (with options like NASA, NOAA, EPA, NPS, USFS) and 'Project Sponsor Type' (with options like Federal Agency, State Agency, Academic Institution, Data Resource, Non-profit Organization). The 'Domain' filter is also visible with options like Air Quality, Wildfire, and Climate Change. The main table displays 26 results for 'Geospatial Data Resources' and 14 for 'Domain'. The table has columns for 'Project Sponsor', 'Resource Name', 'Resource Description', and 'Domain'. Two rows are visible: one for 'National Oceanic and Atmospheric Administration (NOAA)' with 'Hazard Mapping System Fire and Smoke Product', and another for 'Environmental Protection Agency (EPA)' with 'EPA-Smoke Sense'.

Project Sponsor	Resource Name	Resource Description	Domain
National Oceanic and Atmospheric Administration (NOAA)	Hazard Mapping System Fire and Smoke Product	The Hazard Mapping System (HMS) Fire and Smoke Product is a valuable tool for monitoring and predicting the behavior of fires and smoke. The continental United States is monitored 24 hours a day, year round by polar and geostationary satellites that provide near real-time data on fire location, intensity, and extent as well as smoke analysis results based on visual classification of plumes. The system displays active fire detections and smoke information for each 24 hour period, starting with the earliest observations of the day, on a fresh map of North America. This information is essential for assessing the impacts of fires on air quality, public health, and the environment.	Air Quality, Wildfire
Environmental Protection Agency (EPA)	EPA-Smoke Sense	SmokeSense is a program designed to engage citizen participation to raise awareness about the health effects of wildfire smoke, empower individuals to take protective actions, and improve the understanding of how wildfire smoke affects communities. Through the use of a mobile application, participants report their experiences and symptoms related to wildfire smoke exposure, providing valuable data for researchers and public health officials. The program also serves as an educational tool and resource to increase awareness and encourage people to take steps to protect their health from wildfire smoke.	Air Quality, Social Determinants Of Health

# Observations



The Trans-NIH ecosystem is based on Federation



The ecosystem is heterogeneous



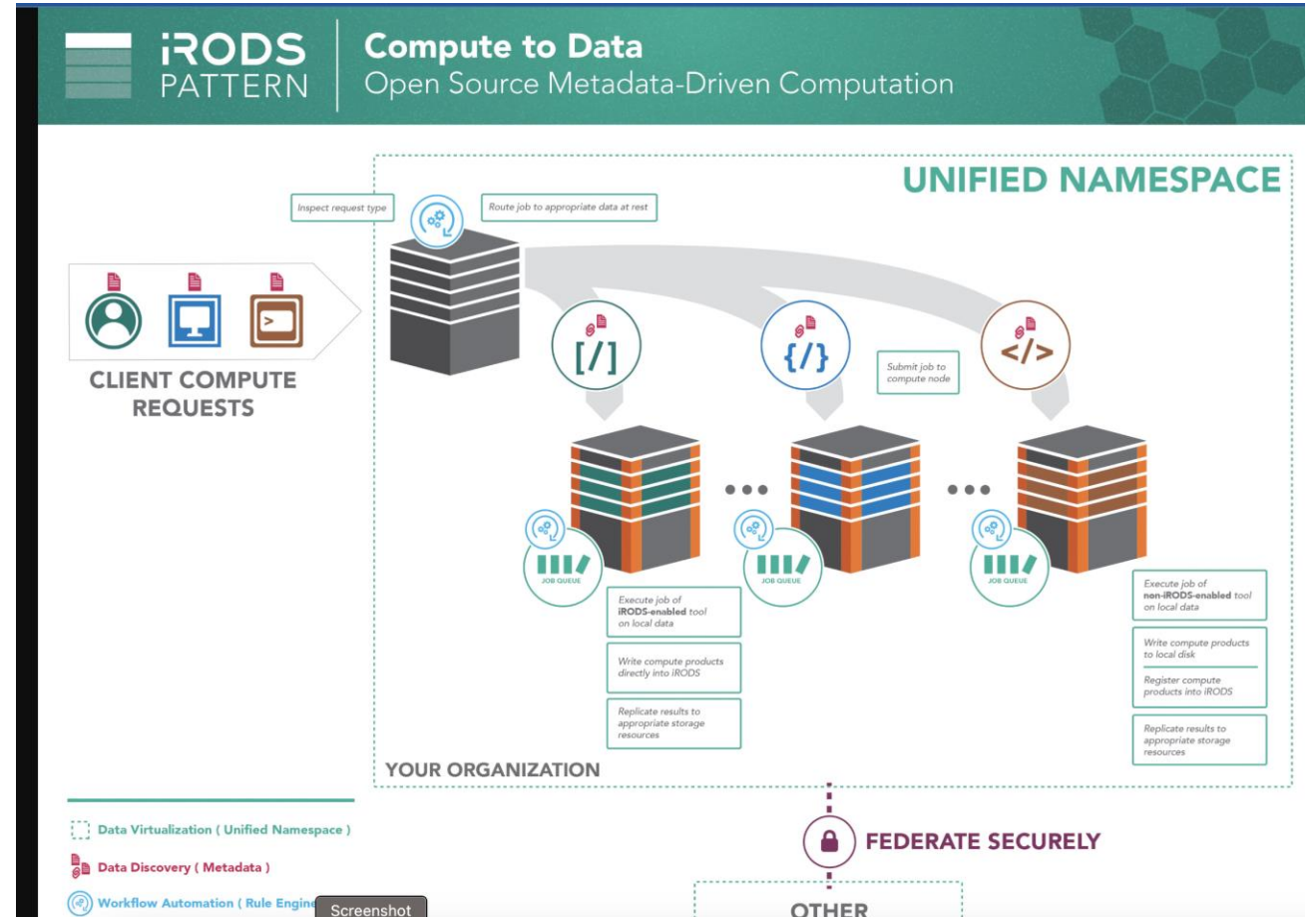
Federation is based on lighter-weight standards than the iRODS concept of Federation (e.g. in DataNet)



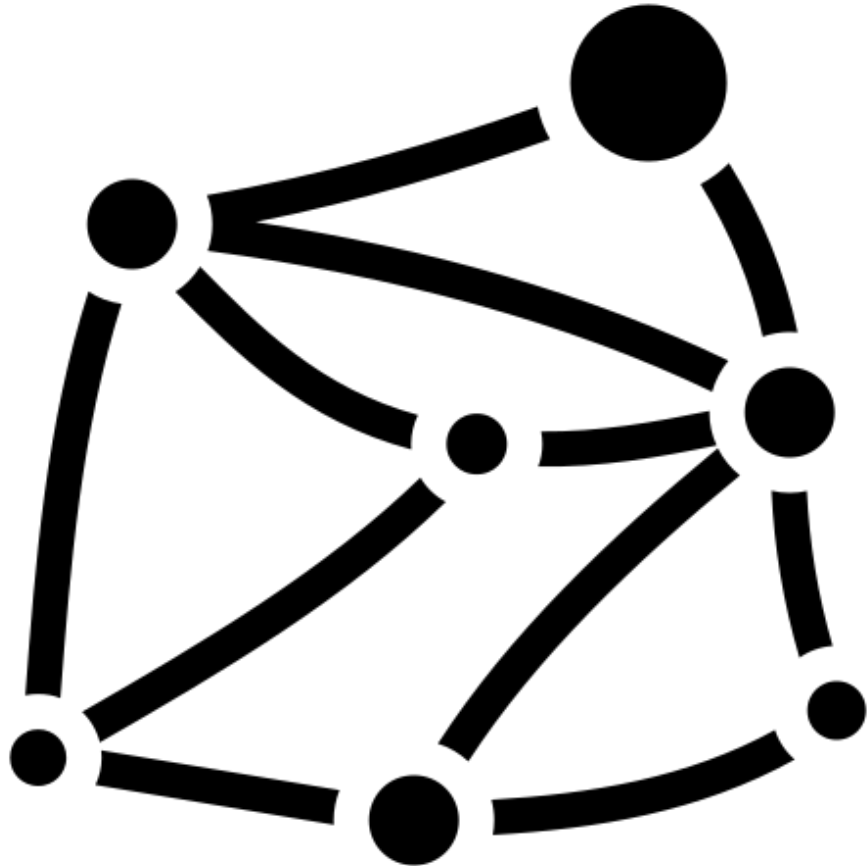
The underlying policy-based mechanisms of iRODS provide many capabilities that can enable federation in this lighter-weight federation model.

# iRODS Core Capabilities

- IRODS and the Policy-Based Data Management Approach remain highly relevant
- Hybrid On-Prem/Cloud is a sweet spot
- Potential to re-cast core capabilities via new interfaces



# iRODS in a Land of Data Meshes



What's a Data Mesh?

1. domain-oriented decentralized data ownership and architecture
2. data as a product
3. self-serve data infrastructure as a platform
4. federated computational governance.

Given multiple data repositories and data commons, how do you search for relevant data and bring it into a workspace to explore and analyze it?

# Light-Weight Federation Mechanisms



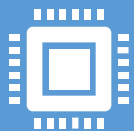
IRODS can already support these things!

Metadata Query and Federated Search

Data Access

Authentication/Authorization

Compute-to-data

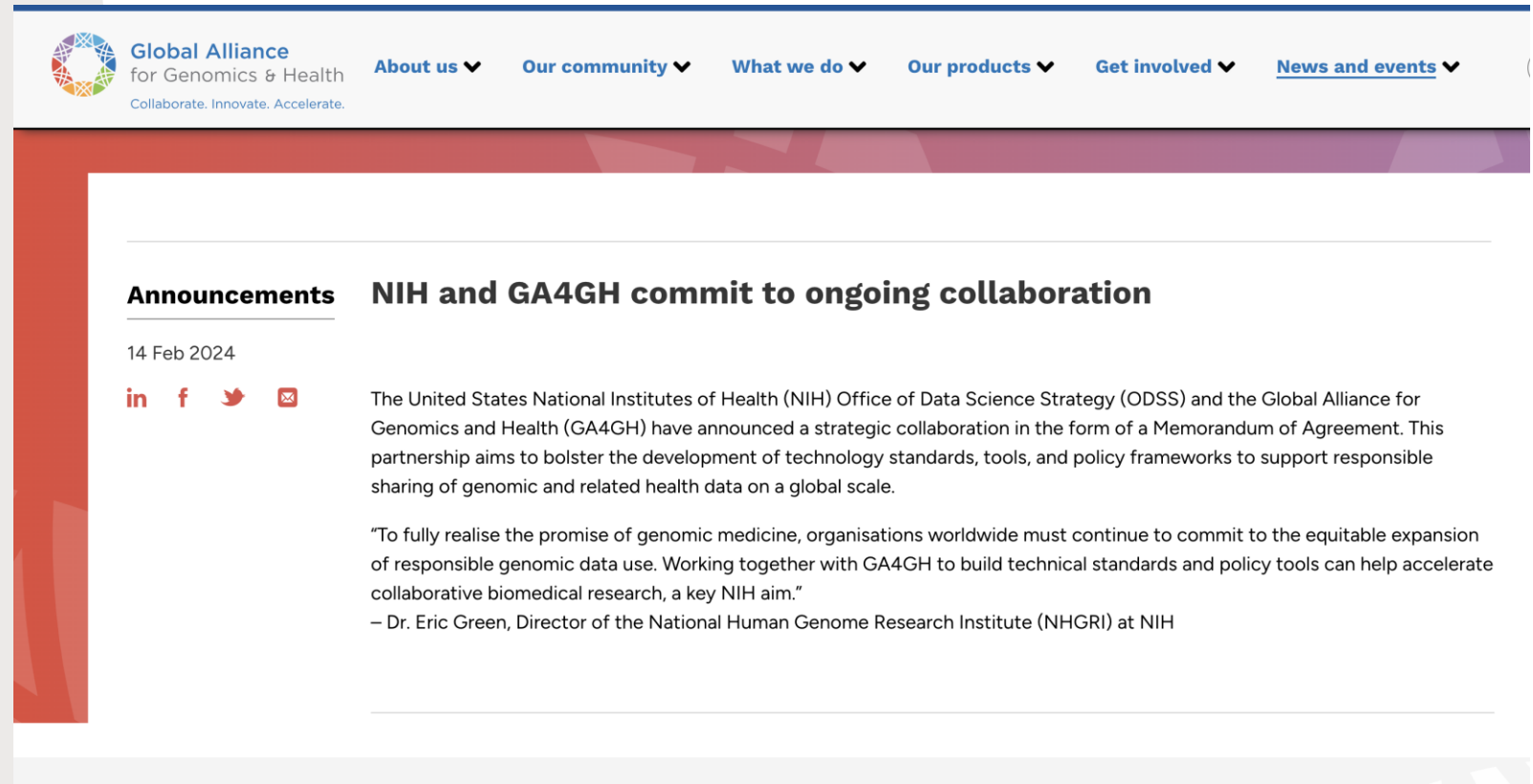


**How can we adopt language (how we talk about iRODS capabilities) and how can we expose these existing capabilities as light weight data mesh services?**



# GA4GH as a Roadmap

- Many of the Trans-NIH standards for data sharing and collaboration are reflected in GA4GH workstreams



The screenshot shows the GA4GH website header with the logo and navigation menu. The main content area features an announcement titled "NIH and GA4GH commit to ongoing collaboration" dated 14 Feb 2024. The announcement text describes a strategic collaboration between the NIH Office of Data Science Strategy (ODSS) and GA4GH to develop standards and policy frameworks for responsible data sharing. A quote from Dr. Eric Green, Director of the National Human Genome Research Institute (NHGRI) at NIH, is also included.

**Global Alliance**  
for Genomics & Health  
Collaborate. Innovate. Accelerate.

[About us](#) ▾ [Our community](#) ▾ [What we do](#) ▾ [Our products](#) ▾ [Get involved](#) ▾ [News and events](#) ▾

---

**Announcements**

## NIH and GA4GH commit to ongoing collaboration

14 Feb 2024

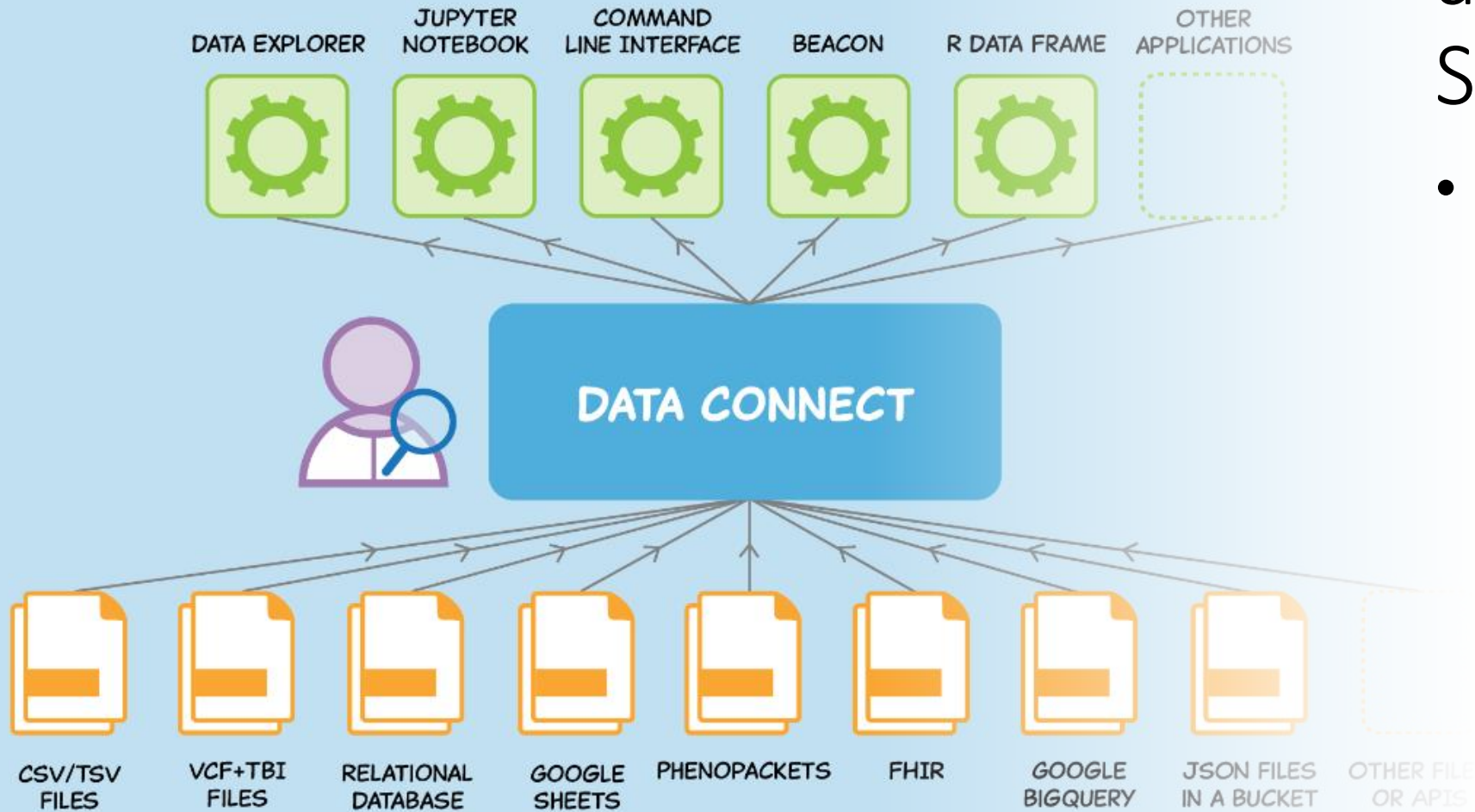
[in](#) [f](#) [t](#) [✉](#)

The United States National Institutes of Health (NIH) Office of Data Science Strategy (ODSS) and the Global Alliance for Genomics and Health (GA4GH) have announced a strategic collaboration in the form of a Memorandum of Agreement. This partnership aims to bolster the development of technology standards, tools, and policy frameworks to support responsible sharing of genomic and related health data on a global scale.

"To fully realise the promise of genomic medicine, organisations worldwide must continue to commit to the equitable expansion of responsible genomic data use. Working together with GA4GH to build technical standards and policy tools can help accelerate collaborative biomedical research, a key NIH aim."

– Dr. Eric Green, Director of the National Human Genome Research Institute (NHGRI) at NIH

## WITH DATA CONNECT...



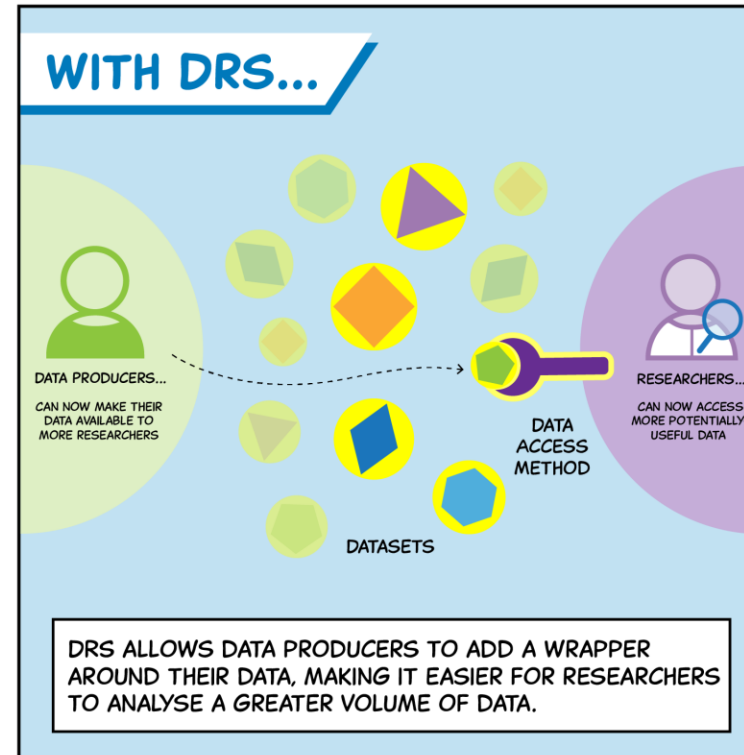
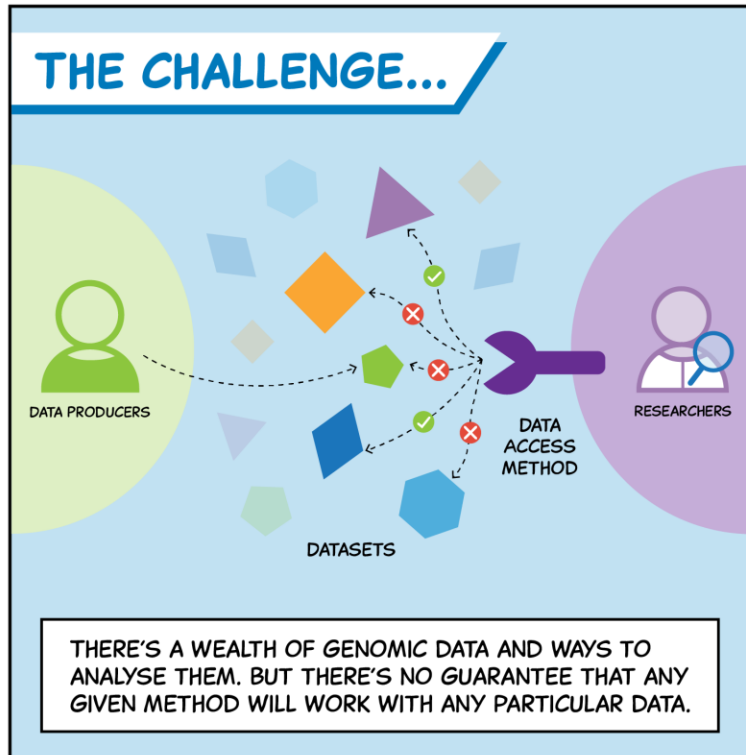
# Metadata Query and Federated Search

- GA4GH [DataConnect](https://www.ga4gh.org/product/data-connect/) Standard
  - Expose GenQuery as DataConnect
  - Expose tabular/JSON data in iRODS files as DataConnect.
  - Data access as structured data
  - Federated search and discovery

DATA CONNECT PROVIDES A SIMPLE AND FLEXIBLE MECHANISM TO CONNECT RESEARCHERS WITH INFORMATION IN VARIOUS DATASETS. FIRST, DATA PROVIDERS DESCRIBE THEIR DATA USING A COMMON MODEL. NEXT, RESEARCHERS PULL THIS INFORMATION FROM DIFFERENT SOURCES INTO THEIR APPLICATION OF CHOICE FOR DATA COMPARISON AND ANALYSIS.

<https://www.ga4gh.org/product/data-connect/>

# DRS – Data Repository Service



*The Data Repository Service (DRS) API provides a generic interface to data repositories so data consumers, including workflow systems, can access data in a single, standardized way regardless of where it's stored or how it's managed. The primary functionality of DRS is to map a logical ID to a means for physically retrieving the data represented by the ID.*

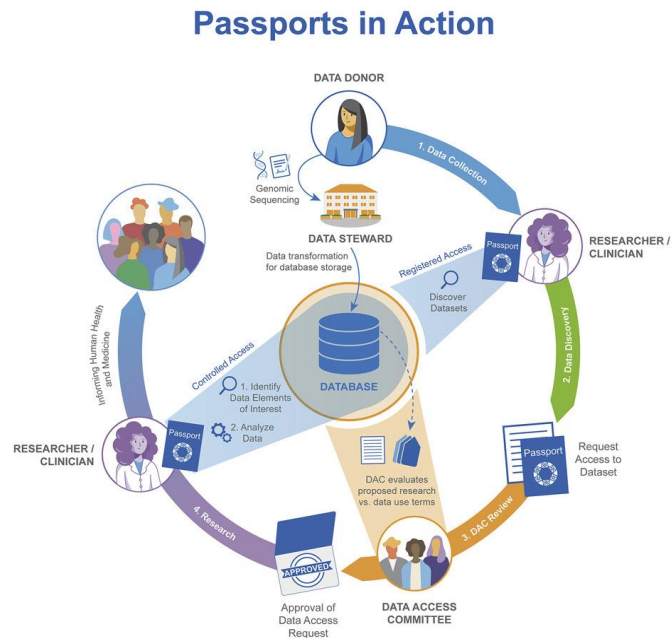
- <https://www.ga4gh.org/product/data-repository-service-drs/>



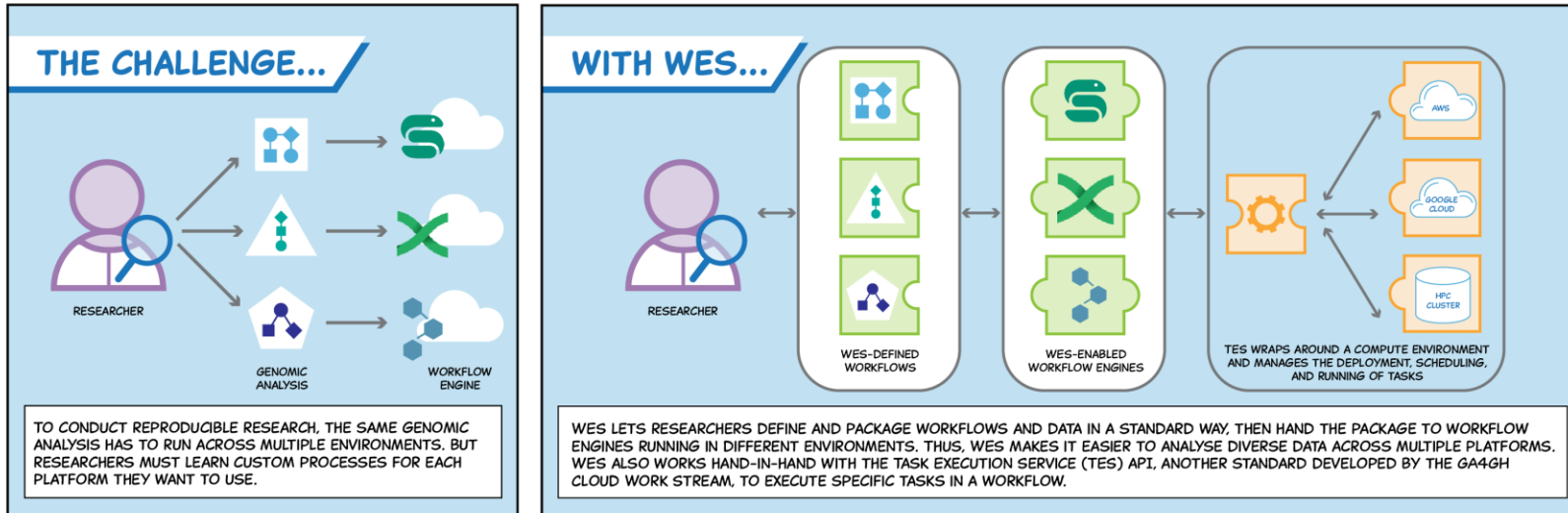
# Authentication/Authorization

## GA4GH Data Passports

*The GA4GH [Passport](#) standard defines a machine-readable digital identity, in the form of Passport visas, that communicates roles and data access rights of a data user granted by a DAC or other authority and enhances data user interoperability between data repositories*



# Federated Analysis



<https://www.ga4gh.org/product/workflow-execution-service-wes/>

## iRODS Capabilities

- Local and cloud storage abstraction
- Metadata management
- Indexing and search capability
- Policy management/Rule engine
- Compute-to-Data

Discover

Authorize

Access

Analyze

# Conclusion

iRODS is a mature system that is built with data abstraction and policy management at its core, it is unique...

BUT

We need to consider the lightweight data mesh interfaces and update the way we talk about federation, policy, and iRODS capabilities.

Thanks! Questions?

