NIH Systems Interoperation Working Group

Intro & Update

2020-03-17

Brian O'Connor
Broad Institute

NIH Systems Interoperation Working Group

NIH Systems Interoperation Working Group was created as an outcome of the NIH Workshop on Cloud-Based Platforms Interoperability held at RENCI Oct 3-4th, 2019

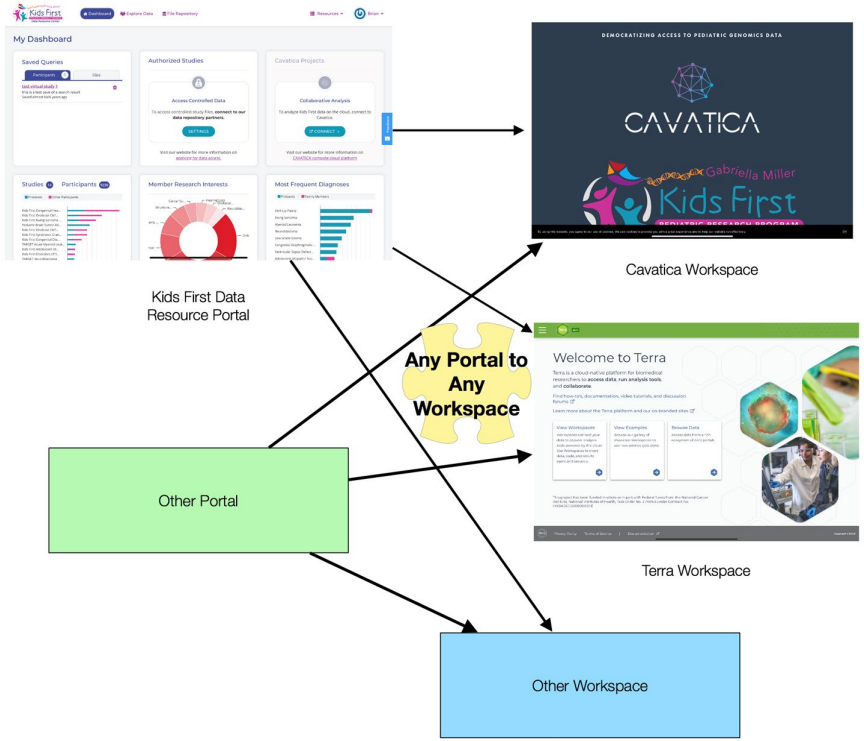
The group's Charter establishes the group's mission, members/teams, high-level scientific and technical goals, and a timeline for our work in 6 month increments.

Goal: the group will spearhead technical improvements to cloud "stacks" created by the Common Fund (Kids First Data Resource Center), NCI (CRDC), NHGRI (AnVIL), and NHLBI (BioData Catalyst) that enable improved interoperability. We will demonstrate progress based on realistic researcher use cases every 6 months.

NIH Systems Interoperation Working Group

The **NIH Systems Interoperation Working Group** goals for the first 6 months:

- Looking at interop between 4 distinct Data Portals and 3 Work Space environments (Terra, SBG, and Gen3)
- Not looking at *all* APIs/points of interoperation initially. Instead, looking for achievable improvements to interoperation between these multiple groups.



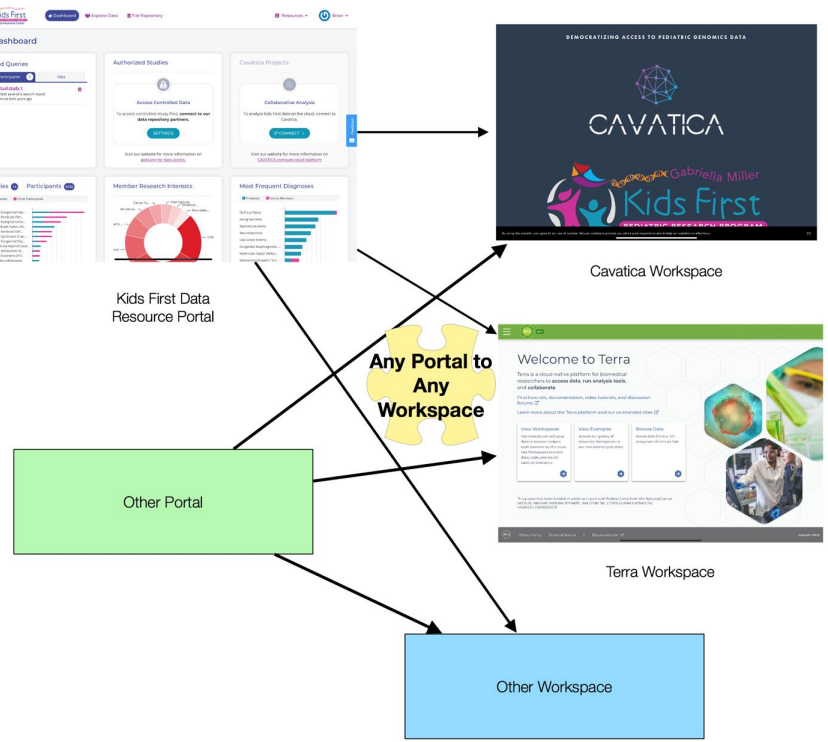
See the Technical Plan

NIH Systems Interoperation Working Group

First effort due by April is to provide a **lightweight mechanism by which Data Portals can "hand off" search results to compute environments**.

Allows researchers to "shop" for data on Kids First, AnVIL, BDCat, and CRDC and compute in Terra, SBG, and Gen3.

PFB and DRS are used to send metadata + data references to compute environments



NIH Systems Interoperation Working Group

All work is grounded in **Researcher Use Cases**

So far we have 6 high-quality use cases that span multiple groups and datasets

1a. NHLBI BioData Catalyst + Kids First DRC

Interop Contact: Allison Heath

Researchers: Bruce Gelb (Mount Sinai)
Analysis Question: The Pediatric Cardiovascular Genetics Consortium (PCGC) is committed to defining the molecular mechanisms for Congenital Heart Disease. They have developed a novel method to identify de novo mutations in clinical probands by post-processing the family genotypes posited by the GATK whole genome sequencing (WGS) pipeline. This method has a precision rate of 95% for de novo SNVs as well as short INDELs (validated by Sanger sequencing of the putative calls). Seven Bridges Genomics, Inc. (SBG) has recently described Pan-genome Graph References for improved WGS analyses and presented the use of personalized genome graphs for more consistent variant calling in family trios (ASHG 2019). This collaboration aims to develop a more accurate pipeline to detect de novo mutations in family trios by utilizing the consistent calls and other graph-related information produced by the SBG graph tools in the PCGC pipeline.

See the group Charter for our use cases

Current Status - Technical

Data Portals

Portal	Group	Export to PFB	DRS URIs/GUIDs
NHLBI Bio Data Catalyst - U. Chicago Windmill	U. Chicago	✓	✓
NHLBI Bio Data Catalyst - SBG data browser	SBG	Likely by April, no DRS	
Common Fund Kids First Data Resource Portal	CHOP	Kids First portal -> Cavatica working now, by April expect a prototype PFB to Terra handoff	
NHGRI AnVIL - U. Chicago Windmill	U. Chicago	Gen3/Windmill likely to be setup and AnVIL data onboarded with PFB handoff to Terra by April. (remains on track as of 3/13)	
NCI CRDC/GDC	U. Chicago	No DRS/PFB support by April	
Others?			

Workspaces

Work Space	Group	Import from PFB	Fence Account Linking
Terra (AnVIL, BD Catalyst, Cloud Resource)	Broad	✓	• ✓ BD Catalyst • AnVIL • ✓ CRDC • Kids First
SBG (BD Catalyst, Cloud Resource, Kids First)	SBG	For Bio Data Catalyst expect PFB but not DRS by April	• ✓ BD Catalyst • ✓ CRDC • ✓ Kids First
Gen 3 Workspace (Notebooks and "apps")	U. Chicago		
Others?			

We are tracking progress in the technical plan document

Current Status - Researcher Use Cases

On the last call we discussed the status of each researcher use case:

- 3 out of the 6 use cases showed good progress for April (1b. NHLBI BioData Catalyst + Kids First DRC, 5. NCI CRDC + NHGRI AnVIL, and 2. NHLBI BioData Catalyst + Kids First DRC), we anticipate others will show progress as well

- Challenges:
  - DRS 1.1 with GUID support a work in progress, PFB also a work in progress
  - Working around systems that don't currently support these common APIs
  - *Balance between doing the analysis once vs. any user being successful on their own*
  - Trying to understand how RAS/auth fits in, should enable nicer user experience in the future

We are tracking progress on Use Cases in the notes document

For More Information...

See the NIH Systems Interoperation Working Group docs.
Consider joining the effort!

- Charter
- Technical Plan
- 2nd and 4th Fridays at 12pm Pacific time
- See the Agenda & Notes document for details on how to join

The **April (online) Interop meeting** will afford an opportunity to assess where we stand 1) *technically* and 2) on *researcher use cases*. It will also give us an opportunity to *plan the next 6 months*.