NCPI Systems Interoperation Working Group -- Charter

This is the Charter document for the NIH Systems Interoperation Working Group, which was created as an outcome of the NIH Workshop on Cloud-Based Platforms Interoperability held at RENCI Oct 3-4th, 2019. It establishes the group's mission, members/teams, high-level scientific and technical goals, and a timeline for our work in 6 month increments.

Version	Date	Description
1.0.0	1/17/2020	Initial version, focused on establishing researcher use cases and work in progress. Approved by: ● CRDC – Tanja Davidsen ● Kids First – James Coulombe ● AnVIL – Ken Wiley and Valentina di Francesco ● BD Catalyst – Jonathan Kaltman (approved 1.0.0 on 1/21)
2.0.0	1/2022	See 2022 NCPI Sys Interop Roadmap [DRAFT]

Mission

This group will spearhead technical improvements to cloud "stacks[1]" 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.

Out of Initial Scope

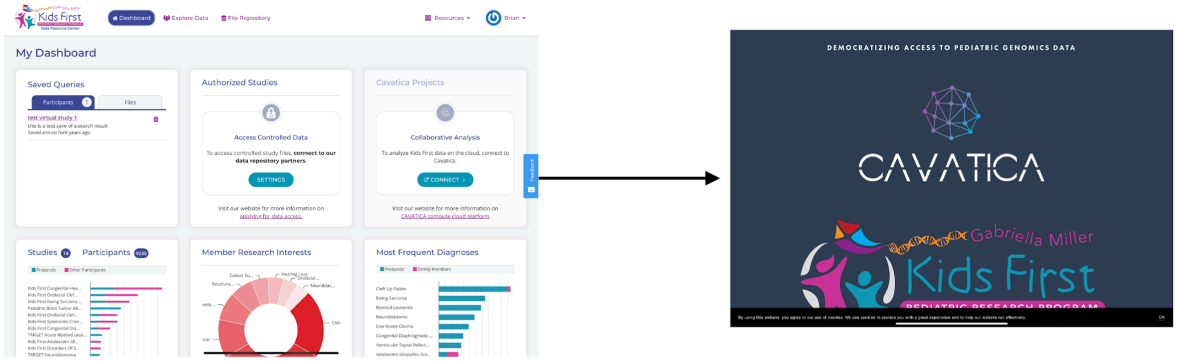
Given the time constraints and interest in showing progress in 6 months, we have purposely constrained the current issue to user functionality and interoperability improvements that are achievable and maximizes value to researchers. In the first 6 months of 2021 we are not requiring harmonization of metadata or large scale data harmonization across projects. These are worthwhile and interesting topics that are likely to be the subject of future areas of focus for this group. But we will stay focused initially on the "lowest hanging fruit" in terms of user functionality and interoperability improvements.

Background

Currently, using GTEx, Kids First, TOPMed, and TCGA cloud-based datasets together or in distinct combinations is difficult. While data portals makefinding data easier in most cases, multiple instances of authentication are often required before a user can query data and subsequently transfer search results from a given data portal to a preferred analysis workspace (cloud compute environment) is not. Currently, researchers cannot browse multiple data portals (Kids First, AnVIL, Catalyst, etc), collect their search results, and take them to a single compute environment of their choosing (e.g., Terra, Seven Bridges, Cavatica, DNAstack, Galaxy, etc). While some data portals can send search results to analysis workspaces (e.g., Kids First DRC to Cavatica for example), this is limited to specific analysis workspace + portal combinations (see **Fig 1**) while other portals do not have the capabilities to interact with analysis workspaces.

New and emerging standards (PFB, GA4GH DRS) can help to make the interface between data portals and analysis workspaces consistent, allowing many portals to send search results (e.g. lists of sample IDs and some metadata) to many different analysis workspace environments, ultimately giving researchers better access to data and increased flexibility in their analysis.

[1] A "stack" is defined as a software platform that may include one or more of the following components: data storage, analysis compute, data browser, and data ingest. A cloud-based stack allow users to access data and compute in the same cloud environment, enabling more rapid research.



Kids First Data Resource Portal

Cavatica Workspace

Figure 1: Currently portals can send search results to only specific compute environments (or no compute environments) making it much harder for researchers to use data across multiple projects in a common compute environment of their choice.

Goal

The primary goal of this activity is to establish a generic and universal handoff mechanism so Data Portal users can further analyze search results on any analysis platform that supports the format (Figure 2). This allows Data Portals to develop and maintain a single “export mechanism” which would be available to Analysis Platforms that invested in supporting the standard format. Importantly, this gives users greater freedom in how and where they compute. An additional goal for multiple use cases involved piloting a single sign on event authentication/authorization workflow through NIH’s RAS effort.

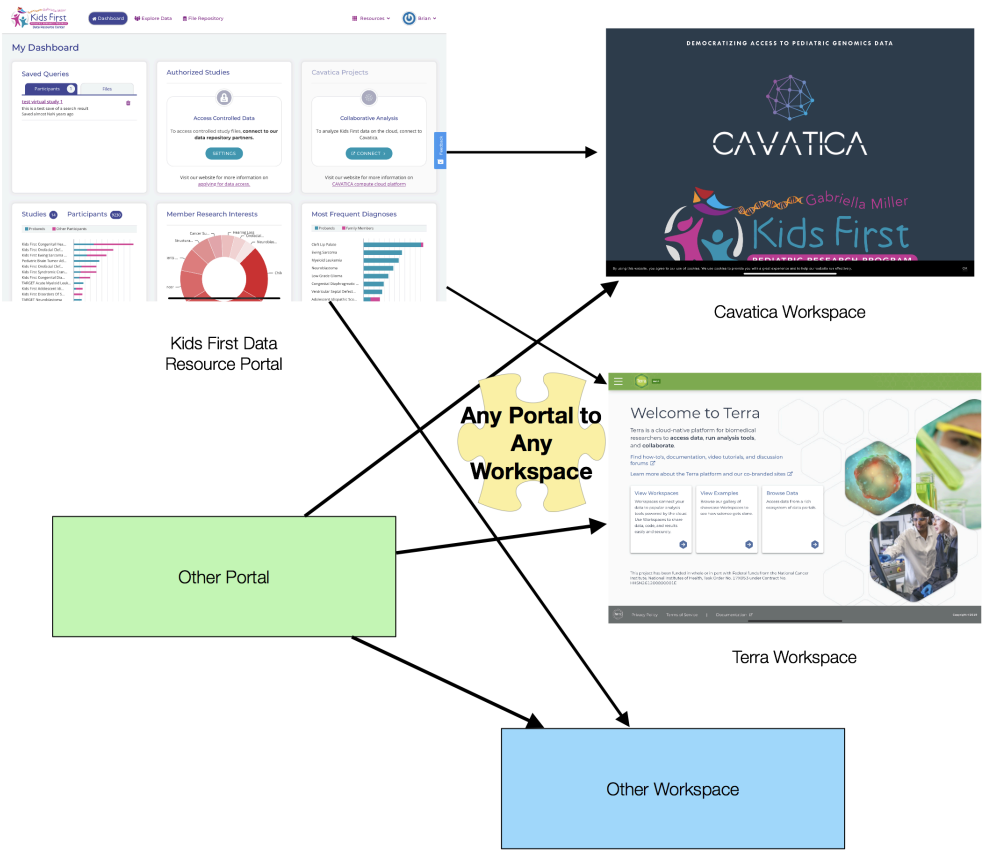


Figure 2: By adopting a common standard to hand off search results from data portals to workspace environments we can 1) give researchers flexibility in how they perform their analysis and 2) we make it possible[2] to work with datasets from different data portals in a common analysis workspace environment, especially when doing joint analysis.

Over the next 12 months, this Working Group aims primarily to improve interoperability between cloud "stacks" created by the Common Fund (e.g.,Kids First Data Resource Portal, NCI (CRDC), NHGRI (AnVIL), and NHLBI (BioData Catalyst)). The goal over the next 12 months is to expand users' (researchers) analysis capabilities by interfacing advanced search capabilities available on portals with data analysis workspace environments and efficiently managing authorization consistent with data sharing policies and data uses to provide a good user experience. By improving the handoff of search results from portals to workspace environments through standardization, **we will enable researchers to query on multiple portals and aggregate their search results to a common cloud workspace of their choosing in order to perform an analysis**. For example, this will let a researcher search for Kids First and TOPMed data on their respective portals and then take the results to the Terra environment where they can perform a joint analysis on these data. Right now, this simple scenario has limited or no support across portals and analysis workspaces, making this type of joint analysis impossible for most users.

[2] Note that data usage Policy is a related, separate issue that must be worked out between workspaces

Deliverables

1. Build the necessary infrastructure to enable the research use cases.
    a. Building on 2020 work, expand the number of portals supporting handoff
2. Document the standards and conventions used so other systems can implement the same approaches
    a. Include comparing and contrasting with other solutions like FHIR
3. Provide user-facing materials (blog post, tutorials, and/or documentation) so researchers can leverage what is built
4. Root effort in real researcher-focused use cases
    a. In 2021 expand our catalog of use cases

To facilitate this effort, each IC resource will:
1) enable users from other IC portals to search their data/tool assets
2) allow users from other IC resources to "bring" data/tools to the other resource for computation, assuming the appropriate data access approvals are in place.

Research Use Cases

We are currently collecting specific research use cases in this section that can drive our proof of concept in 6 months. The current list of use cases above are not exclusive and represent use cases that we anticipate informing us of useful interoperability problems we anticipate to see with many additional use cases. We encourage additional use cases to be added over time.

Please add use cases to this document.

Roadmap

January 2021

- Update Technical implementation plan with IC stack groups
- Charter and implementation plan update approved by IC leadership

July 2021 (6 Months)

1. Provide a mechanism[3] for **specific researchers** (see Dec) to (i) define a cohort; (ii) capture pertinent aspects (e.g. GA4GH DRS URIs/guids, metadata); (iii) import those files to workspace containing the other dataset (we will look at PFB as a potential format)
2. Researchers (2) do the science and report both the results and integration pros/cons.

[3] Note: we aren't going for style points here. This should be considered a pilot with a focus on helping the users get results

3. Identify auth user stories and requirements of RAS and other auth efforts.
4. Demo at F2F meeting **April 16-17th**. Show both the results and integration concept.

December 2021 (12 Months)

5. Refine the mechanism created in (2) to create a global standard manifest (e.g. PFB with GA4GH DRS URIs/guids etc.) and SOP (e.g. download manifests and “pull” into analysis ecosystems via a simple REST API). Approach standards groups as needed.
6. Improve existing integrations to be compatible with (5)
7. Leverage ga4gh passport for auth when/if interfacing resources that are compliant with passport (e.g. Kids First DRC - Cavatica - BDCatalyst).
8. Provide guidance to new portals and analysis ecosystems on (5); including a rough quantification of effort required to create globally usable manifests.
9. Stretch goal - outreach around this effort + results (2) inspire users who previously had not considered interop (believed it wasn’t possible). Continued demos and webinars.

Additional/Future Roadmap Ideas and Documents

This is our space for cataloging additional ideas for possible future work e.g. portals cross indexing, large-scale cross IC recomputes, etc. For ideas see:

- 201910 - NIH Cloud Platform Interop Meeting - Next Steps
- Catalyst/KFDRC User Narrative Sketch
- BDCatalyst/KFDRC Interoperability PCGC

Process

The group will **meet regularly**, likely bi-weekly, with monthly report-backs to the "NIH Interoperability Group + Workshop Organizing Committee".

The group will **present progress in 6 months in a Face to Face meeting** and use this meeting to plan next steps and future work.

Information will be stored in a **Google Drive** folder to maximize transparency: *Google Drive*

Notes and the regular meeting agendas will be made available in *NIH Systems Interoperation Working Group - Agenda & Notes*

Team

Team Co-Leads: Brian O'Connor, Jack DiGiovanna
Oversight Committee: see "NCPI Interoperability Group + Workshop Organizing Committee"

Team Members: Representatives from the Common Fund (Kids First Data Resource Portal), NCI (CRDC), NHGRI (AnVIL), and NHLBI (BioData Catalyst) have been invited based on

suggestions from NIH IC leadership. Additional key team members are invited to join through our **self-service registration form**: https://forms.gle/jFRUBMzc3X9VdyUc9

Future Ideas/Projects

Please put your ideas here!!

- PFB vs. GA4GH Discovery Table format bakeoff