

NIH System Interoperation Working Group -- Technical Plan

This is the Technical Plan document for the [NIH Systems Interoperation Working Group](#). It establishes the specific technical work needed to accomplish the goals of the team in the first 6 months of operation (leading up to a April 2020 interoperability face to face meeting). Over time, future iterations of this technical plan will incorporate additional work/efforts.

Version	Date	Description
1.0.0	1/17/2020	Initial version approved by IC leadership: <ul style="list-style-type: none">• CRDC – Tanja Davidsen• Kids First – James Coulombe• AnVIL – Ken Wiley and Valentina di Francesco• BD Catalyst – pending

Proposed Activity -- Portal to Workspace Handoff	1
Purpose	2
Proposal	2
Prior Work	4
Timeline & Work	4
Summary of Work	4
Portals Implementing Export to PFB with DRS URIs	4
Workspace Environments Implementing Import from PFB and Fence Account Linking	5
Phase 1 -- Standards and Conventions Tasks	6
Phase 2 -- Implementation	7
Data Portals Tasks	7
Workspace Environments Tasks	9
Phase 3 -- Researcher Use Case Demonstrations/Testing	9
Desired Outcome	10

Proposed Activity -- Portal to Workspace Handoff

Create and implement a standard mechanism for Data Portals to "send" search results to cloud-based Workspace Environments, enabling researchers to access datasets from multiple projects in the analysis environment of their choice.

Purpose

A **Workspace Environment** is one where users can take links to data files (genomic, clinical, phenotypic, etc) along with analytical tools and put them together, performing analysis at scale through a simple web interface. Examples include Terra, Seven Bridges Genomics, DNAnexus, DNAnexus, and Galaxy.

Data Portals allows researchers to query across data they make available to the community. Examples include the Kids First Data Resource Center Portal (<https://portal.kidsfirstdrc.org>), the Genomic Data Commons (GDC) Portal (<https://gdc.cancer.gov>), and the GTEx Portal (<https://gtexportal.org>).

Some portals, for example the Kids First Data Resource Portal, have the ability to search for data and **"hand off"** the results to an analysis workspace environment, for example Cavatica (Figure 1).

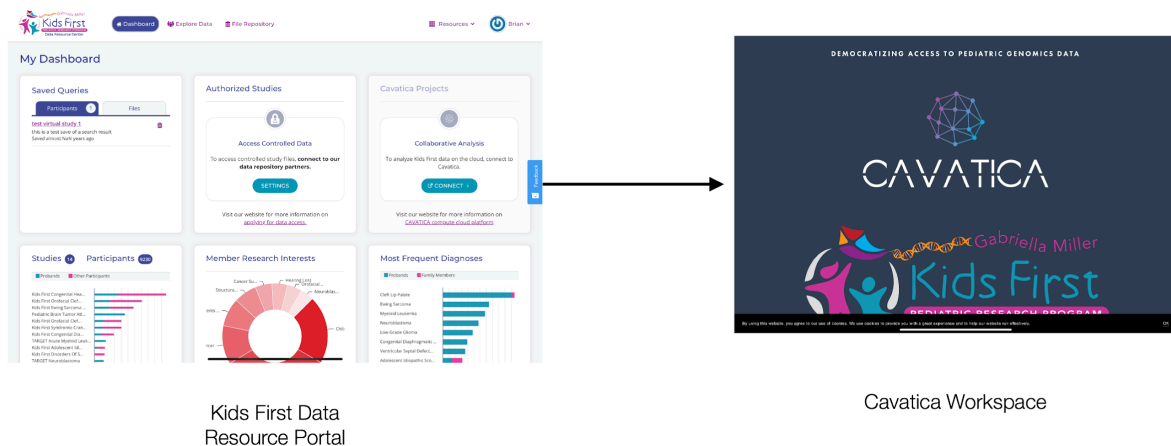


Figure 1: current handoff of search results from the KFDR Portal to Cavatica for analysis.

But current implementations are pairwise, there are *currently no standards around this functionality* so each data portal must implement a custom way to hand results to a workspace environment where researchers can compute. This is cumbersome and time consuming, resulting in researchers having fewer options for compute environments researchers can use for their work.

Proposal

The goal of this activity is to establish a generic and universal handoff mechanism so data portals can send search results to any workspace environment that supports the format (Figure 2). This would allow for data portals to quickly and easily add computational

workspace environments, giving their users much more freedom in how and where they compute.



Figure 2: With a standardized "handoff" mechanism any data portal that supports the mechanism can send search results to any compute workspace environment.

The proposal is in three phases (with some overlap expected between phases):

1. **Phase 1:** Technical standards established and documented
 - a. Develop technical guidelines for a data portal search result handoff mechanism based on the Portable Format for Bioinformatics (PFB) currently used by the BioData Catalyst Portal handoff to the Terra workspace environment.
 - b. Develop technical guidelines for using the Data Repository Service (DRS) GA4GH standard in PFB files for sharing data access URLs

- c. Develop technical guidelines for GA4GH AAI support in the DCFC (U. Chicago) used to power storage/auth for Kids First, GDC, AnVIL, and BioData Catalyst (not critical path but recommendation from U. Chicago and Kids First)
- 2. **Phase 2:** Implementation in multiple Data Portals and Workspaces Environments
 - a. Data Portals
 - i. With existing support (update as needed): AnVIL (coming soon) and BD Catalyst Windmill instances
 - ii. Needing support: Kids First DRC Portal and others (see below)
 - b. Workspace environments
 - i. With existing support (update as needed): Terra
 - ii. Needing support: SBG/Cavatica and others (see below)
- 3. **Phase 3:** researcher use cases demonstrated/tested
 - a. Data portals and workspace environment teams work with researchers to facilitate their use according to the researcher use cases we identified
 - b. Demo a handful of successful researcher use cases at the April 2020 F2F meeting

Prior Work

The needs for this generic handoff mechanism were echoed in site visits by CFDE to Common Fund DCCs, for example, the Kids First Data Resource Center expressed interest in a generic mechanism to hand off search results to workspace environments like Terra. U. Chicago has implemented the PFB format for generic handoff between their Windmill data browser and the Terra workspace environment and this could be an excellent starting point for an emerging standard.

Timeline & Work

Part 1 could be finished in 6 months following the establishment of a working group. Part 2 could be partially finished within 6 months depending on engagement from Data Portal and Workspace Environment groups. We expect that not all portals and workspaces may have this work completed by the 6 month meeting date (April 2020).

Summary of Work

Please add to the tables if I'm missing Portals or Work Spaces

Portals Implementing Export to PFB with DRS URIs

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	
<i>Others?</i>			

Workspace Environments Implementing Import from PFB and Fence Account Linking

Work Space	Group	Import from PFB	Fence Account Linking
Terra (AnVIL, BD Catalyst, Cloud Resource)	Broad	✓	<ul style="list-style-type: none"> ✓ 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	<ul style="list-style-type: none"> ✓ BD Catalyst ✓ CRDC ✓ Kids First
Gen 3 Workspace (Notebooks and "apps")	U. Chicago		
<i>Others?</i>			

Please add to the tables (or additional tables) below if we're missing major items

Phase 1 -- Standards and Conventions Tasks


Expected Due Date	Item	Point Person/ Group	Description
1/31/2020	PFB guidelines	Alessandro (U. Chicago) and Alex Baumann (Broad) and Jack (SBG) and Owen White (Common Fund, pointing to common metadata model resources)	Develop technical guidelines for a data portal search result handoff mechanism based on the Portable Format for Bioinformatics (PFB) currently used by the BioData Catalyst Portal handoff to the Terra workspace environment. This will give guidance for both data portals and workspace environments about how to receive PFB in a standardized way. Sample code needed. <i>Do we need agreement on a minimal list of PFB fields?</i> <ul style="list-style-type: none"> • How do I make a PFB? -> U. Chicago • How do I receive a PFB as a workspace environment? -> Broad (with SBG and others to help)
1/31/2020	DRS guidelines	Brian O. (UCSC/GA4GH) and Garrett/Alessandro/Bob (U. Chicago), Allison (Kids First), Alex Baumann (Broad)	Develop technical guidelines for using the Data Repository Service (DRS) GA4GH standard in PFB files for sharing data access URLs. This includes working with U. Chicago to solidify a GUID strategy and update DRS spec to include this. <i>DRS 1.1 that supports GUIDs</i> Sample code needed. <ul style="list-style-type: none"> • URI schema • Resolving process with GUIDs • AAI • Important API flow is the focus of this simple doc... multiple DRS providers

2/3/2020 -> TBD	AAI guidelines	Allison (Kids First) and Alessandro (U. Chicago) and Alex Baumann (Broad)	Develop technical guidelines for GA4GH AAI support in the DCFC (U. Chicago) used to power storage/auth for Kids First, GDC, AnVIL, and BioData Catalyst (not critical path but recommendation from U. Chicago and Kids First). <i>If we don't have this in place in time for the April F2F workspace environments will need to provide UIs for linking individual Fence instances for AnVIL, GDC, BioData Catalyst, and Kids First to a user account in SBG, Terra, and other compute environments.</i> Sample code needed. RAS workshop happening on the 27th, good time to work together. <i>What can we have done by April? How do we see GA4GH Passports and RAS related to future work?</i>
ongoing	Registry of Workspace environments/portals	Groups, Jack/Brian	We need some very, very lightweight registry (list on GitHub) that is our place to document the available Workspace Environments that support this handoff mechanism

Phase 2 -- Implementation

Data Portals Tasks

Expected Due Date	Done	Item	Point Person/ Group	Description
?		DCFS Indexd instances for all projects represented by our use cases supporting DRS 1.1 w/ GUIDs.	Garrett and Alessandro (U. Chicago)	Right now the Fence instance for BioData Catalyst supports a pre-release DRS for systems to access data bytes. This needs to be implemented across all the DCFS stacks that support AnVIL, Kids First, and CRDC. This will likely be a DRS 1.1 specification that

				supports GUIDs.
?		Also data model support needed for PFB.	Garrett/U. Chicago, Allison/Kids First	Data model needed by each project in order to generate PFB... otherwise need a way to work with flat files.
?		Kids First Portal - PFB support	Allison/Kids First , Gina	PFB handoff generation support. Mechanism for adding "send to" functionality to specific Workspace Environments
?		AnVIL (Windmill-based) - PFB support	Garrett/U. Chicago	PFB handoff generation support. Mechanism for adding "send to" functionality to specific Workspace Environments
Done		BioData Catalyst (Windmill-based) - PFB support	Alessandro/ U. Chicago, Steve Cox	PFB handoff generation support. Mechanism for adding "send to" functionality to specific Workspace Environments (see list above) -- DONE for Terra, pending for Seven Bridges
?		GDC and CRDC - PFB support	Garrett and Alessandro/ U. Chicago	PFB handoff generation support. Mechanism for adding "send to" functionality to specific Workspace Environments (see list above)
?		Proteomics Data Commons, ICDC, Clinical Trial Nodes [CRDC] - PFB support	TBD,	<p>This is what was written, is the goal to move these to the PFB convention?</p> <ul style="list-style-type: none"> • CSV for handoff generation (includes file_id and file name) • Node APIs as data access layers (DRS not implemented on these systems) • Fence is auth layer for controlled data • IndexD is pointing to all data assets

				<ul style="list-style-type: none"> • Pre-indexing from Seven Bridges to bring in key metadata. • <i>Pull mechanism</i> to bring in data from * portals to Seven Bridges
--	--	--	--	---

Workspace Environments Tasks

Expected Due Date	Done	Item	Point Person/ Group	Description
?		Seven Bridges (Cavatica, BD Catalyst, Cancer Genomics Cloud) - PFB handoff support	Jack DiGiovanna/ SBG	<ul style="list-style-type: none"> • Endpoints for user login with Fence and refresh token retrieval for Fence instances for KidsFirst, AnVIL, Catalyst, and GDC -- partial • DRS support for file access -- done • Existing portal integrations -- Kids First Portal, PDC, ICDC, windmill (in progress), i2b2 (in progress)
?		Terra (AnVIL, BD Catalyst, Fire Cloud) PFB handoff support	Alex Baumann/Broad	<ul style="list-style-type: none"> • Endpoints for user login with Fence and refresh token retrieval for Fence instances for KidsFirst, AnVIL, Catalyst, and GDC -- partial • DRS support for file access -- done • Existing portal integrations -- windmill, dockstore

Phase 3 -- Researcher Use Case Demonstrations/Testing

TODO: I think we need to spell out who's doing what testing and when. Let's try to lock this down by the end of Jan 2020 once we have Phase 1 and 2 more clearly articulated.

Desired Outcome

The desired outcome of this activity is an established standard mechanism for data portals to hand off search results to workspace environments.

If successful, this will allow researchers to accomplish two things:

1. Use the analytical workspace environment of their choosing, allowing them flexibility to use many different compute platforms
2. Collect search results across multiple DCC portals and leverage them in a single Workspace environment

This last point is incredibly important since it allows researchers to leverage data from multiple NIH ICs and projects in a common Workspace, allowing them to perform analysis across studies easily. This is something that researchers simply cannot do right now.

If successful, this also allows data portals to focus on developing and maintaining a single “export” mechanism that any workspace environment can ingest.

This is very important as many existing (in development) Data Portals do not have integrations currently included in their budgets or schedules.

The ultimate metric that we've been successful will be if the scientific use cases described in our Charter are feasible and researchers are able to accomplish their scientific goals.

We would like to demonstrate a minimum of 1-2 researcher use cases being completed by the April 16-17th 2020 F2F meeting.