NIH Cloud-Based Platforms Interoperability (NCPI) Virtual Workshop

October 5-6, 2021
11:00am - 4:00pm ET

Working Group Executive Summaries

Community Governance Working Group

Co-Chair names: Robert L. Grossman, Stan Ahalt
Executive Summary highlighting:

- Progress since April Workshop
  - Convergence on key terms and concepts to support cloud platform interoperability
  - Draft of Technical Paper on Cloud Platform Interoperability (clean version)
  - Draft of Technical Paper on Cloud Platform Interoperability (with comments)
  - Draft of White Paper on Challenges and Opportunities for Cloud Platform Interoperability currently being updated.
- Blockers / Outstanding Needs
  - Broader discussion of cross-NIH policy issues needed for broader cloud platform interoperability.
  - Broader socialization and acceptance of key terms and concepts.
- Driving Use Cases
  - See papers above
  - Improving level of NCPI cloud platform interoperability, including formalization of interoperating cloud platform tools and applications
  - Upcoming: RECOVER challenges
- Collaborative Opportunities
- Future Goals / Next Steps
  - Development of APIs to support cloud platform interoperability
  - Machine readable cloud platform interoperability policies
  - Framework for enforcement of interoperability policies

FHIR Working Group

Co-Chairs names: Robert Carroll, Allison Heath

Overview

The FHIR Working Group is focused on providing standards-compliant data and services to enable research use cases with a consistent foundation across studies and platforms. The major foci are clinical and phenotypic data, bridging to DRS, and study metadata.

Progress since April Workshop

We've made strides in three key categories: standing up services, revising and implementing the data modeling, and developing metadata standards.

- For services, the Kids First team has made their FHIR data service available in production to users. AnVIL teams have been working on implementation and security/compliance details for FHIR in Terra.
- A development NCPI server is currently hosting open access row level data from Kids First and TARGET for anyone to utilize
- We have worked with the REDCap team for the first public release of the external FHIR module that supports mapping and export of FHIR resources, which can be used to support data ingest by the platforms.
- For our data modeling, we released a beta version of the row-level data model, emphasizing rare disease and birth defect use cases with some revisions based on feedback from HL7 experts. We have implemented this model for data in KF and AnVIL.
- For the study and metadata standards, we have developed and implemented a pilot approach to more robust study and metadata representation. This model is compatible with existing data as well as transformed FHIR native data and covers common use cases for sharing and searching public resources.
- We presented a webinar workshop on the use of FHIR data, showing how to access the server and retrieve some data. With ODSS support, we also hosted a FHIR training event for NCPI developers.

Blockers / Outstanding Needs

- Security and compliance for public facing services
- RAS integration

Driving Use Cases

- We've begun to formally capture use cases in the Github
- Broadly, we are looking to provide search and data access capabilities using FHIR
- We'd like to make similar data consistently available. For a specific use case, we are looking to represent CMG data and KF data in a similar way. An up and coming use case is looking at UDN data and KF data as well.

- We'd also like to provide structured methods to link data to standards, e.g. to help with search.

Collaborative Opportunities

- Platforms who are interested in adding FHIR services as welcome to participate, new partners include ImmPort and INCLUDE
- We would like to develop tools that can operate on data from different platforms, where shared development can provide value.
- Connecting with other Working Groups to enable workflows, e.g. Portal Search.

Future Goals / Next Steps

- More data! We'll continue to represent data and studies in FHIR, which will help us to refine our processes.
- Iteration on study metadata will be crucial to ensure we can meet needs across the platforms.
- Tools for data transformations: We'll be working on helping ease some tasks like "just give me a table of data".
- We are working on a code-a-thon style event to accelerate development on a couple of use cases.

Systems Interop Working Group

Co-Chairs: Jack DiGiovanna, Brian O'Connor

Our working group spearheads technical improvements to improve interop between eco-systems. We quantify progress via researcher-driven use cases and strive to empower diverse users to broaden their research "across" platforms.

Progress since inception

We have curated over 15 use cases spanning AnVIL, CRDC, Kids First, and TOPMed data. We have worked with NCPI to evolve the use case definition to pair technical personnel and researchers; define clear outcomes; and specify budgets. Importantly, we've established the importance of funding for *researchers*, *cloud cost*, and (if needed) *developers*. We partnered with the Governance WG after our first use case; this helped create the necessary ISAs.

Our WG members were key contributors to advancing the GA4GH DRS standard from 1.0 to 1.1 and subsequent deployment of DRS 1.1 on foundational Gen3 instances. We also served as an organizing force to focus RAS milestones within these ecosystems. We deeply evaluated the PFB format; made performance improvements; and agreed on minimal common attributes.

The *data portals* and *workspaces* have made substantial changes and improvements to enable interop via the standards in the prior paragraph. The Gen3 portals have developed a "send to" button that can be received by Terra, Seven Bridges, or Gen3. Tooling has also been prototyped that makes other manifest formats ingestible in the same flow; this broadens accessible data.

Use Cases

Two cardiovascular (Taylor et al. (PCGC), Marojan et al. (PCGC, JHS, GTEx)) and one LINE1 research (McKerrow et al. (CPTAC, TCGA, GTEx)) use cases successfully completed. There are currently more than four use cases in progress and we are actively recruiting more.

Next Steps

We will partner with the Outreach WG to help drive adoption and work across the NCPI to empower more users; further quantify the impact of our solutions; and continue to refine established *data-to-compute* and ongoing *compute-to-data* capabilities in our ecosystems.

Strengths • Researchers have production-grade methods to analyze AnVIL, CRDC, Kids First, & TOPMed data on Terra and Seven Bridges. • Standards were advanced (DRS, RAS, PFB).	**Weakness** • Adoption has been limited. • Our WG was not able to perform outreach. • Uneven support of standards across NCPI.
Opportunities • Current developments and synergies with RAS will make NCBI data accessible; other "providers" (e.g. INCLUDE) have joined NCPI.	**Threats** • Misaligned funding timing creates gaps. • Evolving policy introduces significant delays. • External dependencies (e.g. RAS, PFB) can

• New ecosystems (e.g. PASC/RECOVER) likely will need *current* and *federated* approaches.	strongly affect WG roadmaps. • Diluting focus via related topics (e.g. search)

Outreach and Training Working Group

Co-Chairs: Anton Nekrutenko, Ashok Krishnamurthy

Mission of the NCPI Outreach WG

The NCPI Outreach WG is tasked with spreading the word of participating platforms, which include AnVIL, BioData Catalyst (BDC), the Cancer Research Data Commons (CRDC), Kids First Data Resource Center (KFDRC) and the National Center for Biotechnology Information (NCBI). Each NCPI platform has a dedicated set of users who are not necessarily aware of what other existing platforms do, what data they have, and what analytical tools they provide. Many data analysis resources in life sciences have existed in relative isolation for a long time. This can lead to the emergence of deeply entrenched cultures within each resource. **The main mission of the Outreach WG is to prevent the formation of silos by providing unified access to key information and training resources associated with each NCPI platform.**

NCPI Outreach WG Achievements

Major efforts have been made in building out the NCPI Portal to present content from each participating platform to reduce barriers to key information and resources.

- **Progress since April Workshop**
    - Extensive updates to NCPI portal: https://anvilproject.org/ncpi
        - Updated description and NCPI fact sheets of the participating platforms: https://anvilproject.org/ncpi/platforms
        - Highlight NCBI/dbGaP as a participating platform: https://anvilproject.org/ncpi#platforms
        - Overview of the key technologies that enable NCPI (RAS, DRS and FHIR): https://anvilproject.org/ncpi/technologies
        - A catalog of NCPI training and outreach materials, including access to user support and social media communication channels: https://anvilproject.org/ncpi/training
        - Added a Topical Search function to the NCPI portal. This searches across not only https://anvilproject.org/ncpi, but also the portals for AnVIL, BDC, CRDC and Kids First. NCBI will be added to the Topical Search shortly.
    - Data Dashboard for searching datasets across participating NCPI platforms: https://anvilproject.org/ncpi/data
    - Launched NCPI twitter account: https://twitter.com/NIHCloudInterop

        - Participated in discussions with STRIDES, CFDE, and other peer environments to coordinate plans, e.g. https://ngs-docs.github.io/2021-august-remote-computing/

- **Blockers / Outstanding Needs**
    - Need written descriptions of scientific use cases to highlight on portal
    - Need guidance on other technologies & resources to highlight

- **Driving Use Cases**
    - NCPI data dashboard enables many clinical and basic research studies, e.g. comparison of healthy gene expression from AnVIL/GTEx to cancer or other diseases studies in the other platforms
    - Other developments largely driven by NCPI-wide use cases

- **Collaborative Opportunities**
    - Participated in discussions with STRIDES, CFDE, and other peer environments to coordinate plans, e.g. https://ngs-docs.github.io/2021-august-remote-computing/

- **Future Goals / Next Steps**
    - Develop training materials, budgeted templates and other guides for managing cloud costs
    - Expand discussions of interoperability technologies, add additional pointers to developer resources