NCPI Workshop (Remote Event)
April 16, 2020
Detailed Meeting Notes

Valerie Cotton (NIH) captured the following detailed notes from our NCPI Workshop.

The “NIH Cloud-Based Platforms Interoperability” (NCPI) effort is a collaboration among NHGRI AnVIL, NHLBI BioData Catalyst, NCI Cancer Research Data Commons (CRDC), and the Common Fund-supported Kids First Data Resource Center. The goal is to enable and promote end-user analyses across these platforms through federation and interoperability. Initially convened in October 2019, the group has established 6-12 month milestones on which they demonstrated progress during an April 16, 2020 web-meeting. Each of these platforms’ initial focus is on human genomic and associated phenotypic data sharing and analyses, although their long-term scope is expected to expand to other data types and biological systems.

Over the last 6 months, the following working groups were established to “divide and conquer” on various layers of the interoperability effort; however, additional activities are still needed (e.g. data harmonization).

WG	Leads	Activities
Coordination	Valentina di Francesco (NHGRI) and Ken Wiley (NHGRI)	Facilitate WGs’ calls; facilitate the organization of the next workshops
Community Governance	Stanley Ahalt (RENCI) and Bob Grossman (UChicago)	Refine the interoperability principles whitepaper presented by Bob Grossman. Understand roadblocks to interoperability across different communities (researchers, developers, funders, etc.)
Systems Interoperation	Brian O’Connor (UCSC) and Jack DiGiovanna (Seven Bridges)	Identify research use cases for interoperability and test/implement standards (e.g. GA4GH APIs). Established “charter” and set a 12-month roadmap.

Training	Ashok Krishnamurthy (RENCI) and Mo Heydarian (JHU)	Hosted a “train your colleague” meeting for developers to become familiar with the other platforms. Consolidating platforms’ fact sheets (to develop public-facing communications materials). Document cloud costs through examples (e.g. running each platform’s key pipelines).
FHIR	Robert Carroll (Vanderbilt), Allison Heath (CHOP)	To understand how cloud based resources can leverage FHIR resources to serve the clinical data sharing needs of their communities

April 16th Zoom Meeting (slides)

Where are we now? – Anthony Philippakis

- The old way was the bring data to the researchers, which comes with security issues and computational limitations. The new way is to bring researchers to the data to improve accessibility and accelerate research (and potential opportunities for audit trails etc.).
- One monolithic entity to handle all genomic and clinical data is not realistic nor desirable, but we want to empower researchers to cross-analyze diverse datasets and data types, **which requires building systems in a “federated and interoperable way**”.
    - NCI was one of the first to do this by funding multiple different “cloud resources”, using common pipeline and data models, that are part of the “Cancer Research Data Commons” led by the University of Chicago.
- The four NCPI platforms have agreed on a cloud-based infrastructure “parts list” that will make an interoperable data ecosystem possible: Repositories (data storage), [data portals/browsers (to search/aggregate)], workspaces (for access/analysis), analytical workflows and tools
    - Adoption of GA4GH standards will help with next steps
- Need:
    - Ability to pull data and tools from various repositories and pull them into the different workspaces.
    - Data “aggregators” to search across various repositories.
    - Single sign-on authentication (verify user identity; “passport”) and authorization (claim to access specific studies/datasets; “visa”)
        - **NIH “RAS” Researcher Auth Services (RAS)** effort ongoing in parallel, led by Rebecca Rosen (NIMH). Federated authentication (and account linking), protocol translation services, federated authorization to simplify controlled access to more data for more researchers. Integrating GA4GH Passport standard.
        - GA4GH is developing “Data Use Ontology” standard (DUO; common data elements describing consent-based data use limitations).

    - Stronger and more consistent phenotypic data collection and harmonization
        - Use of common data models for metadata and phenotypic data (FHIR?)
    - Agreed upon governance models for data and metadata.
    - Continue testing and implementing data exchange standards (e.g. DRS for Data, TRS for Tools, WES for analytics)
- Success looks like the creation of an open and federated data ecosystem. Failure looks like a collection of data silos.

Working Group Updates

Community / Governance – Bob Grossman

- Distinct from “technical standards”, this WG has drafted a set of “Operating Guidelines” for exchanging data across platforms that have formed “trust relationships”. The intention is for NIH and the community to agree to the document as a framework for next steps into data federation.
    - The current draft (Version C) can be found here: https://docs.google.com/document/d/1oGb41XjDIq5fCGGa1FbUSjKEGb_BLleirTYObFQqjkQ/edit?usp=sharing
    - Briefly the principles are:
        - **1. Form trust relationships with other platforms**. “Trust relationships” are based on agreements about security, compliance, and liability that provides the foundation for two or more platforms to interoperate.
        - **2. Follow the golden rule of data resources:** if you take someone else’s data, let them have access to your data. Importantly, this principle calls out that any restrictions on data access or use (including the ability for data to flow across platforms) should be based on transparent and clearly documented policies. For example, limitations may be imposed by patient consent or data security matters, but not investigator or consortia preferences.
        - **3. Support the principle of least restrictive access:** Provide another trusted platform access to your data in the least restrictive manner possible (e.g. expose APIs).
        - **4. Agree on standards, compete on implementations.** Try to keep your system open so applications can compete to provide the best *experience* for researchers.
        - **5. Plan to support patient partnered research:** Support patient partnered research so that individuals can provide their data and have control over it within your system. If you cannot do this today, add this to your platform roadmap.
    - Next steps including developing metrics to evaluate adherence to the principles.
    - These principles should/may also be adopted by emerging COVID-19 research platforms.

Outreach/Training WG – Mo Heydarian

- This group hosted a “train your colleague” session featuring technical presentations on the software tools and infrastructure used by the platforms; including Gen3, Terra, Dockstore, PIC-SURE, ISB-CGC, and Seven Bridges. This knowledge serves as a reference base for identifying paths of interoperability.
- Intend to build and display a public knowledge base of training information, a cloud cost guide, and reference information about all the platforms, initially on the AnVIL portal and eventually transition to an NCPI stand-alone site.

FHIR Working Group – Robert Carroll and Allison Heath

- Background
    - **Problems** Facing the FHIR research community
        - Good progress on EHR interoperability using FHIR (use of release 4 mandated by ONC for EHR vendors) but it is still a serious challenge (a lot of detail is hard to represent in FHIR)
        - Research world, including NCPI, has similar interoperability challenges
        - FHIR offers a lot of promise, but effective tooling, standard model dev and processes not in place
    - **Goals** To help prepare us for solving the problem(s)
        - Take a hands-on approach to learning and prototyping with FHIR through the WG kickoff project (“Project Forge”) for current NCPI participants
        - Learn how to effectively collaborate on FHIR modeling. Example: GA4GH group is trying to model Phenopackets in FHIR.
        - Gain a shared understanding of the problems FHIR solves and its current weaknesses
    - **Objectives**
        - Take a practical approach to learning and prototyping with FHIR
        - Provide feedback to the HL7 FHIR and NIH communities – we have a lot of data to inform standards.
        - Create a roadmap for clinical data interoperability among datasets and platforms
        - Build community and engage the appropriate stakeholders for the longer term.
- Roadmap
    - **Progress** Since October Workshop
        - Group has officially been formed and met
        - Kids First DRC team has built a preliminary FHIR modeling toolchain (see slides)
        - Organization of pilot to support interoperability
    - **Roadmap** (“Project Forge” - 3 months)
        - KFDRC will organize initial infrastructure and dev process; look at structure and FHIR servers, how it is being used today
        - Groups will identify key datasets and familiarize themselves with those and FHIR modeling.
        - Collaborative and iterative process to create a baseline interoperable FHIR for research model

        - Along the way - discuss naming conventions, model release process (i.e. versioning, distribution)
        - Load 2 key datasets into the FHIR test server, review data with the FHIR Data Dashboard.
    - **Challenges**
        - Collaborative FHIR modeling with a fully remote, multi-organizational team
        - Security framework and data sharing among IC-sponsored platforms. FHIR is patient-level and some of the default servers don't necessarily support the access model we are interested in.
        - Long term FHIR server licensing, deployment, and operation
    - **Opportunities**: A lot of misconceptions about FHIR, but robust framework and in-client applications offer a powerful back-end for data modeling and we can iterate on this. Opportunities to converge on this like the early days of genomics. Hospital systems are already moving over to this and there is a huge opportunity to connect but research needs are different.
    - **Potential User Story**: Start with something syndromic as a use case (e.g. Down syndrome intersection with heart disease), different phenotypes and want a context of connecting across studies. Initial data dashboard to navigate and self-describe patient cohort of interest > initial set of lightweight explorer services > navigate to more services > exchanging data is a core of FHIR – interchange and accept data from other FHIR servers even if they are different data models. Sometimes you will want highly curated data but other times you want raw data so you can start working with it. The WG will ramp up to understand the back-end and what is solved here, and from there we can engage with other groups in this space.

NIH RAS (Researcher Auth Service) – Rebecca Rosen

- **VISION**: Develop a unified, efficient, and secure authentication and authorization service that enables streamlined access by researchers to NIH-funded data resources across multiple systems. Provides standardized methods of logging and auditing such access and is compliant with NIST and GA4GH standards. This is an NIH owned and operated service meant to provide streamlined access to NIH datasets to facilitate interoperability. Building enterprise-level services and will build out additional services over time. Collaboration between CIT and ODSS, with NCBI/dbGaP engagement.
    - Community discussion on whether passport authorization might be sufficient to access some data, since passport links to eRA where you can see institutional official has vouched for you.
- **PROGRESS**: CIT is pulling identities from eRA and NIH active directory. V1.0 OIDC AuthN/Z endpoints deployed in testing environment for Phase 1 partners. KFDRC/BDCatalyst and CRDC/AnVIL are able to access CIT testing environment and access tokens for user endpoints. Implemented basic auditing and logging of data and metrics. Internal system monitoring and notifications [https://auth.nih.gov/docs/RAS/serviceofferings.html]
- **Roadmap** (2020) –

    - Summer production deploy of SSO and AuthZ endpoints for Phase 1 interoperability use cases (eRA accounts and dbGaP claims).
    - Use case / requirements gathering and planning workshop for Phase 2 data repository integrations: NIMH Data Archive (NDA), dbGaP/NCBI, Common Fund Data Ecosystem, and All of Us
    - End of year production deploy to support Phase 2 integration use cases and account linking
- **Challenges**
    - Building a foundation system in 6 months is a risk
    - Remotely aligning development timelines with partner systems who have overlapping projects and priorities
    - Facilitating long-running analytic pipelines; IM service is FISMA high (e.g. end-points are configured to get a refresh token without user interaction every 15 days – some pipelines might need longer?)

Systems Interoperation WG - Brian O'Connor / Jack DiGiovanna

- The group's *Charter* establishes the group's mission, members/teams, high-level scientific and technical goals, and timeline.
- Goals (first 6 months): *See the Technical Plan*
    - Analyze & develop interop between 4 distinct *Data Portals* and 3 *Workspace* environments (Terra, SB, and Gen3)
        - Initial technical effort focused on a **lightweight mechanism** by which **Data Portals can "hand off" search results to compute environments**. Enabling researchers to (i) define a cohort; (ii) capture pertinent aspects (e.g. files, IDs, metadata); (iii) import those files to workspace containing the other datasets.
        - Initial candidates to send data references and metadata to compute environments:
            - **GA4GH Data Repository Service (DRS) API** to access data in a standard way regardless of where it is stored or managed.
                - The group discussed using DRS in workspaces as well as repositories
            - **Portable Format for Bioinformatics (PFB)** bulk streaming approach currently used for generic handoff between the Gen3 Windmill data browser and Terra workspaces in BioData Catalyst
                - The group debated the value of PFB for all portals. Standardizing DRS ID handoff might be a better focus or starting point in the roadmap. Potential role for FHIR in clinical data exchange instead of or to complement PFB.
- The group provided updates on the use cases described in the charter.
- Despite use cases, still need to:
    - **Standardize handoff** mechanisms across systems
    - Continue to **Generalize handoff** across systems.
    - **Define security requirements** and agreements across systems, not projects, within NIH

- **Generate User Stories/Journeys**, in addition to use cases
- **Prioritize** activities and use cases across all teams, so resources and timelines are aligned
- **Fund activities** in alignment with their prioritization
- **Reprioritize** funded work within each IC

Overall the groups presented tremendous progress on key use cases, proposed solutions for identified challenges, and continue to formulate a roadmap to move this forward. However, timelines and implementation are difficult to coordinate with lack of funding streams to specifically support these efforts. Additionally, NIH input is needed to prioritize use cases, define timelines, ratify operating principles and evaluate adherence, and provide policy/security guidance.