NIH Cloud-Based Platforms Interoperability (NCPI) Virtual Workshop

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

Workshop 101

The October 2021 NCPI Virtual Workshop will be organized around key topics of interoperability, including PFB, FHIR, RAS, Search, and End-User Cloud Costs. We intend to produce two deliverables as a result of this workshop: a glossary and a list of concrete next steps, with assigned owners and use cases that drive priority next steps.

Workshop invitees have been invited to a private Slack channel within the NIH Cloud Platforms Interop workspace that will serve as an important communication tool for the workshop. Relevant workshop documents will be pinned to that Slack channel by Monday, October 4. If you are not in the NCPI Slack workspace, click here. If you have not received an invitation to #oct_workshop, please email amiller@renci.org.

Zoom Information

https://renci.zoom.us/j/93739330988?pwd=dkoybGorOGxDVkVyZWVJZTU1MGEzQT09
Meeting ID: 937 3933 0988 (bit.ly/NCPI_Oct2021)
Passcode: NCPI
Find your local number: https://renci.zoom.us/u/abB1nyEKpO

Prior to the workshop please take a moment to download the most recent version of Zoom.

- Follow these instructions or
- Watch this how-to video here: https://youtu.be/E7zERcVLUBM

Relevant Documents and Links

- Slides
- Working Group Executive Summaries
- NCPI October 2021 Workshop Evaluation Form
- NCPI Fall Workshop Meeting Recording (available by 10/20/21)
    - Day 1 AM / Day 1 PM
    - Day 2 AM / Day 2 PM

Agenda

Pre-workshop			
Due Date	**Activity**	**Owner**	**Links**
Sept 15	Platforms fill in table on 6 topics with updates and gaps, next steps (sample)	Platform leads	
Sept 20	Based on topic slide, BDCatalyst Coordinating Center (BDC3) identify terms that need definition(s) in common Glossary	BDC3	
Sept 29	Platform leads/WG Chairs fill in Glossary	Platform leads/ WG chairs	
(TBD) Sept 29	(TBD) Platform leads/WG Chairs to create slides to nominate scientific use cases to drive next steps	Platform leads/ WG chairs	
Sept 29	WG executive summaries, including updates on driving use cases (update here)	WG Chairs	
Oct 1	BDC3 share all materials	BDC3	
Day 1: Tuesday, Oct 5			
Time	**Activity**	**Owner**	**Links**
11:00-11:05am	Welcome	Stan Ahalt, Patrick Patton	Slides|Notes
11:05-11:40am	Connecting Data, Enhancing Software…What Does a Data Ecosystem Look Like?	Susan Gregurick	Slides|Notes
11:40-11:50am	Goals Day 1: Calibrate, Catalog, Identify Gaps/Challenges	Stan Ahalt	Slides|Notes
11:50 -12:15pm	Demo of Successful Federated Use Case (from search to FHIR to workspace)	Brian O'Connor, Jack DiGiovanna, Robert Carroll	Slides|Notes
12:15-1:00pm	Updates on Key Topics (Part 1) •PFB (10 min) (Grossman) •FHIR (15 min) (Carroll) •RAS (20 min) (O'Connor)	Moderator: Becky Boyles	Slides|Notes
1:00-1:45pm	Lunch Break		

1:15-1:45pm	Lunch Breakout 1: Discuss Gaps and Decide on Concrete Next Steps •RAS and data access (O'Connor)	Brian O'Connor	Slides|Notes
1:45-2:35pm	Updates on Key Topics (Part 2) •End-User Cloud Costs (20 min) (Schatz) •Search (20 min) (Rogers) •Other Interoperability Efforts (10 min) (Ahalt)	Moderator: Becky Boyles	Slides|Notes
2:35-3:05pm	Breakout Session 2: Discuss Gaps and Decide on Concrete Next Steps •PFB (VanTol) and FHIR (Carroll) •Other Interoperability Efforts (Ahalt)	Robert Carroll, Stan Ahalt	Slides|Notes
3:10-3:15pm	Break Plan for Day 2	Becky Boyles	Slides|Notes
3:10-4:00pm	Breakout Session 3: Discuss Gaps and Decide on Concrete Next Steps •End-user Cloud Costs (Schatz) •Search (Rogers) (EasyRetro)	Michael Schatz, David Rogers	Slides|Notes
Day 2: Wednesday, October 6			
Time	**Activity**	**Owner**	**Links**
11:00-11:10am	Welcome and Goals Day 2: Synthesize next steps, driving use cases, determine NIH/NCPI priorities	Stan Ahalt	Slides|Notes
11:10-12:40pm	Breakout Report Backs and Discussion •PFB (10 min) (Grossman) •FHIR (10 min) (Carroll) •RAS (20 min) (O'Connor) •End-User Cloud Costs (20 min) (Schatz) •Search (20 min) (Rogers) •Other Interoperability Efforts (10 min) (Ahalt)	Moderator: Becky Boyles	Slides|Notes
12:40-12:50pm	GA4GH Relationship	Brian O'Connor	Slides|Notes
12:50-2:00pm	Lunch Break		
1:30pm-2:00pm	NIH Breakout: NIH Coordination Working Group Discussion of Priority Next Steps	NIH Only (via separate invitation)	
2:00-2:15pm	Use Case Overview: The Journey of a NCPI Use Case	Asiyah Lin	Slides|Notes
2:15-3:20pm	Review of Current Scientific Use Cases	Moderator: Valentina Di	Slides|Notes

		Francesco	
2:15-2:30pm	Genetic Sex as a Biological Variable and X-inactivation	Melissa Wilson	Slides|Notes
2:30-2:50pm	Interoperability between Kids First & Undiagnosed Diseases Network (UDN) Data via dbGaP/SRA	Valerie Cotton, Allison Heath	Slides|Notes
2:50-3:05pm	Leveraging Functionally Equivalent Pipelines for Long-Read Data on Different Systems	Owen Hirschi	Slides|Notes
3:05-3:20pm	Conducting reproducible science in PIC-SURE interoperating with Seven Bridges/Terra	Simran Makwana	Slides|Notes
3:20-4:00pm	Synthesize Goals and Next Steps for the next 6 Months, with focus on driving use cases	Stan Ahalt, Jon Kaltman	Slides|Notes

Meeting Notes Day One

Welcome (slides)

- Stan and Patrick overview of logistics, mtg. roles, statement of conduct and community rules of engagement
- Note - future invitation lists are determined using past registration lists (*please be sure to register if you have not already)

Connecting Data, Enhancing Software…What Does a Data Ecosystem Look Like? (slides)

- Susan Gregurick overview of slide deck → *[slide link to be added]*
- Data Management & Sharing Policy - effective Jan. 25, 2023
    - Researchers will be required to plan for how scientific data will be preserved and shared
    - Submission of Data Management and Sharing Plan outlining how scientific data and metadata will be managed/shared
- Alignment with FAIR Principles and TRUST Principles:
    - Transparency
    - Responsibility -
    - User Focus
    - Sustainability

    - Technology - support secure/reliable services
- NIH Data Repositories in different stages of maturity and readiness -- we need to assess and develop metrics for use;
- FAIR and TRUST principles -- 14 awards supported in 2021; just the start for future data management and sharing policies as we approach implementation in 2023
- AI-readiness -- challenging requiring engagement with users and feedback; development of use cases
- Data FAIR and AI/ML readiness in existing NIH supported data; how do we take the data we have and make it more FAIR and AI ready;
- Software considerations -- enhancing tools, working with STRIDES Initiative encouraged but not required; developing best practices for reuse;
- Developing best practices to provide transparency and foster reusability/collaboration; alignment with NIH Data Science principles
- https://datascience.nih.gov/tools-and-analytics/best-practices-for-sharing-research-software-faq
- Improving access to secure data -- NIH and NSF supported programs; RAS implementation;
- MIDRC - medical imaging and data resource center → interoperability efforts
- HEAL Data Ecosystem -- providing the “sandbox”; repository recommendations; core metadata recommendation, etc
- NIAID Data Ecosystem;
- BRICS - Biomedical Informatics Computing System
- What does an Ecosystem look like? Feedback essential in driving progress and next steps following FAIR and CARE Principles;
- Office of Data Science Strategy https://datascience.nih.gov/

Q&A Period

- Valentina Di Francesco - What solutions are being tested in BioData Catalyst?
    - MIDRC pilot program; imaging data, clinical studies → ability to deliver images to different cloud based systems; sharing/linking data requires a lot of work and integration
    - Bob Grossman - MIDRC includes the two latest radiological institutions; committed to making the data FAIR and interoping with authorized workspaces;
    - Jon Kaltman - Pilot of COVID images -- observational studies that focuses on ICU patients; the clinical data will be deposited within BioData Catalyst; MIDRC will ingest through their pipeline and co-localize the data within BioData Catalyst (imagining and clinical data)
- Ben Heavner - What are the training efforts planned to implement these best practices?
    - Susan - Data Scholars Program; Data Interns (under graduates) partnerships with NIMHD training at HBCUs and Hispanic Serving Institutions; Training Coordinating Center within Tribal Communities; STRIDES “Cloud Lab” to allow researchers in underrepresented communities access to Cloud sandbox at no cost;

- Stan Ahalt - Will there be an ongoing effort to continue discussions around policy issues?
    - Susan - Yes, groups meeting on various aspects of policy -- cybersecurity, developer access; policy tends to lag and move slow, however brainstorming is underway for forward looking policies

Goals Day 1: Calibrate, Catalog, Identify Gaps/Challenges (slides)

- Stan Ahalt- overview of goals for day 1; NCPI progress brainstorming potential activities; agenda day 1 and day 2
- 5 working groups; use cases to drive the processes → more updates on Day 2
- Working groups executive summaries → *[to insert link here]*
- Tackling Data *Ecosystem* questions via working groups (addressing 4 of the 6 boxes from Susan's slide)
- Workshop goals: discuss RAS, PFB, FHIR, Cloud Costs, Search, other interoperability efforts
    - Move the needle forward on each topics- focus on communicating with each other, understand where we are and where we need to be; document points to hold ourselves accountable
    - Use case status
    - Look for gaps in topics
    - Look for policy/development blockers
    - Identify next steps to help reach NIH goals
- Meeting deliverable: NCPI Glossary
    - Goal: highlight common definitions and differences/controversy around certain words/definitions
    - Important to be precise around language re: policy

Demo of Successful Federated Use Case (from search to FHIR to workspace) (slides)

Robert Carroll

- Data Life Cycle of clinical/phenotypic/research data
- Important idea of NCPI: platforms and services -- FHIR; getting the information needed
- Research user can perform search across FHIR and multiple datasets
    - Research use case: Can we provide tools that allow researchers to perform common tasks?
- Challenge: not a common foundation of metadata/study sources
- Been working on representing study metadata since the last NCPI workshop → working towards modernizing this representation
- Big picture: there are important categories of things (metadata) that are critical to portray accurately
- Products:
    - FHIR Example
    - Study Summary Tool

        - Study Browser Tool
- Demos
    - Several studies loaded into FHIR servers
        - Four study summaries using row level data
        - Representation of the studies with detailed data; simple search and easy to look through tables and search data across all studies (ex. Which groups/subjects have arrhythmias?)
        - Ability to create a synthetic cohort
    - Research Study summary tool
    - Shiny app interactive content

Brian OConnor

- Data Access & Compute; Systems Interoperation WG
- Doubled dataset to ~11 PB of data accessible on the cloud; cloud environments are ready to go
- Systems Interop: how can we make this process easy/nice for users? How can we let users access data across all systems (all 11 PB)? How do we improve data access/compute/interoperability amongst the platforms?
    - Addressed this by focusing on 11 researcher use cases
- Systems Interop started in Jan of 2020; noticed that platforms were becoming siloed, and the team wanted to break down these barriers and allow people to work across systems
- Vision for Interoperability: consistent over the last 2 years; want users to be identified/authenticated/authorized via RAS, create synthetic cohorts of interest, and then hand this off to a cloud based analysis environment, which should be able to access the data from all of the projects
- 3 areas of focus: search result handoff: PFB, data access: DRS, auth: RAS
- WG progress from 2020 → 2021
    - 2020 progress: worked on MOUs/ISAs for system interconnects, PFBs for data handoff from portals to workspaces, DRS for data access, and progress on researcher use cases
    - 2021 progress: RAS authentication, GA4GH standards, a bridge between FHIR work and importing synthetic cohorts using a PFB bridge, more workspaces for supporting DRS servers, RAS passports, and continuing researcher use cases

Jack DiGiovanna: Use Case Success Stories

- Ex. Researcher analysis on CGC spanning 3 portals
- Ex. Researcher PCGC analysis on CAVATICA and BDC powered by SB spanned the PCGC data governed by Kids First and PCGC data governed by TOPMed; led to 1+ publications
- Vision: take data from any portal and put it in any workspace → but, UX not yet optimal and still needs tweaking
    - Export button to certain workspaces making it easier for users (rather than a download/upload workflow)

    - Improved UX workflow video/demo-- mockup

Brian OConnor

- Use Case Success Stories: #7, Tim M. looking at congenital heart disease. Wanted to work with Kids First data, wrote code to pull data into the platform and download it, but this mean that he had to download it into his own bucket
    - Feedback from Tim: very time effective to use the export button to get data into a workspace; saved many hours
- Use Case #11: Melissa W. examining Sex as a Biological Variable; created a Terra workspace referencing AnVIL, BDC, CRDC, and Kids First datasets. Used DRS to access data on demand without duplicating any data across systems
- Demo video from a researcher perspective: searching data across platforms, selecting data, and using PFB export to a workspace without any manual intervention; same with BDC portal to export PFB over to Terra and create workspace without manual intervention
    - Similar approach on GDC portal and Kids First
    - End result: a cleaned workspace with data from multiple portals; easy to launch analysis; big takeaway: *data aren't being duplicated/copied*
- Priorities for 2022:
    - Finish RAS Milestone 3: move from a user linking many accounts to a single RAS login
    - Connect more portals and data repositories: move from having to modify/script/change the data to get it into a workspace, to adding many different portals with a "send to analysis/workspace" button; also want to add more DRS servers
    - Get the word out to users and let them know what they can do

Stan Ahalt- comments on the impressive progress from the Systems Interop WG

Chat

Jack DiGiovanna: Live demos are cool :)
Robert Carroll: Not quite enough time unfortunately! I'm happy to set anyone with google ID up to run the browser, though.
Jack DiGiovanna: Really important point. All the sys interop demos are *in production*

Updates on Key Topics: PFB (Grossman) (slides)

Provided background/overview regarding (Portable Format for Biomedical Data (PFB)

- Another way to think of it is another alternative of bulk FHIR format
- Bulk format allows for persistent ID in metadata
- Avro format that could be assigned FHIR attributes

Provided Gaps/updates and next steps as it pertains to Gen3, Seven Bridges, NCBI and NCPI Outreach.

Ben H: Would it be useful to have tools for translating between VDB and PFB? (Seems akin to Avro vs. Parquet?) Are there use cases that column stores would be desirable?

- VDB is schema driven that can represent any data of any sort.

Kurt: It could be fruitful to explore the interoperability. The VDB team does this full time.

- Ben H: This is pretty in the weeds technically.
- Becky: Great point Kurt. I would love to hear a more in depth discussion.

Updates on Key Topics: FHIR (Carroll) (slides)

Provided updates and Gaps/next steps as it pertains to AnVIL, BDCatalyst, Kids First, dbGap, NCPI Outreach

Updates on Key Topics: RAS (O'Connor) (slides)

RAS -- lot of coordination with everyone on the call to work together to put together a series of design documents -- milestone 3.

Groups coordinated a 3 milestone plan.

Discussion also included the New Passports approach, lessons learned along the way.

Provided updates and Gaps/next steps as it pertains to Gen3, Seven Bridges,NCBI and NCPI Outreach

Breakout 1: RAS and data access (O'Connor) (slides)

- **Attendees:**
-

Agenda

- gaps/risks in milestone 3
  - Stan
    - What is the future of whitelists from dbGaP?
    - Timeline, what happens if we go beyond the expected timeline? Dependencies between teams
    - Single sign on
  - Becky
    - BDCat… users don't want to see duplicate logins… RAS is a big part of that…
  - Valentina
    - Timelines… the milestone 3 timeline is aggressive… who is doing what and when will testing resources be available

  - Mike Feolo
    - "NCBI will continue to support the whitelists we produce until they are no longer needed."
  - Alex and Peter U. Chicago
    - Feedback in most recent tech plan on implementation details
    - For example, building implementation using existing services
    - RAS docs talk about particular components like clearinghouse… but these are not distinct in Gen3
    - Service vs. responsibility… spec compliant but not 100% the same breakdown into distinct services
    - Architecture behind implementation… emphasis on things not breaking… existing approach, new RAS approach… may cause confusion in the tech plan
    - Performance… something to watch, they will monitor this
      - Storage of auth information… how things are cached
      - Inside vs. outside the Gen3 platform
      - Getting close to running tests on this
    - Have an MVP in place by Dec… won't have all the performance improvements since it's an MVP for testing
  - Manisha and Michele
    - Concern is the timeline
    - Understand what Peter is saying
    - What will be built, when, and the impact for the end user
    - Single sign on
    - Milestone 3 is the right direction
  - Nicole and Michale - Terra
    - Consortium and developer access … need to make sure that's on track
    - Timeline
    - Performance… DRS 1.3
    - Requester pays needs to work
- gaps/risks beyond milestone 3
  - Workspaces being locked down by visas from RAS, for derived data?
  - Securing FHIR and other services with Passports?
  - Performance and scaling?
  - Consortium users
  - Ben H. NCBI will continue to support the whitelists we produce until they are no longer needed.
  - Susan G. "I would like to see the global RAS, IAM discussion. there are opportunities to engage ELIXIR. If there is interest, please articulate that"

Notes:

-

Updates on Key Topics: End-User Cloud Costs (Schatz) (slides)

- Michael Schatz presented slides re: CRDC, AnVIL, BDCatalyst, Kids First & NCBI updates & next steps
- Kurt McDaniel: Meet w/ GCP & AWS regularly; they want to participate in STRIDES; AWS agreed to host full set of SRA normalized data format; moving carefully in moving controlled-access data but hopefully available in early 2022
- Becky Boyles: Users are interested in estimating costs before doing work; how gernealizable is this work?
    - Michael: Some workflows are predictable & some less so; IDing popular tools & running different configurations but there are many parameters; trying out in lots of different environments; concerns that it will be complex and an ongoing challenge; also other efforts underway to help monitor running workflows; recommend users start small
- Ben Heavner: Should we discuss at interop meeting the mechanics of billing & funding, how users get credits & charges being correctly applied
    - Michael: Always recommend users start by exploring STRIDES discounts; will be hard to provide unified billing as some users go through NHLBI and others use personal credit cards; important to clearly explain options to users

Chat

Anne Deslattes Mays: Don't platforms give an ability to estimate costs

- Stan Ahalt: Not really.
- Jack DiGiovanna: Depends on the platform and tools -- Many tools have a benchmark in the description. For inputs of size and type blah, using the default config, it will cost X on this instance config

Anne Deslattes Mays: If we encourage everyone to use GitHub and then use GitHub actions that allows you to estimate costs - GitHub actions has minimal requirements in terms of RAM usage so encourages the use of small test data sets. I have done this with Nextflow but anyone able to do this with CWL ?I bet it could be done. And then with that we have the rule of thumb of what can be done.

- Jack DiGiovanna: The CWL workflows in the BDC, CRDC, and CAVATICA public app have benchmarks already
- Anne: So you have rules of thumb for cost estimates — can you share?And do you have examples for GitHub actions? I'd love to work with you on this :)
- Jack: Here's an example for GATK 4.0 with metrics for Whole Exome Seq https://cgc.sbgenomics.com/public/apps/admin/sbg-public-data/whole-exome-sequencing-bwa-gatk-4-0/41
  Need to expand to "read more" and got to the "Expected Workflow Performance" section
- Anne: Are costs buried in there?
- Jack: $0.40 for a NA12878-135x input; $0.47 for a HG002 Oslo 190x
- Table at the bottom

Updates on Key Topics: Search (Rogers) (slides)

- Dave Rogers presented slides re: CRDC, BDCatalyst, Kids First, AnVIL, NCBI & NCPI-portal updates & next steps
- Jack DiGiovanna: Where do we see search at a high level, e.g., searching across cancer landscape or all of NCPI?
    - Dave: How can users find data to use for their needs? Don’t have metadata on harmonization. Also which cloud is it in? Also policy concerns re: permissions around moving data.
    - Steven Cox: Need to determine near-term priority drivers; taking practical steps to get people together and write down use cases and develop personas
    - Robert Carroll: Hoping to align work different groups are doing
    - Valentina Di Francesco: Can anyone say more about the OTA to publicly launch CDA API and enable controlled-access data query?
    - Ben Heavner: Can anyone say more about the OTA to publicly launch CDA API and enable controlled-access data query?
        - Erika Kim: Working on publicly launching APIs
    - Stan Ahalt: Need to adhere to FAIR principles and have metadata to determine interoperability; continuum of improving data to enable science across disparate datasets
        - Ben: Breaking search down into smaller pieces is pragmatic; there are datasets w/ different levels of harmonization; need to be particular about what we’re searching for; dataset search over metadata attributes can be harmonized for specific research questions
        - Steve: Addressing metadata representation of things that need to be searched is a primary goal
        - Dave: Could surface level of harmonization to help users find what they need; can clarify how interoperable datasets are; could go in dataset catalog

Chat

- Brian: Re. Search: How detailed are the Search use cases?
- Ben: Is harmonization a pre-requisite for search?
- Anne: Maybe harmonization should be a service. I think there are two parts. Where it is doesn’t matter
- Valentina: Can anyone say more about the OTA to publicly launch CDA API and enable controlled-access data query?
- Ben: Intersection with the “Common data elements in RADX” that Susan Gregurick mentioned earlier?
    - Robert: Re harmonization- I don't know to what extent we need to "harmonize strongly", but "standardize weakly" and provide tools to help cover the gaps.
    - Bridging the source data aspects with "some" kinds of standard concepts. Provide a consistent representation of those items for tools like DUG to work on.
    - To Stan- I'm obviously highly opinionated on this, but I think we can make this a continuum and make practical steps forward that we can continue to build over

time.
  - Anne: Bioproject and bio study.
  - Stan: To Robert - I completely agree. Bootstrap.
  - Becky: Agree, but I would agree that not defining personas and identify use cases within search (which Ben has worked on) has gotten in the way of potential progress
- Robert: Doing the work to align the metadata also gets us on the path to enabling users to harmonize the row level data
  - Mike: Prior to request and post approval will have different granularities or harmonization needed
  - Ben: I like that idea a lot - some sort of standardized way to describe the level/state of harmonization or data structure/representation within a study as a metadata element describing the study data…
  - Ben: ie. I find a study with interesting data; is the data a bunch of scanned .pdfs, or is it encoded in JSON? Do clinical measurements align with some standard, or local clinic practices/encoding? -- I feel like Michael DuMontier did work along these lines about “assessing FAIRness"?
  -
  -

Updates on Key Topics: Other Interoperability Efforts (Ahalt) (slides)

- Stan Ahalt presented slides re: CRDC, AnVIL, BDCatalyst, Kids First, NCBI updates & next steps
- Valentina Di Francesco: How to pursue systematic approach for training workshops to spread word of interoperability efforts? Considering resources & funding.
  - Becky Boyles: There’s a need for an NCPI communications strategy to provide complementary messaging; few users of each system are aware of these emerging capabilities
  - Stan: Consider how we train more people & use training as mechanism to expand the number of people exposed to the possibilities; unclear which other venues to discuss this in; need to make accessible to non-experts; encouraged by progress so far but NIH leadership needed for next steps
  - Brian O’Connor: Uses Excel & coding to transform data; users can be very productive with modest skills; want to get that message out; many users don’t know this is possible; next NCPI meeting could be focused on sharing work done so far and generating more users
  - Adam Resnick: Risk in having non-uniform processes; examples & clear instructions needed to avoid disappointing users
  - Dave Rogers: Could have shorter-term goal to get more explanatory info on portal; also working on NCPI Twitter account to help spread the word

Chat

- Valerie: to clarify INCLUDE is focused on Down syndrome research https://www.nih.gov/include-project. --

- Valerie: Anne Deslattes Mays is our new DATA Scholar. She has been very active today :) Please feel free to connect with her or in slack :) https://datascience.nih.gov/data-scholars-2021/amplifying-and-sustaining-the-impact-of-childhood-cancer-structural-birth-defect-and-down-syndrome-data
- Becky: Accessibility often comes from work together to put out a training. This is often how we find and address issues.
- Stan: This risk was what I was trying to allude to.
- Ben: Suggests a goal of each group that participates in NCPI to give user documentation of a use case that does this kind of interoperability
-

Breakout 2: PFB (slides) and FHIR (slides)

- **Attendees:** Ben Heavner, Robert Carroll, Jeremy Costanza, John Cheadle, Becky Boyles, Radhika Reddy, Mike Feolo, Michael Lukowski, Nicole Bolliger, Liz Amos, Tanja Davidsen, Lon Phan, Sai Subramanian, Allie Gartland-Gray, Alexander VanTol, Manisha Ray, Peter Vassilatos, Eric Wenger, Michele Mattioni, Dave Rogers, Brian O'Connor, Jessica Lyons, Adam Resnick, Steven Cox, Ann Van, Anne Deslattes Mays, Brian Walsh, Dave Rogers, Jack DiGiovanna, Jay Ronquillo, Joe Asare, Pauline Ribeye, Ravindar Eskandary, Tim Slade, Tom Madden, Andre Paredes, Peter Vassilatos, Candace Patterson

FHIR - Robert Carroll

- **Gaps/Key Blockers:**
    - Adoption across platforms
    - BDC perspective: What problem does FHIR solve for that ecosystem? What uses are there for FHIR?
        - There may be a need to get EHR data in bulk; we could use FHIR for that. It is indeed federally-mandated for EHR to use FHIR.
        - FHIR will solve future problems even if we don't know exactly what they are right now.
        - It allows for potential standardization - gives us an API that we don't have to invent.
        - Use as a modeling language - provides a vocabulary and data structure that can be adopted
        - Exchange data from disparate systems in a common way
        - Ingest of other data, e.g. with REDCap tools or other CDE representations

    - Need a map to communicate what the goals are RE: FHIR, and where the limits are. It is valuable to us in part because it can serve so many different roles, which is also what makes it unclear. <u>We need to be very clear for us what problems we want FHIR to solve</u>.
    - Since FHIR is an exchange framework, we would need to agree on a common implementation guide to make it interoperable. So how do we do this?
- **Next Steps (next 6 months):**
    - Add information from this discussion and from existing IGs (Implementation Guides) onto the portal
- **Chat:**
    - Manisha Ray: Strongly agree with the point that FHIR is already the standard for EHR, and all the programs will use EHR data eventually. Regardless if additional methods are adopted, FHIR will have to ALSO get accommodated at some point for clinical data
    - Michele Mattioni: you can search data using FHIR API
    - Becky Boyles: Can you say more about what/ when you see FHIR becoming implemented for clinical data? Do you mean clinical trials?
        - Mike Feolo: FHIR is a exchange framework we would need to agree to a common Implementation Guide to make this interoperable.
            - 2nd use is to exchange data from disparate systems in a common way
    - Michele: Also one important bit about FHIR → given the different type of Search Experience users do need, having one API to search opens the ability to create different type of Search experience, while the API below is the same
    - Brian Walsh: I've spoken to several senior researchers that have the same questions Rebecca has. Her comments are very representative

- What are the key use cases of FHIR?
- What are the core milestones and timelines, and what use cases will be supported (and not) at each time point.
-

<u>PFB - Alex VanToll</u>

- **Gaps/Key Blockers:**
    - Models can be different (NCPI standard solves some of this problem but not all)
    - PFB having its own schema can be a pro and a con - adopters don't have to buy into a full data model and it can serve a variety of needs.
    - With PFB - because the data models are so different, it's a lot more overhead on the system consuming the information to write different parsers, etc.
    - Time-insensitive generation of PFBs: It can take a long time to generate a PFB for a large cohort (e.g., wait an hour to generate a PFB, which leads to worse UX)

        - What is the common case for PFB? Selecting whole studies, typically; if we had pre-generated PFBs, this 5-10 minutes might be a few seconds.
        - Could try to solve this with caching as well
    - On-demand creation of PFB does not allow for assignment of a persistent ID.
    - Lack of support on the portals for ‘PFB lite’
- **Next Steps (next 6 months):**
    -
- **Chat:**
    - Ben Heavner: I need to drop to a conflicting meeting, but for PFB discussion, my main question is whether there are column-store use cases that PFB doesn't support that should also be implemented along with PFB? Technical questions about exchange between row-store and column-store serialization approaches
    - Manisha Ray: Can you have generic parsers if the data model can be different for every system?
        - Brian O’Connor: Yes, since PFB includes the model, you can parse that model and then load the data. It’s how the PFB import works on Terra, for example.
        - Candace Patterson: To go over off of Manisha’s question - how is the NCPI Interop model being used for PFB/FHIR? Are we mapping the various DCs fields to the model so we could have a standardized way to export the data to the analysis platform
    - Brian Walsh: FHIR has a fairly deep graph, with optional references. Can PFB match this need for reference depth?
        - Michael Lukowski: Yes Brain, the model can be anything you want
    - Michele: Does the model specify the semantic of the attribute?
    - Brian O’Connor: Ya, you can embed ontologies into a PFB Michele… it’s enough information to create target table structures and fill them in workspaces… it’s just Avro describing tables and their possible column values.
    - Jack DiGiovanna: Are there end-users that leverage the graph relationships in PFB? I found this very attractive when PFB was pitched, but all the usage we’ve seen “flattens” the PFBs, which both i) takes time and ii) removes the self-describing benefit.
        - Brian O: I probably don’t understand the question Jack, when you handoff a PFB for BDCat or AnVIL, for example, you get a multi-table structure with relationships between
        - Jack: Yes, but you fundamentally convert the graph or triple store to a 2d table; Another gap we have is PFB support, e.g. the portals so far don’t support “lite PFB”
        - Manisha: Strong agreement on making it easy for end users regardless of format

Breakout 2: Other Interoperability Efforts (slides)

- **Attendees:** Mike Schatz, Stan Ahalt, Asia Mieczkowska, Valentina Di Franesco, Asiyah Lin, Beth Sheets, Erika Kim, James Coulombe, Jeremy Zollars, Laura Biven, Stephen Mosher, Sweta Ladwa, Valerie Cotton, Vivian Ota Wang, Marcia Fournier

- Prompt/Question: Interoperability for experienced or novice users:
    - Sweta - Would like to see "what's out there" as a novice user. The data that's offered at each ecosystem, functionality of ecosystems as well.
    - Valentina - Both would be important. Goal #1 is bringin users onto the platform. Goal #2 would be to empower users to use data from other ecosystems, such as BDCatalyst. Need a plan crafted for specific types of users to focus on addressing their needs.
        - Stan - Encouraging to hear people see the necessity of platforms working together. I don't think users would have imagined doing this type of work 2 years ago, but think users have seen the promise of interoperability. NIH environment is complicated, hard to fit into it, and is only going to get more complicated with new studies coming on board.
    - Valerie - Scalability issues should be tied to use cases to address things researchers actually want to do. Don't boil the ocean, focus on figuring out and addressing how users want to use the ecosystems.
- Is it a problem that the User Interfaces are not consistent across ecosystems?
    - Michael S. - Not sure it's a problem
    - Stan - I'd agree it's not a major problem, unless we're wanting to scale significantly to many users. You'll run out of people willing to learn multiple UX/UI.
- Vivian - What has been useful from AnVIL's experience to create a central portal? Advantages of having a central portal (vs. each platform having its own).
    - James Coulombe - Each research institute has special needs & preferences for what they'd like to immediately see in a portal. May be a good thing to have a wide variety of options, as long as that design has been informed by that user community. Ex. I'd personally like to have a simpler version that lets me focus on the biological aspect.
    - Stan - Data without brains (people who make sense of it) is useless. Need to empower all types of researchers, not just those who can also wear a bioinformatician hat.
    - James - As the world changes we may appreciate having variability in our portals to reduce the risk of a major change making everything obsolete.
    - Michael S. - That's been our AnVIL philosophy since day 1 - Keep distinct identities in portals. Key standardization is how users are able to bring in tools to AnVIL. Need to empower all levels (simple APIs all the way up to intricate tools).
        - One thing we can decide on is how to deploy tools in a standard way. How to deploy apps in a standard way, so that it can be ported to multiple platforms.
    - Valerie - NCPI has done well basing things on APIs. Should make it clear to the community how we're exposing data via APIs, and in what way.

- Stan - Clarification on a point Valerie made - are you saying we should have a central reference for guidance? Or a central portal?
    - Valerie - Emphasis on having use cases drive our direction. What do users need to get their work done?
- Stan - We're hearing a discussion about standardizing training & documentation. This would let us cross-educate each other and bring in users in a more consistent fashion. Maybe NCPI can devote a greater degree of energy to.
    - Valerie - If every time a user achieves something we have the use case show how it's done I think that can go a long way. Empower researchers to replicate a thing in their own time/space.

Chat

- Wrt portal discussion:
    - James Coulombe: I think having different portals whose design is informed by the users is really a good thing as long as the data can be found and used across the platforms.
-

Breakout 3: End-user Cloud Costs (slides)

- **Attendees:** Michael Schatz, Stephen Mosher, Enis Afgan, Beth Sheets, Jamie Guidry Auvil, Jeremy Zollars, Manisha Ray, Patrick Patton, Michael Baumann, JackDiGiovanna, Nicole Bolliger, Tanja Davidsen, Ravinder Eskandary, Kurt McDaniel, Anne Deslattes Mays

- Michael S - Two/three major buckets: 1) Costs associated with data storage / networking (egress, related costs), 2) Compute / analysis costs. Any other major buckets folks have encountered that should be included?
    - Anne - May be worth splitting out egress/ingress costs
    - Enis - Should we consider implicit costs of paying for personnel? (Ex. of charging from admin handling data)
    - Manisha - We've covered line items, but I don't think that covers gaps of the user's journey. **How** are costs generated? How does one pay costs? (doesn't fall easily in standard grant language) eg personnel vs consumables vs capital expense
        - Michael - NCPI can help get standard language from NIH on how to address concerns like this.
        - Anne - We need to teach people that cloud costs are consumables, like reagents.
    - Jack - We have seen a lot of support from NIH on supporting users with their cloud costs, there just seems to be a lack of documentation.
        - Michael - Seems like something we can tackle within 6 months - get NCPI blessing on a description of costs.

        - Anne - Missing documentation on where things should go.
- Manisha - Some platforms have Google handling billing, which requires a Google account. I don’t have one, and found it hard to follow the demo. Didn’t want to use my own CC and account to demo something.
    - Enis - Would users with light spending be a target audience here? (Someone who pays cents a month).
        - Michael - Part of the appeal of cloud computing is doing things at scale. You may me right, but if we can’t solve something for users who are paying a dollar a month, how are we going to solve it for someone spending thousands a month?
    - Stephen - Wonder if there are people who are just getting their feet wet, and are getting scared off by the difficulty of implementing a $5 account now. Should we focus on simplifying onboarding?
    - Jack - There’s a level of fear for things like egress that users may never need to pay for. We need to address this level of fear that some users may have of racking up large charges.
        - Michael - Some key users will have high costs. Maybe we should talk about estimating costs. What’s the compelling use case for doing work in the cloud that could be done on a laptop?
            - Anne - May be less about scaling the work and more to do with easily sharing work or other reasons. Avoid downloading files.
                - We should also work to get people off laptops so they can more easily reproduce their work by using tools like GitBook.
- How do people estimate their costs? Ex. from Michael of estimating for his own work, but was off by a factor of 3 due to circumstances with scaling.
    - Anne - There’s a lot of unknown when you’re pushing science with fresh ways of computing. Not sure we can prevent those scenarios.
    - Jack - I think that builds on your prior point: I think in the next 6 months we should set user expectations on what we can deliver. Can we develop GWAS in a bottle that people can use to estimate their costs, even if it’s only as close as Michael’s example.
- Michael - Is there more we can do on the technology side to help?
    - What if NIH purchased some reserved nodes that users could use at low cost?
    - Manisha - I wonder if we could get some flexibility in a grant to cover exploratory work. Might be able to fall under T32 training grants?
        - A piece of information that may also be missing are setup costs. It may take more time/money to start up something vs. continue it. (Ex. with wetlabs setting up new equipment).

- **Goals for next 6 months:**
    - Stephen - Improve documentation on billing documentation.
    - Michael - Drawing out user stories to inform or serve as use cases

  - Enis - Maybe a NCPI-built dashboard with some common costs to avoid reinventing the wheel across each ecosystem. Consolidate existing resources.
  - Michael - Wonder if we need a way for users to publish costs in a limited scope to help inform the general public.
    - Jack - Funding perspective - how could we get these things built? Not sure, not sure there is funding for this work.
- **Gap to highlight** - Are there bigger things we should look at? Platforms people can get into with no costs to learn?
  - Adam - Might be worth looking at simple apps that you can just plug in and go, that should in theory have low/predictable costs.
- Hackathon to highlight ways to optimize costs? May be a long-term effort, not the next 6 months.
  - Anne - May be worth looking at - calling out how costs can be looked at in different ways.
  - Jack - Like the idea, could maybe expand to look at developing tutorials.

Chat

- Manisha Ray: I see these at a high level as “what” (egress, compute, etc); “how” (grants, strides etc); “who” (admin time); “where” (which cloud, which platform, each has different billing, etc). What and How are the biggest gaps IMO
- Anne Deslattes: I think we want to help elevate the small research group to integrate with the larger data ecosystem and lowering the threshold of entry to the cloud computing is a great democratizing force
- Jack DiGiovanna: Collaboration, compliance, policy
- Anne Deslattes: Strategies for estimation — use one chromosome instead of the whole genome
- Manisha Ray: I suspect policy alignment from NCPI would be easier than technology alignment
- Manisha Ray: Similar we could push for a centralized documentation around the funding- treat the cloud costs as consumables, use this example language in justification, etc
- Beth Sheets: Dockstore has proposed hosting some info from workflow run logs from workspaces, but there were security concerns associated with this I believe.
-

Breakout 3: Search (slides)

- **Attendees:** Dave Rogers, Robert Carroll, Adam Resnick, Allie Gartland-Gray, Asia Miezkowska, Asiyah Lin, Brian Walsh, Candace Patterson, Deanne Taylor, David Pot, James Coloumbe, Jeremy Constanza, Jessica Lyons, Joe Asare, ken Wiley, Laura Biven, Michael Lukowski, Michele Mattioni, Mike Feolo, Natalie Kucher, Natalie Madero, Peter Vassilatos, Pauline Ribeyre, Rebecca Boyles, Robert Carroll, Sai Subramanian, Stan Ahalt, Steven Cox, Susan Gregurick, Tim Slade,Valentina Di Francesco, Valerie Cotton, Vincent Feretti
- **Easy Retro Discussion**

    - ‘Experimental Metadata’ - defined here as how the data was collected/processed, e.g. what machine it was collected on, how long was tissue stored, etc.
    - Phenotype - what is phenotype, a symptom? A disease? An algorithmically-defined cohort (EHR context)? This is good fodder for the glossary.
    - Prioritization of searchable attributes - within commonly-known entities within a study, how do we prioritize how to search on these?
    - Data oriented impediments must be better itemized. If we know what these impediments are, let’s prioritize them.
    - Envision: a distributed data set catalog that has minimal standardization so you can do distributed queries across them - this is a long-term view. A baby step towards this might be a link back to the data browsers for each dataset.
    - What are our real use cases and who are the real users of search? This helps inform whether we need a dedicated working group/tiger team
        - We have our use cases.
        - Where are other people getting user feedback to drive/improve searches?
    - How do we figure out who our users are (who are these mythical data parasites that will reuse these data?)
- **Gaps/ Key Blockers:**
    - Searching over phenotypes
    - Understanding how data is consented and how to apply for access
    - Finding and gaining access to different search platforms
    - Potential users are unclear what each search does
- **Next Steps (6 month timeframe):**
    - Understanding use cases for search from real users
    - Create a Tiger Team/WG for search
    - Create a list of search components and APIs used in NCPI platforms
    - Taxonomy of Search/Roadmap for Search - name each facet of search and how they relate to each other. Example: In google, there are lots of searches for different types of datasets. We don’t have a diagram that says ‘here are the kinds of search we do, and here is how they are connected to each other’.
    - Portal that demonstrates all our different searches, with instructions, and link to a feedback form.

Chat

- Valentina Di Francesco: A potential use case is to tell a user where all the data for a study is: in some cases legacy data is on dbGAP, and new data can be on both AnVIL and BDC.
- Valentina Di Francesco: The users do not know which samples are on BDC or on AnVIL, and the metadata associated to those
- Ideas:

    - Mike Feolo: dbGaP can add a field indicating data location, so people can see what studies are in what platform. Search could search over FHIR service. Possible to make this a faceted search in our advanced search as well. We are going to push the variable metadata to FHIR format in the next year or so.
    - Rebecca Boyles: Also, DUG uses the dbGaP data dictionary format. So also, if we can adopt that format DUG can search over all the data dictionaries
    - Deanne Taylor (KF) : Kids First has spent some time evaluating our user base and what kind of use cases they're interested in. This has influenced our types of search support.
    - Kids First has spent some time evaluating our user base and what kind of use cases they're interested in. This has influenced our types of search support.
    - Robert Carroll: I'd be interested to see what it could look like to get DUG to take in FHIR versions of variable metadata- gives us some flexibility to grow it over time.
    - Mike Feolo: Use cases and personas driven from customer engagement
- Robert Carroll: Re the use cases for fhir metadata search, I don't think there is a use for BDC users at this time. If multiple platforms provide variable metadata over FHIR, it might be worth it to help your users (and NCPI users) to find variables they are interested in. Especially if these studies do not have the data in flat files in dbGaP

Plan for Day 2 (slides)

Meeting Notes Day Two

Welcome and Goals Day 2: Synthesize next steps, driving use cases, determine NIH/NCPI priorities (slides)

Notes:

1.

Chat:

-

Report Back: PFB (slides)

Grossman

- Gaps and key blockers
    - PFB “light” solves interoperability issues
    - Blocker: don’t carefully distinguish between use cases; would be helpful to distinguish current use cases for interoperability in NCPI and others

- PFB Interop Trade Offs
    - Computing PFB on the fly is time consuming; trade off between flexibility of virtual cohort selection on the fly and consume that data in another platform vs pre-computed PFB
    - Flexibility of having different data models vs supporting arbitrary/fixed models that must be parsed by the cloud platform
- PFB Gaps
    - Confusion about what PFB is and is not
    - Time intensive
    - Clarifying similarities and differences between PFB/VDB/etc. vs FHIR
- PFB Next Steps
    - Document & distinguish different PFB use cases
    - Possible next step: take published AI/ML studies, create a self contained PFB file, and use that as a use case for exporting from one platform and importing into another platform
        - Use case to demonstrate something more than PFB Light

Q&A
Stan Ahalt- can anyone speak to why FHIR is a useful mechanism for search?
Becky- let’s circle back to that question after Robert C. goes through his slides

Chat
Jack DiGiovanna to Everyone: Do we see support coming for both PFB-Lite and PFB-regular on the different portals?
Robert Grossman: Yes, it’s in the roadmap.

Report Back: FHIR (slides)

Carroll

- FHIR Gaps and Blockers
    - Adoption across platforms
    - Lack of clear documentation of the uses of FHIR; easy to get stuck thinking that FHIR is one thing and roles are not clear
    - Need a map to communicate the specific goals and limitations of FHIR services -- a roadmap
- Actionable Next Steps (next 6 mos)
    - Align on research study and metadata version 1 representation; easier to share and collaborate on public data
        - Could help facilitate portal and search activities
    - Develop milestones around broad services; work with platform to develop roadmaps for these opportunities
        - What are the problems that FHIR could help solve?
        - Ex. This is what AnVILs benefit/use case is, how could this be helpful to everyone else?

        - Different priorities in different groups
- FHIR Use Cases
    - FHIR includes a data model, vocab tools, and service layers
    - Ingestion of EHR data and other data (REDCap, CDEs, etc.)
    - Vocab tools - helps empower semantics of the data
    - Represent data in existing structure as well as in a harmonized way
    - Options for server implementations of a global standard
    - Data exchange from disparate systems
    - Options for study summaries, metadata
    - Ability to link out to DRS URIs
- Plan to continue to circle back, build out road maps, and find ways to help people understand FHIR and FHIR use cases

Q&A

Stan Ahalt- is there leverage to be had in either PFB or FHIR in a search setting? Would one or the other format supply us with some ability that we wouldn't otherwise have?
Adam Resnick- misconception that it's a PFB vs FHIR decision; not in the same lane; FHIR has layers of services that support access & query of FHIR based tools in ways that are pre-built into the healthcare ecosystem

- Large scale economic investment in FHIR (eg. Google, AWS, Azure)
- PFB vs FHIR may not be the right framework

Robert Carroll- general agreement with Adam
Robert Grossman- FHIR is a mandate built into the healthcare ecosystem; we have to interoperate with it. The role of PFB is more narrow; it's a file format that doesn't come with servers/APIs/search functionalities, etc. FHIR and PFBs are not comparable. Long term versioning and storage are sometimes easier to do with PFBs. In complicated ecosystems, you want interoperability for importing/exporting bulk data
Stan Ahalt- agreement with Bob G. Need to create use cases and ask what the right tech to deploy is for each use case.
Allison Heath- can do basic search on FHIR servers; can do faster aggregation via elastic search; PFB and FHIR have their own strengths but aren't necessarily comparable
Robert Carroll- there are tons of sources of vocabularies, but not a great bridge

Chat

Ben Heavner: There are some really great FHIR demos (but I don't have links at hand) - do others happen to have links to demos that show what FHIR enables?
Jack DiGiovanna: I think they are in Robert's slides yesterday
Robert Carroll: Ben- on FHIR demos NCPI is getting close to having the quality row-level data to demo the use cases. Hopefully soon. The study summary data I showed yesterday does actually use fhir both as the row-level data source and the study-level data source!
Brian Walsh: +1 as well.   FHIR ~ common data model, staging database, extensibility
Brian Walsh: PFB ~ data transfer
Ben Heavner: Yes, the demo (starting at slide 31), but I was thinking of some videos from L7, perhaps - more general than the great demo you presented yesterday.

Jack DiGiovanna: Ah, I think you can to be added to access the FHIR server @ben.
Then you can *be* the demo :)
Ben Heavner: I mean more to help inform others - things like this, perhaps - https://www.youtube.com/watch?v=OIt0GrCPu8k (though I haven't watched this particular video)
Allison Heath: There's no need to request access, you can login and check it out for anyone here: ncpi-api-fhir-service-dev.kidsfirstdrc.org/
Robert Carroll: For general FHIR demo: If you have an iPhone, connect Apple Health Records to your patient portal!
Ben Heavner: Yep - https://www.apple.com/healthcare/health-records/
Ben Heavner: "Apple is using the SMART on FHIR (Fast Healthcare Interoperability Resources) standard which enables users to download their health records and share available health data with participating organizations."
Allison Heath: A deck we made about the basics of search in the FHIR API (can see both what it can do and if you've worked with these types of APIs before the weaknesses as well): https://docs.google.com/presentation/d/1Vdd1uVitm4H0yx3OkCODJir8dIltki2IGJtZpxddtxw/edit#slide=id.g88f2892937_0_74
And the much more thorough HL7 docs here: https://www.hl7.org/fhir/search.html (we made the deck to get straight to use cases we had)

Report Back: RAS (slides)

O'Connor

- RAS Concerns/risks for milestone 3
    - Timeline concerns; RAS rolled out for auth in systems by the end of QI 2022
        - Concerned about getting a testing environment up and running by early December
        - General concerns about production release timeline by 2022
    - Architectural ideal vs implementing on existing services
    - Performance in parsing potentially very large passports
    - Feedback from projects like BDC wanting to see a single sign on experience (may not be included in milestone 3)
- Beyond milestone 3
    - Performance, batch operations, requester pays -- will require back and forth with implementers
    - Thinking through what happens with workspaces that have derived data
    - Secure other APIs beyond DRS (ex. Using passports to secure a FHIR server) with Passports
    - How will we have developer/consortium access lists that live outside of dbGaP, and how to support packaging of passports that are non RAS based
    - Working with other IAM systems/partners
- Next Steps (6 mos)
    - Top priority is to meet milestone 3 goals
    - Begin planning milestone 4, what it will look like, what we need to work on (ex.

        Performance, derived data, securing other APIs, developer access lists)
    - Great time to start thinking about other (international) groups that use Passports -- are we all using them in the same way? Are implementations of passports compatible with each other?
        - Potential for further data access

Q&A

Ben Heavner- RAS beyond dbGaP permissions model; how will RAS apply to data and user created workspaces on analysis platforms? Currently a trust based model where workspace owner ensures everyone has appropriate access/permission
Brian O'Connor- DRS URIs are secured, but how do we carry the passports/visas through to the workspace environments? Up for further discussion
Stan Ahalt- do we need to think about the way we credential people in a mathematical way?
Brain O'Connor- two areas of growth: growing datasets that people have access to within NIH projects, and collaboration with other large international groups; time to start thinking longer term- how do we open up the whole ecosystem (beyond NIH) to our researchers?

Chat

Ben Heavner: That's a very exciting vision!
Valerie Cotton: I think this sharing of derived topic intersects with governance. I.e., which data go back into the repository for public distribution (via dbGap). Certainly something to think about.
Brian OConnor: I love that idea Mike!!!   Google cloud shell is another good example of a free tier item that's really, really useful
Ben Heavner: I really like this idea of NCPI being a place where we can connect with broader industry and governmental resources - to learn from their work, consider aligning with it, and engage with broader tech standards and practices even beyond life sciences research.
Brian OConnor: mybinder.org but with NCPI tools and data access ready to go

Report Back: End-User Cloud Costs (slides)

Schatz

- Gaps/Key Blockers
    - Cloud cost model is a cultural shift for end users -- anxiety over runaway costs, budgeting and planning, etc.
        - Solutions: education, training, etc.
    - Be mindful of both direct and indirect/overhead costs (ex. storage vs admin time burden/administrative costs)
        - Free credits are still costly in terms of manpower and time
    - A consumable model for thinking about costs -- estimation for popular workflows, etc.
- Next Steps
    - Standardize budget templates and guides; standardized language for grants endorsed by NCPI

- ○ Document end-to-end case studies of the entire lifecycle; balance of positives and potential drawbacks/considerations
- ○ Aggregate cost/resource modeling efforts into a spreadsheet/DB
- ○ Longer term idea: potentially have NCPI support a “free tier” for students/analysts
- ○ In-person codeathon to look at popular workflows and discuss optimization
- ○ Supplemental funding, dedicated R01s

Q&A

Stan Ahalt- agreement that this is a big social change and cloud cost questions often arise in BDC; emphasis that more work needs to be done in this space; idea of providing people with workspaces to “dabble” in might be a way to kick start this -- transition to cloud won’t be a success until we figure out these issues

Michael Schatz- agreed; want to support all researchers all the time. Will have to speak to all different levels of researchers and research projects

Becky Boyles- loved the reflection back that the conversations are happening in individual projects, but recognize the need to bring this up to the NCPI level and see what can be done collaboratively

Ben Heavner- aware of broader efforts around building cloud infrastructure; excited about NCPI being a place that internal NIH focus of intra-institute cooperation is aligned

Adam Resnick- idea of linking free tier to well established non-experimental settings so that users can create shareable products that are immediately interoperable with other NIH supported datasets

Michael Schatz- agreed; free tier would probably be very popular. In addition, full access should also be a free tier option -- drop a researcher into a workspace that’s already ready to go

Ken Wiley- like the idea; could be helpful for both established and resource limited institutions & seems like a win-win. Would take coordination across NIH

Becky Boyles- “free” means NIH paid

Stan Ahalt to Ken- is your point that internally, getting people to make the shift to using the Cloud is challenging?

Ken Wiley- there are some institutions that are more comfortable with their local systems and don’t see a need to transition to cloud; but there are other institutions that don’t have the infrastructure to do large scale genomic analysis and they are being left behind -- providing incentives/training/already established pipelines and workflows could help them get established

- Political, cultural, infrastructural barriers that all need to be addressed

Stan Ahalt- agreed; want to be able to set students up in already established workflows; seeing a similar set of challenges

Valentina Di Francesco- recent discussions with council members on this topic; how to bring genomic data science to undergraduates? Key issues from council members: Do you really need the cloud? If we are serious about teaching undergrads how to do genomic data science, this is a barrier we need to overcome. To generate a genomic data scientist, need so much educational background that it’s not clear that the teaching infrastructure in some of these universities is in place to teach all of this

James Coulombe- really like this idea though not sure how NIH could manage it; difficult to get resources for this due to conservative budgets. Important to have pilot projects so people can demonstrate feasibility and then get grant support to enable them to use the Cloud.

Michael Schatz- important question: who is our target audience?

Chat

Steven Cox: Have we explored hybrid as an alternative to the cloud/on-prem dichotomy? Seems like "both" is the likely outcome.
Stan Ahalt: I honestly think that we may look back in 10 years and conclude that, in addition to the great science that platforms accomplish, an equally significant outcome is permitting students to have access to workflows that they could not build.
Ken Wiley: We should also work on bringing in university administrators to better learn what barriers they have in being able to support students and faculty in using cloud resources
Adam Resnick: We've wondered if part of the anachronism re cloud and teaching is that these new students should have a curriculum that actually includes "learning how to use the cloud" — this is where they will have to work in the not so distant future (if we do our job right in NCPI)! Separately, while there is recognition that you can learn to do bioinformatic on "dummy data" — learning how to do SCIENCE requires access to "real" data — and Finally, we see there is a tight linkage between what you "learn" and what you end up pursuing as a scientist - which has hampered pipelining in the pediatric and rare disease landscape in many ways given the historically more limited access to such data/resources . . .
Stan Ahalt: One additional point - giving students access to pre-configured workflows in the cloud, and preferably on platforms with data, dramatically speeds up the process of learning. You side-step a LOT of time that we currently spend in getting students' *computers* configured with the software, the libraries, etc., etc. When we teach in the cloud, we can have them using Jupyter notebooks in the first class! And they love it!
Michele Mattioni: there is also the important part on democratizing access to resources -- the cloud does provide that
Asiyah Lin: @Ken in this aspect, working together with Google or Amazon, Azure to bring the general public about using cloud platforms for science maybe a strategy to go.
Jonathan Kaltman: BDC fellows are convenient but perhaps not the right target since we really want to know how to serve those naïve to these systems.
Stan Ahalt: @Adam, I agree (I actually agree with all of these!) that we are going to need to accept, and therefore teach, that we are going to be doing science differently in the future. But we professors are sloooooow to change!
I love the idea of educating the administrators!
Sai Subramanian: Great point about making the cloud platforms available for students. We have some excellent use cases where some universities are using the Seven Bridges CGC to train students at the Masters level. We did a user interview with one such team (more details and interview transcript is here https://www.cancergenomicscloud.org/teach-and-train-using-cgc) who highlighted the barriers of using cloud resources for teaching and training
Stan Ahalt: Being able to search, find, assemble both data and code is a dramatic reduction in the "friction" of doing science.
+1 Sai !

Report Back: Search (slides)

Rogers

- Gaps
    - Understanding/validating/documenting search personas and use cases
    - Gaps in finding studies-- understanding how data is consented and how to apply for access; how to search over data for specific characteristics
    - Gaps in building cohorts-- findings/gaining access to different search portals; difficulty in sending search results to a workspace
- Next Steps
    - Form a Search WG that would...
        - Conduct UX research to determine specific use cases & recruit users
        - Create a list of search components and documentation on how to use each piece
        - Develop a way to collect feedback for troubleshooting
        - Create a search taxonomy to inform a search roadmap
        - Link back to studies in NCPI dataset catalogue
        - Generate input for upcoming RFI: NOT-OD-21-187
        - Gain an understanding of what FHIR is and integrate this into the search strategy -- potential integration with FHIR WG

Q&A

Stan Ahalt- do we need a deeper discussion about other searches (not just the searchers that we're used to)? -- need to describe when you might use one versus the other

Steve Cox- agreement; document what the different searches are and how they differ from each other; also, wanted to point out interest in hearing from people that use search in different/interesting ways

Dave Rogers- being able to index and build upon results would be an interesting area to explore

Becky Boyles- what NCPI opportunities can we leverage?

Steve Cox- important conversation about metadata and curation; using FHIR but also going beyond FHIR to address semantic interoperability

Robert Carroll- FHIR is nice to be able to use semantics that are interoperable as well as create new ones

Steve Cox- the burden here is placed on curators

Stan Ahalt- if we use FHIR, hoping to be able to define standards by use

Robert Carroll- architecture question: how useful will this be if everyone can use their own vocabularies? But how rigid can we be?

Dave Rogers- use a standards based approach to submit to certain repositories

Steve Cox- incumbent on us to at least have that strategy; a lot of people are actually looking for guidance

Report Back: Other Interoperability Efforts (slides)

Ahalt

- Gaps and key blockers
    - Need to understand/define clear use cases from real world applications in order to identify next steps -- there is demand for interoperability
    - Need to be able to search across platforms -- use cases are emerging here
- Next Steps
    - Actively seek out researchers that want interoperability features
    - Standardize how tools/apps are deployed across the ecosystem so we can run the apps that we want in order to achieve our computational goals, regardless of platform
    - Develop clear methods for publishing completed use cases -- use for replication as well as training (youtube videos)
    - Look for opportunities to train others in the space
- Need national- and university-level leadership to think about what it will take to get this “right”

Q&A

Valerie Cotton- discussed the value of having multiple portals vs a centralized portal; hearing from communities that community-specific portals are valuable
Brian O’Connor- +1 idea of publishing beyond a white paper and turning these into tutorials (ex. Publishing working environment as an artifact) -- can clone someone’s workspace and see exactly how the analysis was done (bonus: this would be a relatively minor lift that would add a lot of value)
David Pot- this is being done in some places; bringing interactivity to end users
Becky Boyles- closing comments; need to acknowledge that many of these issues are “people” issues. Reflect on making cultural change.

Chat

Robert Carroll: And to build and refine over time.
Stan Ahalt: An emergent standard is possible, I think.
Jonathan Kaltman: What is the resource requirements for retrospectively fitting data into FHIR standard? Most data we have is already collected, is not EHR derived, and is not standardized to FHIR.
Robert Carroll: On the row-level data, two layers:
1. "standardizing" it is mainly a reformatting option- I believe it's possible to lift up CSVs with dbGaP DDs into FHIR. You don't gain any semantic information here, but it's consistent and accessible.
2. "Harmonizing" to standard vocabs, trying to establish timelines of events, etc, is a lot of manual work, though it can be aided by tools. I know LHC at NLM is interested in some of this, too.
Robert Carroll: I think we can get #1 that works for extant dbGaP data and puts it at least on a structural level comparable with fully harmonized data. That's something they are working on from what I understand.
Stan Ahalt: Executable papers!
Anne Deslattes Mays: People do that -

Use of Zenodo - JupyterLab notebooks — literate programming
Stan Ahalt: Yes, let’s all do it this way!
Stan Ahalt: These emergent tools are very cool!
Anne Deslattes Mays: And community based standard workflows such as cwl, wld, and Nextflow —
Dave Rogers: Executable papers = reproducible papers!
Anne Deslattes Mays: It works very nicely -

GA4GH Relationship (slides)

Notes:
Brian O’Connor presented a general summary of GA4GH (Global Alliance for Genomics and Health) as they relate to interoperability & NCPI

1. GA4GH Organization
    a. Work Streams - develop standards, tools, frameworks to overcome international genomic data sharing. Works closely with driver projects
    b. Driver Projects - Using GA4GH tools/standards/frameworks to enable real projects, thus providing guidance on GA4GH standards development
2. Current Driver Projects:
    a. NCI CRDC (National Cancer Institute Cancer Research Data Commons)
    b. NCI Genomic Data Commons
    c. NHLBI Trans-Omics for Precision Medicine (TOPMed)
3. What GA4GH Standards are Used by NCPI today?
    a. Passports and Authentication & Authorization Infrastructure (AAI)
    b. Data Repository Service (DRS)
    c. Tool Registry Service (TRS) used by workspaces - Terra/SB bring in workflows from Dockstore using TRS
    d. Various file formats maintained by GA4GH e.g., CRAM, SAM/BAM, VCF/BCF
    e. Others from the Audience: None mentioned
4. What new opportunities are possible with GA4GH?
    a. New API possibilities
    b. Other things: Starter Kit; Technical Alignment Sub-Committee (TASC), Federated Analysis Systems Project (FASP)
    c. What new standards might we want to propose? None mentioned.
5. Overview of GA4GH FASP (Federated Analysis Systems Project) - demonstrate GA4GH standards using actual researcher use cases. Use case #7 (Tim Marjorian’s use case)
    a. FASP is a great conduit to get feedback for GA4GH products
6. Engagement Opportunities with GA4GH
    a. GA4GH Connect OCT 12-14
    b. GHIF Nov 16-17
    c. FASP Regular Bi-weekly Meetings
    d. GA4GH EDI Advisory Group: info@ga4gh.org

Questions:

- Valentina Di Francesco: “GA4GH has been trying to engage NIH, but is running up against having to deal with 27 institutes. What is the proper engagement interface of GA4GH with NIH?
    - NCPI is a fantastic driver since it brings together so many NIH folks.
    - Next steps: reach out to GA4GH leadership
- Asiyah Lin: “What does it take to become a driver project for GA4GH?”
    - Driver projects need some level of commitment, consistent engagement with one or more workstreams, dedicated staff.
    - Take-away: Have a 1-on-1 with GA4GH folks and figure out what next steps are.

Chat:

- Anne DeslattesGithub: Reproducibility note: how can I reproduce the Jupyter Notebooks analysis link
- Valntina Di Francesco: We/NHGRI has been testing DUO extensively
- Valerie Cotton: thankfully the NIH Genomic Data Sharing Policy Institutional Certification already aligns pretty well with DUO. NIH Data Use Limitation definitions: https://osp.od.nih.gov/wp-content/uploads/NIH_PTC_in_Developing_DUL_Statements.pdf
- Ben Harver: Are there use cases around study registration that might lead to a need for new standards?
    - Melissa Wilson: I'm involved in one that might.
- Valentina Di Francesco: NHGRI is exploring DUO during study registration, As well as for data access requests
- Melissa Wilson: Due to small scale population data use approval on a case by case with the community population. But maybe this already exists?
- Meen Chul Kim: FYI, GA4GH Pedigree WG presents a new pedigree ontology (OWL) and a new pedigree model, and their implementations in FHIR on 10/12.
- Valerie Cotton: +1 Valentina. When we spoke to GA4GH about KF being a driver project, they were clearly not interested in more NIH drivers. Which is fine - we will coordinate where it makes sense.
- Rebecca Boyles: I do wonder if we can be move effective by coordinating our representation and interest in GA4GH across NCPI
- Rebecca Boyles: Maybe we don’t need to be a driver project if we can better leverage the seats at the table we have.
-

Use Case Overview: The Journey of a NCPI Use Case (Lin) (slides)

Notes:

- Asiyah Lin overview of linked slides →
- NCPI Use Case:
    - Access and analyze of data from NCPI platforms (n>=2) is needed to answer a scientific question
    - Ultimate goal: to drive the development of NCPI interoperability technology

specification

- Phases (see slides for full content): Proposal → Implementation → Dissemination
- Training video example: https://anvilproject.org/ncpi#demo-of-search-result-hand-off
- Use Case Tracker: https://github.com/NIH-NCPI/NCPI_use_case_tracker/issues
- Q/A:
    - Valentina - Structure around use cases much needed;
    - Asiyah: Welcomes additional feedback; all necessary code/tools will be made available to users

Chat:

- Valerie: Valentina. This is extremely important moving forward for coordination and transparency
- Anne: The code that was used to implement are they in GitHub as well?
    - Asiyah: All codes will be available for users to experiment
- Michael: Is there dedicated funding available to support cloud costs for interoperability use cases?
    - Jack: This funding has come from a couple different sources to date. e.g. CRDC, KF, ODSS, BDC, etc I _think_ AnVIL

Slack Chat

-

Current Scientific Use Case: Leveraging Functionally Equivalent Pipelines for Long-Read Data on Different Systems (Plon/Hirschi) (slides)

Notes

- Owen Hirschi overview of linked slides;
- Long-read sequence analysis tools uploaded on these platforms exist in different coding languages
- We have set up a functional long-read sequencing analysis pipeline on the CAVATICA platform
- We have been able to identify *de novo* variants previously found via pipelines on the Terra platform
- We have also identified a 5 to 10% difference in raw and merged structural variants across the two platforms
- Ongoing work:
    - Understand differences in called *de novo* events and aligned sequence files in HG002 trio on both platforms
    - Determine if there is an larger data set we can process on CAVATICA and Terra respectively to test full functional equivalence
    - Perform long-read sequence analysis on BASIC3 cohort using the pipeline on CAVATICA to identify novel *de novo* structural variant
- Q/A: *Session Recorded*
    -

Chat:

-

Slack Chat

-

Current Scientific Use Ce DRSAase: Interoperability between Kids First & Undiagnosed Diseases Network (UDN) Data via dbGaP/SRA (Cotton/Heath) (slides)

Notes

- Valerie Cotton and Allison Health overview of linked slides; wide spectrum of involvement
- Key goal to utilize RAS and DRS
- Build DRS links into the project (SevenBridges proposal)
- Q/A:
    - Ben Heavner: End-to-end state -- end state “for now” vs. forward looking analysis (reproducibility considerations);
    - Adam Resnick: “House MD” work not using Jupyter notebooks;

Chat:

-

Slack Chat

-

Current Scientific Use Case: Genetic Sex as a Biological Variable and X-inactivation (Wilson) (slides)

Notes

- Melissa Wilson (Associate Professor at ASU) overview of linked slides
- Q/A:
    - Michael Schatz:
        - Single cell data -- is there such data within the NCPI universe that would add such power?
    - INCLUDE Project/down syndrome samples;
    - Stan Ahalt - supports calling the x chromosomes;
    - Adam Resnick - further advocates from the Kids First side;

Chat:

-

Slack Chat

-

Current Scientific Use Case: Conducting reproducible science in PIC-SURE interoperating with Seven Bridges/Terra (Makwana) (slides)

Notes

- Simran Makwana and Paul Avillach (HMS/PIC-SURE) overview of linked slides
- Anyone with an eRA Commons ID can access PIC-SURE
- dbGaP challenge in working with multiple studies and consent groups; PIC-SURE API simplifies the user journey
- PIC-SURE/Seven Bridges interoperability use-case in analysis of the ORCHID Study
- PIC-SURE/Terra interoperability use-case in Sickle Cell Disease Study
- PIC-SURE option as search tool across NCPI platforms; clinical, sequencing, biosamples, index files
- Q/A:
    - Paul - Easiest technical approach would be centralized / most difficult policy-wise and unlikely to occur; a mixed/hybrid solution may be most viable path forward
    - Asiyah - Status of the use case, if a mature use-case, how would be promote as training material
        - Jessica Lyons - PIC-SURE wants to move beyond just interoping with Seven Bridges and Terra (as in BDCatalyst) -- we are not there yet (e.g. AnVil)

Chat:

-

Slack Chat

-

Synthesize Goals and Next Steps for the next 6 Months, with focus on driving use cases (slides)

Notes

- Stan Ahalt presented slides re: PFB, FHIR, RAS, End user cloud costs, Search, Other interop effort, GA4GH & Use cases
- Jon Kaltman: Kudos to organizers; Growing level of maturity of NCPI group is evident & it's an exciting time
- Discussions from NIH-only breakout:
    - Chris Kinsinger: Is it within NCPI's scope to provide a service where NCPI could check and confirm implementation of best practices?
    - Valentina Di Francesco: RAS is a bottleneck we should fix ASAP; discussed establishing Search working group, training for students & admins, and making cloud space available free for students
    - Alastair Thomson: ODSS could provide regular touchpoints to maintain focus across platforms

    - Asiyah Lin: Search is a community need; need to determine what type of group to form & who can lead it; compare tools; set priorities; NCPI community can be more integrated; could collaborate with big vendors for training
    - James Coulombe: Should make transitioning to cloud easier through increased training across NIH
    - Q/A:
        - Valerie Cotton: With establishment of search WG, should we document potential use cases & disseminate them before scheduling meetings?
            - Stan Ahalt: Should adopt principle for all WGs & TTs that we start with defining use cases which should reduce competition; want to demonstrate legitimate differences
        - Becky Boyles: We have an opportunity to collect resources on overlapping activities like training & GA4GH and make them more impactful & efficient
            - Jon Kaltman: Agree, and this will be key to all our successes; need to continue innovating and expanding user base while conserving resources where possible
        - Stan Ahalt: NCPI work is an addendum for many folks but has resulted in great progress; how can this be communicated to NIH leadership?
            - Jon Kaltman: Focusing on use cases is very helpful and speaks to NIH leadership; promoting discovery through scale or speed or variety of data is a tremendous selling point; standardized use metrics can be developed to show we're growing user base; there are discussions of cultural barriers getting onto the cloud which need to be addressed to grow user base which is critical; need to find users who are having trouble finding data and understand their needs and address them through technology; showing growth through metrics and showing that science is getting done that couldn't be done otherwise are significant selling points
            - Mike Feolo: Existential justification needed by leadership will take the form of use metrics; we need expert trainers that aren't necessarily from each platform but knowledgeable about cross-platform use; NIH could make suggestions
            - James Coulombe: The value of large data efforts is deeply ingrained in NIH leadership

Chat:

-

Slack Chat

-