NCPI Spring 2021 Workshop Agenda

Monday, May 3, 2021 from 11:00am to 4:30pm EDT
Tuesday, May 4, 2021 from 11:00am to 4:30pm EDT
Hosted Remotely by the NCI Center for Cancer Data Harmonization (CCDH)
WebEx Link

Helpful Links

Recording	**Day 1**	**Day 2**
Slide Decks	**Day 1**	**Day 2**
Meeting Notes	**Day 1**	**Day 2**
Chat Transcript	**Day 1**	**Day 2**

End of Meeting Feedback Poll
May 2021 EasyRetro Board

Result of Fall 2021 Meeting Dates Poll: **Fall meeting will be Oct 5-6, 2021**

Background

First organized in 2019, the NCPI's goal is to establish and implement guidelines and technical standards to empower end-user analyses across the four participating platforms and facilitate the realization of a trans-NIH, federated data ecosystem. NCPI is a collaborative project between five NIH Institutes and Centers (NCI, NHGRI, NHLBI, NIH Common Fund, and NCBI) as well as external partners comprising five Working Groups: Coordination, Community Governance, FHIR, Outreach and Training, and NIH Systems Interoperation.

This workshop will be held online, free of cost, on May 3rd and 4th, 2021, 11:00am-4:30pm EDT. With updates of NCPI Working Group progress, Community Interoperability talks, and six Breakout Sessions over two days, we hope to engage participants in conversation and draw insight from everyone's expertise.

Please feel free to reach out to us with general questions or comments by emailing Digant Shah at digant.shah@nih.gov. If you have any issues accessing the NCPI BOX folder, please contact Asiyah Lin at asiyah.lin@nih.gov.

All 20 min topics are to be 15 min presentation and 5 min for questions, all 15 min topics are to be 12 min presentation and 3 min for questions

Day 1: Monday, May 3

11:00am-12:30pm – 1.5 hours: Welcome and Working Group Updates

10 min Welcome

Sam Volchenboum (UChicago) and Tanja Davidsen (NCI)

20 min NIH Coordination Group

Valentina Di Francesco (NHGRI)

20 min Community/Governance

Bob Grossman (UChicago) & Stan Ahalt (RENCI)

20 min Systems Interoperability

Jack DiGiovanna (Seven Bridges)

20 min FHIR

Allison Heath (CHOP) & Eric Torstenson (Vanderbilt)

12:30-1:00pm – 30 min break

1:00-1:20pm – 20 min: Working Group Updates continued

20 min Outreach and Training

Anton Nekrutenko (PSU)

1:20-2:30pm – 1 hour, 10 min: Three *Concurrent* Breakout Groups

10 min State-of-the-topic Introduction by Facilitator

60 min Group Discussion

Topic 1: Data harmonization and interoperability, including models, terminologies, mapping, provenance - *Chris Chute (JHU) & Tricia Francis (JHU)*

Topic 2: Search – *Kathy Reinold (Broad) & Steven Cox (RENCI) & Jay Ronquillo (NCI)*

Topic 3: RAS interoperability – *Andre Paredes (UChicago) & Brian O'Connor (Broad)*

2:30-3:00pm – 30 min break

3:00-3:20pm – 20 min: NCBI's Journey in Support of a Federated Cloud Data Sharing Ecosystem

Mike Feolo (NCBI)

3:20-4:20pm - 1 hour: Breakout Groups Report Back

20 min Topic 1: Data harmonization and interoperability, including models, terminologies, mapping, provenance - *Chris Chute (JHU) & Tricia Francis (JHU)*

20 min Topic 2: Search – *Kathy Reinold (Broad) & Steven Cox (RENSI) & Jay Ronquillo (NCI)*

20 min Topic 3: RAS interoperability – *Andre Paredes (UChicago) & Brian*

O'Connor (Broad)

4:20-4:30pm - 10 min: Conclusion of day

10 min Summary/Wrap Up

Sam Volchenboum (UChicago) and Tanja Davidsen (NCI)

Day 2: Tuesday, May 4

11:00am-12:30pm – 1.5 hours: Welcome and Community Interoperability Talks

10 min Welcome

Melissa Haendel (U of Colorado) and Tanja Davidsen (NCI)

15 min NHLBI External Speaker

Proof of concept of interoperable approaches for improving outcomes of pediatric diseases

Tim Majarian (Broad Institute)

15 min Kid's First External Speakers

Kids First and Multi-Cloud BASIC3

Sharon Plon (Baylor College of Medicine) & Owen Hirschi (Baylor College of Medicine)

15 min NCBI External Speaker

Analyzing Gene Fusions on NCI and St Jude Cloud

Jinghui Zhang (St. Judes)

15 min NHGRI External Speaker

Cloud-Based Whole Genome Sequencing Analysis Workflow

Xihong Lin (Harvard)

15 min NCI Host Speakers

NCI CRDC Center for Cancer Data Harmonization efforts

Sam Volchenboum (UChicago) & Melissa Haendel (U of Colorado)

12:30-1:00pm – 30 min break

1:00-1:20pm – 20 min: Community Interoperability Talks Discussion

Group discussions on topics covered in morning Community Interoperability talks led by *Adam Resnick (CHOP)*

1:20-2:30pm – 1 hour, 10 min: Three *Concurrent* Breakout Groups

10 min State-of-the-topic Introduction by Facilitator

60 min Group Discussion

Topic 1: Data/Tools/Workflow/Compute Interoperability and Functional Equivalence - *Jack DiGiovanna (Seven Bridges) & Michael Schatz (JHU)*

Topic 2: Cloud Costs & Benchmarking – *Alex Baumann (Broad) & David Pot (GDIT)*

Topic 3: Governance – *Stan Ahalt (RENCI) & Bob Grossman (UChicago)*

2:30-3:00pm – 30 min break

3:00-3:20pm - 20 min: The Future of Interoperability - featuring results of the pre-meeting EasyRetro Board activity

Brian O'Connor (Broad)

3:20-4:20pm – 1 hour: Breakout Groups Report Back

20 min Topic 1: Data/Tools/Workflow/Compute Interoperability and Functional Equivalence - *Jack DiGiovanna (Seven Bridges) & Michael Schatz (JHU)*

20 min Topic 2: Cloud Costs & Benchmarking – *Alex Baumann (Broad) & David Pot (GDIT)*

20 min Topic 3: Governance – *Stan Ahalt (RENCI) & Bob Grossman (UChicago)*

4:20-4:30pm – 10 min: Conclusion of meeting

10 min Summary/Wrap Up (announcement of fall meeting dates, poll for feedback on the meeting)

Melissa Haendel (U of Colorado) and Tanja Davidsen (NCI)

Day 1 Breakout Group Descriptions

1. **Data harmonization and interoperability, including models, terminologies, mapping, provenance** - *Chris Chute (JHU) & Tricia Francis (JHU)*
   Data interoperability is in the eye of the beholder. Realizing discovery and analytics across different sources, domains, and data modalities is one of the most challenging aspects of modern data science. We will discuss addressing tensions in the use of common data models for observational and research data, terminologies and challenges in harmonizing them, provenance and how to encode it, and identifier management to support it.
2. **Search** – *Kathy Reinold (Broad) & Steven Cox (RENSI) & Jay Ronquillo (NCI)*
   Searching and finding relevant datasets is a challenge across all parts of NIH. Who is searching for data? What are they searching for? What are your favorite search features? How do your systems currently enable searching, and how would you like your systems to evolve? What "facets" are important? What percentage of your users do this through Graphical Interfaces, versus programmatic APIs? The intent of this session is to envision the future of search; your active participation is requested!
3. **RAS interoperability** – *Andre Paredes (UChicago) & Brian O'Connor (Broad)*
   Users of various platforms have to juggle multiple logins and identities and these platforms currently have to use a variety of login approaches and mechanisms for determining user authorization to access datasets. The NIH RAS project promises to provide a modern way to log in users as well as describe their dataset authorizations using GA4GH Passports. Many (most) of the NCPI systems now support user login with RAS yet user authorization information from RAS is not yet being leveraged by these systems. The reasons for this are varied but one area needing agreement is how passports should be used by systems to access data over the GA4GH DRS API. This breakout will examine this, and other, topics in order to drive consistency across NCPI systems (AnVIL, BDCatalyst, CRDC, and Kids First).

Day 2 Breakout Group Descriptions

1. **Data/Tools/Workflow/Compute Interoperability and Functional Equivalence** - *Jack DiGiovanna (Seven Bridges) & Michael Schatz (JHU)*
   The key use-case we'd like to address here is "as a researcher, I need to analyze data in two or more Platforms". This inverts the current NCPI Systems Interoperability approaches of "make it easier for researchers to bring all the data they need to a single platform". It also introduces a foundational challenge as Platform's description language support and functionalities are not equivalent (e.g., a Terra user leveraging WDL pipelines cannot run them directly on CAVATICA and vice versa). Here we need to focus and operate within reality. This should **not** be a tool bake off, converter bake off, or

brute-force search of possible solution space. **As an outcome of this breakout, we should identify (i) a limited set (≤ 3) of analyses (e.g. long-reads, transcriptomics) to focus on; (ii) two environments for each analysis; (iii) a limited set of humans to serve as driving use-cases; (iv) potential metrics/metadata to evaluate pipeline/workflow similarity.**

2. **Cloud Costs & Benchmarking** – *Alex Baumann (Broad) & David Pot (GDIT)*
   Users of our various platforms many times come with fixed budgets and time limits, often tied to grant funding. These users would like to be able to predict their costs and compute times before writing grants, want to track the burndown of time and money along the way, and often have very little flexibility to handle time or (especially) cost overruns. In this breakout, we'd like to hear from you about what aspects of cloud costs and compute benchmarking have been the biggest hurdles to users, and what solutions you have found that have been most successful. How have you found success/challenges in educating users about best practices and how to use the cloud cost-effectively?
3. **Governance** – *Stan Ahalt (RENCI) & Bob Grossman (UChicago)*
   The goal of the Governance working group is to come to a common understanding that supports the interoperability of the cloud platforms participating in NCPI. In this breakout group we will discuss two proposed considerations for interoperating cloud platforms and the underlying frameworks that support them, including 1) the underlying NIH policies, such as Genomic Data Sharing; 2) security frameworks, such as NIST 800-53; and, 3) and, additional considerations, such as the role of data trustees/stewards, requirements in Interconnection Security Agreements (ISAs), and interpretations of egress requirements and/or constraints. The goal of this meeting is to discuss the two considerations so that we can take a trial vote about them.