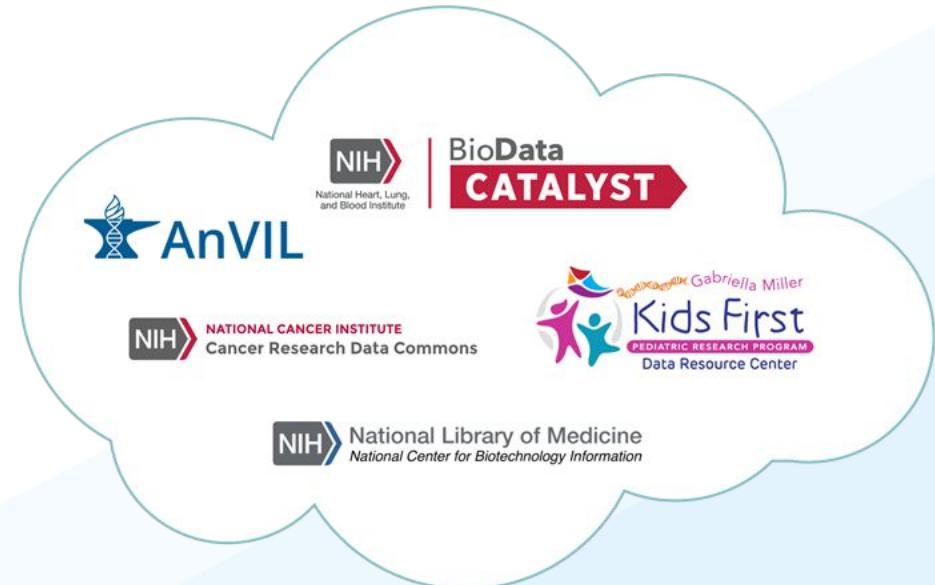


Welcome to the...

NIH Cloud Platforms Interoperability Spring 2021 Workshop

We'll be starting shortly!

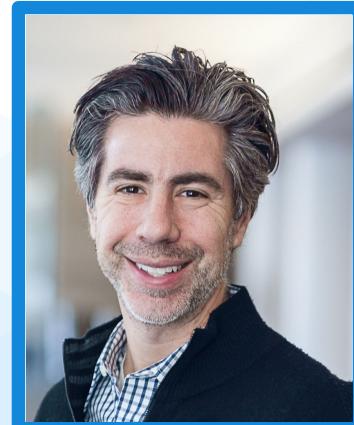


May 3 & 4, 2021 11:00am-4:30pm EDT

tinyurl.com/NCPIagenda

Welcome – NCPI Spring 2021 Workshop Day 1

Samuel Volchenboum
University of Chicago



Tanja Davidsen
National Cancer Institute



Logistics

- Please use the **WebEx application** and not a browser
- Please mute when not speaking
- We will be recording all the sessions except the breakout sessions
- Notes will also be taken during the sessions
- Speakers please turn your camera on when speaking
- If you have not registered, please do: **tinyurl.com/NCPIregistration**
- Agenda: **tinyurl.com/NCPIagenda**
- Fall 2021 Workshop poll: **tinyurl.com/NCPIfallpoll**

Agenda

Day 1: Monday, May 3

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

12:30-1:00pm – Break

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

1:20-2:30pm – Three Concurrent Breakout Groups

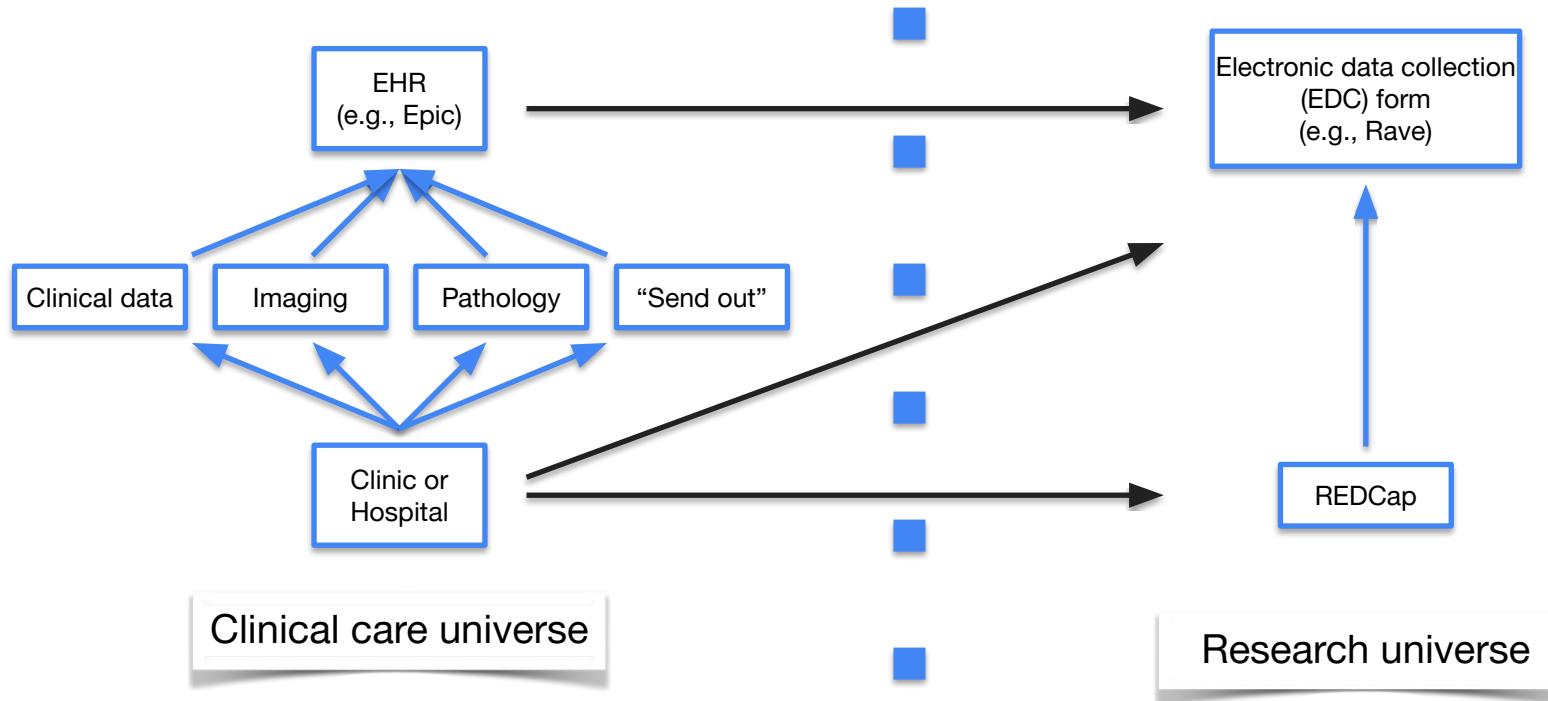
2:30-3:00pm – Break

3:00-3:20pm – NCBI talk

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

4:20-4:30pm – Wrap Up

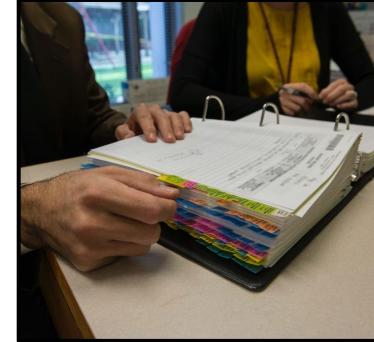
Parallel universes



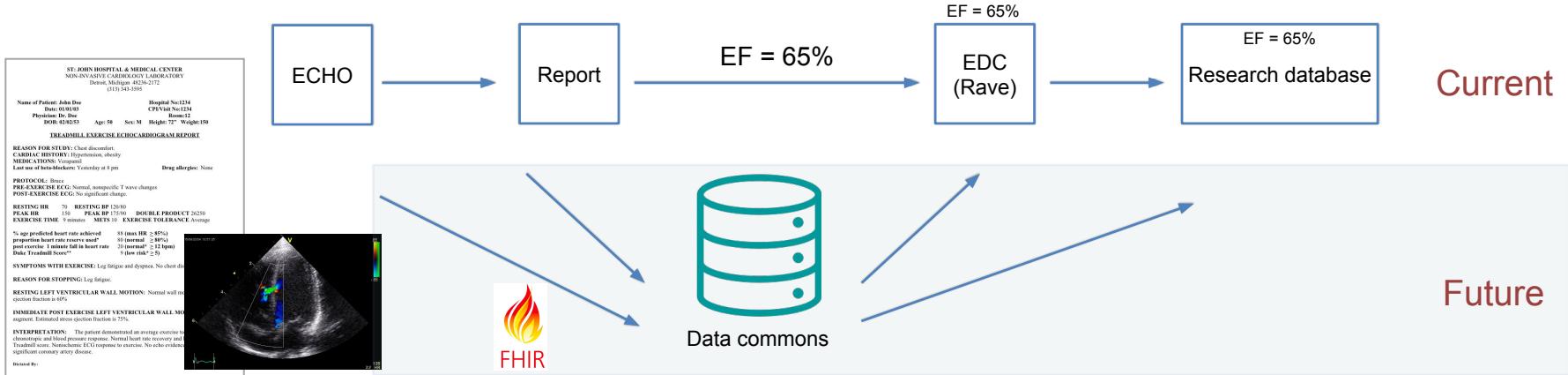
Arcane manual processes

Activation Date: 8/8/01 Version Date: 9-4-01
NANT 99-02
TITLE: MODULATION OF INTENSIVE MELPHALAN (L-PAM) BY BUTHONINONE SULFOXIMINE (BSO) WITH AUTOLOGOUS STEM CELL SUPPORT FOR RESISTANT/RECURRENT HIGH-RISK NEUROBLASTOMA (IND # 63,112)

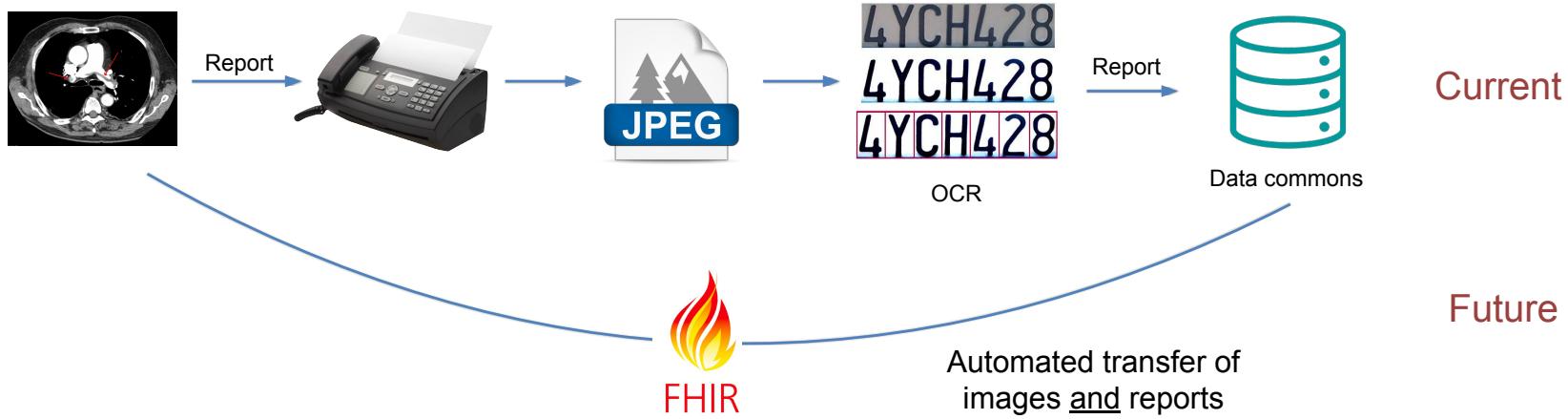
CTCAE CODE (per protocol)	CTCAE SHORT NAME (per protocol)	Current Grade	Maximum Grade This Course* (for occurrence)	Maximum Grade This Occurrence **	Attribu- tion	Date Onset	Resolved Yes / No	Date Resolved
10020943	Hypo albuminemia	0	1	1	3	05/23	Yes	06/02
10020943	Hypo albuminemia	0	1	1	3	06/05	Yes	06/09
10021038	Hypo natremia	0	1	1	3	05/26	Yes	05/27
10021038	Hypo natremia	0	1	1	3	05/28	Yes	06/02
10021059	Hypo phosphatemia	0	1	1	3	05/09	Yes	05/15
10021059	Hypo phosphatemia	0	1	1	3	05/23	Yes	06/05
10018236	Fatigue	0	1	1	3	05/23	Yes	05/26



The information funnel



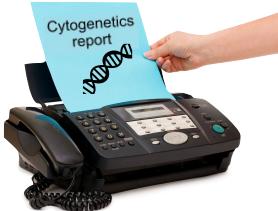
Lossy information transfer



Legacy data transfer methods



Cytogenetics lab



Fax



Medical center



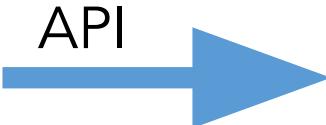
Clinical research assistant

PARTICIPANT INFORMATION	
Participant Number	_____
Study Group	_____
Study Site (Health Centre Name)	_____
Inclusion/exclusion criteria	<input type="checkbox"/> Met all <input type="checkbox"/> Not met <input checked="" type="checkbox"/>
Date of Informed Consent	_____
Date of Birth	_____
Gender	<input type="checkbox"/> Male <input type="checkbox"/> Female
Pregnant	<input type="checkbox"/> Yes <input type="checkbox"/> No <input type="checkbox"/> Unknown
If pregnant, Estimated Gestational Age _____ weeks	
Date of Enrollment	_____
Has malaria in the last 28 days	<input type="checkbox"/> Yes <input type="checkbox"/> No <input type="checkbox"/> Unknown
Has antimalaria in the last 28 days	<input type="checkbox"/> Yes <input type="checkbox"/> No <input type="checkbox"/> Unknown

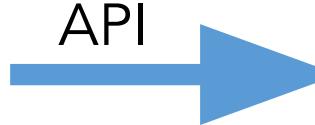
Case report form



Cytogenetics lab



Data commons



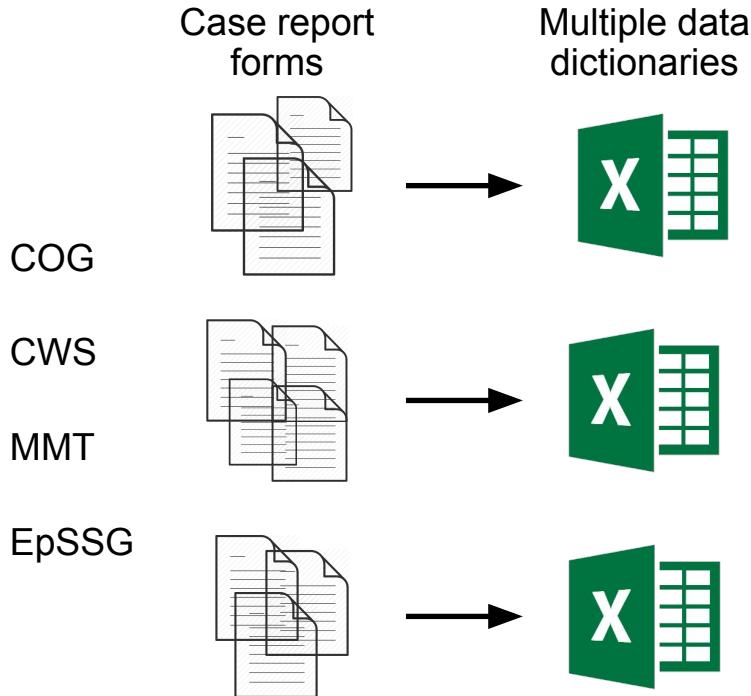
Research database

Manual field mapping

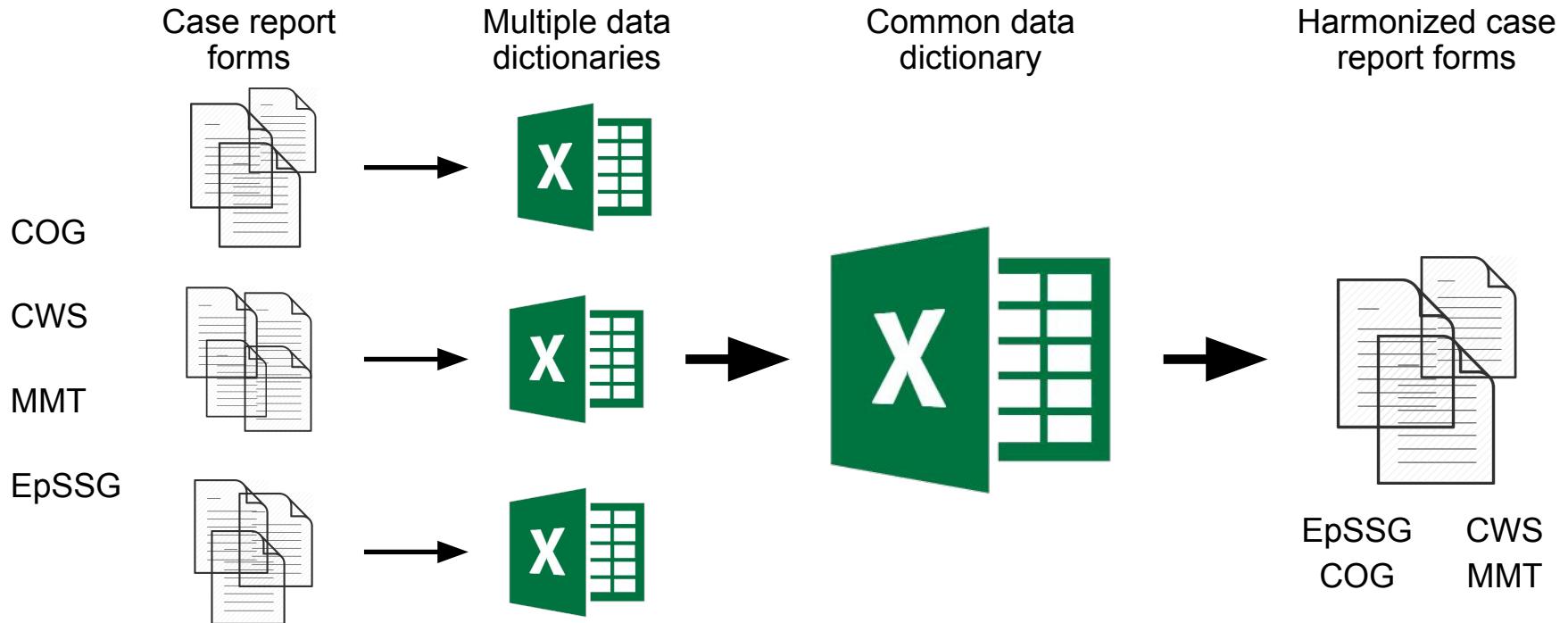
Row	LOINC #	Component	System	Ex. Units	Method	%99.+. . .	Long Common Name
19	30350-3	Hemoglobin	BldV	g/L;g/dL			Hemoglobin [Mass/volume] in Venous blood
18	30351-1	Hemoglobin	BldMV	g/dL			Hemoglobin [Mass/volume] in Mixed venous blood
16	30353-7	Hemoglobin	BldCoV	g/dL			Hemoglobin [Mass/volume] in Venous cord blood
17	33025-8	Hemoglobin	BldCoV	g/dL	Calculated		Hemoglobin [Mass/volume] in Venous cord blood by calculation
14	30354-5	Hemoglobin	BldCoA	g/dL			Hemoglobin [Mass/volume] in Arterial cord blood
15	33026-6	Hemoglobin	BldCoA	g/dL	Calculated		Hemoglobin [Mass/volume] in Arterial cord blood by calculation
13	40719-7	Hemoglobin	BldCo	g/L;g...			Hemoglobin [Mass/volume] in Cord blood
12	30352-9	Hemoglobin	BldC	g/dL			Hemoglobin [Mass/volume] in Capillary blood
11	14775-1	Hemoglobin	BldA	g/L	Oximetry		Hemoglobin [Mass/volume] in Arterial blood by Oximetry
10	30313-1	Hemoglobin	BldA	g/dL			Hemoglobin [Mass/volume] in Arterial blood
21	61180-6	Hemoglobin	Bld^fetus	g/L			Hemoglobin [Mass/volume] in Blood from Fetus
20	54289-4	Hemoglobin	Bld^BPU	g/dL			Hemoglobin [Mass/volume] in Blood from Blood product unit
8	20509-6	Hemoglobin	Bld	g/dL;... .	Calculated	0.2679%	Hemoglobin [Mass/volume] in Blood by calculation
7	718-7	Hemoglobin	Bld	n/dl : .		2.3221%	Hemoglobin [Mass/volume] in Blood
9	55782-7	Hemoglobin	Bld	g/dL	Oximetry		Hemoglobin [Mass/volume] in Blood by Oximetry
22	41995-2	Hemoglobin A1c	Bld	g/dL			Hemoglobin A1c [Mass/volume] in Blood

Which hemoglobin maps to the one requested in the clinical trial?
(spoiler: don't know - protocols rarely utilize standardized codes)

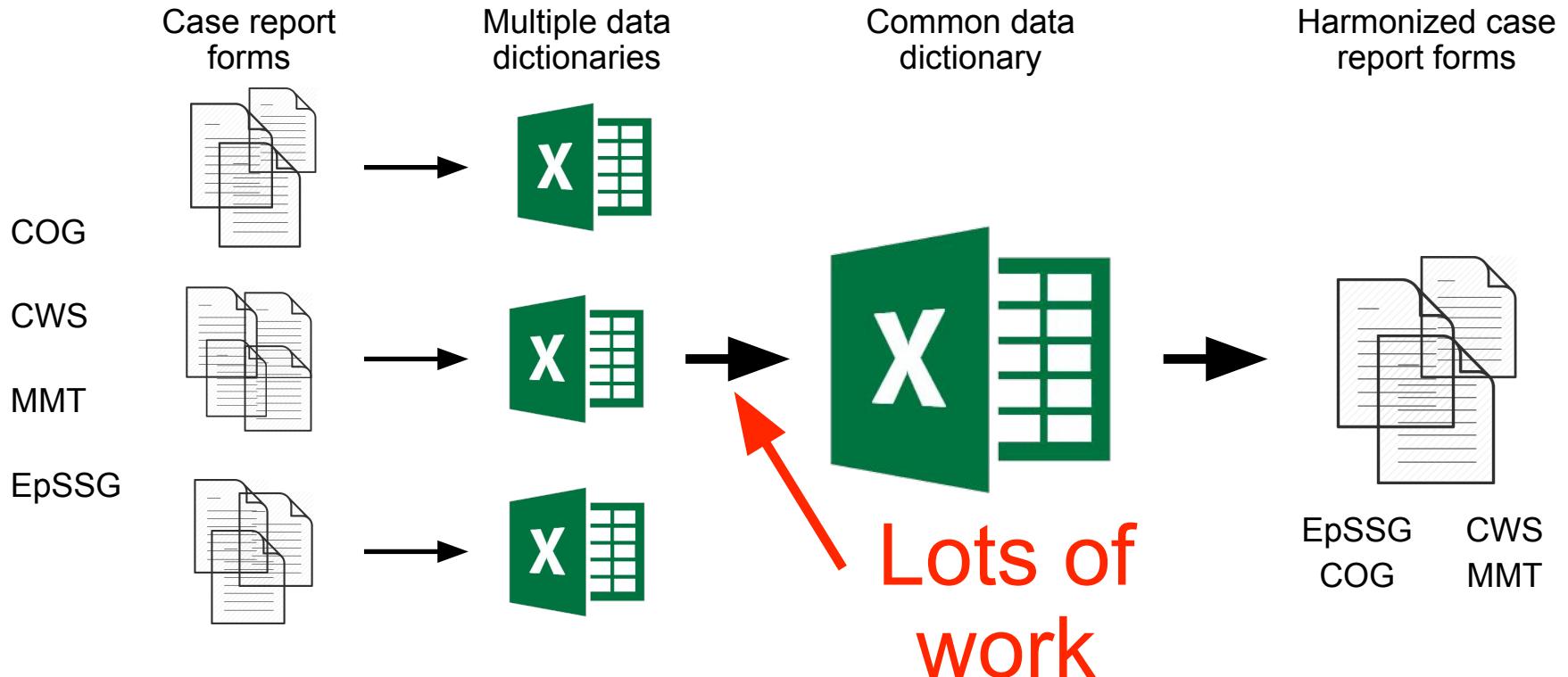
Lack of harmonization across groups



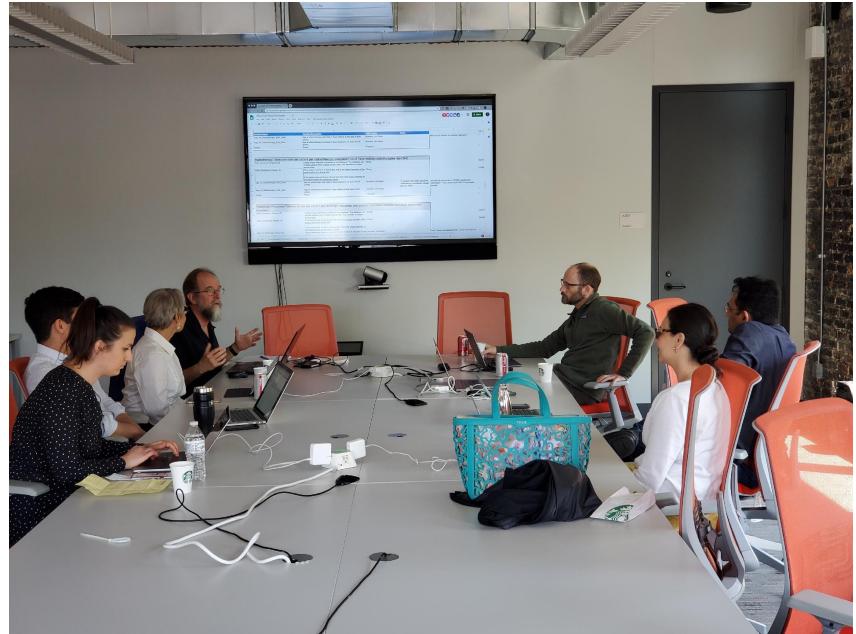
Lack of harmonization across groups



Lack of harmonization across groups



Data dictionary development



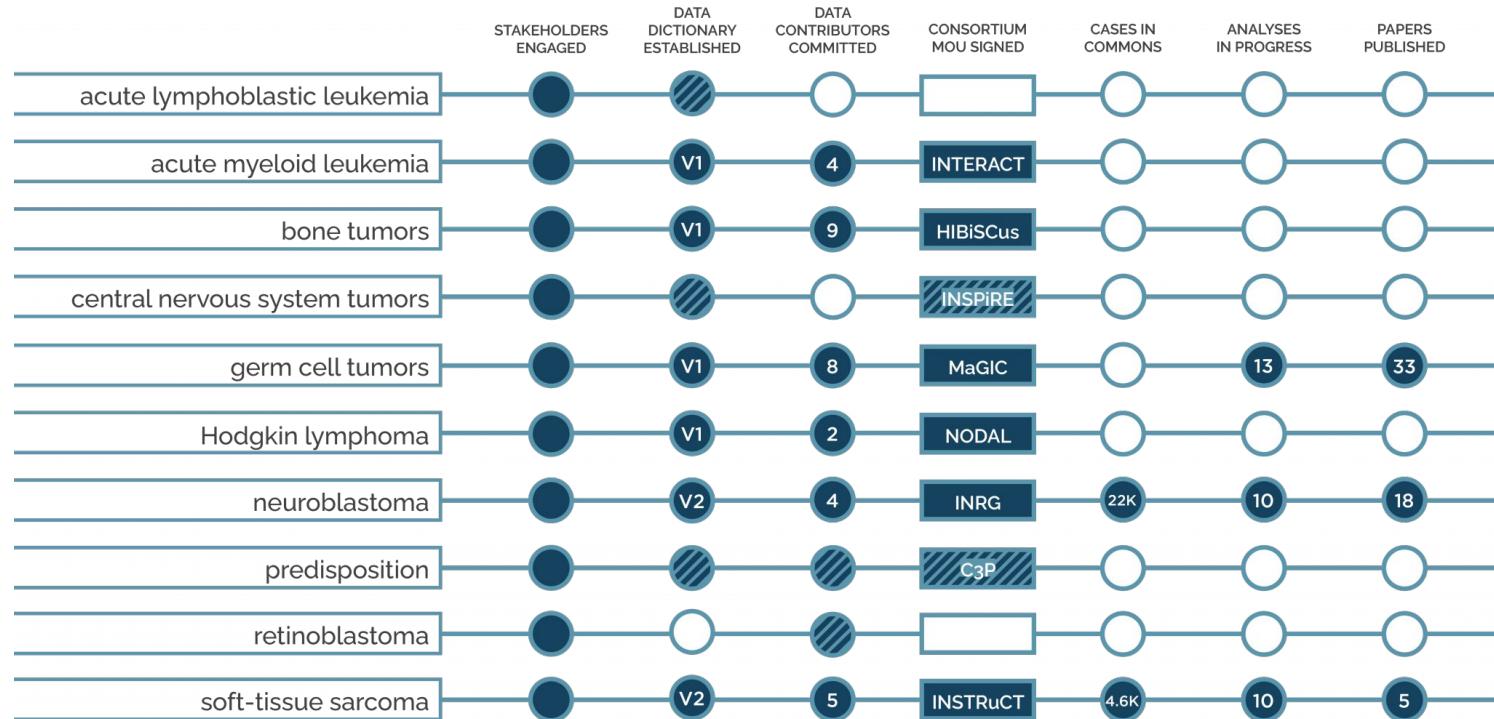
Example - RMS site of disease

Major Primary Site	CWS	COG	EpSSG/MMT Name
HEAD & NECK	Orbit	2=Orbit	Orbit
	Scalp	10=Scalp	Soft tissue of scalp External auricular canal Ear soft tissue, external ear Temporal muscle
	Parotid	9=Paratoid	Parotid, soft tissue
	Oral Cavity	7=Oral cavity	Gum Base of tongue Lip Lower lip Upper lip Tongue
	Larynx	5=Larynx	Larynx
	Oropharynx	8=Oropharynx	Oropharynx Lingual tonsil Mandible soft tissue Bone of face (Maxillary) Masseter Oral cavity
	Cheek	3=Cheek	Cheek
	Hypopharynx	4=Hypopharynx	Hypopharynx
	Thyroid & Parathyroid	11=Thyroid & Parathyroid	Thyroid
	Neck	6=Neck	Neck Neck Supra-clavicular soft tissues Neck, nodes Nos
		12=Other Head & Neck	Chin Soft tissue face (non specified region) Face specified region Nasolabial fold (skin) Nostril

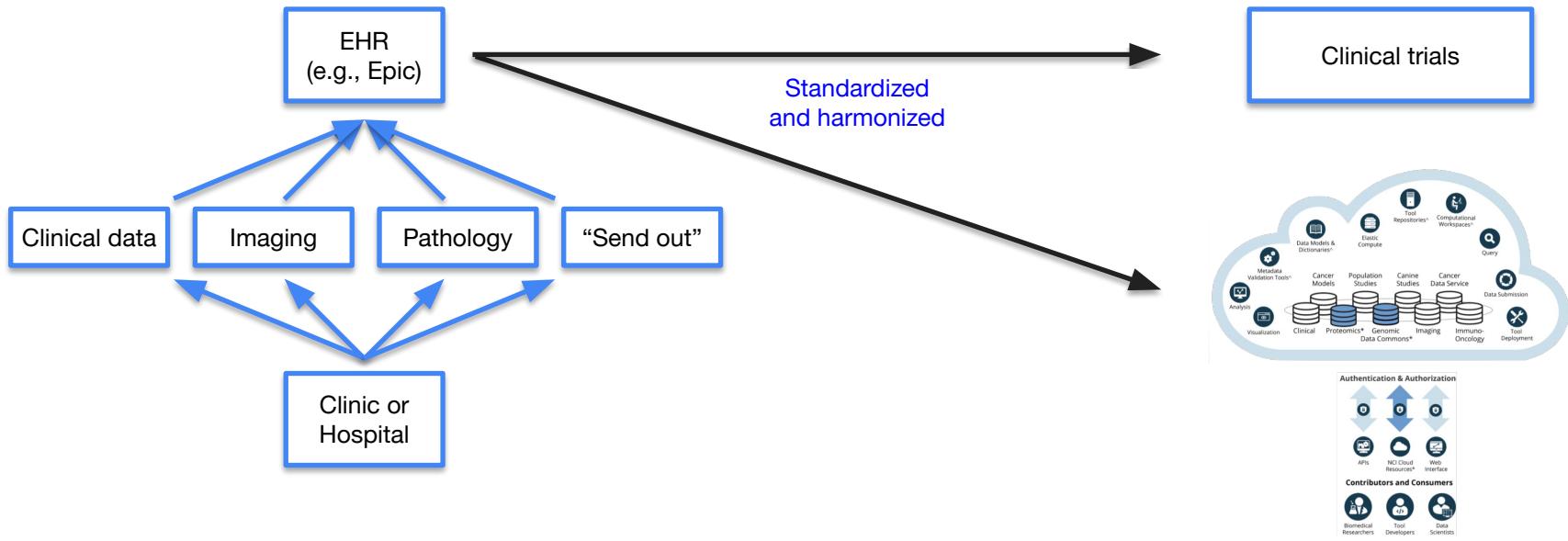
Harmonized dictionary

Instruct Variable Name	Instruct Permissible Values Term	Mapped Standard Code
HEAD AND NECK	Eyelid	C0015426
	Orbit	C0700042
	Other orbit	C0700042
	Cheek	C0007966
	Hypopharynx	C0020629
	Larynx	C0023078
	Neck	C0027530
	Oral cavity	C1711367
	Oropharynx	C0521367
	Parotid	C3272625
	Scalp	C0036270
	Thyroid and parathyroid	C0574117
	Other face	C0015450
	Other head and neck	C0460004
	Middle ear	C0013455
	Nasal cavity and paranasal sinuses	C0027423
	Nasal cavity	C0027423
	Paranasal sinuses	C0030471
	Nasopharynx	C0027442

Progress in the Pediatric Cancer Data Commons



Standards can help us achieve one universe



Research data commons

Cloud-Based Biomedical Data Storage and Analysis: Implications for Trustworthy Governance

Sarah Nelson (University of Washington)



Working Group Updates

NIH Coordination

Valentina Di Francesco NHGRI/AnVIL





Current NCPI Coordination WG Members

**NHGRI AnVIL**

Valentina Di Francesco
Ken Wiley
Natalie Kucher

NHLBI BioData Catalyst

Jon Kaltman
Alastair Thomson
Chip Schwartz
Sweta Ladwa

CF Kids Firsts

Valerie Cotton
James Coulombe
Huiqing Li

NCI CRDC

Tanja Davidsen
Allen Dearry
Erika Kim
Zhining Wang
Jamie Guidry Auvil
Jay Ronquillo
Marcia Fournier

NCBI

Kurt Mac Daniel
Kim Pruitt

CFDE

Lora Kutkat
Haluk Resat
Chris Kinsinger

NIH Office of Data Science Strategy

Asiyah Lin
Laura Biven
Vivian Ota Wang



Coordination WG's Responsibilities



- Serve as the NCPI Governance body
- Stewardship of the NCPI WGs activities
- Liaison with ODSS, OSP and other parts of the NIH



Updates since the
Fall 2020 workshop

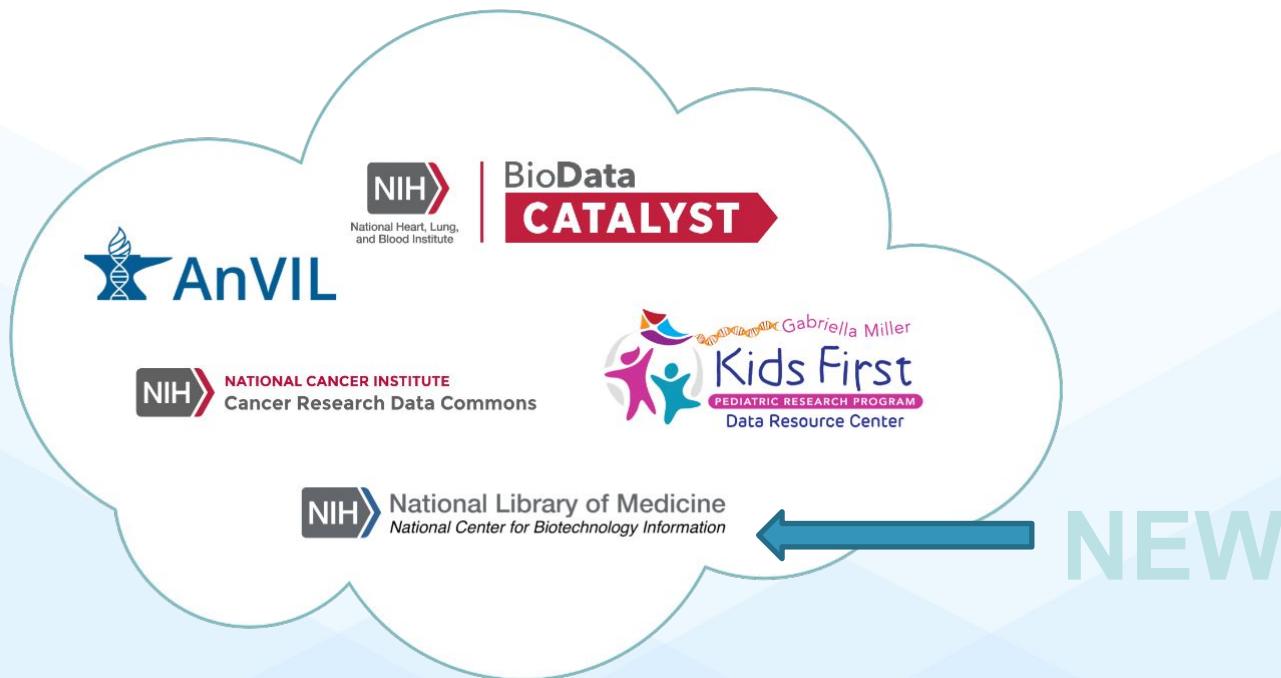


Asiyah Yu Lin, MD. PhD.

Asiyah Lin joined NIH as an ODSS supported DATA Scholar to work on the NCPI project. She has a background of Pediatrics, Immunology and Medical Informatics. Having worked for the FDA, start-ups and an NGS lab, Asiyah has +10 years experience in ontology-based data integration, analysis for biological and health data. She advocates leveraging ontologies for data interoperability and establishing knowledge eco-system for science and regulatory communities.

C	G	T	A	C	G	T	A
A	C	G	T	A	C	G	T
C	G	T	A	C	G	T	A

NCPI Onboarding New Members





2021 Objectives – Supported by ODSS



- **Search and aggregate data across platforms.** Enabling search of clinical data, studies, subjects, and samples through tools such as APIs to assemble cohorts across multiple sources for cross-dataset analysis.
- **Perform outreach activities** (portal, training, data dashboard) to ensure alignment with related efforts, engage users, and foster collaboration (internally across NCPI and with external efforts).
- **Cloud costs estimation** for analyses to enable researchers to budget for cloud costs and perform cost optimization.
- **Cross-NCPI-platform workflow execution.**
- **Define guiding principles for technical interoperability** and overcoming operational barriers.
- Ensure **RAS/GA4GH Passport** implements a common authentication and authorization mechanism across NCPI.



NCPI Developers Access and NIH OSP



Question

- What mechanism the cloud platforms should employ to allow access by NCPI developers across the 4 platforms?

Issues Discussed

- Developer definition
- Mechanisms of developer access (request vs whitelist)
- Data use restrictions
- Upholding participant protections and privacy
- Upholding transparency on who access the data
- Publication restriction

Next Steps

- OSP to draft proposal for developers access to send to NCPI Coordination WG for feedback

From 2020 Principles to 2021 Considerations

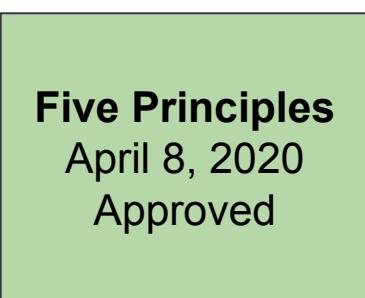
Five Principles for Interoperating Data Platforms

Version C
April 8, 2020

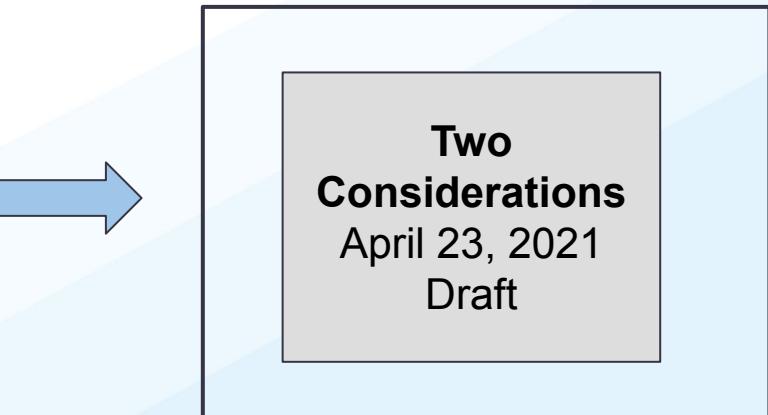
Over the last few years, a growing number of cloud-based data platforms have been developed that provide the research and translational community with access to data that is integrated with computational resources, services and workspaces, as well as knowledge resources, semantic services and AI services.

As the number of these platforms grows, it is becoming critical to establish some operating principles so that platforms can interoperate, allowing researchers to access, explore and integrate data from multiple platforms.

	dbGaP Model	GDC Model	CRD	BDC	AnVIL	KF
Status	reviewed	reviewed	under review	reviewed	reviewed dbGaP & white list GDCOS	reviewed dbGaP & white list
User Auth	dbGaP	dbGaP	dbGaP	dbGaP & white list		
Environment Authorization	Sixty-Five (65) official who has the legal authority to sign off on behalf of the organization's COO or CIO's data security assessment "dbGaP Model"	Sixty-Five (65) official who has the legal authority to sign off on behalf of the organization's COO or CIO's data security assessment "dbGaP Model"	Sixty-Five (65) official who has the legal authority to sign off on behalf of the organization's COO or CIO's data security assessment "dbGaP Model"	Sixty-Five (65) official who has the legal authority to sign off on behalf of the organization's COO or CIO's data security assessment "dbGaP Model"	Sixty-Five (65) official who has the legal authority to sign off on behalf of the organization's COO or CIO's data security assessment "dbGaP Model"	Institute CIO Research organization's IT Director
Data access ("a "Agree" by another cloud platform")	Any platforms authorized by researcher's organization (via dbGaP "dbGaP Model")	Any platforms authorized by researcher's organization (via dbGaP "dbGaP Model")	to be determined	Data cannot leave BDC Platform.	Restricted to platforms with an ISDN with AnVIL.	Any platform authorized by researcher's org. (via dbGaP)
Data Egress - "download"	Any platforms authorized by researcher's organization (via dbGaP "dbGaP Model")	Any platforms authorized by researcher's org. (via dbGaP "dbGaP Model")	Any platforms authorized by researcher's org. (via dbGaP "dbGaP Model")	Data cannot leave BDC Platform.	dbGaP model for downloaded data	Any platform authorized by researcher's org. (via dbGaP)
API	entire can be downloaded, but no API to data	All data is available via an API	Data objects available via API; CCDB and CDA clinical data	API within BDC for data objects and harmonized data in the future API for multiple data models (e.g., GDC, dbGaP API for clinical/phenotype data models)	API within AnVIL for data objects and harmonized data in the future API for multiple data models (e.g., GDC, dbGaP API for clinical/phenotype data models)	All data is available via GDC/portal
Trust relationships	NA	open to any auth. env.	need to determine	need to determine	need to determine	API Gen3 for clinical/phenotype data models



White Paper
with Table of
Platforms &
their Auth. Env.
Oct 23, 2020



From Grossman & Ahalt

focus since the last meeting



NCPI “Rules of Engagement”

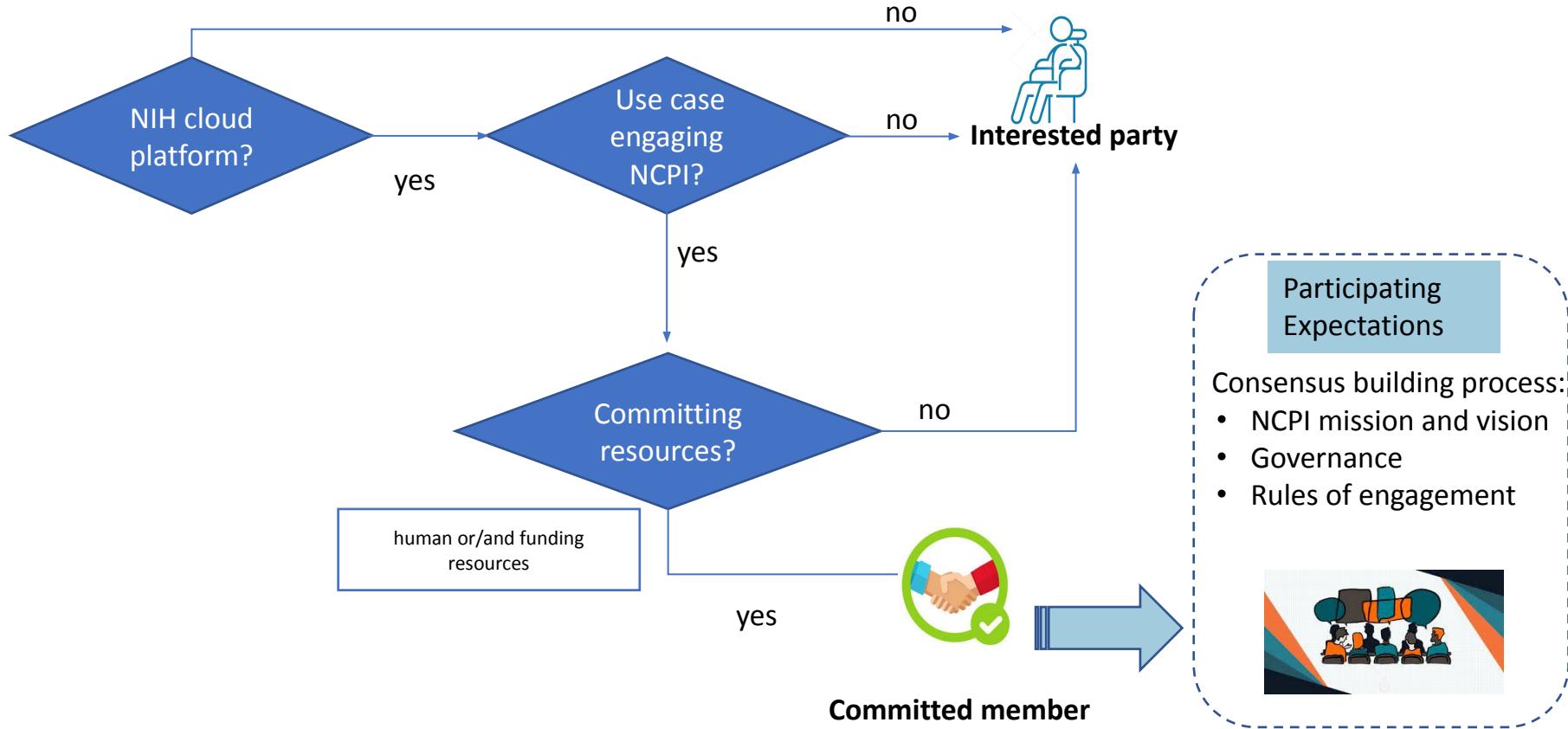


5 proposed criteria:

1. Agree to the NCPI principles and interoperability “considerations”
2. Willing to test, adopt and/or extend NCPI technology specifications
3. Identify interoperability use case when entering the collaboration
4. Commitment to participate in WGs
5. Share, open communication, transparency



A Decision Tree for Initial Engagement





Status of Y2 Goals (from Oct 2020 Wrkshp)



- Host NCPI all hands workshops every 6 months
- Pursue additional funding support for NCPI activities



- Identify and agree upon next year's priorities and milestones
- Implement interoperability principles
- Continue collaboration with RAS
- Solidify collaboration with GA4GH work streams



- Offer training opportunities for outside investigators
- Share best practices for platforms interoperability across NIH

Goals of this meeting

What?

Identify 2-4 use cases/collaborative projects to demonstrate interoperability among 2 or more resources

- 6 - 12 month timeframe
- Concrete
- Support real science
- Solve low hanging fruit issues
- Identify specific asks of NIH (How does NIH want to do X or handle Y?)



"Perfect is the enemy of good."

Commonly attributed to Voltaire

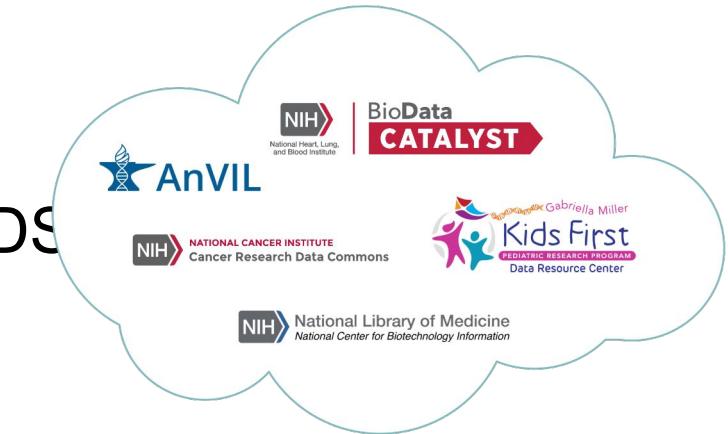
How?

For each use case identify responsible working group and individuals

Jonathan Kaltman
NCPI Oct 2019

THANK YOU

- NIH NCPI Coordination WG
- NCI, NHGRI, NHLBI, CF, NLM, ODS
- All NCPI Members



Working Group Update

Community and Governance Working Group

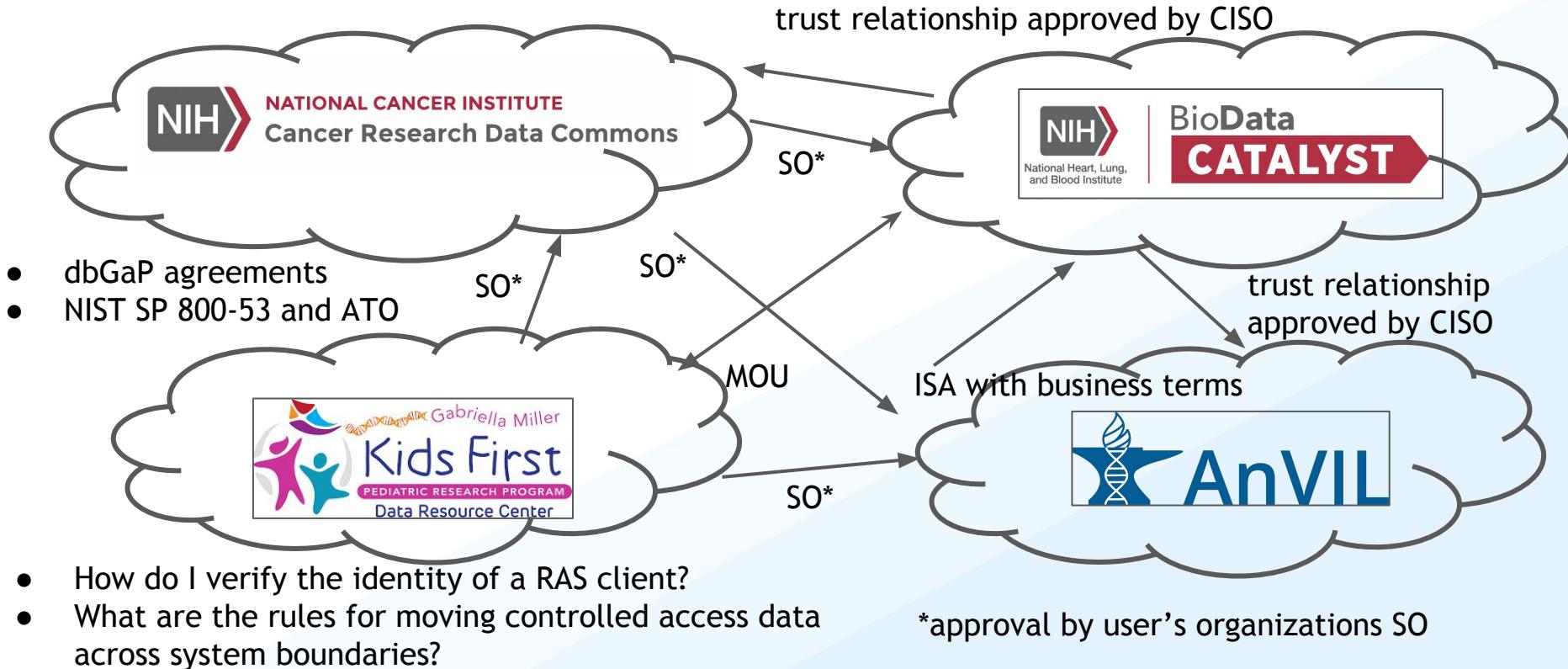
Robert L. Grossman
University of Chicago



Stan Ahalt
RENCI



Where We Are Today



From 2020 Principles to 2021 Considerations

Five Principles for Interoperating Data Platforms

Version C
April 8, 2020

Over the last few years, a growing number of cloud-based data platforms have been developed that provide the research and translational community with access to data that is integrated with computational resources, services and workspaces, as well as knowledge resources, semantic services and AI services.

As the number of these platforms grows, it is becoming critical to establish some operating principles so that platforms can interoperate, allowing researchers to access, explore and integrate data from multiple platforms.

	dbGaP Model	GDC Model	CRD	BDC	AnVIL	KF
Status	reviewed	reviewed	under review	reviewed	reviewed dbGaP & white list GDC	reviewed dbGaP & white list
User Auth	dbGaP	dbGaP	dbGaP	dbGaP & white list		
Environment Authorization	Sixty Official who has the legal authority to attest to the organization's COO or CIO's data security assessment "dbGaP Model"	Sixty Official who has the legal authority to attest to the organization's COO or CIO's data security assessment "dbGaP Model"	Sixty Official who has the legal authority to attest to the organization's COO or CIO's data security assessment "dbGaP Model"	Sixty Official who has the legal authority to attest to the organization's COO or CIO's data security assessment "dbGaP Model"	Sixty Official who has the legal authority to attest to the organization's COO or CIO's data security assessment "dbGaP Model"	Institute CIO Institute CIO Research organization's IT Director
Data access ("a "agree" by another cloud platform	Any platforms authorized by researcher's organization (via dbGaP Model)	Any platforms authorized by researcher's organization (via dbGaP Model)	to be determined	Data cannot leave BDC Platform.	Restricted to platforms with an AnVIL with AnVIL.	Any platform authorized by researcher's org. (via dbGaP)
Data Egress - "download"	Any platforms authorized by researcher's organization (via dbGaP Model)	Any platforms authorized by researcher's organization (via dbGaP Model)	Any platforms authorized by researcher's org. (via dbGaP Model)	Data cannot leave BDC Platform.	dbGaP model for downloaded data	Any platform authorized by researcher's org. (via dbGaP)
API	entities can be downloaded, but no API to data	All data is available via an API	Data objects available via API; CCDB and CDA objects available via API; clinical data	API within BDC for data objects and harmonized data in the future API for multiple data models; API for GDC; API for clinical data models	API within AnVIL for data objects and harmonized data in the future API for multiple data models; API for clinical data models	All data is available via GDC portal API; Gen3 for clinical data; AnVIL API for clinicalphen Q1 2021.
Trust relationships	NA	open to any auth. env	need to determine	need to determine	need to determine	need to determine

Proposed Considerations for Interoperability of Cloud Platforms Draft B-2-2

Five Principles
April 8, 2020
Approved



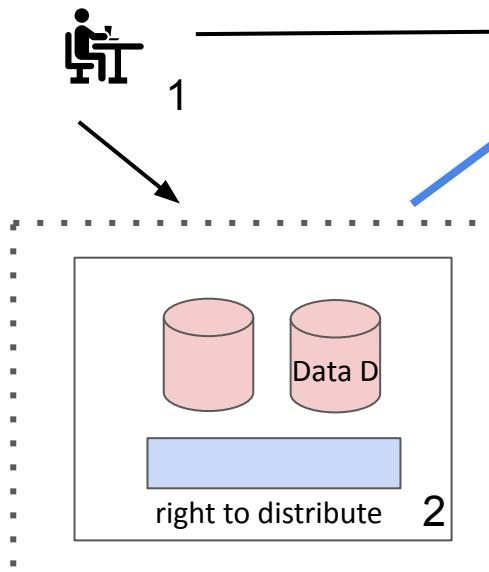
White Paper
with Table of
Platforms &
their Auth. Env.
Oct 23, 2020



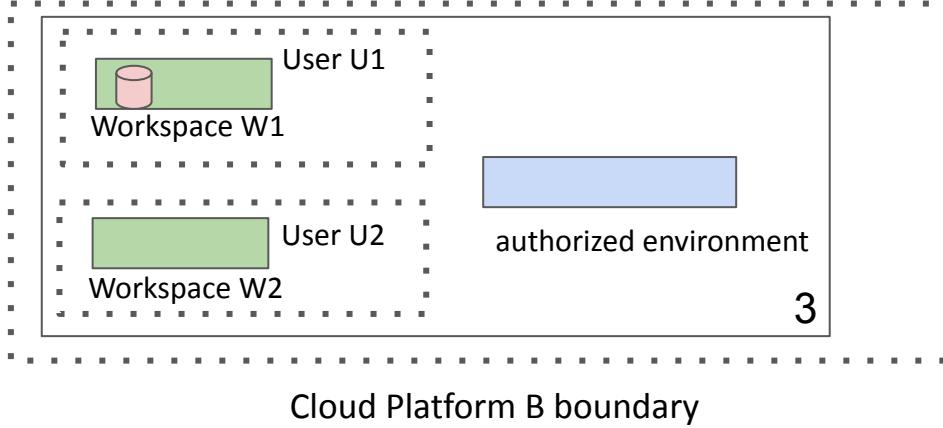
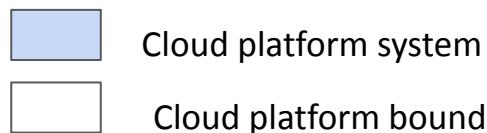
Two
Considerations
April 23, 2021
Draft

focus since the last meeting

Four Key Concepts



Cloud Platform A boundary



Cloud Platform B boundary

1. A **user** is **authorized** to access a dataset
2. A cloud platform A has the **right to distribute** a particular dataset.
3. A cloud platform B is an **authorized environment** for a particular dataset.
4. Each dataset has a **data trustee** (aka **data steward**) that makes decisions about 1), 2) and 3)

We have **interoperability** when an authorized environment can access data from two or more cloud platforms..





Authorized Environments

- **Authorized environment -**
 - New concept in our October 2020 White Paper
 - Example, for a cloud platform, the Institute's CISO can authorize an environment, say by approving an ATO for FISMA Moderate environment
 - Example, with dbGaP, the organization's IT Director through the organization's SO authorizes an environment for data downloaded from dbGaP
- Decisions about authorized environments can be based on the **sensitivity of the data**.
- **Authorized Environment Principle** - authorize environments and authorize users and trust the authorizations
- We have **interoperability** when an authorized environment can access data from two or more cloud platforms.

Platform	Data Auth Determination	Data Trustee	System Trustee	Right to Distrib Gov	Auth Env Gov	Data Egress
NCBI dbGaP	NIH DAC	NIH	NIH	NIH Owned / Operated	End user's Signing Official	Yes
NCI CDRC	NIH DAC	NIH	NIH	NCI ATO & NIH Trusted Partner	NCI ATO and/or End user's Signing Official	Yes
CF Kids First	NIH DAC	NIH	NIH	NCI ATO & Trusted Partner	End user's Signing Official	Yes
NHLBI BioData Catalyst	NIH DAC	NIH	NIH	NHLBI ATO	NHLBI ATO or NHLBI trusted env.	Yes, but not encouraged
NHGRI The Anvil	NIH DAC	NIH and Awardee	Awardee	Awardee via NHGRI Coop. Agreement & NIH Designated Data Repo.	Awardee ATO requires ISAs w/ business terms	Yes, but not encouraged

Table 5. This table shows the proposed basis for granting a cloud platform the right to distribute controlled access datasets.

We are close to interoperability for several of the NCPI cloud platforms:

1. The Working Group participants have all agreed on key terms and concepts, such as right to distribute, authorized environments, and data trustees/stewards.
2. The data steward/trustee (NIH or grant awardee) must simply agree that two or more cloud platforms are authorized environments. We have included a sample memo for this purpose.
3. There are still differences being discussed i) approval by SO and/or ATO; ii) specific security requirements; iii) standard ISAs; iv) what about inclusion of liability & related business requirements.

To:

From:

Date:

Re:

This is to recognize the following cloud platforms as authorized environments so that users who have been authorized by dbGaP, RAS, or other approved authorized mechanism to access a dataset can explore and analyze the data in the authorized environment and **[fill in with cloud platforms that have the right to distribute data]** has approval to distribute the data to the authorized environment.

Authorized Environments:

Authorized Environment	Type	Date
	ATO issued by [fill in]	
	ATO issued by [fill in]	
	Approved by [fill in]	
	Approved by [fill in]	

Authorized environments

Institute/Center	Proposed Authorized Environments	Proposed Basis for Approval
NCI	SBG, Terra, ISB Cloud Platform, Gen3 + Any platform approved by the end-user's Institutional Signing Official, <u>per the terms of the DUC & these guidelines</u>	IC-CISO-FISMA-Moderate-ATO and/or SO-approved
NHGRI	Terra, Gen3	Org-CISO-NIST-800-53-approval; specifically, approval by Broad CISO, with the requirement of an ISA between AnVIL and the platform
NHLBI	Terra, SBG, Gen3	IC-CISO-FISMA-Moderate-ATO
Kids First Program	Any platform approved by the end-user's Institutional Signing Official, <u>per the terms of the DUC & these guidelines</u>	SO-approved

Table 2. This table shows the proposed basis for approving an environment as an authorized environment.

Right to distribute

Institute/Center	Proposed Platforms that can distribute data	Proposed Mechanisms
NCI	Approved CRDC platforms	Approval as a NIH Trusted Partner to distribute controlled access data.
NHGRI	AnVIL/Terra	Approval by Broad CISO, with the requirement of an ISA between AnVIL and the platform
NHLBI	BioData Catalyst	Approval by NHLBI CISO
Kids First Program	Bionimbus Gen3 for controlled access data	Approval as a NIH Trusted Partner to distribute controlled access data.

Table 3. This table shows the proposed basis for granting a cloud platform the right to distribute controlled access datasets.

Active Discussion Issues

- Framework for authorizing environments:
 - dbGaP Data Use Certification with User's Signing Official (SO) with recommendation from IT Directory as formalized by dbGaP
 - or NIST SP 800-53 Moderate ATO
 - ATO from Institutes / Centers
 - ATO from third party
 - or, presumably, both?
- Working on standardized ISAs.
- How do we interoperate USG and third-party systems operated by awardees?
- Can we start with decisions about less sensitive data?



Next Steps

- We have broadened the NCPI Community / Governance discussion to include security specialists, which have started to discuss specific NIST 800-53 security requirements.
- We are looking forward to feedback about our draft considerations from a broader audience to gain additional feedback and identify any additional concerns.

Questions

Systems Interoperation WG updates

Jack DiGiovanna*

Seven Bridges

Brian O'Connor

Broad Institute



... ; John Cheadle (BDC) ; Nicole Bolliger (Broad) ; Jiaqi Liu (UChicago) ; Gina Kuffel (UChicago) ; Mai Nguyen (CHOP) ; Bill Longabaugh (ISB-CGC) ; Anton Nekrutenko (AnVIL) ; Jay Ronquillo (NCI) ; Jonas Almeida (NCI) ; Michael Lukowski (U Chicago) ; Binam Baijacharya(UChicago) ; ; ; Kate Herman (Broad) ; ; ; Sai Lakshmi Subramanian (SB) ; Garrett Rupp (SB) ; Jessica Lyons (HMS) ; Danielle Pillion (HMS) ; Michael Baumann (Broad) ; Alex VanTol (UChicago) ;

; ; ; Robert L. Grossman (UChicago); Gina Kuffel (UChicago) ; ; ; ; ; ; ; Amanda Charbonneau (CFDE) ; ; ; John Cheadle (BDC) ; Bill Longabaugh (ISB); Valentina Di Francesco (NHGRI); Asiyah Lin (NHGRI) ; ; ; Sai Lakshmi Subramanian (SB); Stephen Mosher (JHU) ; ; Natalie Kucher (NHGRI); Michael Lukowski (U Chicago); Nicole Bolliger (Broad); Jay Ronquillo (NCI); Jonas Almeida (NCI); Kurt Rodamer (NCBI) ; ; Pauline Ribeyre (UChicago); Lynette Lilly (UChicago) ; Alex VanTol (UChicago); Brian Walsh (OHSU); Danielle Pillion (PIC-SURE), Jason Stedman (PIC-SURE); Jessica Lyons (PIC-SURE)

; ; ; ; ; Mark Jensen (FNLCR); Bill Longabaugh (ISB); John Cheadle (BDC) ; ; ; Jason Stedman(HMS) ; Valentina Di Francesco (NHGRI) ; ; Michael Lukowski (UChicago) ; ; Sai Lakshmi Subramanian (Seven Bridges) ; David Pot (GDIT) ; ; ; Nicole Bolliger (Broad) ; Allen Dearry (NCI) ; Jay Ronquillo (NCI) ; Jonas Almeida (NCI) ; Pauline Ribeyre (UChicago) ; Alex VanTol (UChicago) ; Garrett Rupp (Seven Bridges) ; Jiaqi Liu (UChicago) ; Binam Bajracharya(UChicago) ; Jessica Lyons (HMS) ; Danielle Pillion (HMS);



OVERVIEW

Connected Data

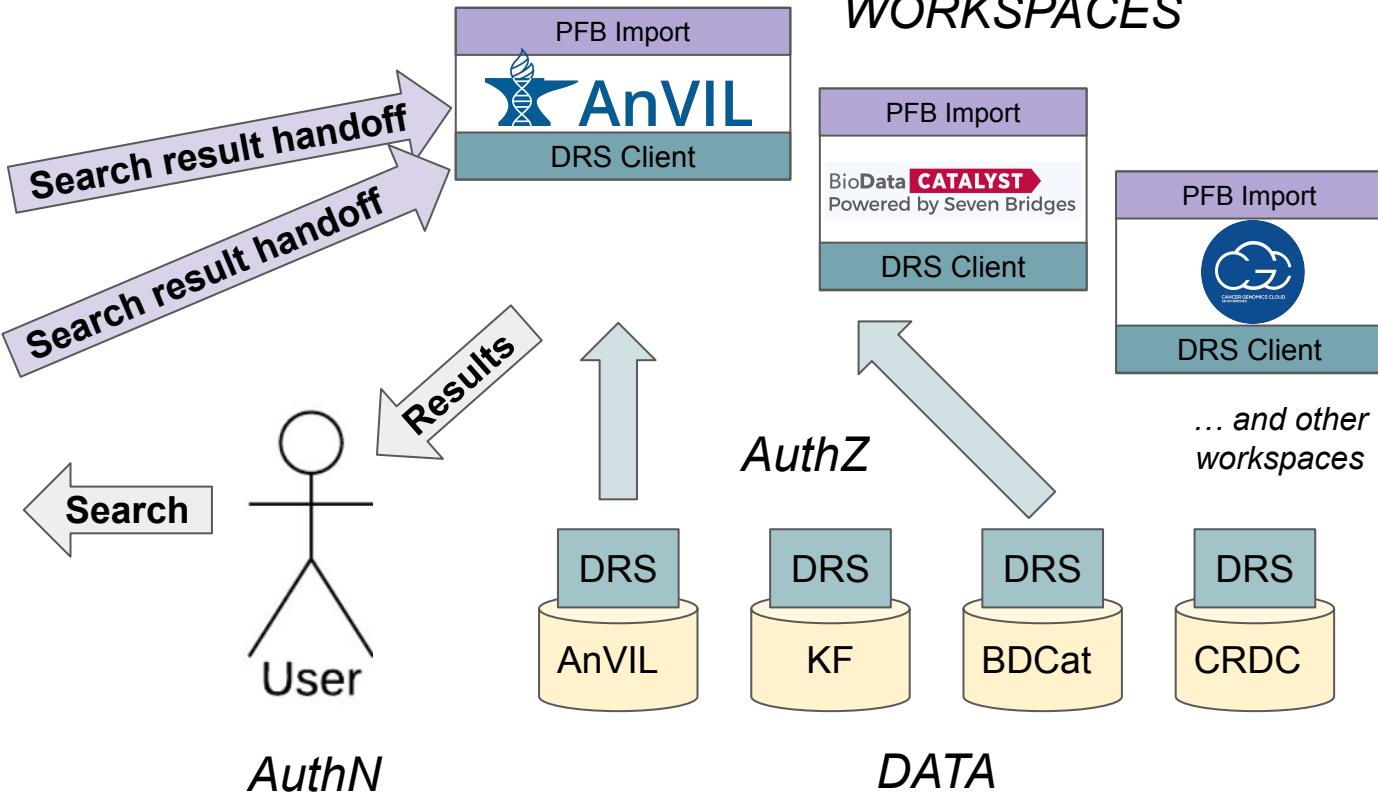
Use Cases

Tech Successes

Lessons Learned & Next Steps

Diverse users can co-analyze data to drive science

PORTALS



dbGaP is the source of truth for authorization

NCBI Resources How To Sign in to NCBI

dbGaP dbGaP Search Help

Limits Advanced



dbGaP

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.

Access dbGaP Data

- [Advanced Search](#)
- [Controlled Access Data](#)
- [Public FTP Download](#)
- [Collections](#)
- [Summary Statistics](#)

Resources

- [dbGaP Data Browser](#)
- [Phenotype-Genotype Integrator](#)
- [dbGaP RSS Feed](#)
- [Software](#)
- [dbGaP Tutorial](#)

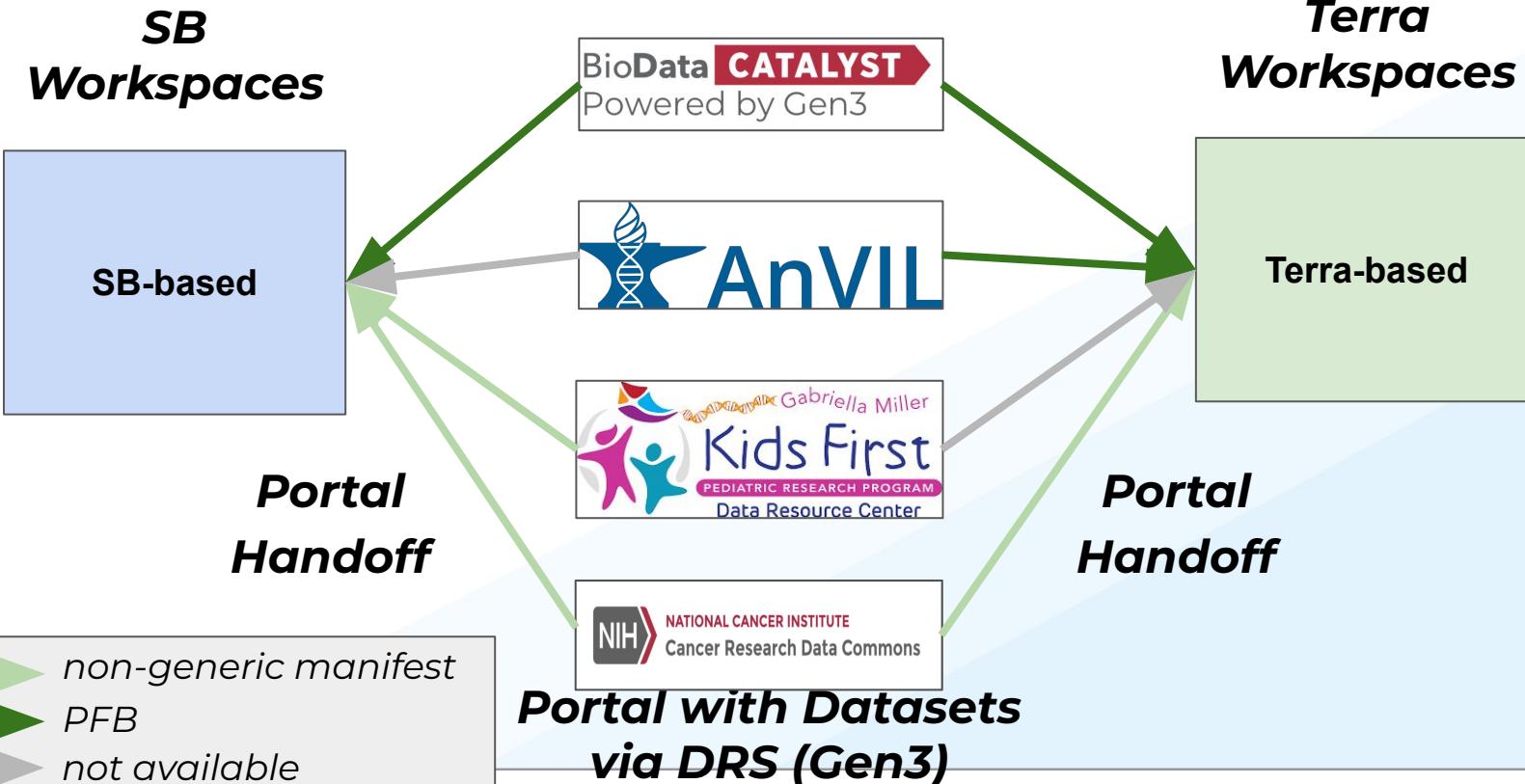
Important Links

- [How to Submit](#)
- [FAQ](#)
- [Code of Conduct](#)
- [Security Procedures](#)
- [Contact Us](#)

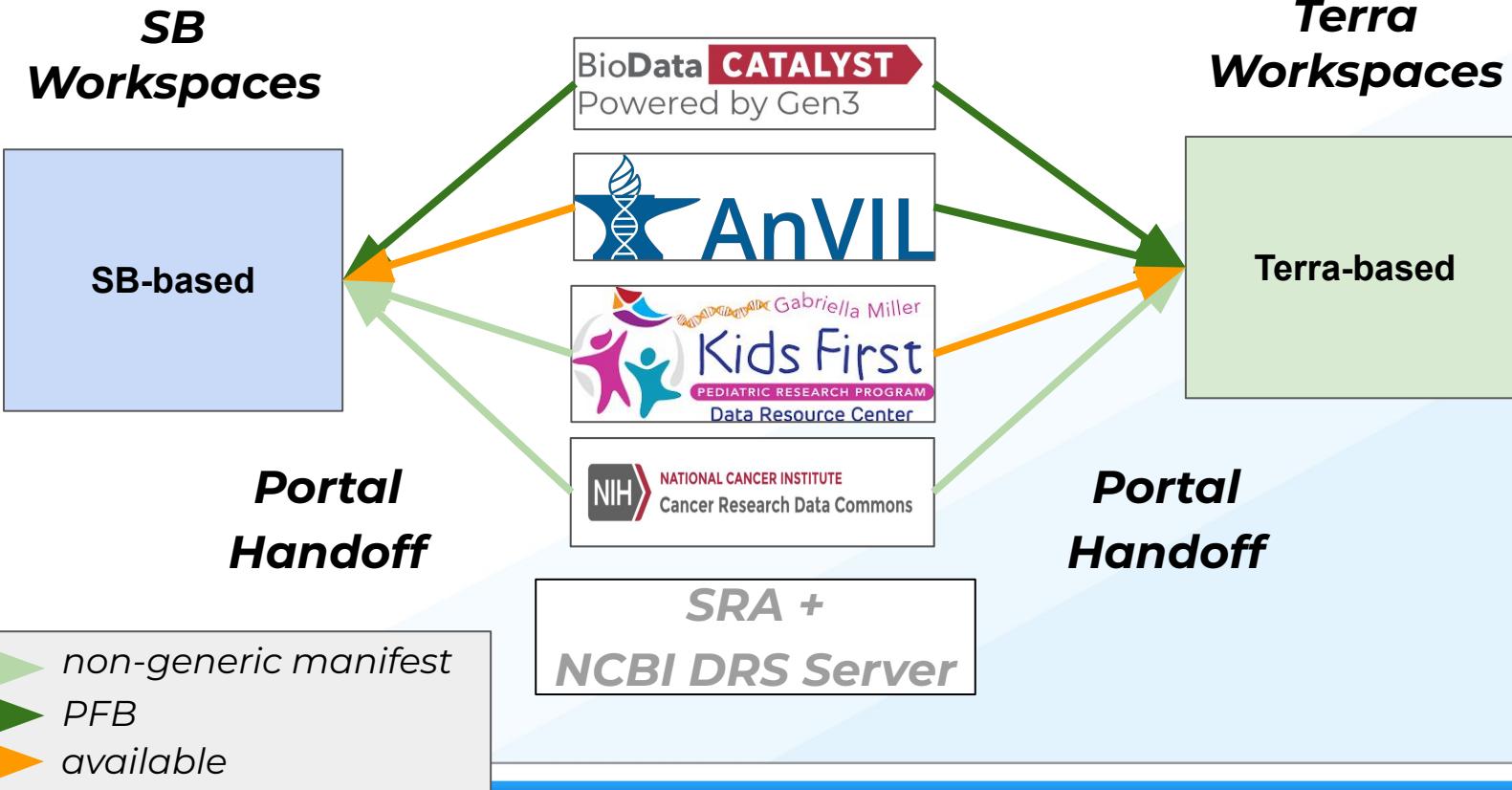
Latest Studies

Study	Embargo Release	Details	Participants	Type Of Study	Links	Platform
phs001997.v1.p1 Kids First: Genomics of African and Asian Orofacial Clefts Triads	Version 1: passed embargo	V D S A	791	Cohort, Parent-Offspring Trios	Links	
phs001987.v1.p1 Kids First: Esophageal Atresia and Related Malformation Trios	Version 1: passed embargo	V D S A	79	Parent-Offspring Trios, Cohort	Links	

Connectivity: Fall 2020



Today all four portals connect to Terra & SB workspaces





Overview

CONNECTED DATA

Use Cases

Tech Successes

Lessons Learned & Next Steps

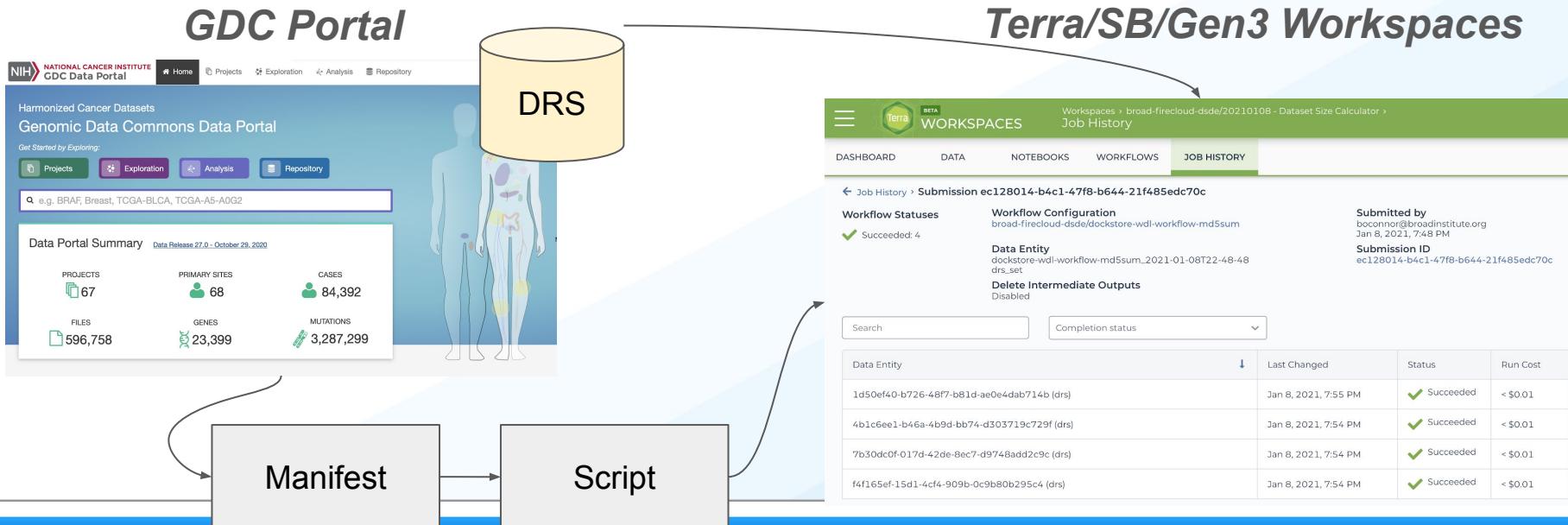
Cancer Research Data Commons

Genomics Data Commons Portal -> Workspace

Prototyped a process to convert GDC manifests to workspaces

Interest from **GDC** to develop a PFB-export functionality

Also started discussions with **other CRDC** Data Portals





Gabriella Miller Kids First Kids First and GTEx analysis in CAVATICA (CFDE)

Goal: Evaluate the functional equivalence for the RNA-Seq Pipeline between Kids First and GTEx.

Steps:

- Define testing GTEx cohort
- Download raw data to CAVATICA via AnVIL Gen3
- Run Kids First RNA-seq pipeline
- Compare with V8 results

Gabriella Miller Kids First

Kids First and GTEx analysis in CAVATICA (CFDE)

The screenshot shows the AnVil interface with a search results table for GTEx subjects. The table includes columns for sample_id, bam_path, md5_hex, Subject, Projects, Sex, and Ancestry. A large arrow points from the text "Get testing file manifest from AnVil" to the "Download Manifest" button at the top of the table.

sample_id	bam_path	md5_hex	Subject	Projects	Sex	Ancestry
GTEx-111CU-0126-4	gs/fc-secure-f8156; e2728562de6bf64b0	98	GTEx-111FC-3326-5	gs/fc-secure-f8156; eb7dfee72d50e37a1		
GTEx-111YS-0126-4	gs/fc-secure-f8156; 5b2ae962c8597c74		GTEx-1120-0326-4	gs/fc-secure-f8156; f5d44070b5c35b73		
GTEx-11220-0326-4	gs/fc-secure-f8156; 7f044070b5c35b73		GTEx-1128S-3286-5	gs/fc-secure-f8156; 729b9e3d091a870e		
GTEx-1128S-3286-5	gs/fc-secure-f8156; 729b9e3d091a870e		GTEx-117X-0326-4	gs/fc-secure-f8156; 9f9066ee563beaa42		
GTEx-117X-0326-4	gs/fc-secure-f8156; 9f9066ee563beaa42		GTEx-117YX-0326-4	gs/fc-secure-f8156; f5db176c3035f4456		
GTEx-1189X-3226-5	gs/fc-secure-f8156; 3c3b1a57eef275eb		GTEx-1192X-3226-5	gs/fc-secure-f8156; 3c3b1a57eef275eb		
GTEx-11DXW-1026-4	gs/fc-secure-f8156; 981c12d030308182de		GTEx-11DXW-1026-4	gs/fc-secure-f8156; 981c12d030308182de		
GTEx-11DX-0326-4	gs/fc-secure-f8156; 5c3ca85cd6d8098b		GTEx-11DX-0326-4	gs/fc-secure-f8156; 5c3ca85cd6d8098b		
GTEx-11DXY-1626-4	gs/fc-secure-f8156; 6752d6e8d777634		GTEx-11DXY-1626-4	gs/fc-secure-f8156; 6752d6e8d777634		
GTEx-11DXY-1626-4	gs/fc-secure-f8156; 6752d6e8d777634		GTEx-11DX-0326-4	gs/fc-secure-f8156; 78cfdf49871d18f747		
GTEx-11DX-0326-4	gs/fc-secure-f8156; 78cfdf49871d18f747		GTEx-11DXZ-0226-4	gs/fc-secure-f8156; 26ca107dbecc6f63c		
GTEx-11DXZ-0226-4	gs/fc-secure-f8156; 26ca107dbecc6f63c		GTEx-11DYG-2926-6	gs/fc-secure-f8156; bedd76bf35c98b07		
GTEx-11DYG-2926-6	gs/fc-secure-f8156; bedd76bf35c98b07		GTEx-11DZT-2926-5	gs/fc-secure-f8156; 8f2de9f603248f505e		
GTEx-11DZT-2926-5	gs/fc-secure-f8156; 8f2de9f603248f505e		GTEx-11Ei6-2926-8	gs/fc-secure-f8156; c3835b2c4990f49c6		
GTEx-11Ei6-2926-8	gs/fc-secure-f8156; c3835b2c4990f49c6		GTEx-11EM3-0326-6	gs/fc-secure-f8156; 5509012aaa73f1f10		
GTEx-11EM3-0326-6	gs/fc-secure-f8156; 5509012aaa73f1f10		GTEx-11EMC-0526-6	gs/fc-secure-f8156; fb4a0c3e5b171272f		
GTEx-11EMC-0526-6	gs/fc-secure-f8156; fb4a0c3e5b171272f		GTEx-11EMC-3326-6	gs/fc-secure-f8156; 8c94ab3f068883ect		
GTEx-11EMC-3326-6	gs/fc-secure-f8156; 8c94ab3f068883ect		GTEx-11EQ9-0126-4	gs/fc-secure-f8156; c50981d63c67d9e5		

Get testing
file manifest
from AnVil

Gabriella Miller Kids First

Kids First and GTEx analysis in CAVATICA (CFDE)

The AnVIL

Data File Downloadable

Data Access ▾

- Data with Access
- Data without Access
- All Data

Filters

- Projects
- Samples
- Sequencing

GTEx.v8

- Project Id: CF-GTEx (selected)
- Project Short Name: GTEx
- Project dbGaP Accession Number: phs000724.v3.p2
- Project dbGAP Consent Text

Showing 1 - 20 of 981 subjects

Project Id	Anvil Project Id	Ancestry	Sex	Age Value	Phenotype Group	Samples Count	Sequencing Count

sample_id bam_path md5_hex

Subjects:

sample_id	bam_path	md5_hex
GTEX-111CU-0126-4	gs://fc-secure-f8156/e2728562de6bf640c	98
GTEX-111FC-3326-5	gs://fc-secure-f8156/eb7dfee72d50e37a1	
GTEX-111YS-0126-4	gs://fc-secure-f8156/5b2a9f62c8597c74	
GTEX-11220-0326-4	gs://fc-secure-f8156/f5d44070b5c335b73	
GTEX-1128S-3282-5	gs://fc-secure-f8156/729b9e3d091a870e	
GTEX-117X-3-126-4	gs://fc-secure-f8156/9f9066ee563beaa42	
GTEX-117YX-0126-4	gs://fc-secure-f8156/f5db176c3035f4456	
GTEX-1192X-3226-5	gs://fc-secure-f8156/3c3b1a57eef275ebe	
GTEX-11DXW-1026-4	gs://fc-secure-f8156/981c12d030308182de	
GTEX-11DX-0126-4	gs://fc-secure-f8156/5c3ca85cded8d9098b	
GTEX-11DY-1626-4	gs://fc-secure-f8156/6752d6e8d77677634	
GTEX-11DXY-1626-4	gs://fc-secure-f8156/78cfdf49871d18f747	
GTEX-11DX-Z-2226-4	gs://fc-secure-f8156/26ca107dbcccf63c	
GTEX-11DYG-2926-4	gs://fc-secure-f8156/bedd76bf35c98b07	
GTEX-11DZ-2926-4	gs://fc-secure-f8156/8f2de9f603248505s	
GTEX-11EiB-2926-8	gs://fc-secure-f8156/c3835b2c4990494c	
GTEX-11EM3-0326-6	gs://fc-secure-f8156/5509012aaa/73ff110	
GTEX-11EMC-0526-6	gs://fc-secure-f8156/fb4a0c3e5b171272f	
GTEX-11EMC-3326-4	gs://fc-secure-f8156/8c94ab3068883ect	
GTEX-11EQ9-0126-4	gs://fc-secure-f8156/c50981d63c67d9e5	

Get testing
file manifest
from AnVIL

README.md

gen3-client

build passing release v2021.05

gen3-client is a command-line tool for downloading, uploading, and submitting data files to and from a Gen3 data commons.

Read more about what it does and how to use it in the [gen3-client user guide](#).

gen3-client is built on Cobra, a library providing a simple interface to create powerful modern CLI interfaces similar to git & go tools. Read more about Cobra [here](#).



COMMON
WORKFLOW
LANGUAGE

Make Gen3-client as
an CAVATICA app

Gabriella Miller Kids First

Kids First and GTEx analysis in CAVATICA (CFDE)

The AnVil

Data File Downloadable

Data Access ▾

- Data with Access
- Data without Access
- All Data

Filters

- Projects Subject Sample Sequencing
- Project ID CF-01Ex (Selected) Q
- Project Short Name Q
- GTEx Q
- Project dbGaP Accession Number Q
- ph20024a.v2.c1 Q
- Project dbar Consent Text Q

Showing 1 - 20 of 981 subjects

Project Id	Anvil Project Id	Ancestry	Sex	Age Value	Phenotype Group	Samples Count	Sequencings Count



sample_id	bam_path	md5_hex
GTEx-111CU-0126-4	gs://fc-secure-f8156/e2728562de6bf64b0c	
GTEx-111FC-3326-5	gs://fc-secure-f8156/eb7dfee72d50e37a1	
GTEx-111YS-0126-4	gs://fc-secure-f8156/5b2a9f62c8597c74	
GTEx-1122O-0326-5	gs://fc-secure-f8156/f5d44070b5c35873	
GTEx-1128S-2826-5	gs://fc-secure-f8156/729b9e0d91a870e	
GTEx-117X-3126-4	gs://fc-secure-f8156/9f9066ee553eaa42	
GTEx-117YX-0126-4	gs://fc-secure-f8156/f5db176c30354468	
GTEx-1192X-3226-5	gs://fc-secure-f8156/3c3b1a57ee275ebe	
GTEx-11DXW-1026-4	gs://fc-secure-f8156/981c12d03108b2de	
GTEx-11DXY-0126-4	gs://fc-secure-f8156/5c3ca85cd8d098b	
GTEx-11DXY-1626-4	gs://fc-secure-f8156/6752d6e8d7767734	
GTEx-11DXY-3126-4	gs://fc-secure-f8156/78cfdf49871d1f7747	
GTEx-11DXZ-0226-4	gs://fc-secure-f8156/26ca107dbcccf63c	
GTEx-11DYG-2926-4	gs://fc-secure-f8156/bedd76bf3f5c98b07	
GTEx-11DZT-2926-4	gs://fc-secure-f8156/8f2de9f63248f505e	
GTEx-11EiB-2926-8	gs://fc-secure-f8156/c383852c499049fc	
GTEx-11EM3-0326-4	gs://fc-secure-f8156/5509012aaa73f1110	
GTEx-11EMC-0526-4	gs://fc-secure-f8156/fb4a0c3e5b171272f	
GTEx-11EMC-3326-4	gs://fc-secure-f8156/8c94ab3068883ect	
GTEx-11EQ9-0126-4	gs://fc-secure-f8156/c50981d63c67d9e5	

Get GTEx data via CAVATICA workflow

CAVATICA Projects Data Public Apps Public Projects Developer Controlled projects yuankun ▾

CFDE download Interactive Analysis Notes

Dashboard Files Apps Tasks

Search task names Status: COMPLETED

Task Name	Submitted by	Submitted on	App	Duration	Status	Action
download_files run - GTEx-111CU-0126-SM-5GZGW	zhangh1	Apr. 12, 2021 09:30	download_files	1 minute	COMPLETED	C
download_files run - GTEx-111CU-0126-SM-5GZGW	zhangh1	Apr. 12, 2021 09:58	download_files	4 minutes	COMPLETED	C
download_files run - GTEx-111CU-0126-SM-5GZGW	zhangh1	Apr. 12, 2021 10:19	download_files	4 minutes	COMPLETED	C
download_files run - GTEx-111FC-3326-SM-5GZYV	zhangh1	Apr. 12, 2021 10:50	download_files	10 minutes	COMPLETED	C
download_files run - GTEx-111YS-0126-SM-5987T	zhangh1	Apr. 12, 2021 10:50	download_files	25 minutes	COMPLETED	C
download_files run - GTEx-1122O-0326-SM-5H124	zhangh1	Apr. 12, 2021 10:50	download_files	5 minutes	COMPLETED	C
download_files run - GTEx-1128S-2826-SM-5N90I	zhangh1	Apr. 12, 2021 10:50	download_files	5 minutes	COMPLETED	C
download_files run - GTEx-117XS-3126-SM-5GIDP	zhangh1	Apr. 12, 2021 10:50	download_files	5 minutes	COMPLETED	C
download_files run - GTEx-117YX-0126-SM-5EGH5	zhangh1	Apr. 12, 2021 10:50	download_files	5 minutes	COMPLETED	C
download_files run - GTEx-1192X-3226-SM-5987D	zhangh1	Apr. 12, 2021 10:50	download_files	6 minutes	COMPLETED	C
download_files run - GTEx-11DXW-1026-SM-5H11K	zhangh1	Apr. 12, 2021 10:50	download_files	24 minutes	COMPLETED	C
download_files run - GTEx-11DXY-1626-SM-5H12L	zhangh1	Apr. 12, 2021 10:50	download_files	20 minutes	COMPLETED	C
download_files run - GTEx-11DXY-3126-SM-5N9BT	zhangh1	Apr. 12, 2021 10:50	download_files	27 minutes	COMPLETED	C
download_files run - GTEx-11DXZ-0226-SM-5EGQZ	zhangh1	Apr. 12, 2021 10:50	download_files	6 minutes	COMPLETED	C

Refresh Showing 1-100 of 501 < >

README.md

gen3-client

build passing release v2021.05

gen3-client is a command-line tool for downloading, uploading, and submitting data files to and from a Gen3 data commons.

Read more about what it does and how to use it in the [gen3-client user guide](#).

gen3-client is built on Cobra, a library providing a simple interface to create powerful modern CLI interfaces similar to git & go tools. Read more about Cobra [here](#).



COMMON WORKFLOW LANGUAGE

Make gen3-client as an CAVATICA app

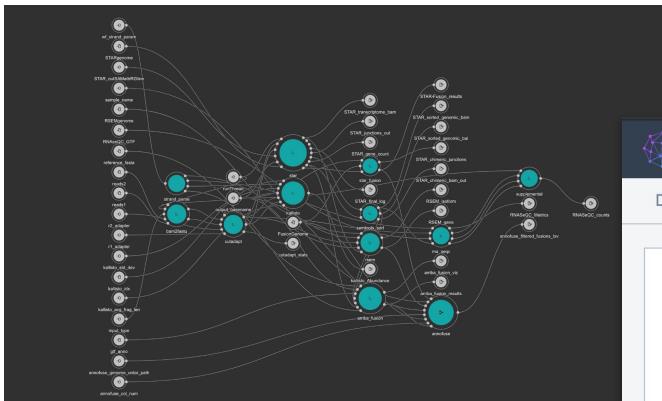


Gabriella Miller Kids First

Kids First and GTEx analysis in CAVATICA (CFDE)

Kids First RNA-Seq Pipeline

STAR-2-Pass → RSEM/STAR-Fusion/Arriba

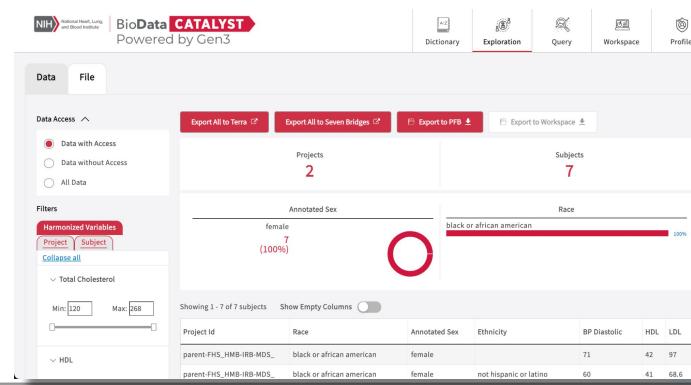
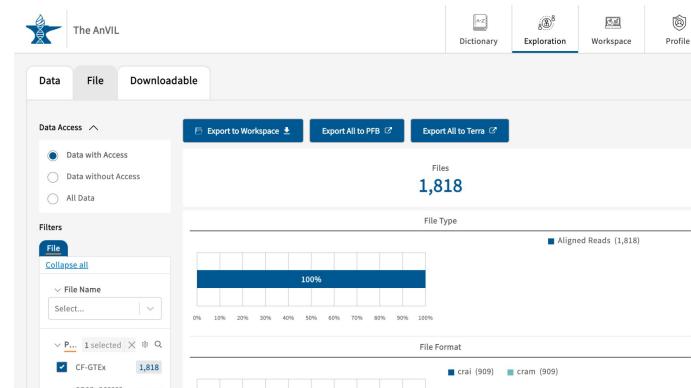


The screenshot shows the CAVATICA platform's interface for a project titled "CFDE GTEX RNAseq analysis". The top navigation bar includes links for "Projects", "Data", "Public Apps", "Public Projects", "Developer", and "Controlled project". Below the navigation is a secondary menu with "Dashboard", "Files" (which is selected), "Apps", and "Tasks". The main content area displays a list of files under the heading "Files". A search bar contains the text "rsem.genes". Filter options include "Type: All", "Sample ID: All", "Task ID: All", and "Tags: All". A "Clear filters" button is also present. The file list shows three entries:

	Name	Type	Size	Last Modified
<input type="checkbox"/>	1a0704b9-a361-4613-8c05-79dba8bf0a15.rsem.genes.results.gz	Files	RESULTS.GZ	-
<input type="checkbox"/>	7edeb201-fd6e-497d-b9f2-42ef7a420e03.rsem.genes.results.gz	Files	RESULTS.GZ	-
<input type="checkbox"/>	6e4070e0-ff01-4648-9f4e-73d8e860b319.rsem.genes.results.gz	Files	RESULTS.GZ	-

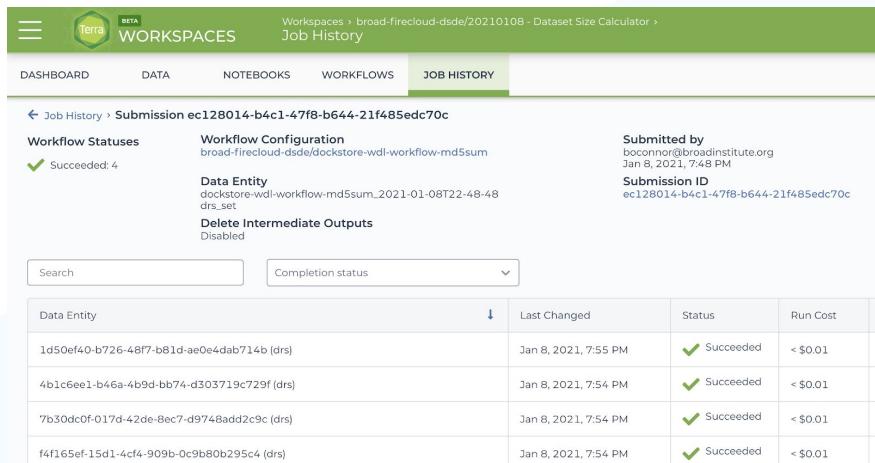
AnVIL & BioData Catalyst

Gen3 Portal push to Terra, SB, or Gen3 Workspaces



Production push to Terra and Gen3

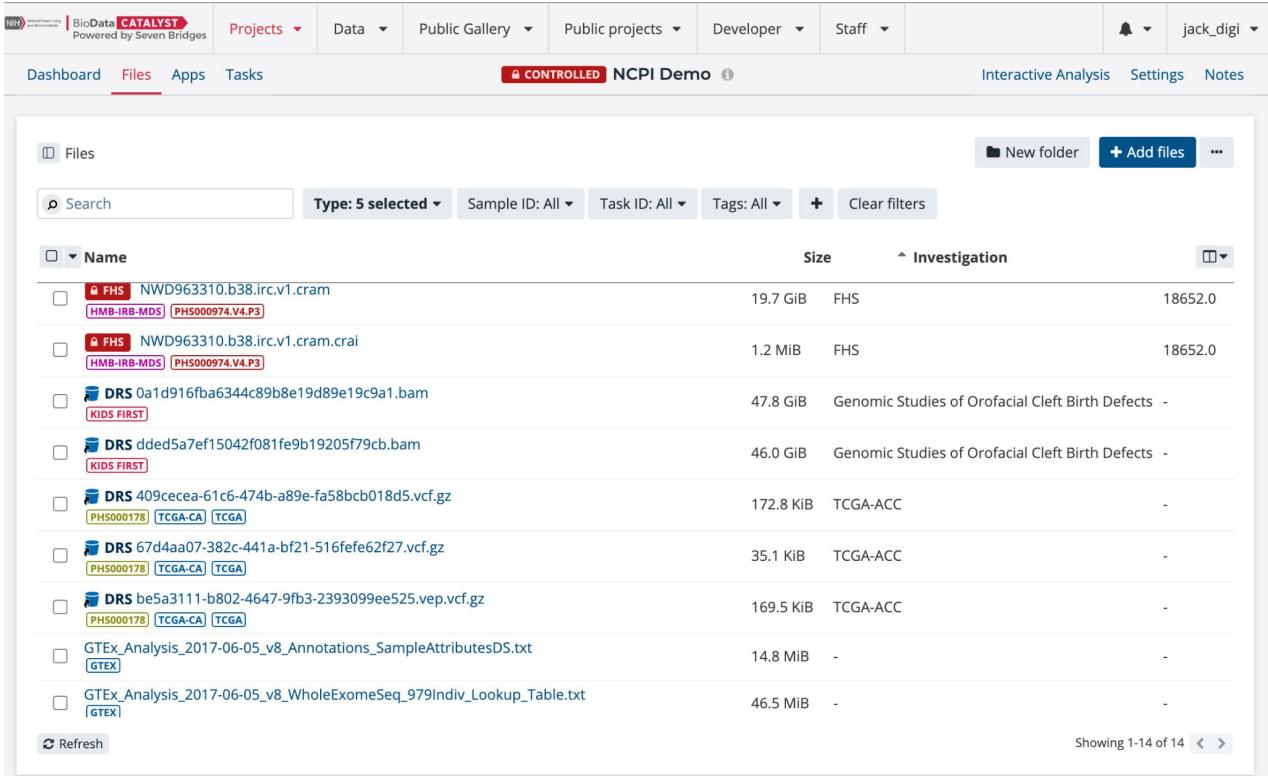
BDC push to SB, GTEx via manifest





All the data!

TOPMed, KidsFirst, TCGA, GTEx all in one workspace



The screenshot shows the BioData CATALYST workspace interface. At the top, there are navigation links: Projects, Data, Public Gallery, Public projects, Developer, Staff, a notification bell, and a user dropdown for 'jack_digi'. Below the header, there are tabs for Dashboard, Files (which is selected), Apps, and Tasks. A red banner indicates 'CONTROLED' access for the 'NCP1 Demo' project. To the right of the banner are links for Interactive Analysis, Settings, and Notes.

The main area displays a list of files. The columns are labeled: Name, Size, and Investigation. The 'Name' column includes checkboxes and small icons (blue square with white checkmark, red square with white FHS, blue square with white DRS, etc.). The 'Size' column shows file sizes like 19.7 GiB, 1.2 MiB, etc. The 'Investigation' column lists study names like 'FHS', 'Genomic Studies of Orofacial Cleft Birth Defects', 'TCGA-ACC', and 'GTEx'. There are also small colored squares next to some investigation names (red for FHS, blue for DRS, green for TCGA, orange for GTEx). Filter options at the top include 'Type: 5 selected', 'Search', and 'Clear filters'.

Name	Size	Investigation
NWD963310.b38.irc.v1.cram HMB-IRB-MDS PH5000974.V4.P3	19.7 GiB	FHS
NWD963310.b38.irc.v1.cram.crai HMB-IRB-MDS PH5000974.V4.P3	1.2 MiB	FHS
D RS 0a1d916fba6344c89b8e19d89e19c9a1.bam KIDS FIRST	47.8 GiB	Genomic Studies of Orofacial Cleft Birth Defects -
D RS dded5a7ef15042f081fe9b19205f79cb.bam KIDS FIRST	46.0 GiB	Genomic Studies of Orofacial Cleft Birth Defects -
D RS 409cecea-61c6-474b-a89e-fa58bcb018d5.vcf.gz PH5000178 TCGA-CA TCGA	172.8 KIB	TCGA-ACC
D RS 67d4aa07-382c-441a-bf21-516fefef62f27.vcf.gz PH5000178 TCGA-CA TCGA	35.1 KiB	TCGA-ACC
D RS be5a3111-b802-4647-9fb3-2393099ee525.vep.vcf.gz PH5000178 TCGA-CA TCGA	169.5 KiB	TCGA-ACC
GTEX_Analysis_2017-06-05_v8_Annotations_SampleAttributesDS.txt GTEx	14.8 MiB	-
GTEX_Analysis_2017-06-05_v8_WholeExomeSeq_979Indiv_Lookup_Table.txt GTEx	46.5 MiB	-

At the bottom left is a 'Refresh' button, and at the bottom right is a page number 'Showing 1-14 of 14' with navigation arrows.

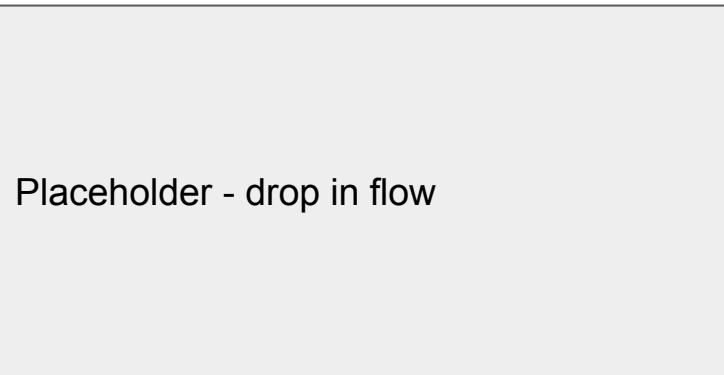
Tim will also show this tomorrow.

Quick Demo!



Current user experience: TOPMed, KidsFirst, TCGA, GTEx all in one workspace

All the data, but all
the auth!



Placeholder - drop in flow

- eRA Commons (6x)
- Gen3 AnVIL API key
- BDC + CAVATICA
auth_tokens

Fortunately, all of
these systems also
talking to RAS



Overview

Connected Data

USE CASES

Tech Successes

Lessons Learned & Next Steps



Systems Interop WG mission



The group will spearhead technical improvements to cloud "stacks" created by the Common Fund, NCI, NHGRI, and NHLBI that enable improved interoperability. We will demonstrate progress in realistic researcher use cases every 6 months.

Want more info? Check out the WG [charter](#). Iff you are interested, please [join](#).

Goals of these updates

Project info is **fresh**

Blockers identified

Outcomes **curated** for NCPI biannual meetings

New WG members **onboarded**



Image credit: Adrian Black on [Flickr](#)

Current Use Cases

(additional detail)

LEAD	ONE-LINE SUMMARY	STATUS	
Gelb	PCGC (BDC, KF) <i>de novo mutations</i> with graph callers	Inactive	
Grossman	PCGC (BDC, KF) & Vandy AFib joint calling, annotation, and GO enrichment; <i>interop/tech focus</i>	Active	
Gharavi	GTEx (AnVIL, KF, BDC) find datasets as healthy controls	Active	
Lyons	User journey from PICSURE-API to Platform (TOPMed) for variant level info	<i>In Prep</i>	
Stranger	TCGA, GTEx (CRDC, AnVIL) sex-DE on normal & tumor	Inactive	
Manning	PCGC, GTEx, F/JHS (BDC, KF, AnVIL) genetic factors in CHD	Active	
Almeida	IDC (CRDC) tile server for autoML image analysis; bearer token auth	Active	
Goldmuntz, Taylor, et al.	PCGC (BDC, KF) joint calling, harmonization, gene set analysis + ML	Active	



Overview

Connected Data

Use Cases

TECH SUCCESSES

Lessons Learned & Next Steps



PFB, FHIR, or other approaches?

Many different manifests/ mechanisms to describe cohorts

Simple manifest → optimized & performant (great!), not self-describing so difficult to generalize

FHIR/Bulk FHIR → standard (great!) and future direction but will take time for systems to build full FHIR clients. Works well if your data maps

PFB → generic manifest, self-describing (great!), based on an open standard (Avro, great!), and easy step up to support from manifests

QUESTION: Sweet spot between manifests and FHIR?

Future Question: where does the ga4gh selection-object fit?

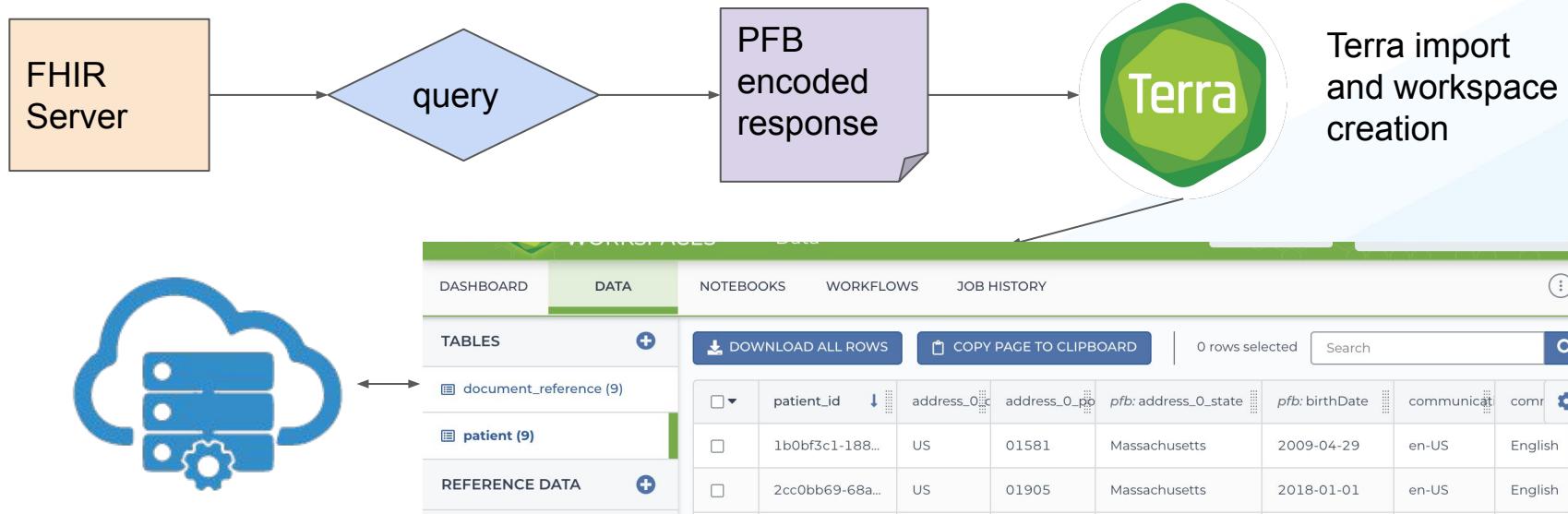


Goals for FHIR -> PFB experiment

- **Goal:** provide a connection from AnVIL, BDCat, CRDC, and GMKF portals to workspaces like Terra, SBG, and Gen3. Provide an interim path to use other FHIR servers without developing FHIR clients first.
- AnVIL and BDCat portals support PFB → Workspace handoff
- Kids First Data Portal has a FHIR server
- Can we use FHIR → PFB as a handoff mechanism?
 - Make this generic
 - Useful for multiple FHIR servers beyond GMFK (dbGaP, AnVIL...)
 - Ensure this is scalable, deployable by others, web service based

Prototyped a bridge from FHIR -> PFB

We prototyped a bridge between FHIR and PFB



Useful since it bridges the queries FHIR affords to the workspace environments (AnVIL, BDCat, CRDC) that offer compute on data.

Patient table

Terra BETA WORKSPACES Workspaces > nimbus-pfb-test/PFB Test > Data COVID-19 Data & Tools Cloud Environment NONE

DASHBOARD DATA NOTEBOOKS WORKFLOWS JOB HISTORY

TABLES + DOWNLOAD ALL ROWS COPY PAGE TO CLIPBOARD 0 rows selected Search

	patient_id	address_0_id	address_0_pp	pfb:address_0_state	pfb:birthDate	communicat	communicat	communication_0_id	communicat	pfb:gen	⋮
document_reference (9)	1b0bf3c1-188...	US	01581	Massachusetts	2009-04-29	en-US	English	urn:ietf:bcp:47	English	female	
patient (9)	2cc0bb69-68a...	US	01905	Massachusetts	2018-01-01	en-US	English	urn:ietf:bcp:47	English	male	
REFERENCE DATA +	3caf8b35-8f6...	US	02138	Massachusetts	2010-03-30	en-US	English	urn:ietf:bcp:47	English	female	
OTHER DATA	421f6c7e-bbb...	US	02129	Massachusetts	2015-10-20	en-US	English	urn:ietf:bcp:47	English	male	
Workspace Data	50903d89-d1...	US		Massachusetts	2009-05-02	en-US	English	urn:ietf:bcp:47	English	male	
Files	7357464b-97...	US		Massachusetts	2017-05-13	en-US	English	urn:ietf:bcp:47	English	male	
	7da367de-ad...	US	02141	Massachusetts	2006-02-22	en-US	English	urn:ietf:bcp:47	English	female	
	8a4c5099-7dc...	US		Massachusetts	2011-12-12	en-US	English	urn:ietf:bcp:47	English	female	
	8d6b4d4d-c2...	US		Massachusetts	2010-05-01	en-US	English	urn:ietf:bcp:47	English	male	

Reference table

The screenshot shows the Terra Workspaces interface. The top navigation bar includes 'WORKSPACES' (BETA), 'Data' (selected), 'Notebooks', 'Workflows', and 'Job History'. On the right, there are links for 'COVID-19 Data & Tools', 'Cloud Environment NONE', and a gear icon. The main area displays a table titled 'document_reference (9)'. The table has columns for 'document_r...', 'entry_0_full...', 'pfb:entry_0_resource_content_0_attachment_url', 'pfb:entry_0_resource_id', 'entry_0_resource_i...', and a settings gear icon. Each row contains a checkbox, a URL, and a detailed URL. The rows correspond to the following data:

document_r...	entry_0_full...	pfb:entry_0_resource_content_0_attachment_url	pfb:entry_0_resource_id	entry_0_resource_i...	
<input type="checkbox"/>	05dfd1c8-6ba...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	
<input type="checkbox"/>	0e19ecaf-4c2...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	
<input type="checkbox"/>	3205d2fd-1e8...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	
<input type="checkbox"/>	34ff1c2e-aeca...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	
<input type="checkbox"/>	37a84061-ba...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	
<input type="checkbox"/>	598a2353-24...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	
<input type="checkbox"/>	7b0f1671-36...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	
<input type="checkbox"/>	83268d9a-f9b...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	
<input type="checkbox"/>	a39aed0-60...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	https://kf-api-dataservice.kidsfirstdrc.org/dr...	



RAS is providing authentication



NIH Researcher Auth Service (RAS) Sign In

Username

Password

[Forgot Password?](#)

View consent options upon login

[Sign In](#)

Smart Card Holder? [Sign in with PIV Card.](#)

[Trouble signing in?](#)

All systems completed
Milestone 1

Milestone 2 in progress

Spirited discussion and efforts on Milestone 3 design



DRS for CONTROLLED DATA

- DRS 1.2 is an upcoming public interface standard
 - RAS standard for authorization is the GA4GH passport
 - Passport authorizes authenticated user to access content
-
- No common token system (OAuth/OIDC/GA4GH/RAS) authenticates client system
 - DRS uses a Clearinghouse that is tightly bound to source of authority



DRS for CONTROLLED DATA

- User authenticates through RAS
 - Passports issued by RAS or known brokers contain RAS visas
 - Passports delivered to DRS via POST to increase success rate
-
- DRS validates passport via controlled-access Clearinghouse
 - DRS ultimately returns a URI to access resource



Overview

Connected Data

Use Cases

Tech Successes

LESSONS LEARNED & NEXT STEPS

Lessons learned - technical

We started with a very strict user definition to **build a solution for the largest audience**. We had to relax this assumption temporarily

PFB/Avro manifests are promising, but **there's no free lunch**

A **single AuthN/Z** would simplify development and improve UX



Lessons learned - humans

It's **extremely difficult** to engage the Sys Interop audience

- Attending *defensively* to ensure things don't go off course but lacking funding / resources / time to *drive the boat*?
- Some other blocker?

We are going to reach out to individual groups to present - increase information flow, spark collaborations

Request to understand funding; make contributing to NCPI Sys Interop a deliverable of future funding; help researchers get credit for success



Calls to action

1. Alignment on **RAS Milestone 3** and to get there as quickly as possible. This is currently blocking widespread use of DRS.

1. **We need active use cases** now that policy blockers have been removed and technical blockers are reducing
 - a. If Data Portals have active user communities seeking additional analysis capabilities, **help build functionality** to participate.



Summary

All Portals *have a path* to all workspaces

Resolved most technical concerns identified in last meeting

Two use cases have completed successfully, other in development - we need more engaged researchers.

What's next:

- Near-term:
 - Using (equivalent) tools on multiple platforms
 - Connect with NCBI DRS Server
- Mid-term: Stay tuned for the *Future of Interop* talk tomorrow
- **Please provide your feedback - it will influence our roadmap**



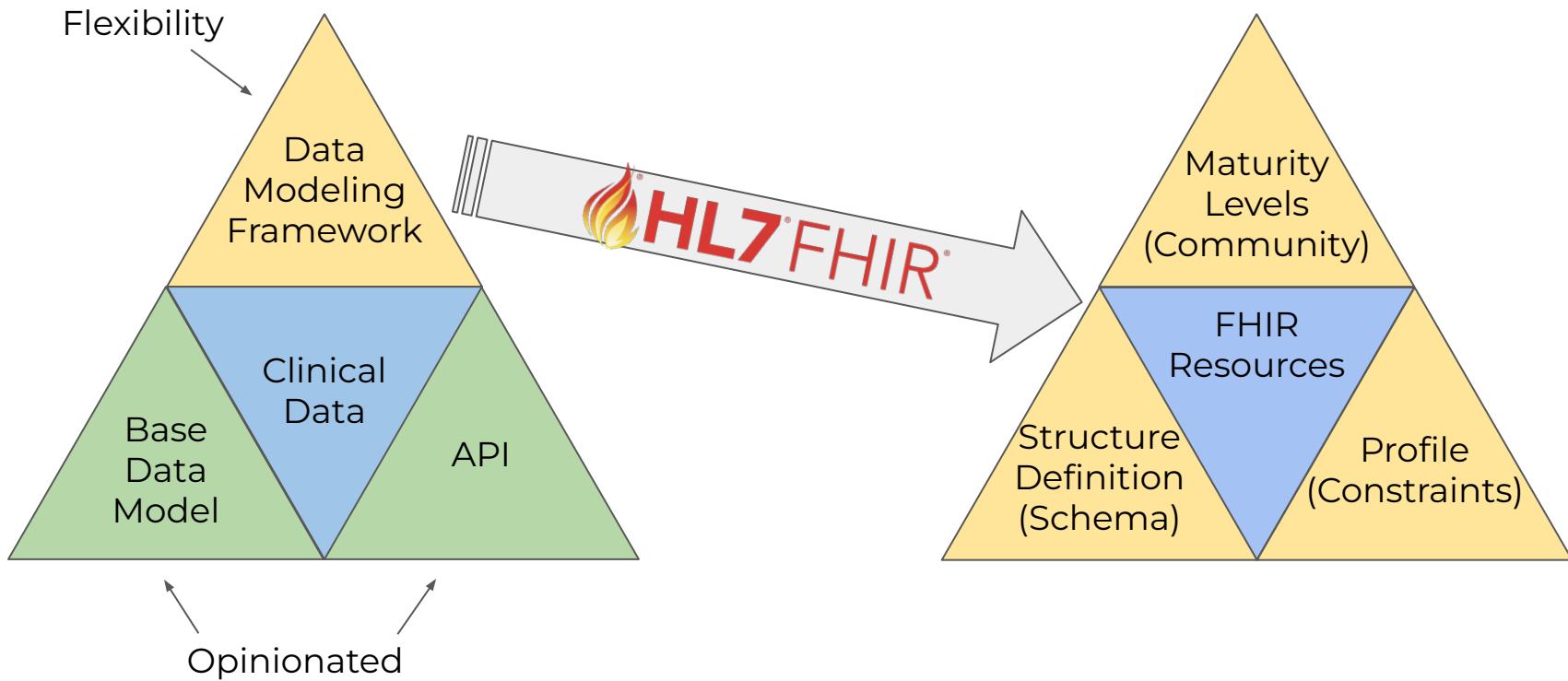
NCPI FHIR WG Update

NCPI Spring 2021 Workshop

May 3, 2021

Allison Heath (CHOP)
Eric Torstenson (VUMC)

Where We Left Off Last Time: Framework for Clinical Data Interoperability





Focus of Last Six Months

- **NCPI Implementation Guide Development**
 - Use Case Gathering
 - Profiling
 - Utilizing FHIR Shorthand (FSH)
- **Path Towards Production**
 - Server evaluation
 - RAS and Controlled Access
- **Tooling and Initial Utilization**
 - PIC-SURE bulk FHIR import
 - PFB to FHIR
 - NCPI Dashboard



What are FHIR Implementation Guides?

- **Implementation Guide (IG)**: set of rules about how FHIR resources should be used to solve a particular problem
- FHIR describes a general set of capabilities to solve many data exchange problems
- FHIR IGs describe how FHIR is used in particular contexts
 - Jurisdiction Base:
 - US Core FHIR Profiles: <http://hl7.org/fhir/us/core/>
 - Application Solution:
 - Bulk Data Access (Flat FHIR): <http://hl7.org/fhir/uv/bulkdata/>
 - Domain Guide:
 - Clinical Genomics Reporting: <http://build.fhir.org/ig/HL7/genomics-reporting/>
- Registry of IGs: <http://www.fhir.org/guides/registry/>



NCPI FHIR IG v0.1.0 - Key Use Cases

- **Representing Research Studies**
 - ResearchStudy, ResearchSubject
 - DRS Document Reference
- **Rare Diseases**
 - NCPI Phenotype, NCPI Disease, NCPI Family Relationship
- **Childhood Cancer**
 - NCPI Phenotype, NCPI Disease, NCPI Family Relationship
- **Existing Study Data**
- **EHR Data**
- **Draft**

Existing Study Data - CARING Example (POPS)

The screenshot shows a Microsoft Excel spreadsheet titled "Dosing Information" with a search bar containing "Remdesivir (GC8)". The "Home" tab is selected in the ribbon. A red box highlights the search bar, and a red arrow points from it to a JSON tooltip displayed over the data table. The tooltip contains the following JSON:

```
{  
  "id": "med0301",  
  "code": {  
    "coding": [  
      {  
        "code": "2284958",  
        "display": "remdesivir Injectable Product",  
        "system": "http://www.nlm.nih.gov/research/umls/rxnorm"  
      }  
    ]  
  },  
  "resourceType": "Medication"  
}
```

The table in the Excel spreadsheet has columns: Protocol, Site (Redacted), Participant ID, Project ID, and Site (Redacted). Rows 1-4 show data corresponding to the JSON entry.

- Information captured in file names (not easily accessible) is transformed into clear, explicitly stated data
- Normalized to a controlled vocabulary
- Medication resource is then linked to all cases where it is used, via the ID



Existing Study Data



Project ID	Segment	FLACC predose not assess	FLACC predose dt (Redacted to)
1491	POP02- Active	Not assessed	1
1491	POP02- Active	Not assessed	1
1506	POP02- Active	Not assessed	0

- Pain scale assessment as a Observation
- Choose appropriate controlled vocabulary
- Provide context for the measure (reference)

```
"referenceRange": [
  {
    "high": {
      "code": "{score}",
      "system": "https://ucum.org/trac",
      "value": 10
    },
    "low": {
      "code": "{score}",
      "system": "https://ucum.org/trac",
      "value": 0
    }
  ],
  "status": "final",
  "subject": {
    "reference": "Patient/001"
  },
  "valueQuantity": {
    "code": "{score}",
    "system": "https://ucum.org/trac",
    "value": 0
  },
  "code": {
    "coding": [
      {
        "code": "38215-0",
        "display": "Pain severity total Score FLACC",
        "system": "http://loinc.org"
      }
    ]
  },
  "resourceType": "Observation"
```

Terminology Usage in NCPI IG

7.4.1 Resource Profile: NCPI Phenotype

Defining URL:	https://ncpi-fhir.github.io/ncpi-fhir-ig/StructureDefinition/phenotype
Version:	0.1.0
Name:	Phenotype
Title:	NCPI Phenotype
Status:	Draft as of 2021-04-29T15:00:34-05:00
Definition:	Representation of phenotypic observations (present or absent)
Publisher:	NCPI FHIR Working Group
Source Resource:	XML / JSON / Turtle

[Text Summary](#)

[Differential Table](#)

[Snapshot Table](#)

[Snapshot Table \(Must Support\)](#)

This structure is derived from [Condition](#) ↗

Name	Flags	Card.	Type	Description & Constraints	?
Condition		0..*	Condition	Detailed information about conditions, problems or diagnoses	?
code		0..1	CodeableConcept	Identification of the condition, problem or diagnosis Binding: Phenotype Codes (required)	

7.10.1.2 Expansion

This value set contains 1880 concepts

Expansion based on [Human Phenotype Ontology v0.1.0 \(CodeSystem\)](#)

FSH for IG Development

- NCPI FHIR IG development using **FSH** and **SUSHI**
 - Easier to read, write, validate, and curate FHIR resources than with JSON/XML
 - Allows rapid and collaborative development with accessible tracking changes

```
{  
  "resourceType": "StructureDefinition",  
  "id": "ncpi-phenotype",  
  "url": "http://fhir.ncpi-project-forge.io/StructureDefinition/ncpi-phenotype",  
  "version": "0.1.0",  
  "name": "ncpi-phenotype",  
  "title": "NCPI Project Forge Human Phenotype",  
  "status": "draft",  
  "fhirVersion": "4.0.0",  
  "kind": "resource",  
  "abstract": false,  
  "type": "Observation",  
  "baseDefinition": "http://hl7.org/fhir/StructureDefinition/Observation",  
  "derivation": "constraint",  
  "differential": {  
    "element": [  
      {  
        "id": "Observation",  
        "path": "Observation"  
      },  
      {  
        "id": "Observation.code",  
        "path": "Observation.code",  
        "binding": {  
          "strength": "required",  
          "valueSet": "http://fhir.ncpi-project-forge.io/ValueSet/phenotype-codes"  
        }  
      },  
      {  
        "id": "Observation.valueCodeableConcept",  
        "path": "Observation.valueCodeableConcept",  
        "binding": {  
          "strength": "required",  
          "valueSet": "http://fhir.ncpi-project-forge.io/ValueSet/phenotype-observation-codes"  
        }  
      },  
      {  
        "id": "Observation.interpretation",  
        "path": "Observation.interpretation",  
        "binding": {  
          "strength": "required",  
          "valueSet": "http://fhir.ncpi-project-forge.io/ValueSet/phenotype-interpretation"  
        }  
      }  
    ]  
  }  
}
```

JSON Profile (Project Forge)

Profile: Phenotype

Parent: Condition

Id: phenotype

Title: "NCPI Phenotype"

Description: "Representation of phenotypic observations (present or absent)"

* ^version = "0.1.0"

* ^status = #draft

* code from phenotype-codes (required)

FSH Profile (NCPI IG)



IG Development on Github

- A pre-release IG is available via [GitHub Pages](#)
 - Profiling
 - [Condition](#) >> [Disease](#) and [Phenotypic Feature](#)
 - [Observation](#) >> [Family Relationship](#) (Pedigree)
 - [DocumentReference](#) >> [Data Repository Service \(DRS\) Document Reference](#)
 - Use Cases
 - Research Representation
 - Rare Disease
 - Childhood Cancer
 - EHR Data
 - Background
 - FHIR Relevance
- Feedback is welcome at the [repository](#) (issues, PR requests, etc.)
- Hands on IG development group meets every other week



FHIR Server/Platform Evaluation

- Multiple Servers/Platforms to be Tested
 - HAPI/Smile CDR, Google Healthcare API, Azure API for FHIR
- Test Suite Objectives
 - Common set of tests to run against any available FHIR platform allows clear differentiation between different platform offerings
 - Use case driven test suite and test data
 - Weighted test score provides easy mechanism to compare all tested platforms
- Status
 - Framework exists at [github](#)
 - Tests/test stubs are laid out to follow the [google doc](#)
 - Test data can be “imported” from bulk-export or be hand generated
- Reports
 - High-level (summary) [overview](#)
 - Detailed/test level [overview](#)

Test Suite - Example Summary Report

Module_Name	Test_ID	Score	Total_Possible_Score	Perc	#Tests_Passed	#Tests_Failed	#Total_Test_Count
Cap. Statement	2.1.1 - Resource Interaction	1.7777777777777777	2	0.8888888888888888	832	104	936
Cap. Statement	2.1.2 - Conditional Create, Upd & Del	4.4444444444444445	5	0.8888888888888888	104	13	117
Cap. Statement	2.1.3 - Search Includes	1.1111111111111112	2	0.5555555555555556	65	52	117
Cap. Statement	2.1.4 - Resource Search Params	0.717948717948718	2	0.358974358974359	42	75	117
Cap. Statement	2.1 - Cap. Statement (Summary)	8.051282051282051	11	0.8104118104118104	1043	244	1287
Search	2.2.1.1 - Core FHIR Search	15.0	15	1.0	1	0	1
Search	2.2.1.2 - Search Modifiers and Prefix	0.0	15	0.0	0	4	4
Search	2.2.1.3 - Hierarchical Search	0.0	15	0.0	0	2	2
Search	2.2.1.4 - Chaining	0.0	8	0.0	0	2	2
Search	2.2.1.5 - Reverse Chaining	0.0	5	0.0	0	2	2
Search	2.2.1.6 - Missingness	0.0	8	0.0	0	1	1
Search	2.2.1.7 - Composite Search	0.0	9	0.0	0	2	2
Search	2.2.1.8 - _query	0.0	8	0.0	0	2	2
Search	2.2.1.9.1 - Sorting	0.0	3	0.0	0	2	2
Search	2.2.1.9.2 - Paging	0.0	8	0.0	0	2	2



Summary and Next Steps

- Refining NCPI IG
 - Use case and background documentation
 - Guidelines on using existing FHIR resources
 - Terminology selection
 - GA4GH pedigree cross-informing
- Platform Specific FHIR Servers
 - Kids First DRC (end of May, similar timeline for CARING)
 - dbGaP
 - AnVIL
 - Continue to support NCPI FHIR “testbed” servers with KFDRC and synthetic data
- Tooling and API Usage
 - Interchange, Search, Mapping, and Provenance
 - Prioritize based on emerging needs
 - Integrations using Jupyter Notebooks and Shiny Apps in cloud workspaces

Thank You to All Working Group Members

Running Agenda

Members Across:
Kids First DRC
AnVIL
BDC
NCI CRDC
NLM/NCBI
CFDE

NCPI IG Contributors

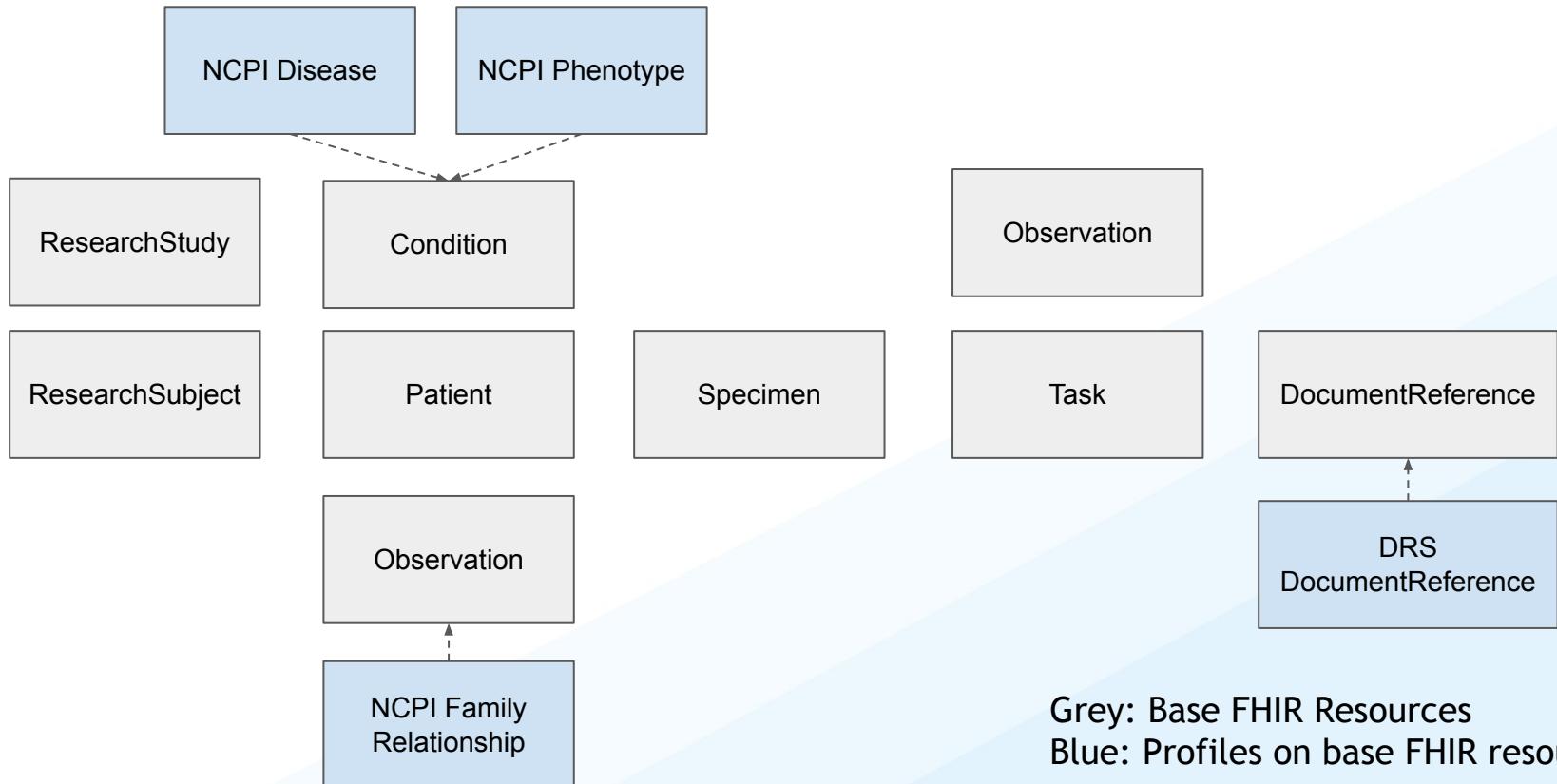
2.1 Authors

Author	Contact
Robert Carroll	RobertJCarroll ↗
Shahim Essaid	ShahimEssaid ↗
Allison Heath	allisonheath ↗
Avital Kelman	fiendish ↗
Meen Chul Kim	liberaliscomputing ↗
Nicholas Van Kuren	nicholasvk ↗
Natasha Singh	znatty22 ↗
Eric Torstenson	torstees ↗
Brian Walsh	bw Walsh ↗



Questions?

Current NCPI IG Profiles



30 Minute Break #1

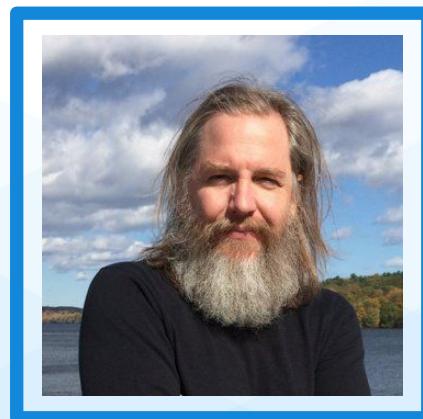
We will resume at 1:00 pm EDT

Announcements

- Fall 2021 Workshop poll: **tinyurl.com/NCPIfallpoll**
- If you have not registered, please do: **tinyurl.com/NCPIregistration**
- The NIH Office of Data Science Strategy recently announced four Notices of Special Interest for supplemental funding: **tinyurl.com/ODSSfunding**

Working Group Update NCPI Outreach

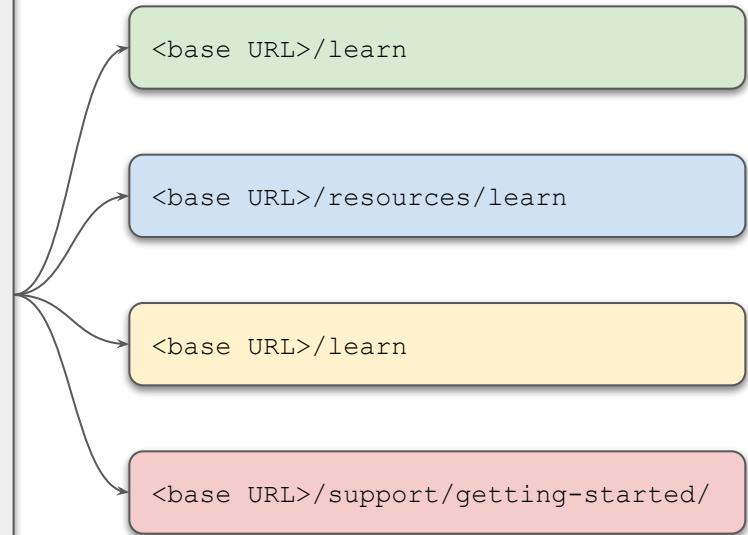
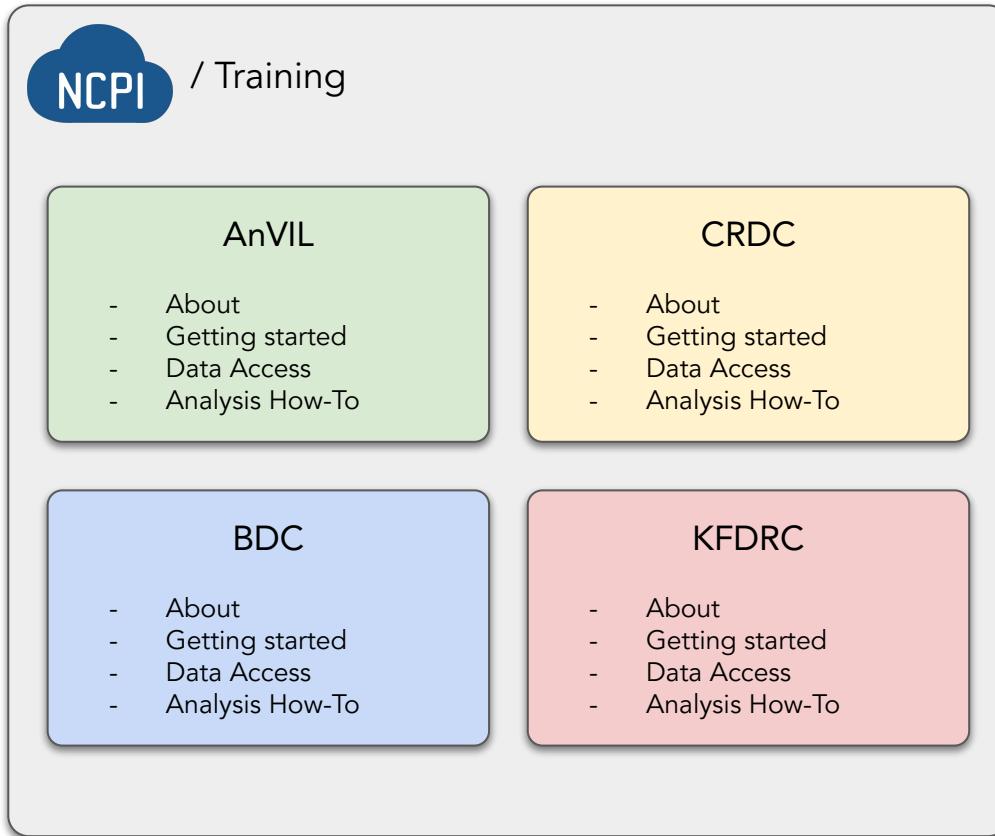
Presenter #1 Dave Rogers
UCSC



Presenter #2 Anton Nekrutenko
Penn State / Galaxy



Training docs for each platform



Outreach objectives

- Landing page for documentation
- Data dashboard

Training docs for each platform | Steps

- Work with outreach person from each platform
- Identify common types of materials
- Develop tagging scheme à la:

Introduction

Basic

Intermediate

Advanced

Data access

Data analysis

...

- Documentation by technology (Galaxy, Terra, Gen3, Jupyter, RStudio, 7B)
- PR #1036

Anton's Demo Here

Global Data Dashboard | Current Status

- We received a list of datasets (a spreadsheet) from all resources
- At this point we are focused *only* on datasets with dbGaP identifiers
- Metadata about these datasets can be fetched via calls against dbGaP FHIR interface

dbGaP FHIR

cc-dbgap-data-types.ipynb ★

File Edit View Insert Runtime Tools Help Changes will not be saved

Share 

Table of contents 

+ Code + Text 

Connect  Editing 

dbGap Data Types from dbGap FHIR Server     

Study Overview

Study JSON

Step 1. Install client & imports

Step 2. Make a request to server and parse JSON

Step 3. Print results.

Section 

dbGap Data Types from dbGap FHIR Server

A quick comparison of data types from dbgap study overview page and study data available from <https://dbgap-api.ncbi.nlm.nih.gov/fhir/x1/ResearchStudy> using study phs001395 as the test case.

Study Overview

The study NHLBI TOPMed - NHGRI CCDG: Hispanic Community Health Study/Study of Latinos (HCHS/SOL) (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001395.v1.p1) shows the following data types:

Molecular Data					
Type	Source	Platform	Number of Oligos/SNPs	SNP Batch Id	Comment
Whole Genome Sequencing	Illumina	HiSeq X Ten	N/A	N/A	Sequencing was performed at the Human Genome Sequencing Center at Baylor College of Medicine

Global Data Dashboard | The idea

OverviewDatasetsAnVIL

Search Summary

Platform	Studies	Subjects
AnVIL	21	59,325
BioData Catalyst	95	421,497
Kids First Data Resource Center	4	3,523
Cancer Research Data Commons	16	86,749
	136	571,094

Search Results

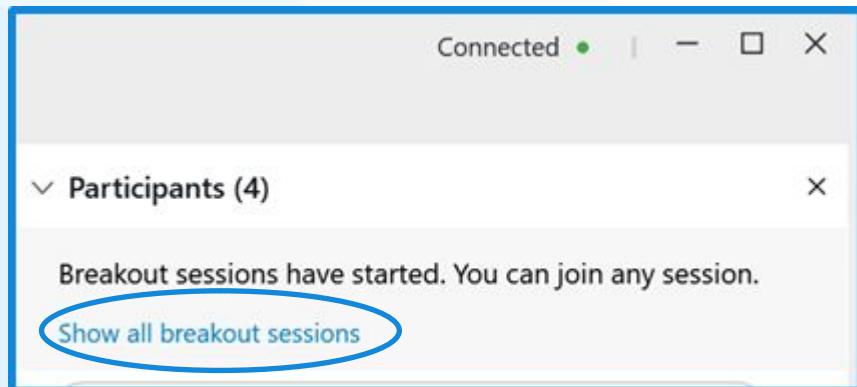
Platform	dbGap ID	Title	Diseases	Data Types	Consent Codes	Subjects
AnVIL	phs001272.v1.p1	Broad Institute Center for Mendelian Genomics	Genetic Diseases, Inborn; Bardet-Biedl Syndrome...	Genotype, SNP/CNV Genotypes (NGS)	HMB-MDS, GRU, DS-KRD-RD, DS-NIC-EMP-LENF	1,031
AnVIL	phs001913.v1.p1	CCDG - Cardiovascular: eMERGE - Northwestern Cohort	Cardiovascular Diseases	--	GRU-IRB	277
AnVIL	phs001502.v1.p1	CCDG-Cardiovascular: University of Pennsylvania Cohort	Cardiovascular Diseases	Genotype, Legacy Genotypes, SNP Genotypes (NGS)	HMB-IRB-PUB	1,373
AnVIL	phs001259.v1.p1	CCDG CVD: VIRGO - Variation in Recover-Role of Gender on Outcomes of Young Acute Myocardial Infarction (AMI) Patients	Myocardial Infarction; Inferior Wall Myocardial...	Genotype, SNP Genotypes (NGS)	DS-CARD-MDS-GSO	2,149
AnVIL	phs001894.v1.p1	CCDG-Neuropsychiatric: Autism- Genetics of Human Developmental Brain Disorders	Autism Spectrum Disorder	--	DS-EAC-PUB-GSO	724
AnVIL	phs001676.v1.p1	CCDG- Neuropsychiatric: Autism - Simons Simplex Collection (SSC)	Autism Spectrum Disorder	--	DS-AONDD-IRB	9,201
AnVIL	phs001740.v1.p1	CCDG- Neuropsychiatric: Autism- Study of Autism Genetics Exploration (SAGE)	Autism Spectrum Disorder	Genotype, SNP/CNV Genotypes (NGS)	DS-ASD-RD-IRB	580
AnVIL	phs001741.v1.p1	CCDG- Neuropsychiatric: Autism- The Autism Simplex Collection	Autism Spectrum	Genotype, SNP/CNV	DS-ASD-IRB	905

Dave's Demo Here

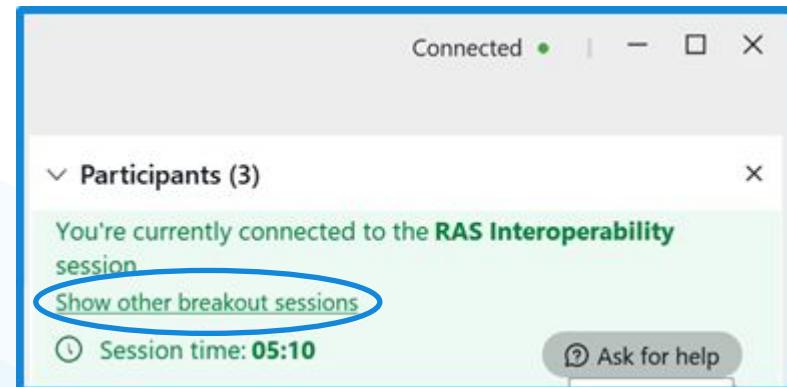
Many thanks to Outreach Group
members and Dr. Asiyah Lin

Breakout Groups: 1:20-2:30pm EDT

Please choose a Breakout Group: You must use the WebEx application



From the main session



From within another breakout group

30 Minute Break #2

We will resume at 3:00 pm EDT

Announcements

- Fall 2021 Workshop poll: **tinyurl.com/NCPIfallpoll**
- If you have not registered, please do: **tinyurl.com/NCPIregistration**
- The NIH Office of Data Science Strategy recently announced four Notices of Special Interest for supplemental funding: **tinyurl.com/ODSSfunding**

NCBI's Journey in Support of a Federated Cloud Data Sharing Ecosystem

Mike Feolo (NCBI)



NCBI Resources in Support of a Federated Cloud Data Sharing Ecosystem

Mike Feolo
Team Lead, dbGaP

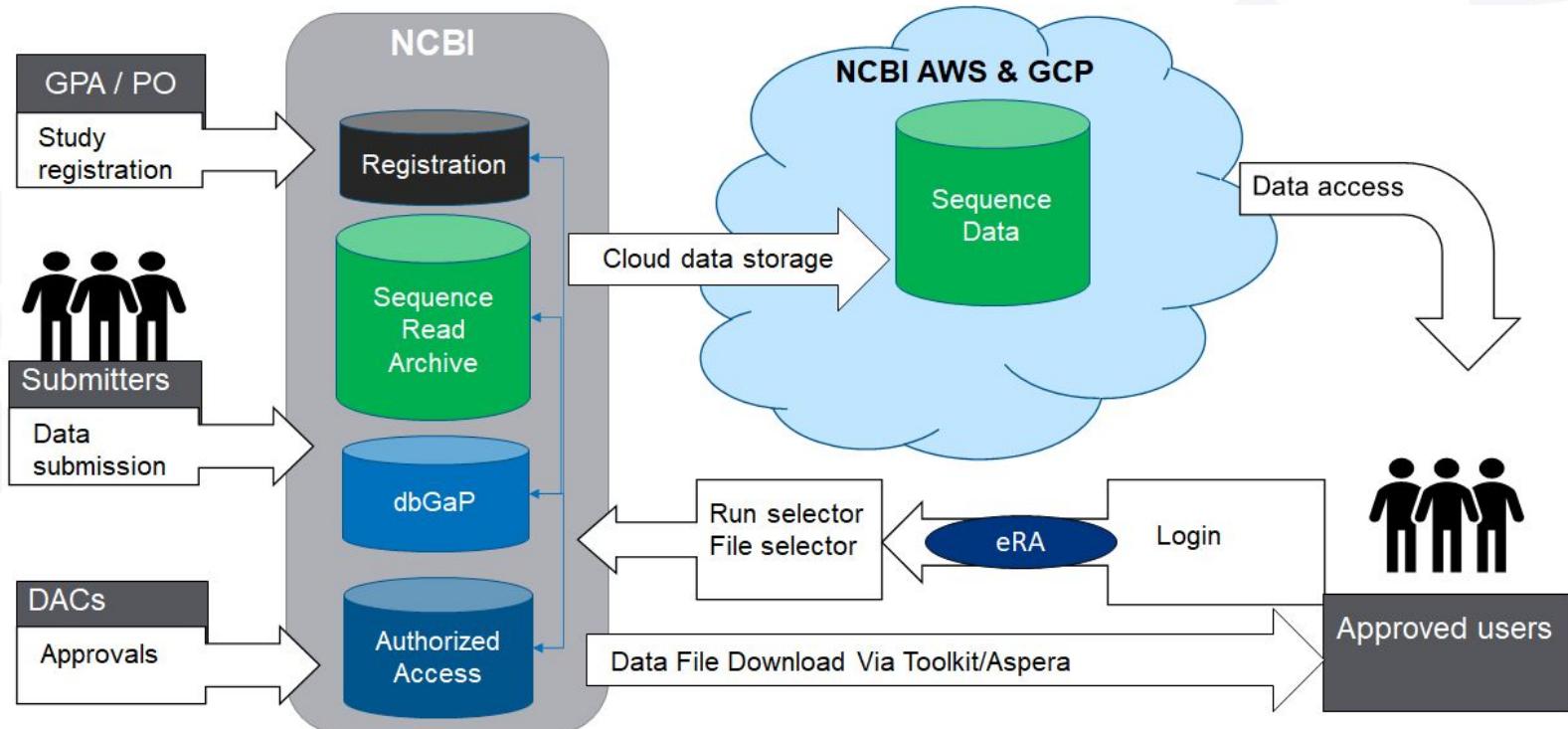


National Library of Medicine
National Center for Biotechnology Information

Overview

- NCBI's Controlled Access Data Sharing Architecture
- Study Registration
- Submission and Processing (dbGaP)
- Sequence Read Archive
- Request and Approval
- Data Access Tools/Services

NCBI's Data Sharing Architecture (current)



Study Registration

Who: NIH Genomic Program Administrators (GPAs), PIs

What:

- Instantiation of study at NCBI
- OMB / PRA Approved form
- Certification
- Consent / Use Restrictions
- Genomic Summary Results
- Data Access Committee designation
- Top Level Data Storage Access Information



Study Registration

Current Interactions with NCPI:

- Consent groups are established in this system
- Configuration of Data Storage details
 - GPA configures each study on registration
 - Curation path
 - Approval letters



Future Interactions with NCPI:

- API Access to system information
- Grant Compliance Reports

Submission and Processing (dbGaP)

Who: Study Investigators, Data Coordinating Centers (DCCs), Sequencing Centers

What:

- QA/QC, Study Accessions, Configures Release for
 - Study Metadata
 - Subject/Sample ids
 - Phenotype Data
 - Molecular Data
 - Analyses, Documents, and Images



Submission and Processing (dbGaP)

Current Interactions with NCPI:

- Study Metadata and Sample Accessioning
 - BioProject and BioSample are shared in INSDC
- Various Existing Telemetry Reports
- dbGaP-on-FHIR See: <https://anvilproject.org/ncpi/data>



Future Interactions with NCPI:

- API for programmatic access to metadata, data and Information
- Build out FHIR sever to deliver "observation" level phenotype data
- Configure **all** data on the Cloud with "RAS enabled" access

Authorized Access System

Who: Requesting Investigators, Signing Officials (SO), Data Access Committee (DAC) members

What:

- System to Request Data
 - Research Use Restrictions (consents)
 - Annual Reporting / Closeout
- Data Access Request (DAR) Review
- Gatekeeper of the NCBI-managed authorizations



Authorized Access System

Current Interactions with NCPI:

- Access Telemetry Reports (aka whitelists)

Future Interactions with NCPI:

- Researcher Auth Service (RAS; more about this later)
- Coordination of versioning and release signals

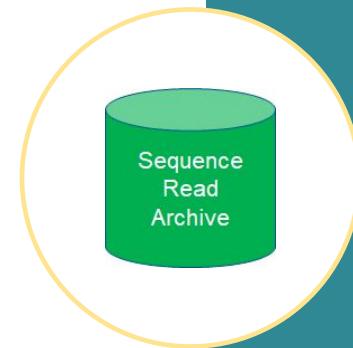


Sequence Read Archive

Who: Study Investigators, Data Coordinating Centers (DCCs), Sequencing Centers

What:

- Controlled Access Archive for sequencing data
- On-prem Storage: ETL of BAM, FASTQ
 - Configured for SRA Toolkit
 - Samples coordinated with dbGaP using BioSample
- Submitted data provisioned on the [Cloud](#) through STRIDES
- Run and Experiment level accessions for On-prem and cloud storage



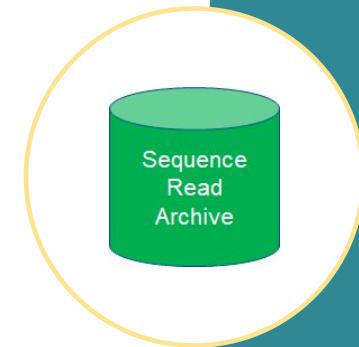
Sequence Read Archive

Current Interactions with NCPI:

- Run Metadata with cloud locations
- SRA Telemetry Reports
- INSDC identifiers in SRA, BioSample and BioProject level

Future Interactions with NCPI:

- API Access to Metadata?
- Direct submission of metadata from NCPI platforms?

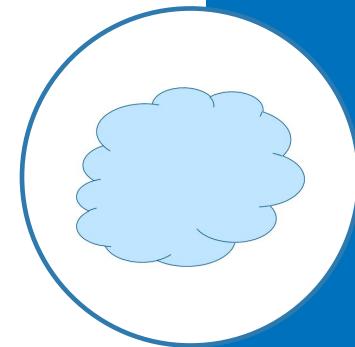


dbGaP Sequence data in the cloud

Who: Any dbGaP Authorized User

What:

- STRIDES funded provisioning of dbGaP sequencing files (4.8 PB of normalized data) into the AWS and GCP
- The oldest half of the data in cold Storage
- Files submitted by users (source files) are available in AWS & GCP cold storage through our new [Cloud Data Delivery](#) service that leverages the SRA Run Selector.



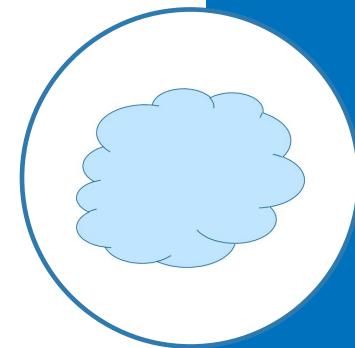
dbGaP Sequence data in the cloud

Interaction with NCPI partners

- Cloud locations are included in SRA metadata submission and are known to both SRA Run Selector and SRA Toolkit

Future Interactions with NCPI:

- Tutorials for NCPI users on how to get to NCBI-configured and cloud-accessible controlled-access data
- Integration of SRA Toolkit and other SRA services with RAS toward federated access of controlled-access datasets



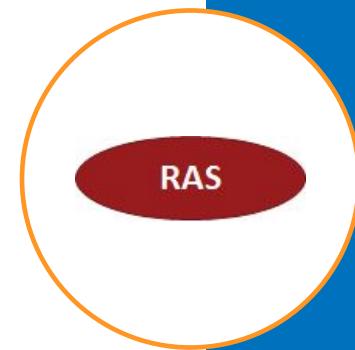
NCBI RAS Development

- GA4GH WG that develop specs for basis of RAS passports
- Piloting use of RAS Auth-Z tokens as part of RAS Phase-2
- NIH DAC authorizations are updated in RAS every 15 minutes
- DRS server supports STRIDES and is piloting use of GA4GH passports as authorization mechanism

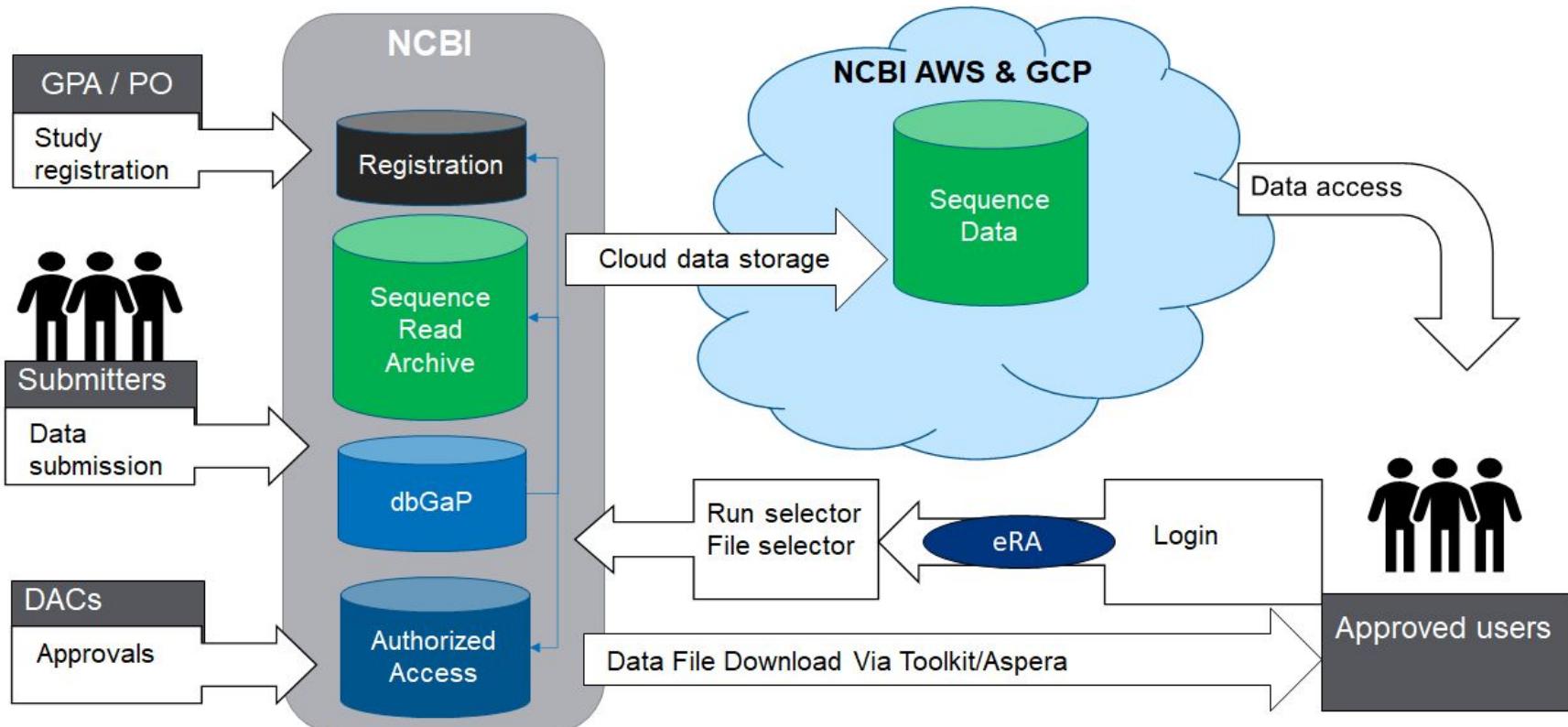


NCBI RAS Development

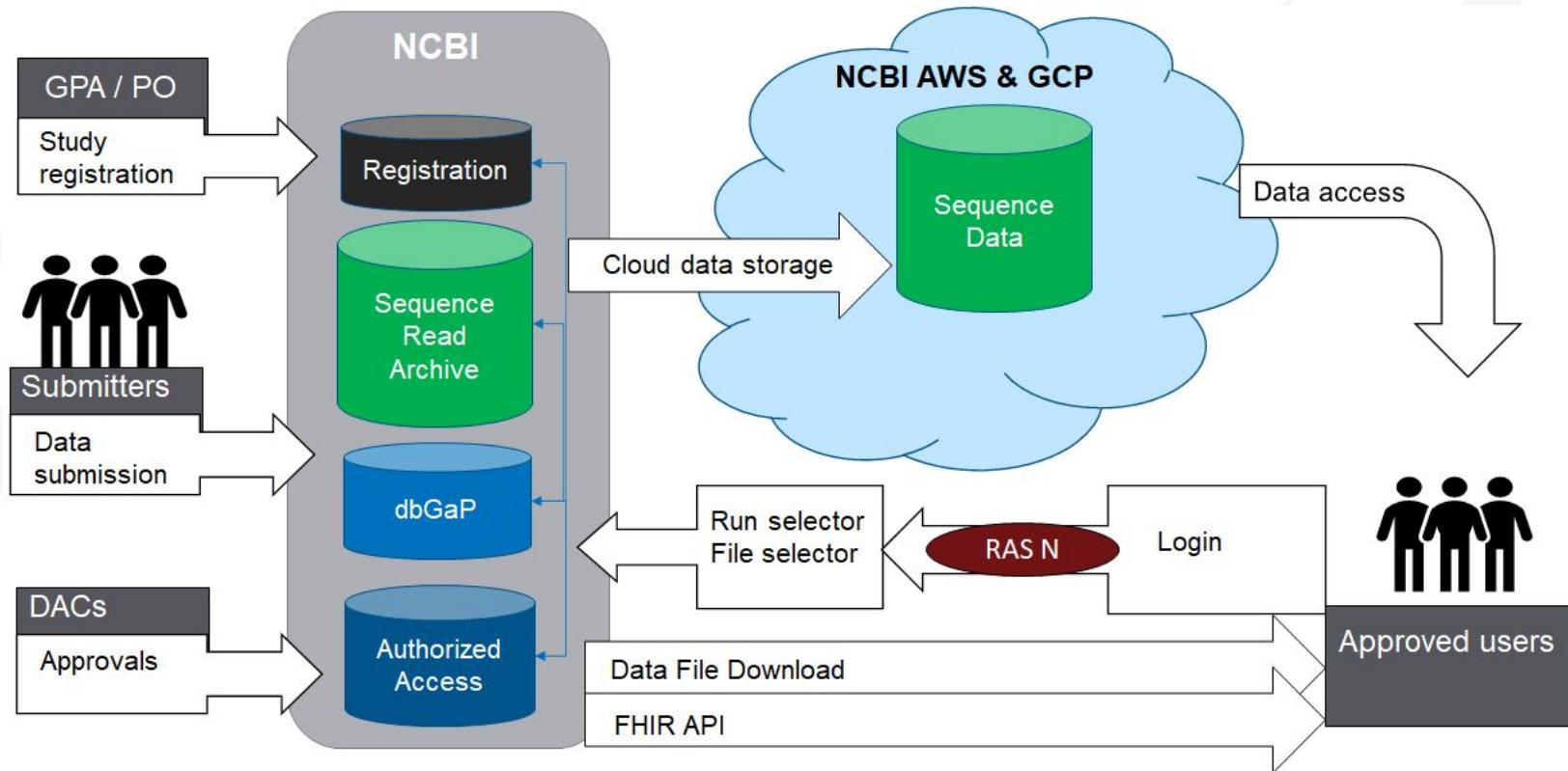
- SRA Run Selector: login through RAS and obtain passport, select files
- Data Repository Service accepts IDs and processes RAS passport through internal (NCBI) clearinghouse
- INSDC accessions translate to DRS through the IDX service
- URLs generated into AWS & GCP cloud buckets



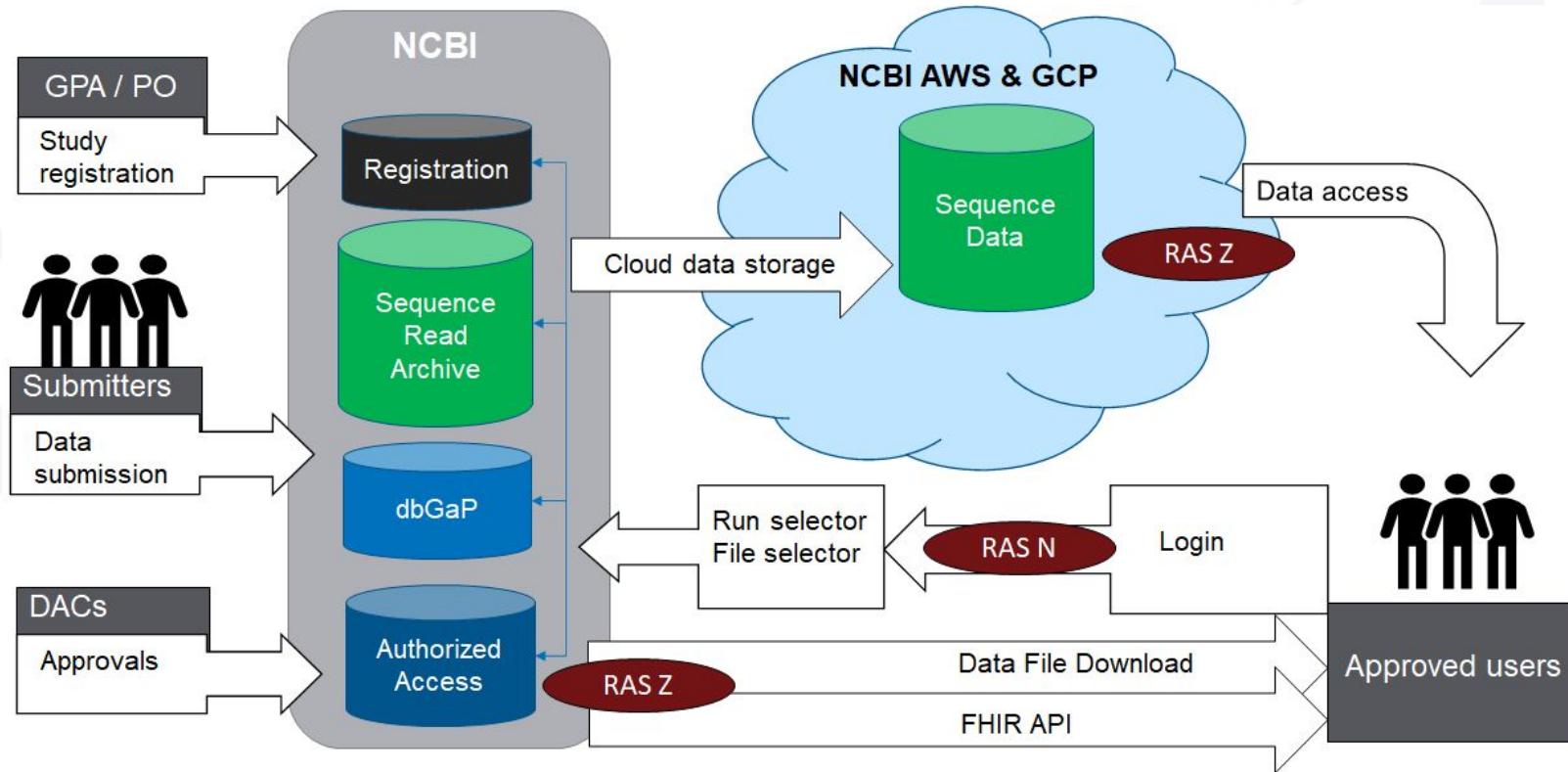
NCBI's Data Sharing Architecture (current)



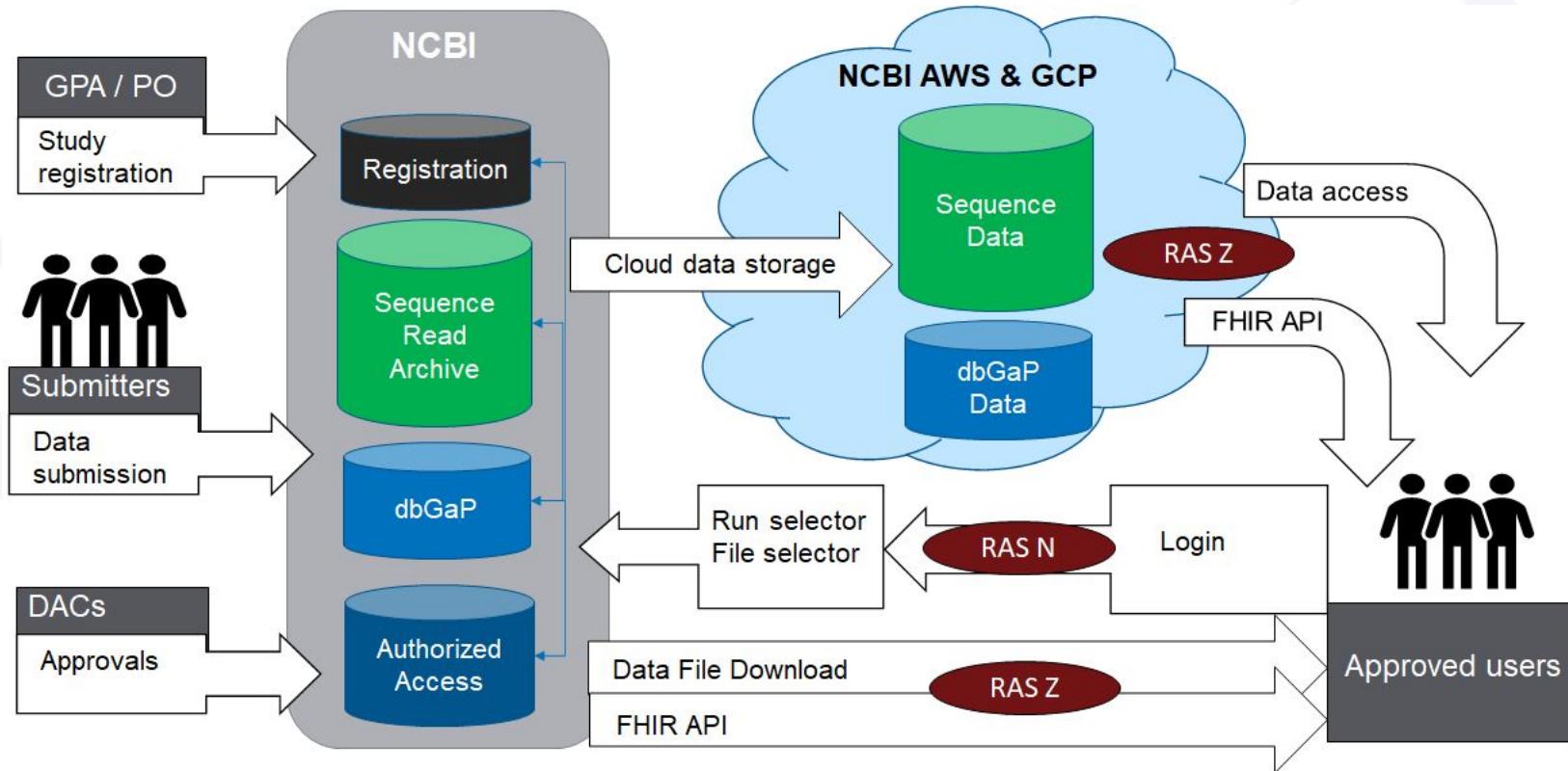
NCBI's Data Sharing Architecture (RAS N)



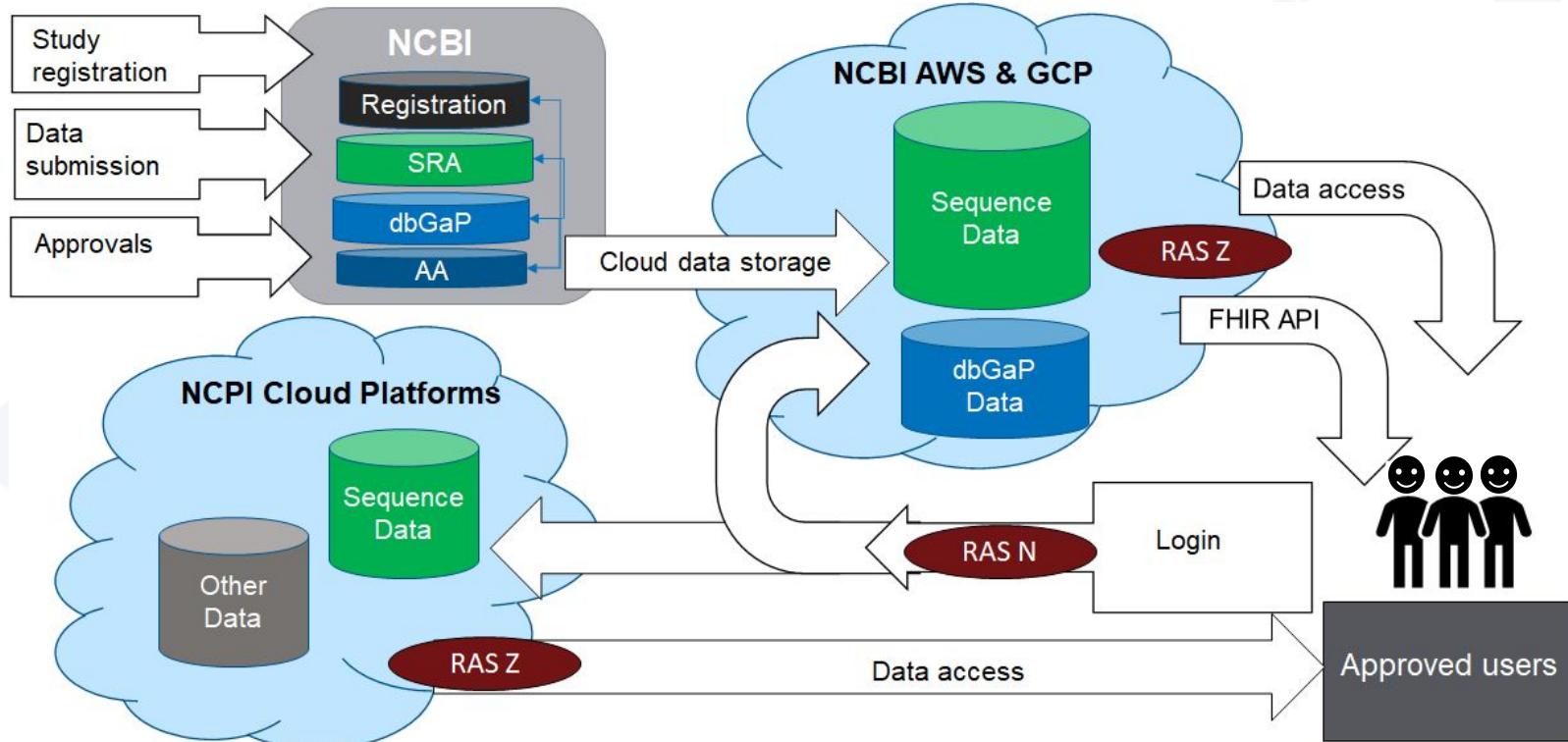
NCBI's Data Sharing Architecture (RAS Z)



NCBI's Data Sharing Architecture (dbGaP on Cloud)



NCBI's Data Sharing Architecture (Multiple Stores)



NCBI Points of Contact

Points of Contact

- dbGaP – Mike Feolo (feolo@ncbi.nlm.nih.gov)
- dbGaPonFHIR – Lon Phan (lonphan@ncbi.nlm.nih.gov)
- SRA – Chris O’Sullivan (osulliva@ncbi.nlm.nih.gov)
- RAS – Kurt Rodarmer Sr. (rodarmer@ncbi.nlm.nih.gov)

NCPI Working Group Participants

- Community Governance Working Group - ***Valerie Schneider***
- Coordination Working Group - ***Kurt McDaniel***
- FHIR Working Group - ***Mike Feolo***
- Outreach and Training Working Group - ***Ravinder P. Eskandary***
- Systems Interoperation Working Group - ***Kurt Rodarmer Sr.***



Acknowledgements

dbGaP Team

Rinat Bagautdinov	Anne Sturcke
Carol Bastiani	Masato Kimura
Monika Bihan	Ashok Komaragiri
Dale Conklin	Moira Lee
Daniil Deriy	Natalia Popova
Svetlana Dracheva	Andrew Russette
Ray Dunivin	Nataliya Sharopova
Adil Faisal	Stefan Stefanov
Mike Feolo	Jack Wang
George Godynkiy	Wendy Wu
Neha Gupta	Zhuoxi (Joe) Wu
Luning Hao	Jewen Xiao
Yumi Jin	Ming Xu
Kuljeet Kaur	Lora Ziyabari

SRA Team

Zinaida Belaia	Kurt McDaniel
Colleen Bollin	Christopher O'Sullivan
Anatoliy Boshkin	Sergey Ponomarev
Kenneth Durbrow	Wolfgang Raetz
Alexandre Efremov	Kurt Rodarmer Sr
Lydia Fleischmann	Robert Sanders
Svetlana Iazvovskaia	Oleg Shutov
Alexey Iskhakov	Yuriy Skripchenko
Kenneth Katz	Adam Stine
Michael Kimelman	Jonathan Trow
Andrew Klymenko	Mike Vartanian
Andrey Kochergin	Eugene Yaschenko
Richard Lapoint	Vadim Zalunin

Breakout Groups Report Back

Data harmonization and interoperability,
including models, terminologies, mapping, provenance

Chris Chute (JHU) & Tricia
Francis (JHU)

Data harmonization and interoperability Breakout

38 participants

9 slide authors: Chris Chute, Sam Volchenboum, Melissa Cook, Allison Heath, Asiyah Lin, Subhashini Jagu, Brian Walsh, Tricia Francis, Deanne Taylor

Large-scale topics

- System interoperability and data harmonization are synergistic
 - The better data harmonization, the easier system interop
- Both are needed for multiple use cases
 - Search, query, analyses
- Much discussion on harmonization topic
 - Clinical world contrasted with basic science and omics world
 - Different starting places
 - Include genomics as well as clinical data in discussions about data harmonization
- NCPI : Hub and spoke model
 - May showcase how federated data from specific programs may interoperate with each other
 - The data harmonization happen more at spoke (platform level) than hub level.
 - Need the programmatic level intervention for full scale effort, but it is out of NCPI scope

Levels of Interoperability

- Semantic
 - Data context
 - Examples - Mondo, HPO, Snomed, ICD-O, NCI
- Syntactic
 - Data language
 - OMOP, BRIDG, FHIR, LinkML
- System
 - Data presentation
 - RDF, PFB, FASTA, VCF
- Structural
 - Data architecture
 - APIs, Docker
- Administrative
 - Authentication, authorization, access mechanisms

Clinical Harmonization

- Historically driven by CMS and ONC for administrative purposes
- Resulted in coherent US Core for Data Interoperability standards
- Spawned the emergence of FHIR, following earlier HL7 specification
 - Support modeling language and terminology binding
 - Development of the NCPI Implementation Guide as an example

Resulted in opportunities for clear harmonization “target” models and semantics

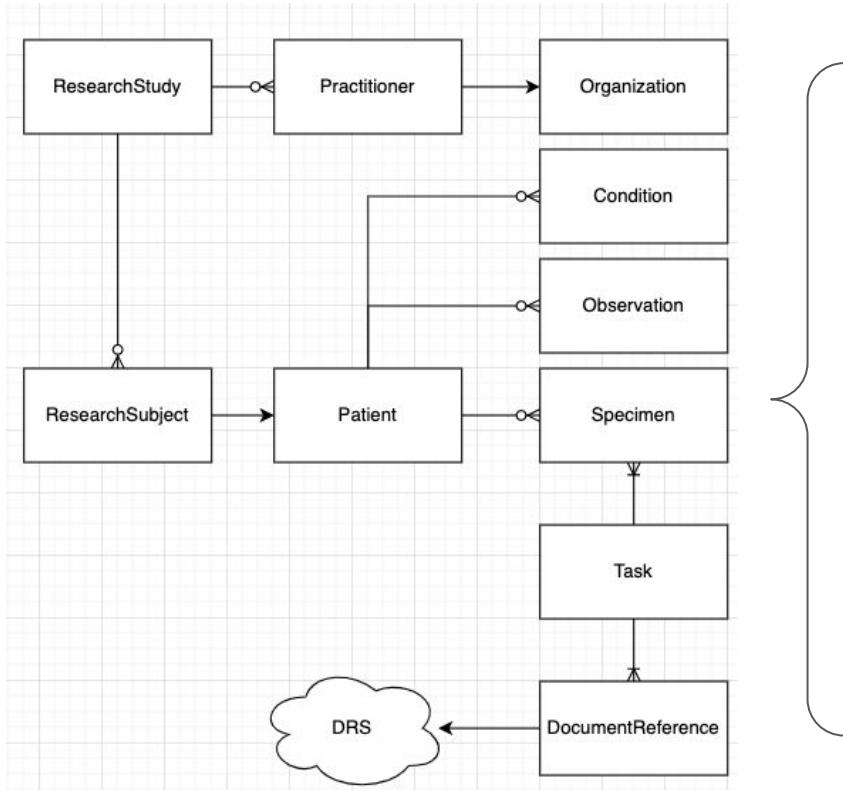
Still presents challenges for binding semantics for research:

OMOP, NCIIt, UMLS,

Basic Science/OMICs data

- More volatile than clinical data
- Still same conversation, though larger spectrum of domains
 - Genomics, proteomics, pathway, etc
- In some domain (genomics) emerging proliferation of vendors and systems
 - Divergent while overlapping data structures and annotations
- Some OMICs and experimental metadata standards have been developed in the past ([MIAME \(2001\)](#), [MINISEQE \(2012\)](#) -- [NCBI GEO](#) used them in submission practices.

Consensus Ontologies



Syntactic Consensus: Key FHIR Resources

- Semantic Consensus needed:
 - How to identify “consensus” Ontologies
 - How to incentivize adoption
 - Evangelize mapping toolkits

Layers of Data Harmonization and Provisioning

- Object: System / platform /application exchange (e.g. FHIR resources)
- Relational: Analytic capabilities (e.g. OMOP)
- Spreadsheet: Data matrices for analysis (where most researchers work)
- “Language of the data”

Who determines/decides on Best Practices

- How to get the right stakeholders in the discussion?
 - Convene communities?
- How to incentivise? NIH concerned about compliance at program level without being too prescriptive.
 - NCPI can be an example and a forum showing how to harmonize across multiple, large programs

Search Breakout Report

Breakout Session Report Back

Search

Kathy Reinold

Broad



Steven Cox

RENCI



Jay Ronquillo

NCI





Discussion Overview

- 34+ participants
- Representation
 - NHGRI, NHLBI, NCI, CF Kid's First, NCBI, ODSS, academia, FNL, RENCI, ISB, SB, Broad, and others
- Questions
 - Who is searching? What are they searching for?
- Topics
 - Discovery vs cohort building vs. results-based
 - Search facets: variants, subject characteristics, clinical variables, study-level, dataset-level, data-level, by modality
 - Hypothesis generating vs. validation
 - Harmonization



Summary

Types of Search:

1. Cohort building
2. Data set discovery
3. Delivering data to analysis workspaces
4. Find specific cases/samples
5. Dataset metadata (availability, access, etc)

Two broad dimensions:

1. Hypothesis generation: visual interfaces preferred.
2. Hypothesis validation: programmatic interfaces preferred

Multiple types of search are required



Next Steps

- Survey
 - Types of search -- what is the highest need?
 - Favorite search features
 - Facets - which facets do you use? What additional facets do you need?
 - What are the most useful aspects of your favorite search tools?
 - What % of use do you see for GUI vs. API vs. SQL?
- Consider agreeing a common data dictionary format
 - Expert sourcing of format through NCPI community
- Decide on specific use cases
- Document search requirements for NCPI
- Consider an initiative to define the core terms we can agree on
 - sex, ethnicity, race, biosample types, ...
- **Is there a Working Group to follow up on these?**



Takeaways from Post-Breakout Report Back Discussion



- Search very timely because of increased interoperability
- Strong desire for practical demonstration of use cases
 - More than simply integrating datasets, can users search across these datasets?
 - Concrete use cases in next 6 months to demonstrate ability search/extract data across platforms
- Impact of data access (open vs. controlled) on ability to search
 - Before applying for access/authorization, can user find out how many samples are in dataset or which studies are applicable?
 - How to engage investigators while getting/waiting for data approval?
- Data harmonization and identifier creation vital for search as well

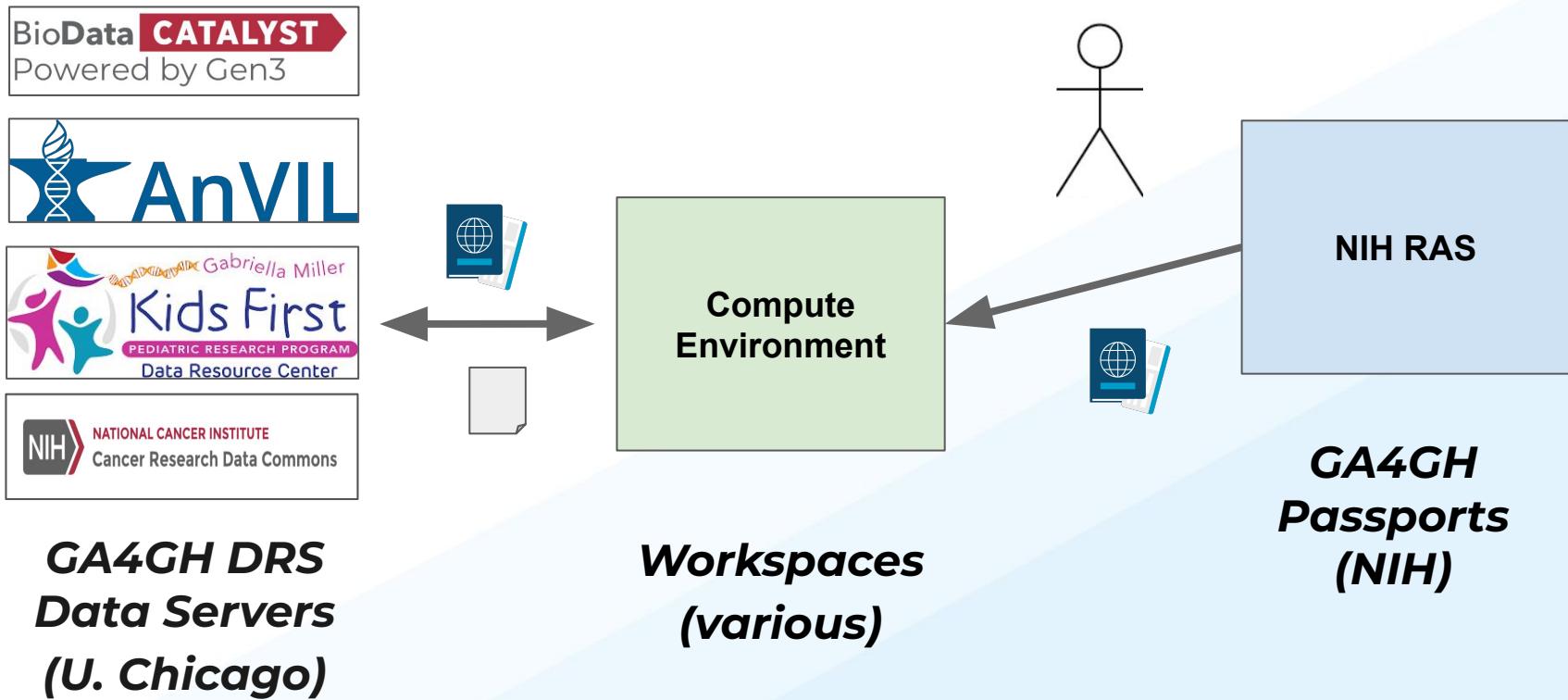
Breakout Group: RAS

Andre Paredes
U. Chicago

Brian O'Connor
Broad Institute



RAS Breakout: High Level Background



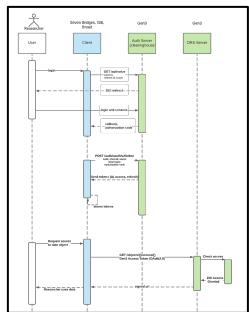


Background

- **RAS design work** across a variety of teams and projects to date:
 - See [RAS Integration Guide 1.4](#) & [Milestone 3 Technical Guide](#)
 - Latest document: [Summary of two preferred approaches](#)
- Groups loosely coordinated a 3 milestone plan:
 - **Milestone 1** : Login with RAS.
 - **Milestone 2** : Gen3 uses RAS Visas as the authorization information instead of dbGaP telemetry files.
 - **Milestone 3** : RAS Passport Visas can be used directly to access data resources, Central Fence is enabled by consistency across IC stacks.

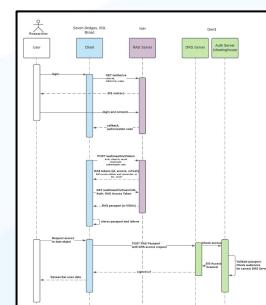
Summary of Milestone 3 Options

- We've worked with CRDC, [AnVIL](#) and [BDCat](#) to converge on a common approach for Milestone 3
- We've tried to help by putting together a [summary of two preferred approaches](#) and collaboratively address concerns... *goal is to add ability to access data with passports rather than taking away previous approach*



1: Current Gen3 Approach

&



2: New Passport Approach



Technical Issues To Discuss

- 1) How does a Data Server ensure the RAS Passport with Visas is coming from a trusted client?**
 - a) Repackaging Passport → Client Passport with signature?
 - b) Mutual SSL certs approach?
 - c) *Does it matter if the client is trusted if RAS trusts it?*
- 2) How do we ensure data access with Passports is performant?**
 - a) POST of Passports?
 - b) Caching strategy?
 - c) Downscoping of Visas? Requires future releases of specifications.
- 3) Others?** Most/All addressed in Summary of two preferred approaches?



Breakout Schedule

1. Trust → 40 minutes
2. Performance → 20 5 minutes
3. ~~Policy or Other Issues~~ → 15 minutes
4. Next steps (NCPI) → 5 minutes

Findings: Trust

How does a Data Server ensure the RAS Passport with Visas is coming from a trusted client?

- 1) Mutual SSL... **yes**, do this for at least BDCat
 - AI: which other systems require this?
- 2) Repackaged, signed Passports are not sufficient to identify a **client**
 - But systems may implement full Passport Brokers that repackage and add new visas in addition to RAS visas... that's OK and satisfies some use cases (like consortium data access)



Findings: Performance

Possible performance issues, caching strategy, and verification of passports

- 1) we need to support POST of Passports+Visas given their size
 - AI: DRS spec needs to be updated, PR available
 - AI: DRS implementations need to be updated
- 2) downscoping is of interest and being actively worked on but is not the solution to passport size restrictions per se
 - AI: GA4GH continue to work on downscoping approach
 - AI: systems ultimately to implement...



Findings: Policy

Policy issues

- Can repackage a passport if your system is a **full GA4GH Passport Broker**
- AI: Need clarification from projects if they require SSL client/server verification



Next Steps Timeline

1. Address any additional concerns from the Passport proposal → ***finalize as whitepaper***
 - a. Consensus on Trust approach → **which systems require mutual SSL** → Q2
 - i. Policy & Governance group?
 - b. Consensus on the proposed DRS POST update to support Passports → Q2
 - i. GA4GH & Gen3 DRS implementations
 - c. Consensus on the proposed downscoping support for Passports + DRS → Q2?
 - i. GA4GH, Client Systems, RAS & Gen3 DRS implementations
 - d. **Use NCPI Sys Interop working group to reach consensus across platforms?** → ***Yes***
2. Adoption of Milestone 3 by DRS servers, RAS (if any changes are needed) and various analysis workspace clients (as well as Signed URL support) → Q3-Q4
 - a. Anything blocking this? Any remaining issues?

NCPI Spring 2021 Workshop Day 1 Wrap Up

- Speakers please send us your presentations from today
- If you have not registered, please do:

tinyurl.com/NCPIregistration

- Please use the **WebEx application** and not a browser
- Fall 2021 Workshop poll: **tinyurl.com/NCPIfallpoll**