2020 GA4GH Connect Virtual Meeting

Federated Analysis Systems Project (FASP) Breakout Agenda

Details subject to change.

Main GA4GH Meeting Agenda:
https://broadinstitute.swoogo.com/2020-ga4gh-connect/agenda?uid=5e7aa4f7ce219

Links to other breakout sessions (zoom links and agendas):
https://docs.google.com/spreadsheets/d/1HD_9wJL1eBq89GAh5whKWrJFIpOXb8crORFWGqbhIQw/edit?ts=5e795c57#gid=0

Meeting Goals: Ongoing Project Work

Relevant Work Streams: Discovery, Cloud, DURI, Security

Chairs: Craig Voisin, Brian O'Connor, Max Barkley

Notetaker: TBD, see notes after the agenda, please feel free to add notes there as we go

Report Back Slide

Wednesday, March 25, 2020

Start Time	Discussion Topic	Related Materials	Speakers
01:00 PM Pacific 04:00 PM Eastern 08:00 PM UTC	Introduction to FASP	Intro Slides	Craig
01:10 PM Pacific 04:10 PM Eastern 08:10 PM UTC	The state of the "Golden Demo" 1. Show demo (20 min.) 2. Discussion of demo and underlying technologies (20 min.)	Slides	Max
01:50 PM Pacific 04:50 PM Eastern 08:50 PM UTC	Implementing GA4GH APIs in Systems Michele @ Seven Bridges: GA4GH DRS Interoperability Demo - Cavatica & Cancer Genomics Cloud 10 minutes Kurt @ NIH (NCBI): GA4GH API implementation at NCBI (tentative) 15 minutes Ian @ NIH (NCI): build this together with Michele & Kurt, what was learned and issues exposed on the journey towards GA4GH API implementation? Group discussion.	Slides	Michele Mattioni Kurt Rodarmer Ian Fore

	30 minutes		
02:40 PM Pacific 05:40 PM Eastern 09:40 PM UTC	FASP in 2020 What are our key goals for the next 6 months leading to the Oct Plenary? - iterate on the "golden demo" - n>1, more groups/implementations participate in the federated analysis systems project "golden demo" - formal feedback to Cloud, Discovery, and DURI APIs - are we missing any key goals?	Slides	Brian, Max, Craig
02:50 PM Pacific 05:50 PM Eastern 09:50 PM UTC	Wrap up - future FASP calls - possible deep tech dive/hackathons to make progress on specific topics (similar to the virtual coffee earlier today on DRS and GUIDs)	Slides	Brian, Max, Craig

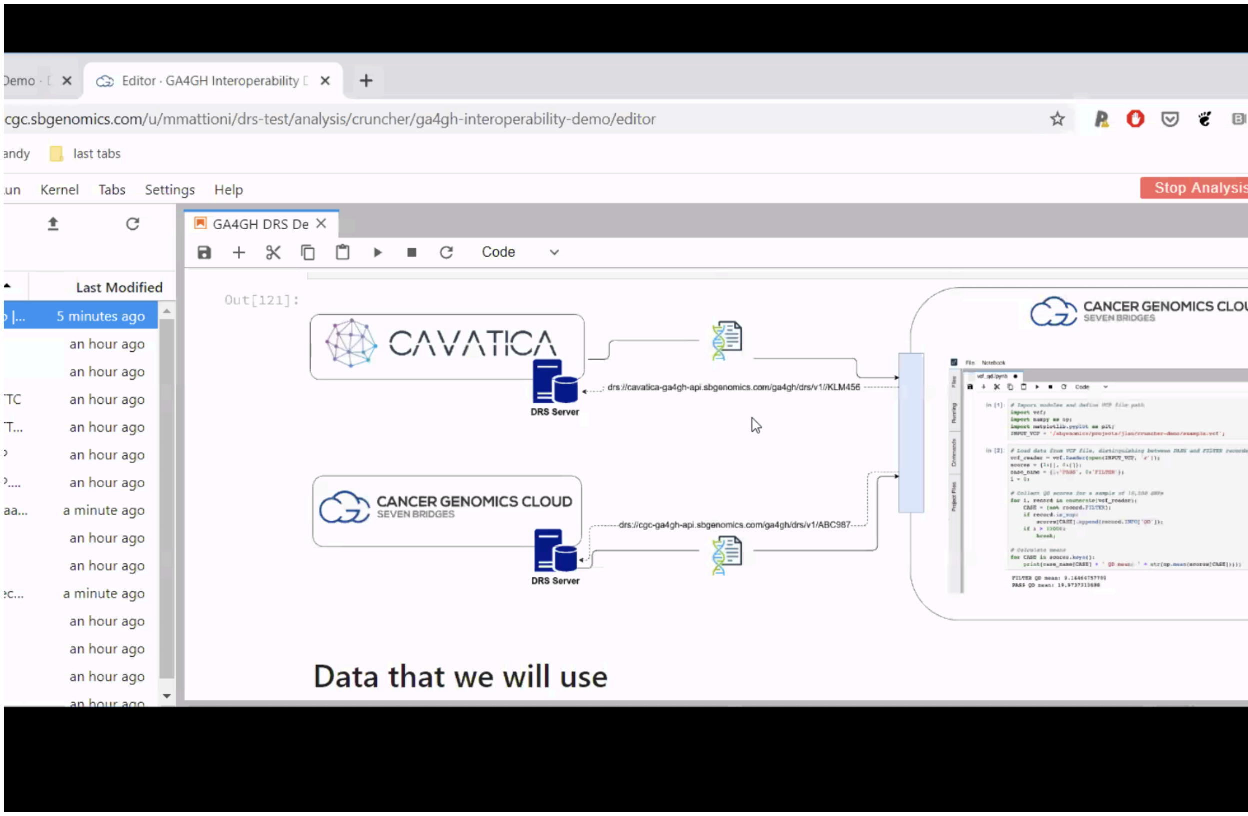
Notes:

- Intro by Craig
    - AI: Please provide feedback on the charter, contributors table, and API feedback
    - See the slide for links
- Update on the Golden Demo by Max
    - Enhancements from previous demo at Plenary 2019
        - DRS
        - UI improvements
        - Federated auth with trust boundaries
    - More federated auth solution
        - more logins with authorization servers
    - Outline
        - search API to select cohort
        - search result has DRS URIs for blob inputs
        - UI form for WES
        - auth between 2 orgs: DNAstack and MSSNG
            - DNAstack has WES
            - MSSNG has dataset
    - DavidG: don't have a line going from WES to DRS, is that right?
        - Correct, the way structured is a web UI that does heavy lifting
        - Discovery and analysis UI calls search API, and then resolves DRS URIs before calling WES endpoint
        - Good thing to discuss
    - Demo video
        - collections available: MSSNG + WES are the two
        - MSSNG contains data, DNAstack for WES
        - Search:
            - auth with consent

    - two endpoints:
      - tables I can see in BigQuery for MSSNG project, but could be anything
    - query gets results table with URIs of DRS servers (gap: not DRS URIs, but only because out of time to do so, no blocker here)
    - select rows as cohort
    - DRS, right now HTTPS URL but could easily be DRS URIs.
  - Run:
    - pick workflow
    - map in columns to workflow variables
    - pick WES instance
    - Auth via 2 flows:
      - WES server
      - dataset
    - Execute workflow
    - Monitor via WES API in GUI
    - WES sees gs URIs not DRS right now
    - future: allow selection of where to run
    - [couple hours later]
    - Output gives task log links
- Next steps:
  - others to integrate?
    - proper authorization
    - one of the biggest blockers
    - don't cut corners on auth
  - one area with bespoke API
    - obtaining tokens
  - Discovery network
  - Improve DRS support
  - Multiple clouds
- Questions for Max
  - Kurt: with all the user interaction, how does someone work at scale? Human intervention scale?
    - some built in already, more needs to be done
    - could have selected all the rows at once
    - if search results contained different org boundaries, I would need one login flow per organization
  - Kurt: what happens when you resolve your IDs -> URLs… and there's a timeout (signed URL or access token)
    - how should WES and DRS work together and hand-off? that helps know when
    - because cromwell as implementation detail, we transfer files to a staging area
    - not ideal, but iterating and have something working
    - don't need to reauth days later
    - alternative is tighter integration between WES and DRS for refresh
  - Kurt: managing egress charges, are egress charges involved in these scenarios?
    - How do we help the researcher to avoid unnecessary egress charges
  - Kurt and Alex: DRS URIs handed to WES
    - Security issues when you provide WES with DRS refresh tokens that are very powerful
    - Max says checkout API key service could be very useful here to get access tokens as needed.
  - David B: how does the cart token service work?

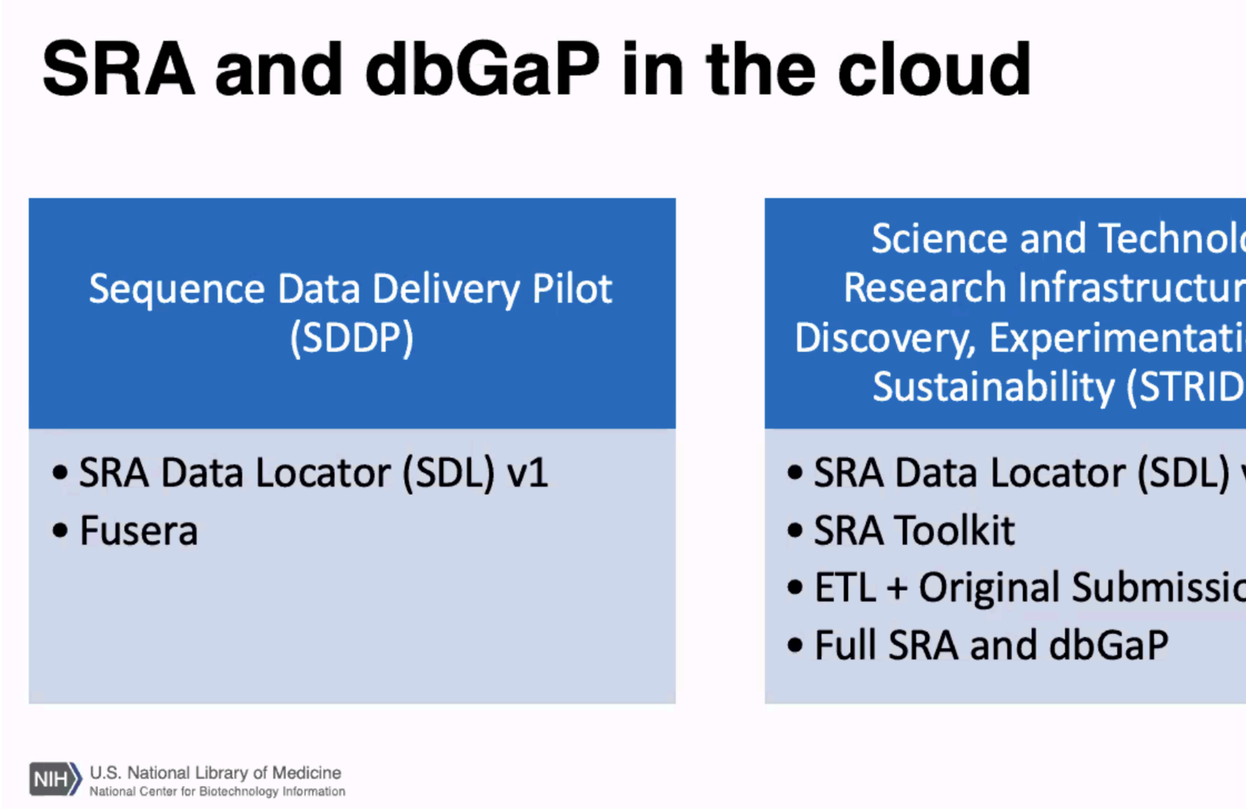
      - Max… two parts… 1) oauth2 flow for a user, authorization flow… authorize for a collection of resources… JWT OAuth2 standard with parameter telling what resources you want. After the dance, you get the bearer token that you can use to get access tokens.
      - OAuth scopes? Max says key is flow where you can specify resources and get access to token dispenser for that particular service.
      - David: are the carts federated? Max could see aggregating carts across multiple orgs.
    - Alex: Is this just passports and visas? Max says it's like OAuth2 vs OIDC. Passport give identity, cart service would build on this to get access tokens for platform resources.
    - Brian: is this gap with checkout auth for cloud a blocker for more onboarding?
      - locally optimal to make this a standard
      - Max feels it has a lot of merit
      - like to see a volunteer to iterate on an API similar enough to iterate and standardize on that
    - Brian: compare and contrast what others are doing (e.g. Auth0) to see what can be reused
      - Agree
    - What were the most difficult API limitations?
    - Anything that had to be done outside of the GA4GH APIs?
      - A: One authorization endpoint was custom… AAI standards was covered by specs but the Access Manager and client trying to get tokens is a slight extension of OAuth
    - Credentials and gs URI for data access, is that the right approach?
      - A: Yes, this is how it's done now
      - Future would use DRS passed in
    - Search
      - Is the schema of how data is stored part of the discovery spec or custom to each system and the discovery API provides a way to search
    - Auth
      - How do you know which token to use for which resource? You showed getting tokens from different sources.
      - A: Max showed his flow for get authorization tokens for resources and tracking which API they go with.
    - How can others walk through the flow? It would be extremely useful for other developers to kick the tires on a working system, examine tokens, etc. Is that possible?
      - Point out where folks can find software they used (GitHub?)
      - A: all items are open source. He'll double check. Auth parts are open source
      - Experience of setting this up needs to be improved.
      -

- Experiences from
  - Michele

  - 
Demo
Editor · GA4GH Interoperability
cgc.sbgenomics.com/u/mmattioni/drs-test/analysis/cruncher/ga4gh-interoperability-demo/editor
last tabs
Kernel
Tabs
Settings
Help
Stop Analysis
GA4GH DRS De
Code
Last Modified
5 minutes ago
an hour ago
a minute ago
Out[121]:
CAVATICA
DRS Server
drs://cavatica-ga4gh-api.sbgenomics.com/ga4gh/drs/v1//KLM456
CANCER GENOMICS CLOUD
SEVEN BRIDGES
DRS Server
drs://cgc-ga4gh-api.sbgenomics.com/ga4gh/drs/v1/ABC987

Data that we will use

  - What they did in SBG to make this work
  - 2 DRS servers in production
  - show how to achieve 2 DRS files to do work on it
  - 2 datasets
  - Dev token from cavatica and sbg stacks, loaded via a JSON
  - Want to see:
    - Waiting for **WES implementation that supports DRS**
    - Complex workflows… need a way for DRS to point to a set of **metadata** that is attached to the file.
    - Need the metadata link passed back via DRS (metadata doesn't need to be stored in DRS)
  - Passports and auth, and pass that in as the auth server could be a good idea
  - Questions for Michele
    - How do you see metadata on DRS working?
      - Clin/pheno group URL to that.
      - Don't embed within DRS response
    - Ian brought up changes in metadata over time
- Kurt
  - NCI created a pilot DRS
  - https://github.com/ncbi/ncbi-drs for their code

- 

SRA and dbGaP in the cloud

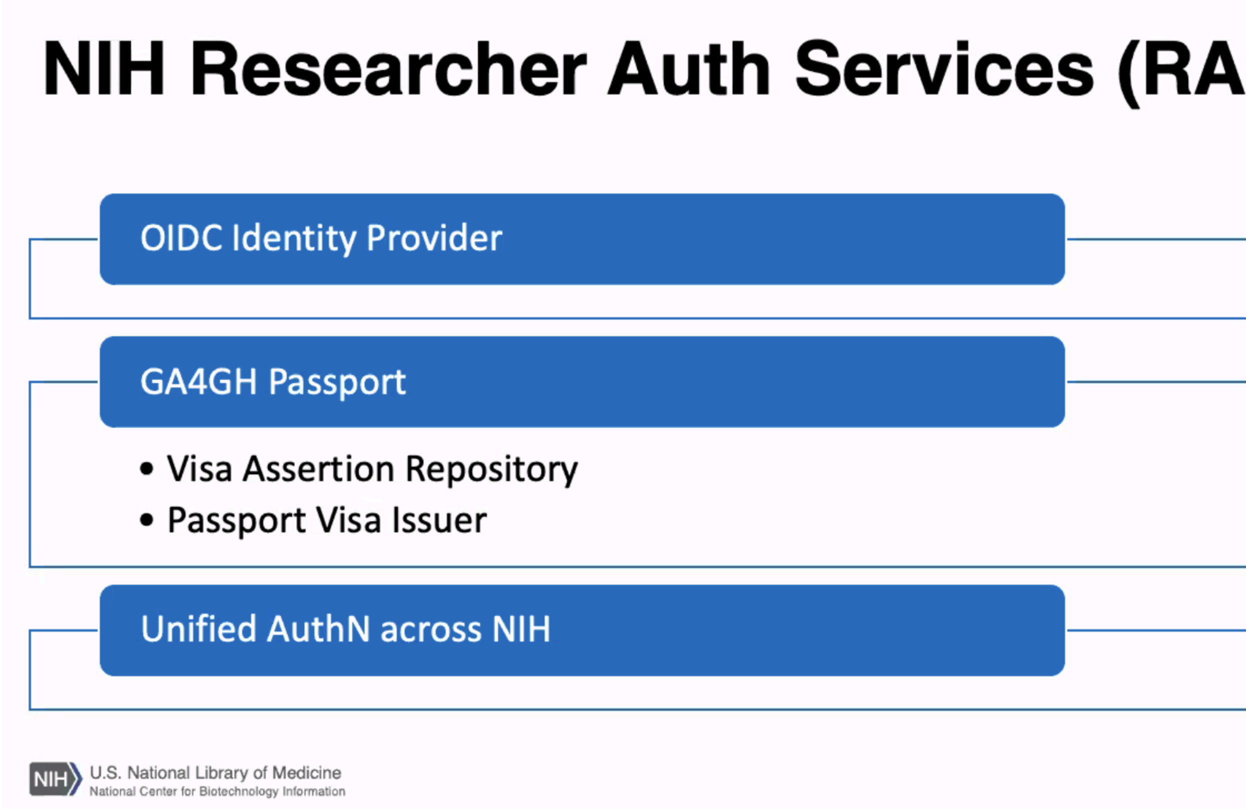
Sequence Data Delivery Pilot (SDDP)

- SRA Data Locator (SDL) v1
- Fusera

Science and Technolo
Research Infrastructur
Discovery, Experimentati
Sustainability (STRID

- SRA Data Locator (SDL) v
- SRA Toolkit
- ETL + Original Submissio
- Full SRA and dbGaP

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

- 

NIH Researcher Auth Services (RA

OIDC Identity Provider

GA4GH Passport

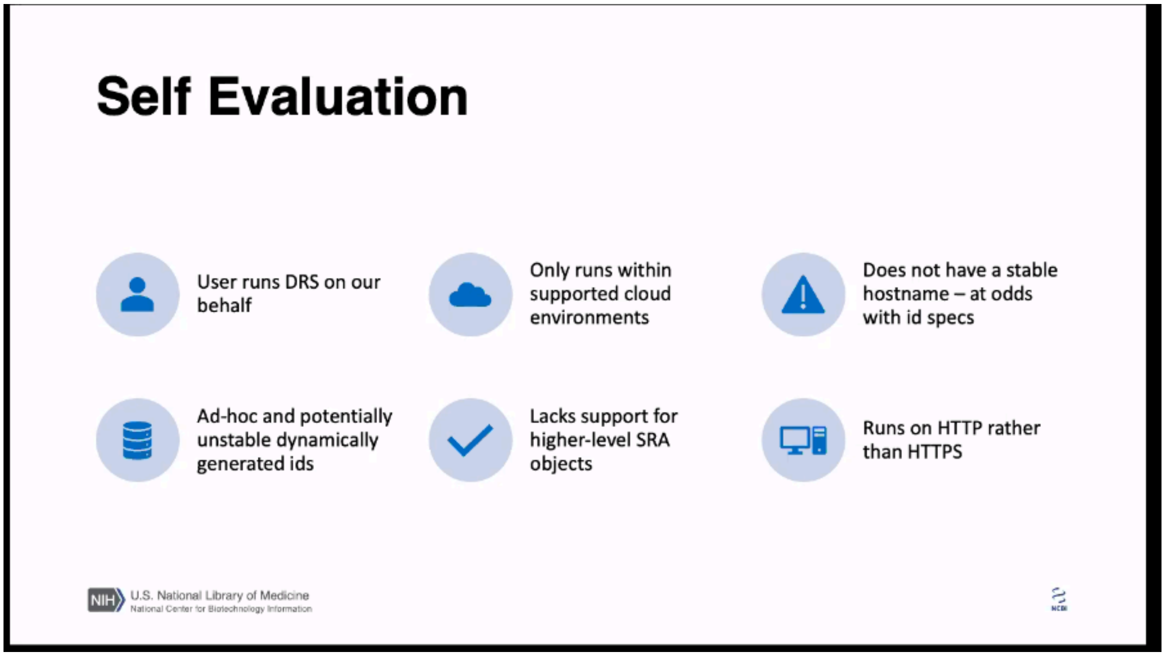
- Visa Assertion Repository
- Passport Visa Issuer

Unified AuthN across NIH

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

- STRIDES has:
  - SRA Data Locator (SDL) v2
  - SRA toolkit
  - ETL + original submissions
  - Full SRA and dbGaP

- NIH Researcher Auth Services (RAS)
    - IdP
    - Passports (includes dbGaP visas)
    - Unified AuthN across NIH
- DRS Pilot
    - access to original submission files in cloud
    - Worked with partners at CHOP
    - rapidly developed in Python
    - Sources on GitHub
- Starting conditions
    - only available on cloud now
    - egress charges
    - no support yet for DRS ids
    - require user-pays or signed URLs
- issue 1: avoid egress charges
    - run DRS service in cloud under user's account
- issue 2: problems with solution 1
    - no stable host name
    - runs over HTTPS with signed certificate (right now running without HTTPS)
- Issue 3: ID namespace quite different than DRS IDs
    - Bundles, for example, can't be added to over time so that's a limitation
- Issue 4: SRA run accessions
    - Treat SRA run accessions as snapshot bundles
- Issue 5: selecting desired blobs
    - DRS requires some other mechanism for generating and listing IDs
    - solution: filter out the ETLs and prioritize what format to send out
- Issue 6: token is GA4GH passport… RAS not generating this yet.
- Issue 7: compute environment
    - access by signed URL
    - evidence that client won't generate charges
    - Access token bound to environment (cool!)
- Issue 8: access token
    - is a URI with an embedded JWT which is bound to the compute environment that is being used
    - solution: return URI proxy endpoint on same VM to handle access token



Self Evaluation

User runs DRS on our behalf

Only runs within supported cloud environments

Does not have a stable hostname – at odds with id specs

Ad-hoc and potentially unstable dynamically generated ids

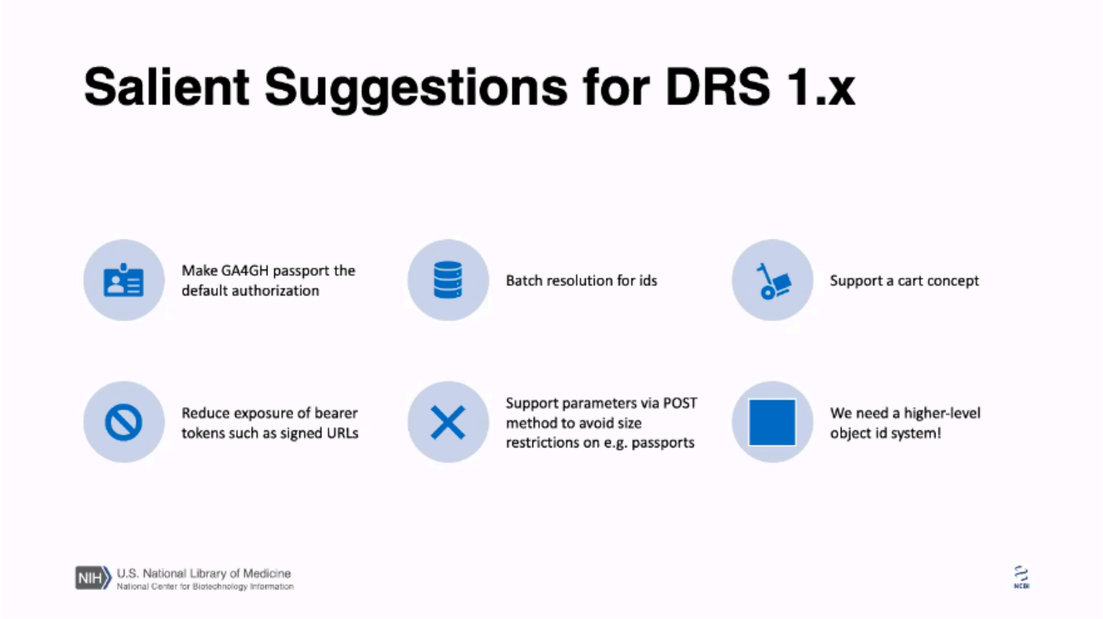
Lacks support for higher-level SRA objects

Runs on HTTP rather than HTTPS

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

NCBI

- Hurdles and TBDs for Real DRS
  - Must have some means of preventing egress charges -- need assistance from cloud providers
  - would like to reduce exposure of bearer tokens
  - expand SRA system to support real DRS IDs
  - Let dust settle on DRS and GUIDs
  - Full integration with RAS
- Suggestions for DRS 1.x:



Salient Suggestions for DRS 1.x

Make GA4GH passport the default authorization

Batch resolution for ids

Support a cart concept

Reduce exposure of bearer tokens such as signed URLs

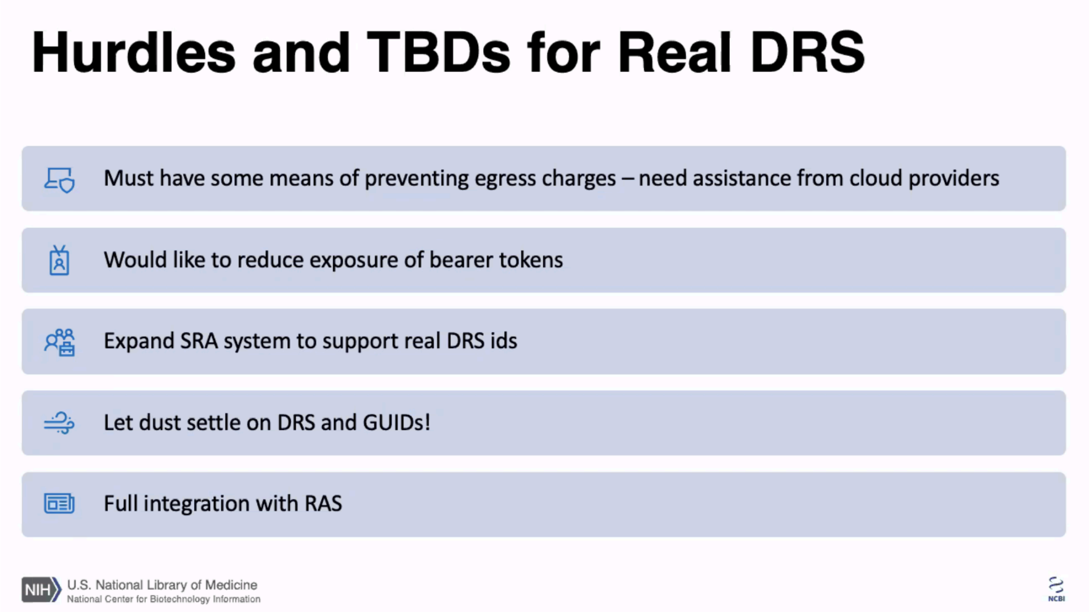
Support parameters via POST method to avoid size restrictions on e.g. passports

We need a higher-level object id system!

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

NCBI

- Questions for Kurt
  - Of these issues, which are you most concerned about when it comes to the workarounds you used?

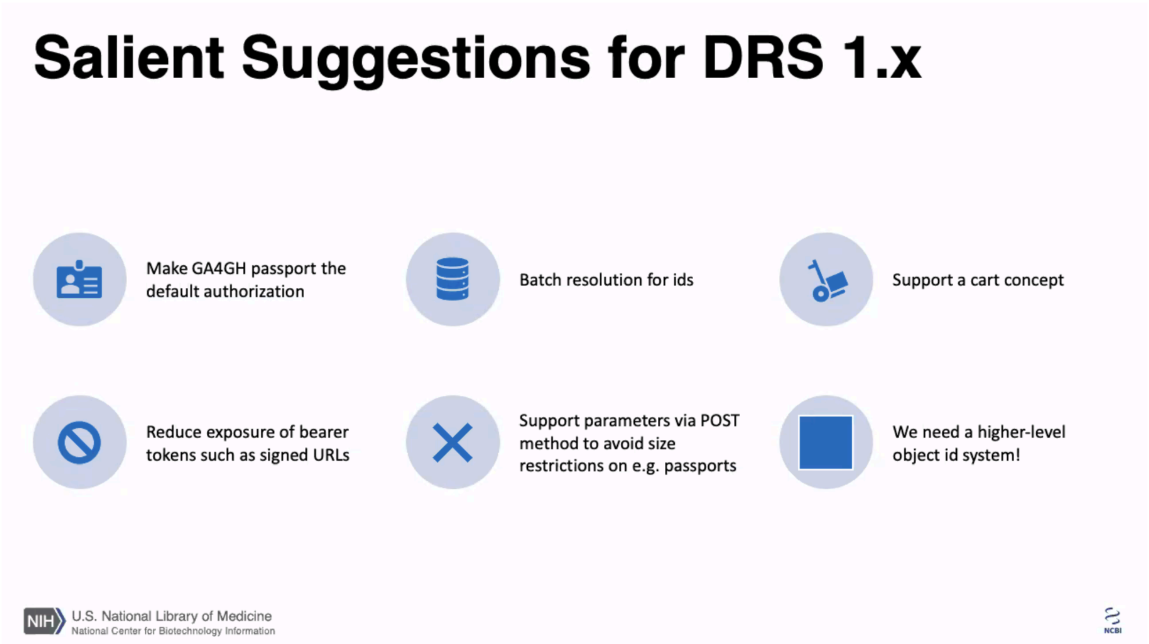
    - 

      ## Hurdles and TBDs for Real DRS

      - Must have some means of preventing egress charges – need assistance from cloud providers
      - Would like to reduce exposure of bearer tokens
      - Expand SRA system to support real DRS ids
      - Let dust settle on DRS and GUIDs!
      - Full integration with RAS

      NIH U.S. National Library of Medicine
      National Center for Biotechnology Information

      NCBI

  - 

    ## Salient Suggestions for DRS 1.x

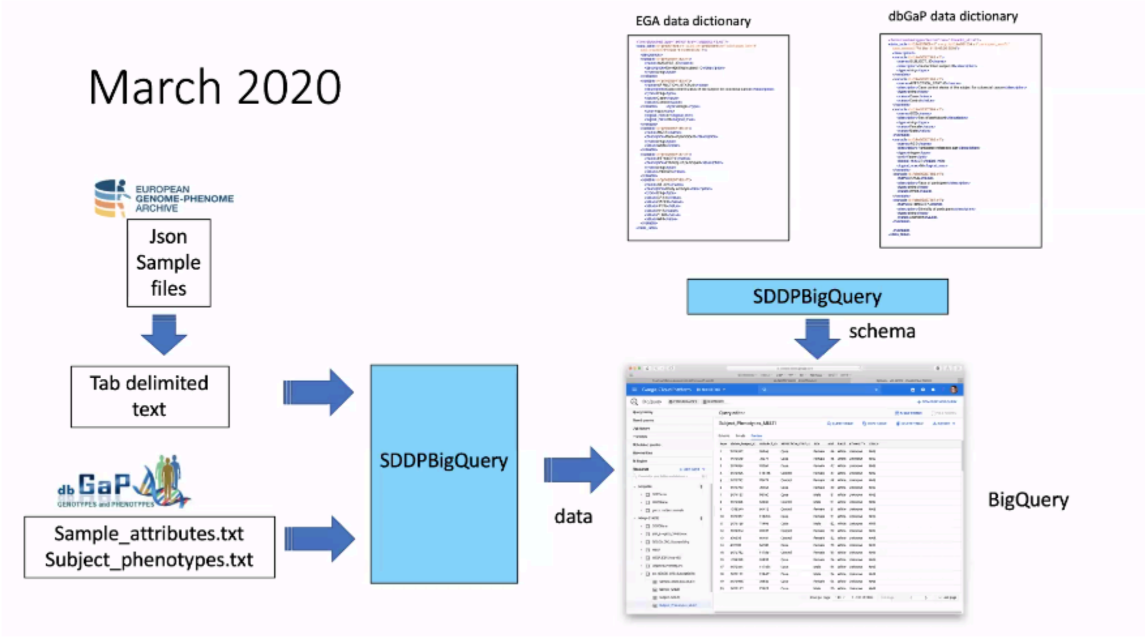
    - Make GA4GH passport the default authorization
    - Batch resolution for ids
    - Support a cart concept
    - Reduce exposure of bearer tokens such as signed URLs
    - Support parameters via POST method to avoid size restrictions on e.g. passports
    - We need a higher-level object id system!

    NIH U.S. National Library of Medicine
    National Center for Biotechnology Information

    NCBI

    - POST suggested since passports can be very large
    - Cart concept… came up in Max's
    - See above slide for areas to improve
- Ian
  - Implementers but also shopping for how to get cancer research done with this toolkit
  - What does NCI and CRDC bring?
    - Human need
    - Data
    - GA4GH Implementations
  - Federation
    - Across technology
      - Platforms
      - Algorithms
    - Across organisations
      - Funder
      - Data provider

    - Across national boundaries
    - Across scientific disciplines
- NCI is a good example of global federation
    - Ecosystem with NIH
    - NCI funded data in EGA
    - US research in EGA
    - combining data from EGA and NCI makes sense
- This example federates:
    - technology
    - data provider
    - national boundaries
- Applies to non-cancer use cases as well
- Hand-off to compute for federated analysis
- Have the ability to link discovery
- Can also do it via Kurt's SRA Locator
- Moving from DOS to DRS
- GECCO and federation
    - consortium
    - federates funder and national boundaries within a single dataset via this consortium



March 2020

EGA data dictionary

dbGaP data dictionary

EUROPEAN GENOME-PHENOME ARCHIVE

Json Sample files

Tab delimited text

dbGaP GENOTYPES and PHENOTYPES

Sample_attributes.txt
Subject_phenotypes.txt

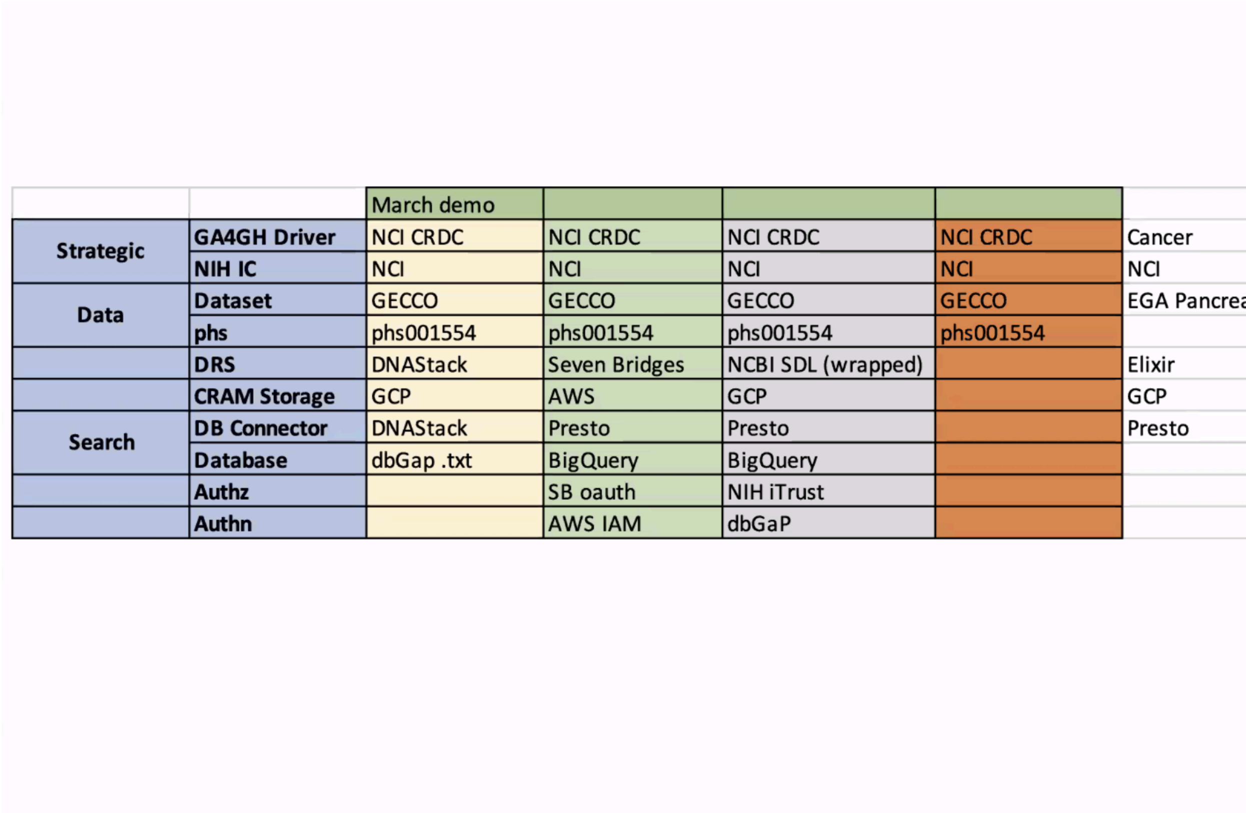
SDDPBigQuery

data

SDDPBigQuery

schema

BigQuery



		March demo				
Strategic	GA4GH Driver	NCI CRDC	NCI CRDC	NCI CRDC	NCI CRDC	Cancer
	NIH IC	NCI	NCI	NCI	NCI	NCI
Data	Dataset	GECCO	GECCO	GECCO	GECCO	EGA Pancrea
	phs	phs001554	phs001554	phs001554	phs001554	
	DRS	DNAStack	Seven Bridges	NCBI SDL (wrapped)		Elixir
	CRAM Storage	GCP	AWS	GCP		GCP
Search	DB Connector	DNAStack	Presto	Presto		Presto
	Database	dbGap .txt	BigQuery	BigQuery		
	Authz		SB oauth	NIH iTrust		
	Authn		AWS IAM	dbGaP		

        - Key points:
            - base work in actual data (EGA and dbGaP for example)
            - INSDC
        - strategize on configurations of components
    - Alex
        - DRS and Passport
            - What visa was used to give you the access for something
            - Lets you track how to protect that data later.
            - If you could tell from a DRS URI which visa was associated, would help in tracking for access to derived results
        - David G. auditing and provenance ...
        - Alex... useful for derived results
- FASP in 2020
    - What are our key goals for the next 6 months heading to Plenary? Overall:
        - Add other DRS endpoints to golden demo, maybe a full stack if possible
        - Show passports use e.g. from RAS
    - Do we have the right specific goals?
        - Value in setting up the golden demo as devs wanting to learn
            - Helm for deploy on Kubernetes
        - Token cart -- Max
        - Feedback on DRS going to cloud -- Kurt
        - Adding in NCI as another stack in the golden demo?... use RAS passports? See the table above -- Ian
    - Who will make the second, third, fourth, etc stack? Freshen the contributors table
    - Feedback on APIs -- API feedback doc
        - DRS
            - See excellent feedback from Kurt
        - WES

        - Passports
            - Max and the checkout API for getting third party access tokens… is this something missing that should be added to our API standards?
                - See what other stacks think… do we need this? Is the approach for the first golden demo work for the second group?
        - Discovery

FASP Next Steps Virtual Coffee

- …
- Max: Want to add one person
    - IF: Can we invite loads?
    - BO: Need to make clear that we attempting to refine APIs
- Max: New collaborators - needs to make clear that the missing GA4GH API clear for 'checkout' endpoint is something they would need to work on.
- BO: Some foundational work needs to happen before we expand
- BO: NIH Interop has a very specific agenda.
    - Maybe goal is to get auth piece working with RAS
    - Data access stacks with auth
    - Getting a couple of the above is good
- IF: There is a difference in European uptake - very NIH-centric. Must have more international engagement on this. Europe, Aus candidates are needed
- BO: DRS and Auth (with Passports) can be our integration end goals.
    - IF: Authz and Authn has to happen, but other things need to happen in parallel
        - What are EGA Data Dictionaries. Search spec is coming up - data model and how these are conveyed is something to move on with. Maybe this is more Search than FASP. Can we drive DRS Spec to use these.
- DG: Don't want to overreach with FASP. I like the scope - takes quarters to make progress on. This is a good core track. If we can do parallel, good
    - To improve - 1. **Cart needs definition** (either “on stds track” or “we know how to solve w/o a new std”
        - Max: Data Security Gap analysis - they did add an item to follow up on this, big progress there.
    - 2. **Have a path demonstrated with 1 new person to show 'here is how to add a new player'** (e.g. a new dataset, wrapped with Search, DRS and Passports)
    - 3. Getting more comfortable on **DRS/SEarch/WES/Discovery processes - where do the arrows go between them, where does DRS URI resolution happen.** Need to formalize the motivations/reasoning behind these.
- DG: Next step in search can happen in parallel
- IF: In my last slide had some thoughts laid out. Who do we want to be compelling to?
    - DG: My 3 points would make us compelling to people like us (which is necessary but not sufficient). Some things you are talking about making it compelling to the scientific community, which is important, but don't want them to get in the way.
    - Max: I think this community just wants something laid out for them on Auth front. Makes it more compelling to the orgs to adopt these standards.
- IF: When we say another player - DNAStack and MSSNG project currently. Could add Seven Bridges in?
- MM: We are working with Broad, ICGC, UChicago, figuring out using the standards in these Driver Projects.

    - re: Cart - I don't know if good or bad, but we do need to review
    - **GA4GH Passport + DRS needs to be nailed down**
    - Things why I feel we are not compelling - complex, even we are figuring out how to build this
- MM: **Want to have more discussions in big crowd to not miss use cases**
- BO: Expedited route to get to the end - Google / DNAStack example as starting point.
    - Next step Ian with RAS and CRDC gen 3 instance to bring in feels like a good point to widen the interactions with users
    - MM: Yes, this makes sense. Need to document clearly.
    - IF: I think these are valid concerns from Michele, I have been advocating that Bob/UC team needs to interact with FASP to get improvements to DRS considered. Want to have NIH interactions
- Max: We already have Data Security action item on the call - maybe even a couple of meetings with Data Security + Bob Grossmans team on their architecutres might be a way ahead
    - IF: Want to make GA4GH Compliant architectures be accommodated as well
    - DG: Don't want to miss use cases.
    - Max: Can only take one shot a time, line up next shot. Feel first shot is collaborators who are implementing a real system. 80% cover by commonalities. Feel 20% can be brought in with the first shot made
- IF: **We need to plan for Security at Search level** This should be built in from the security angle
- IF: Maybe smaller tent being carried faster - can we see if the DNAStack thing can be run by other people?
- MM: CWL related - we made a point not to be a reference implementation - wanted to ensure there were multiple of these, GA4GH concentrates on the specs.
    - IF: Could Seven Bridges are the second player with a working WES implemetation using Bobs DRS implementation (providing Gecko data). I can then run somethung on your WES with the objects I pick
- MM: **WES needs to be upgraded for this**
    - **Max:** We **extended WES endpoint** slightly with multi-part form with a special **tokens.json file mapping access tokens and inputs. Staging area used** before workflow starts.
    - DG: Are the tokens passed to gs:// or drs://?
        - Max: pre-resolved for Cromwerll GCP backend
    - DG: So small alteration to resolve inside WES wrapper
    - Max: Biggest challenge or some method to exchange for Raw access method (checkout endpoint). Multiple permutatons. SOme can be used side-by-side.
    - MM: WES version needs updating. I feel we're not advancing the problem. My issue is if I send the WES with DRS to DNAStack, does DRS resolution get handled this correctly?
    - Max: I think we can get recommendations by giving Cart token to WES and then getting this be resolved. If we can get this right it becomes something that can be passed across
    - IF: Is Michele situation helped by prioritisation if we can leverage this work getting done? Can MM use Kurt, Bob's service is the issue
    - BO: Guid support is a place needing resolution. Bob service needs to support RAS.
    - IF: Alex@Broad wanted to resolve RAS
    - BO: Terra want to support GA4GH Passport from RAS.
- DG: I believe Bob can build DRS today, and whenever the GUID stuff settles, it will be a tiny add
- BO: We need to work with people willing to try things out, Maybe Kurt DRS / RAS Passport becomes the next thing to use, for wokring out the Cart interface?
    - KR: Gearing up to put everyone on RAS. Going all in
    - BO: That sounds like this is something you are ready to move on
    - Max: DRS server for MSSING - resloves DRS to access methods, checkout mechanism does the auth.
- KR: Visa allows for Pre-auth in the passport. dbGap works this way. In addition to making decision at the site
    - **RAS is planning to tightly incorporate GA4GH Passports** - dbGap pilot to add authorisation to it. If successful NIH will move forward with this.
- IF: Can we take this to Steering Committee level? Would be good for non-US focus, to get a couple
    - BO: Slightly wider net, but make sure people tolerant of this being in development

- What could next steps look like? -- BOC
    - Proposal for how passports, cart, and DRS work together - Max DNAstack
    - Ask another group to stand up a DRS instance that supports Passports (say from RAS) -- Bob Grossman?, Kurt? Michele?
    - Once we have a clear understanding of how passports work with DRS, we can invite more teams to implement since we can clearly tell them how to do this
    - WES support for DRS (topic 1: how to pass in drs: URIs; topic 2: how to pass auth through;)
    - Start adding full "stacks" that include search, workflow execution, and DRS

Action Items are indicated with "AI" above