GA4GH Virtual Connect 2022 Meeting Report

The 2022 GA4GH Virtual Connect meeting brought together more than 100 individual contributors from GA4GH Work Streams, Driver Projects, implementers, and more to collaborate and advance standards development work, with a focus on cross-Work Stream initiatives. This document includes an overview of all sessions, with links to recordings, slides, and meeting agendas and minutes.

The next GA4GH Connect meeting will take place in London in April of 2023. GA4GH is actively accepting proposals for working sessions at that meeting. Submit yours here.

Table of Contents

Organised alphabetically by session name

Bad actors in research environments: who's bad?
Beacon v2.0: "beaconise" your data type
Beacon v2.0: migration workshop
Best practices development
Cancer Community workshop
Cloud Work Stream Session
Crypt4GH: enhancing the use of encrypted data
DaMaSC: got schema?
Data access requests: where can we standardise?
Experimental metadata
FASP builder session
GA4GH Cloud APIs in 2025: where do we want to be?
GA4GH community updates
GKS maturity & VA knowledge model v1.0
Mapping REWS deliverables
Opening remarks
Paediatric pharmacogenomics secondary findings workshop
Passports/AAI v1.2 Q&A
QC of WGS study group
Resources for getting started with GA4GH APIs
REWS genetic discrimination
Schema registry service: exploring the uses of a standardised data model sharing API
Sequence Annotation integration with the GKS Work Stream
VCF 4.4 release candidate 2 public feedback

---

14 November

Opening remarks

Slides • Recording

Round-robin updates from the GA4GH Work Stream leads followed by a fireside chat between GA4GH CEO Peter Goodhand and long-time contributor Meg Doerr, who leads the applied ELSI research team at Sage Bionetworks.

Fireside Chat with Megan Doerr

Megan Doerr, an associate director at Sage Bionetworks, spoke with GA4GH CEO Peter Goodhand about new research in ethical, legal, and social implications (ELSI), the future of open science, and the crucial nature of community.

Key Takeaways

- Implementing ELSI scholarship requires building from universal principles into each of our unique contexts.
- In the next decade of open science, communities and collaborative solving will be very important.
- Thinking of a society on a continuum between individualism and communitarianism will allow exploration of new pathways in healthcare and data consent.
- New efforts focus on retaining the unique cultural environments from which data are coming, which can assist researchers in more precisely interpreting that data.

Cancer Community workshop

Agenda & Slides • Recording

After a few months hiatus, the Cancer Community is back to inform members of refinements to the group's structure and approach. Salvador Capella-Gutierrez from the Barcelona Supercomputing Centre will provide a presentation on the new EOSC4Cancer Project. Areas of synergy between the Cancer Community and the EOSC4Cancer Project will be identified and followed up on.

Key Takeaways

- Over the last few months, leadership has been meeting to rethink the Cancer Community meeting format and structure. Key decisions/points from these discussions have been laid out in this diagram. See actions below for next steps
- Salva presented on the EOSC4Cancer Project. The project, launched in September 2022, will aim to prepare EOSC (European Open Science Centre) services for cancer research

centres. The cancer community will be helpful as the project team works to determine which GA4GH standards they can use to address issues of interoperability.
- Future Cancer Community meetings will dive deeper into the EOSC4Cancer Project and areas where we can help. A particular group of interest is squad #3 which aims to connect omics data from multiple sources to a Clinical Decision Support System (CDSS) for precision treatment of metastatic CRC.

Mapping REWS deliverables

Agenda & Slides • Recording

The REWS Mapping Workshop aims to develop a 'REWS Lifecycle' Diagram that can be used for both internal and external communications. The development of this lifecycle will be supported by identifying common themes among subgroups and promoting discussion among work group participants. REWS has been successful in publishing an array of community-developed standards. We will map these through the genomics lifecycle to understand how these standards can work together, and whether there are common themes we can promote throughout. This will clearly demonstrate where REWS deliverables can be adopted, how these can work together, and help identify any gaps that REWS should seek to fill in future work.

Key Takeaways

- Through a fun interactive brainstorming, the REWS community generated several common themes across REWS deliverables and projects, as well as areas for future development. Themes included internationalisation, public opinion / knowledge, data governance, consent, equity, diversity and automation.
- Many REWS deliverables can be applied at multiple points in a data lifecycle - and they can't necessarily be "put in a box"

REWS genetic discrimination

Agenda & Slides • Recording

This workshop will review progress on the perspective article and findings from the Delphi study and further explore the topic of stigmatisation at the level of population groups.

Key Takeaways

- **The Delphi survey** on policy options for GD law is now closing-in on its 3rd and last round of recruitment. A small group of collaborators will be contacted for data analysis in due course. The larger collaborating group will contribute to reviewing the draft publication and guidelines.

- **GD commentary**. Attendees agreed on a ‘draft definition’. Co-authors will have one more opportunity to review the paper once ready to provide small edits before submission. Journal TDB (some suggestions were provided).
- **GD in population group**. There is interest within the group in pursuing this topic. It is broader than just discrimination: misuse of data from equity seeking population groups. Still topic can be pursued by GD workgroup. Yann to work on a first very preliminary draft instrument.

VCF 4.4 release candidate 2 public feedback

Agenda & Slides • Recording
The upcoming VCF version 4.4 overhauls how structural variants and copy number variants are handled and introduces support for STRs and VNTRs. This session will present an overview and justification for all changes. This public presentation of the draft VCFv4.4 specifications gives the public a final chance to provide feedback and request changes before the specification is finalised.

Key Takeaways

- Final chance for public feedback on VCF 4.4 before finalisation

15 November

QC of WGS study group

Agenda & Slides • Recording
Working meeting to continue progress on the development of standardised quality control metrics and reference implementations. Preparing for formal proposal to the GA4GH Steering Committee.
Key Takeaways

- When comparing metrics, we want to ensure metrics are tool agnostic and enable a functionally equivalent implementation. The current metrics draft needs to be updated to include implementation details
- The QC of WGS working group will be preparing to propose/formalise their work to the GA4GH Steering Committee in January. Work is underway to prepare the documentation required.

Best practices development

Agenda & Slides • Recording

This session will provide an opportunity for you to review best practices that have been gathered so far, brainstorm ways to apply them to your own contexts, and identify gaps where best practices are still needed. It will also feature a walkthrough of the interactive best practices platform on which the EDI Advisory Group has been working.

Key Takeaways

- The best practices interactive platform is now available!
- The table is a living document and will be refined over time with input from WS and subgroup leads.
- New best practices can be submitted through this form. These submissions will be reviewed by the EDI Advisory group and, once approved, added to the table for public viewing.
- All WS and subgroup members that would like assistance in implementing the best practices are encouraged to attend the EDI Advisory group meetings.

Data access requests: where can we standardise?

Agenda & Slides • Recording

What do data access committee members want to ask from data access applicants? This session hopes to pick up on several discussions and efforts across GA4GH, including the Ethical Provenance Toolkit work, DACReS, and the Computable Cohorts DARathon. The purpose of this session is not to solve any or all of the challenges in this space but to agree on a set of tasks that the group believes could be achievable in a medium timeframe.

Key Takeaways

- Guidance for Data Access Request forms would be useful
- Particulars of universal rules/requirements, such as how to make a PI accountable, would have the biggest impact before any technical solutions
- Considerable overlap between DAR forms, so identify a core set of “universal” rules
- There is a link between MRCG, DUO and DAR templates
    - What are the human readable questions that map to DUO and/or MRCG?
- This project needs to connect DURI and REWS efforts
- Is there a parallel with the Universal MTA effort that is in widespread use?
- Need to clarify differences/similarities between DAA, DAR, DAC etc.
- Standard DARs can create incentives for data sharing through adoption by funding agencies see https://doi.org/10.1007/s11192-022-04361-2

Schema registry service: exploring the uses of a standardised data model sharing API

Agenda & Slides • Recording

Promoting schema alignment across multiple datasets owned and maintained by various institutes is necessary to perform large-scale federated search and analysis of biomedical data. In order to obtain this alignment, data providers must ensure their schemas are easily discoverable and accessible. Here, we propose the idea for a new GA4GH standard, a “Schema Registry Service API,” which provides standardised methods for investigating, obtaining, and sharing biomedical schemas with the research community. We envision this API being adopted by multiple data curation institutes to promote a decentralised network of schema repositories.

Key Takeaways

- General consensus that a standardised Schema Registry Service API would provide benefit to research community for both technical and social reasons, similar to how TRS benefits the workflow analysis domain
- Focus on few features for an initial version, e.g. work on supporting only one schema language well before thinking of supporting too many conflicting technologies
- Schema Registry will be useful in bringing common access approaches to the complex and distributed landscape of where schemas are currently registered.

Crypt4GH: enhancing the use of encrypted data

Agenda & Slides • Recording

Crypt4GH is the GA4GH’s encrypted file format standard, designed to enable secure distribution by allowing easy and quick custom encryption and direct random access to encrypted data. We are now looking to enable new use cases and better integration with other GA4GH standards.

Key Takeaways

- Monthly meeting time poll should be distributed via mailing list to give participant from all time zones a chance to respond.
- Capabilities listed in /service-info would be a good way to add Crypt4GH support without requiring implementation by all projects htsget usage patterns may not warrant need for transcoding, but more conversation with
- htsget is required to address best solution for Crypt4GH support
- There is interest in Crypt4GH support in cloud APIs. Further discussion is needed, also including AAI and Passports

Experimental metadata

Agenda & Slides • Recording

The session aims to review metadata descriptions and organisational methods collected to date, identify overlaps between groups' methods of metadata collection and organisation, and determine next steps towards potential minimal set of metadata standards.

Key Takeaways

- **Securing Buy-in**: Need to identify and reach out to early stage initiatives that will commit to implementing the standard/guideline when it is created
- State purpose:
- Propose a rough draft of what could be worked on
    - Using ENA as a starting point

Paediatric pharmacogenomics secondary findings workshop

Agenda & Slides • Recording

This workshop will bring together the GA4GH Regulatory & Ethics community to review and discuss the points to consider for the Pediatric Pharmacogenomic Secondary Findings whitepaper currently in development.

Key Takeaways

- Paediatric pharmacogenomics return of results has many unique and nuanced policy issues that can be challenging to address. With feedback from the REWS community, the points to consider and policy recommendations will be revised and include additional text to improve clarity and broader application

Resources for getting started with GA4GH APIs

Agenda & Slides • Recording

This session aimed to get a better understanding of the needs of projects that want to make data available by deploying implementations of GA4GH APIs as well as users that wish to perform tasks that access multiple sources via GA4GH APIs. The session included a presentation on the role and purpose of the GA4GH starter kit and an opportunity for users to identify areas where they need more support in implementing standards within their environments.

Key Takeaways

- Today's discussion contributed towards the identification of "blue bars"

- Potential “blue bars” include annotation and sequence variants, synthetic cohort handoff, use and availability of APIs in a health care context at Philips, eLwazi

Passports/AAI v1.2 Q&A

Agenda & Slides • Recording
The updated specification has now been reviewed by four different reviewers from Verily, BioData Catalyst, ELIXIR Finland, and Datadex, independent of the development team. Most of their feedback has already been addressed and the finalised specification should be ready for release in weeks. This session provided attendees the opportunity to learn more about the specification changes.

Key Takeaways

- Need to increase visibility/participation/integration of AAI in other subgroups, especially Cloud WS
- Discussions needed on defining parts of Passports/AAI that are policy, not technical
- What does a global PP/AAI network look like, 100s of Clearinghouses or few central ones?
    - There may be a role of ORCID here
    - Suggest Biohackathon to build/test an international PP/AAI environment
- As a <data security> API, what is the expected rate of review/update/release?
- Where do we collect and share use cases for future implementers?

Cloud Work Stream Session

Agenda & Slides • Recording
The Cloud Work Stream will strategize with API leads and the community on its goals for 2023 for each API.

Key Takeaways

- API plans were well-received
- WES/TES have items to co-develop (callbacks and authN/Z for data access)

16 November

The GA4GH Cloud APIs in 2025: where do we want to be?

Agenda & Slides • Recording

In this session, the Cloud Work Stream will take out its crystal ball and look a few years ahead into the future of the Cloud APIs: How will each of the current APIs develop? How will they interact with each other? How will they interact with existing or developing standards outside the Cloud Work Stream or even outside of GA4GH? What gaps do we see? Do we need additional standards? And most importantly: *What use cases and user stories do we envision*?

For this session, we will deliberately encourage participants to think outside of the constraints of implementation details, potential breaking changes, legal barriers or technical impracticability to share with us their "In a perfect GA4GH world, ..." stories. Please come and share your vision with us, or just listen in and get inspired!

Key Takeaways

- Collaboratively compiled list of themes for future development
- For prioritising, themes can be voted on by the community for at least 2 weeks:
    - https://easyretro.io/publicboard/Zq5HrNVDVgfGoGFKj4htJTgg8mn2/16647f81-14ad-4106-acc8-3ee570c6be1c
- Themes ranking can be used to inform strategy decisions by WS leads & API champions

Beacon v2.0: migration workshop

Agenda & Slides • Recording

The goal is to map out the upgrade path for migrating a version 1 Beacon to version 2, preferably using real world examples. As an alternative solution, we can discuss the technical feasibility of and interest in a "translation" middleware that would allow v1 Beacons to join a v2 Beacon network, without direct upgrade.

Key Takeaways

- Beacons need a way to share or harmonise their query/response types
- Should gather documentation, example data, implementations, use cases, filters, mappings and tooling centrally

Beacon v2.0: "beaconise" your data type

Agenda & Slides • Recording

Beacon version 2 is a much more powerful platform for data discovery due to its modular design, which allows almost any data type to be "beaconized", including Structural Variant information and newer VCF formats. This is a working session where attendees can bring their own data types and discuss methods for "beaconizing" their data. The goal is to boost uptake of the Beacon v2 and demonstrate its potential to the community.

Key Takeaways

- Beacon does not report on missing filtering terms, only “0” results, this needs to be addressed somehow
- Community would benefit from a “safe” Beacon concept with input from REWS

FASP builder session

Agenda & Slides • Recording

This session aims to explore how LSG WS tools may make use of DRS objects and how to scale access of DRS and related topics, including bundling, bulk requests, metadata, and Passports.

Key Takeaways

- After reviewing the notebook, it is clear that there are some remaining issues to be resolved. This can be followed up on in the context of FASP (e.g. hackathon)
- There are a variety of builder priorities for 2023 put forward by DNAStack, Terra/Verily, Seven Bridges, NCBI, and ELIXIR. Many of these priorities relate to the use and implementation of DRS.

GA4GH community updates

Agenda & Slides • Recording

Discussion of the outcomes of the GA4GH Strategic Refresh and proposed updates to organisational structure, activities, and processes in 2023.

Key Takeaways

- GA4GH will update its matrix to better reflect the current ongoing activities within the organisation, which include study groups and implementation forum in addition to work streams, as well as the stakeholders that participate in those activities, which include organisational members, strategic partners, and individual contributors, in addition to Driver Projects
- The Genomics in Health Implementation Forum (GHIF) will be renamed the National Initiatives Forum, while the Federated Analysis Systems Project will be renamed the GA4GH Implementation Forum (GIF)
    - The National Initiatives Forum will be focused on translating genomics into real world clinical practice, through engagement with initiatives that are working to implement genomics across a health system, often at a national scale
    - The GA4GH Implementation Forum will be focused on advancing real-world use cases for GA4GH standards implementation to facilitate interoperability between initiatives as well as improvements to the standards
- GA4GH will announce an open call for new Driver Projects in late 2022 / early 2023.

30 November

GKS maturity & VA knowledge model v1.0

Agenda & Slides • Recording

This meeting aims to continue progress on integrating VA and VR subgroup outcomes into a single environment using consistent operational processes for defining the maturity of various components and areas being developed, as well as clarity around product releases. The long-term aim is also to incorporate the SA subgroup work into this environment as well.

Key Takeaways

- There was broad support (and no push back) on the fundamental argument behind the need for a maturity model and release/versioning process that would work consistently across the VRS, VRSATILE and VA specifications and schemas.
- The specific proposals made by the co-leads were well received in general. There was a significant amount of complexity in informing the attendees about the complexities in versioning artefacts at a granular component level and then associating their maturity levels and versions within an overall and potentially regular specification release.
- Concerns were raised by Daniel Cameron about whether there was a real need to version components individually, but it was agreed that we would continue diving into the use cases to make sure all parties understood the issue clearly so we could make the best resolution (being the least complex resolution that also enabled the efficient and transparent development of new and developing components).

DaMaSC: got schema?

Agenda & Slides • Recording

Data model harmonisation is a perennial problem. Some community members will present their perspectives on this issue, which we hope will spark a discussion on what the challenges people see and how we might tackle them. This will develop into a more formal proposal for the DaMaSC group to build on.

Key Takeaways

- There are many points that harmonisation could be tackled, from individual elements in a defined schema, to a complete schema, to a data model, to an informational model, or metamodel
- Many newer languages exist such as LinkML and JSON-LD to help harmonise at the higher levels, but this only moves the target
- Most of the community believe that “top-level” guidance from GA4GH on models and schema will facilitate internal harmonisation quicker

- Members will draft a detailed example use case as a point of discussion for a potential data/informational model or metamodel registry product proposal

Sequence Annotation integration with the GKS Work Stream

Agenda & Slides • Recording

Sequence Annotation is developing a set of extensible data models for representing common genomic features (e.g. transcripts, genes, and regulatory regions). In this session, working group facilitators will present the draft data model and discuss how to integrate current progress with existing specifications (e.g. Value Objects/Value Object Descriptors), and develop future tools and processes in-sync with the rest of the GKS Work Stream.

Key Takeaways

- A sequence feature is a piece of knowledge associated with a specific location on a biological sequence. See “junction” (SO accession SO:0000699 http://sequenceontology.org/browser/current_release/term/SO:0000699) and “region” (SO accession SO:0000001 http://sequenceontology.org/browser/current_release/term/SO:0000001).
- “Sequence Feature” at its core can be represented as a VRS Location with a type or subtype field identifying what sequence feature is being represented. Type or subtype in this context is basically what you’d find in column 3 of GFF3 (“type”). “Genome Build” might be a helpful addition to the SequenceFeatureDescriptor entity so users don’t have to pull that information out of the VRS Location. It is still not entirely clear, however, if this is the right approach for this stage of modelling. SA needs concrete data and use cases of what it does to improve whatever areas need improvement. Based on those cases and arguments for the aim and application of it, whether or not SA needs the same modelling approach as VRS and VA.
- SA needs to solidify our representation of gene including a workable definition. Some useful operations relating to sequence features tagged with a gene could include: returning all transcripts related to a gene, returning all regulatory regions associated with a gene, returning the entire sequence region associated with a gene (containing any and all annotations tagged with a specific gene or set of genes from beginning to end).
- A solid transcript model would be most useful to the other GKS working groups at first along with ensuring consistent representation of concepts between VA and VR (pay particular attention to the portability of concepts between specs).

Bad actors in research environments: who’s bad?

Agenda • Recording

It is often stated that data use will be audited, but no practical ways exist to track or understand this use. “Controlled access” effectively transfers full control of data to an approved user, as their ability to analyse and download all or “interesting” parts of it is not restricted in any way. Exfiltration controls are currently rudimentary, usually relying on counting bytes of data transferred to stop attempts to copy an entire data set. This BARE workshop will brainstorm the behaviours of a bad actor if given access to a dataset, genomes in particular and discuss the potential ways this could be tracked/logged and detected better.

Key Takeaways

- Need to define a small and practical test case for this kind of BARE fingerprinting
- TRE/Cloud style platforms allow some control and could be a first ecosystem within which to study these fingerprints
    - Risks in allowing arbitrary code to run on a foreign system
    - “Golden” copies of vetted binaries limits cutting edge research and analysis
- GA4GH Cloud standards represent an exciting opportunity for this kind of analysis tracking, e.g. WES/TRS/TES/DRS
    - DRS with universal IDs calls to data owner’s Clearinghouse
- Future Work Order Token will provide better audit points for detecting these fingerprints, but Passports/AAI already provides information on who is accessing which data and when
- Understanding “risk” is important, a single VCF is more than enough to identify someone
- Many ethical considerations for this kind of user tracking and auditing
- A **new BARE Google Group** has been created to share and discuss test cases