GA4GH September Connect 2023 Meeting Report

Members of the GA4GH community were invited to attend GA4GH Connect 2023 in San Francisco, USA, and online from 19 to 20 September. The GA4GH Connect sessions provided an opportunity for leadership and contributors to advance the GA4GH Road Map and gather feedback on product development and needs. The two-day meeting included Work Stream and Driver Project updates, breakout sessions, and themed events.

The next GA4GH Connect meeting will take place in Ascona, Switzerland, from 21 to 24 April 2024.

Table of contents

Tuesday 19 September

Beacon extensions
Cancer Community
Crypt4GH in the wider scope of GA4GH
GKS: the fellowship of the VRS Passports
DRS: addressing scalability and cold storage
Experiments Metadata Study Group
Long-term maintenance of GA4GH products
REWS General Meeting
VA data, from alpha to beta
Sequence Annotation (SA): raiders of the lost schema
Beacon filters workshop
Ethical Preparedness for Future Pandemics and Epidemics
GKS: the schemarillion
Sequence Collections: a standard for unique identifiers and compatibility of reference genomes
DRS v1.5 road-mapping and discussion
Equity, Diversity, and Inclusion (EDI) Strategic Plan session
Federated variant-level matching
GKS: the two toolkits
Preventing group harms in genomic research — brainstorming session

Wednesday 20 September

Data Models and Schema Consensus (DaMaSC) Study Group — interoperability across Work Streams
Genetic discrimination subgroup working session
GKS: return of the road map
Neuroscience Community Study Group
The Task Execution Service
Tracing ethical provenance for genomic and related health data across the research life cycle: concepts, strategies, and next steps
Cohort building from diverse phenotypic data
Rare Disease Community: challenges in representing, acquiring, and sharing rare disease phenotypes

Tuesday 19 September 2023

Beacon extensions

Agenda and minutes • Slides • Recording

After the acceptance of Beacon v2.0 as a GA4GH standard in April 2023, many domains have expressed interest in deploying a Beacon that suits their needs. This session focused on how Beacon models can be created in order to add more flavours to domain-specific Beacons.

Cancer Community

Agenda and minutes • Slides • Recording

This session welcomed a significant number of new invited members. To provide context, the session included an overview of GA4GH, the Work Streams, and the Cancer Community. Following that, members of standards working groups highlighted GA4GH standards that are applicable to the field of oncology. There was an opportunity for participants to provide feedback to inform the direction and content of future Cancer Community meetings.

Crypt4GH in the wider scope of GA4GH

Agenda and minutes • Slides • Recording

We presented the latest advancements in the Crypt4GH standard as well as proposed security enhancements. Our discussion included the integration of Crypt4GH into other GA4GH standards, particularly Cloud API standards, with the aim of incorporating Crypt4GH as an optional capability to relevant API standards directly. We showcased a prototype resulting from a collaborative PoC with Indigenous Genomics Australia, and also heard from GENXT who showed its pioneering work — where Crypt4GH facilitates secure data delivery to Trusted Execution Environments in a GA4GH-compliant way via TES and DRS.

Key takeaways and action items for April Connect 2024:

- we are working on v1.1 before next Connect;
- unified way of integration with other cloud APIs/AAI;
- encourage Work Streams/leads to view requirements for support across multiple standards;
- think about GIF project to come up with best solutions to integration.

GKS: the fellowship of the VRS

Agenda and minutes • Slides • Recording

The GKS Variation Representation Specification (VRS) has gone through several minor revisions since its initial release as a GA4GH standard in 2019, most recently incorporating genotypes and additional structures for representing copy number variation in VRS v1.3. Implementations of VRS and the extensions from VRSATILE in production systems have revealed key considerations for the development of the second major version of VRS, a primary focus of the VR team in the coming year. This session reviewed implementations of VRS in major genomic evidence resources such as MaveDB and gnomAD, highlighting lessons learned and how those have informed development of the VRS v2.0a model. We discussed the design of this model in the context of a pilot implementation with the ClinVar and CIViC datasets.

This session is for people who want an in-depth preview and discussion of implementation of the GKS VRS specification.

Passports

Agenda and minutes • Slides • Recording

AAI and Passports v1.2 provide a means for federated identity to enable easy and secure access to data. They support systems performing data discovery, providing interactive analysis services, and executing workflows. Implementation of these specs allows for better collaboration without implementing one-off access control mechanisms; by reusing well-recognized authentication and authorisation mechanisms; and while meeting security requirements of the most restricted data stewards. Wider adoption of AAI and Passports will reach more users and speed up collaboration and research.

Key takeaways and action items for April Connect 2024:

- what steps GA4GH projects can take to adopt or deepen adoption of AAI and Passports;
- how the AAI/Passport working group can help with adoption, including future enhancements to the AAI and Passport specifications or other advances.

DRS: addressing scalability and cold storage

Agenda and minutes • Slides • Recording

We addressed DRS scalability and cold storage issues.

Key takeaways and action items for April Connect 2024:

- finalise bulk PR365;
- cold storage support in DRS v1.5
    - issue/feature branch for how to show an object is “accessible/not accessible” or “cold/hot” in the access mechanism listing;
    - issue/feature branch for how to trigger thawing via access mechanism, keep informed of the status (maybe callback).

Experiments Metadata Study Group

Agenda and minutes • Slides • Recording

The Experiment Metadata Study Group aims to define a set of metadata properties to be captured to describe sequencing experiments, as those are crucial to make sense of -omics results made available to the community. In this session, we first had a look at the use cases to be submitted to the Standards Steering Committee to officialise this Study Group. Then, we began the technical work to prepare a draft properties checklist, based on the ENA Experiments Checklist. We got an overview of the work done at EMBL-EBI on this, and discussed with multiple Driver Project representatives, and other attendees, to evaluate how well this is fitting their needs in the context of real projects.

Long-term maintenance of GA4GH products

Agenda and minutes • Slides • Recording

This community discussion worked to develop product maturity-level guidelines and best practices for maintaining and promoting stable products.

Key takeaways and action items for April Connect 2024:

- outline areas of consideration for establishing a formalised approach to standard maintenance across GA4GH Work Streams (oversight, evaluation, governance, alignment with Product Development and Approval Process, etc.);
- establish working group across Work Streams and potentially with GA4GH staff to support regular discussions for developing recommendations for a GA4GH product maintenance framework;
- develop recommendations for a framework to maintain stable GA4GH products across Work Streams that can be either discussed at April Connect and/or presented to leadership by that time.

REWS General Meeting

Agenda and minutes • Slides • Recording

The REWS General Meeting brings together all contributors across all REWS subgroups to discuss topics relevant to the full Work Stream. This session included a vote by REWS members on REWS approval of the Diversity in Datasets subgroup's policy document, subgroup updates, and a discussion around road map directions.

Key takeaways and action items for April Connect 2024:

- Diversity in Datasets policy “Considerations for how genetic and genomic researchers should approach thinking about diversity in data” is approved by REWS!
- New draft REWS road map 2024
    - continuing projects;
    - reinvigorated projects
        - Newborn Sequencing;
        - AI/ML Bias;
    - new project proposals
        - measuring the uptake of ethics and data governance policies in genomics;
        - data visiting;
        - global ELSI group;
- We are seeking interest in the new proposals — if you are interested, please add your name to the +1: on the road map document, or email rews-coordinator@ga4gh.org. Please also add your name to any ongoing or reinvigorated initiatives you would like to be involved in!

VA data, from alpha to beta

Agenda and minutes • Slides • Recording

The Variant Annotation specification provides a framework to define schema for different types of variant knowledge statements that are rooted in a common conceptual model. The VA team has focused on an implementation-driven development approach wherein initial Statement models are defined directly by Driver Projects to support working applications with real data. Using the VA Framework, ClinGen and VICC projects have developed VA-informed alpha models for Variant Pathogenicity, Therapeutic Response, and Population Frequency Statements — which will support exchange of ClinVar, CIViC, and gnomAD data between their variant curation platforms. As these implementation models are tested

and refined in real-world settings, the VA team will oversee their coalescence and hardening into standard models that can be used by the broader community. This process will require harmonisation between VA-informed implementation models and the standard models that VA aims to provide to the community for out-of-the-box interoperability.

This session explored how this harmonisation process will work using real models and data — with the goal of bringing alpha implementations into alignment with the VA specification ahead of its v1.0 release.

This session is targeted towards projects that have or plan to implement the VA Framework to develop application models for their systems, and want to understand and/or help define the process through which standard models emerge in an implementation-led paradigm.

Sequence Annotation (SA): raiders of the lost schema

Agenda and minutes • Slides • Recording

The Sequence Annotation (SA) group has been working to identify concordance among the data models of major annotation providers (e.g. RefSeq, Ensembl, Alliance of Genome Resources). In this session, we presented our findings along with the first iteration of our proposed core conceptual model (v0.1). Discussion focused on identifying potential Driver Projects leading towards a project proposal for a v1.0 GA4GH SA data model standard.

Beacon filters workshop

Agenda and minutes • Slides • Recording

Filters are a key part of Beacon v2.0. Thanks to the feedback provided by implementers, we have identified areas of improvements, which were discussed during this session.

This session was divided into three parts, all equally important to improve the Beacon v2.0 specification:

1. discussion on filtering terms (Tim Beck);

2. discussion on implementing the "OR" operator (Anthony Brookes);

3. discussion on using a Phenopacket as a query (Peter Robinson and his team members: Baha El Kassaby, Francisco Castellanos, Daniel Danis, Michael Gargano,

Camille Liedtka, Chelsea Ferretti).

Ethical Preparedness for Future Pandemics and Epidemics

Agenda and minutes • Slides • Recording

The Ethical Preparedness for Pandemics and Epidemics (EPPE) Study Group launched in May 2023 after a session attended by the Infectious Disease Community and Regulatory & Ethics Work Stream at GA4GH Connect. During that session, Anja Bedeker (PHA4GE) facilitated a discussion on the need for ethical preparedness for future pandemics and epidemics. This need was underscored by a presentation provided by investigative journalist Emmanuel Freudenthal outlining the unethical handling of blood samples during the West Africa Ebola epidemic. Freudenthal and members of the group drew comparisons to the recent COVID-19 pandemic response.

This spurred a conversation regarding the role that standards development organisations, such as GA4GH and PHA4GE, can play in ensuring ethical conduct for future pandemics and epidemics. Research conducted during pandemics and epidemics is crucial for an effective response to disease outbreaks, but also gives rise to new and complex ethical challenges.

The EPPE Study Group is a PHA4GE and GA4GH collaborative initiative attempting to create a protocol/framework/policy that will address the ethical challenges that may occur in genomic research during pandemics and epidemics. The team has dedicated the last few months to completing a landscape analysis to better understand the nature and extent of the problem and resources that may possibly assist in addressing the problem. This is also assisting us in refining the scope of the project. In this session, we reviewed project progress and outlined steps for moving forward.

GKS: the schemarillion

Agenda and minutes • Slides • Recording

The GKS Variation Representation Specification (VRS) has evolved significantly since its first release. In parallel, the HL7 FHIR standard for genomic data has undergone extensive revision. To improve interoperability between these standards, the newest version of FHIR resources borrowed extensively from VRS data structures. Test implementations are being developed to demonstrate structural transformations that permit the expression of genomic variation data in either schema. Ongoing and future work is needed to continue alignment between the specifications and improve

interoperability among research systems that implement GKS standards, including public knowledge-bases such as ClinVar and CIViC, and clinical systems that support FHIR APIs.

This session reviewed areas of close alignment between the two specifications and highlighted areas where further work is needed to improve mappings so that data can flow more seamlessly. The results of these discussions may inform future road-mapping, encourage cross-pollination among the two organisations, and ultimately lower barriers for implementers.

This session is for people who are deeply familiar with GKS VRS and/or clinical genomic testing, and who want to understand the current and potential future state of harmonisation between VRS and FHIR.

Key takeaways and action items for April Connect 2024:

- identify gaps in VRS that should be supported by future versions;
- identify primary touch points between VRS and FHIR that should be maintained to support interoperability between the two specs.

Sequence Collections: a standard for unique identifiers and compatibility of reference genomes

Agenda and minutes • Slides • Recording

The goals of the session were to advertise and invite collaboration on the Sequence Collections draft specification.

Key takeaways and action items for April Connect 2024:

- tooling needed:
  - compute an identifier for a reference fasta;
  - main human references available on a server somewhere (who would host this, since these references would span providers? AnVIL?).

DRS v1.5 road-mapping and discussion

Agenda and minutes • Slides • Recording

This session focused on discussing the direction and updates that will be included in the DRS v1.5 release. Attendees brought their thoughts and ideas for the upcoming year of working on the DRS standard.

Key takeaways and action items for April Connect 2024:

- list of priorities will be turned into a Cloud road map.

Equity, Diversity, and Inclusion (EDI) Strategic Plan session

Agenda and minutes • Slides • Recording

This session presented the first draft of the EDI Strategic Plan. Participants provided real-time feedback and engaged in discussions about the anticipated impact of stated goals.

Preventing group harms in genomic research — brainstorming session

Agenda and minutes • Slides • Recording

Most research ethics frameworks focus on potential harms to participants of research studies. There is, however, increasing focus on potential harms to those not directly involved in research, but who might nonetheless be harmed indirectly by the research because of scientific claims made about the groups those individuals are members of. This session consisted of dynamic brainstorming about the 2024-2025 activities of the REWS Preventing Group Harms subgroup, which proposes to develop tools to assist in making the genomic research environment less vulnerable to occurrences of group harms.

Key takeaways and action items for April Connect 2024:

- considering potentially a discussion document first before a policy document;
- information gathering stage: next steps to plan meetings where people with experience addressing group harm give presentations on their approaches.

GKS: the two toolkits

Agenda and minutes • Slides • Recording

High-throughput annotation using VRS requires high-performance tools for sequence retrieval, variant normalisation, and variant identification. In this session, we described recent updates to the Biocommons Toolkit for high-throughput sequence retrieval and alignment operations, including a discussion of the use of the LSG Work Stream's refget v2.0 and draft Sequence Collections specifications for building a federated ecosystem of sequence collections. We discussed how these updates improve high-throughput, federated VRS annotation of variant collections using the GA4GH VRS-Python Toolkit, and how recent changes to VRS-Python further improve variant representation performance.

This session is targeted at implementers seeking to apply GKS in large-scale

genomics applications using open-source community tools.

Federated variant-level matching

Agenda and minutes • Slides • Recording

This session furthered progress toward the completion of the GIF goals: to establish a federated variant-level matching platform with registered access, using GA4GH products, and with an initial set of nodes within the network.

Key takeaways and action items for April Connect 2024:

- launch one or more federated variant matching platform(s);
- define ways to expand the federated variant matching platform.

Wednesday 20 September 2023

Data Models and Schema Consensus (DaMaSC) Study Group — interoperability across Work Streams

Agenda and minutes • Slides • Recording

GA4GH Work Streams gather experts from particular areas of interest to develop standards, frameworks, and tools. Work Streams operate independently and often define vocabulary and data representation (schema) standards. This workshop explored some of the challenges encountered when bringing together products from different Work Streams, assessed whether core vocabulary across Work Streams would help ease interoperability challenges across products/Work Streams, and brainstormed ideas for accomplishing this.

Genetic discrimination subgroup working session

Agenda and minutes • Slides • Recording

The meeting brought together members of the Genetic Discrimination subgroup of the GA4GH Regulatory & Ethics Work Stream and of the Genetic Discrimination Observatory (GDO). Presenters updated both groups on the work of the Genetic Discrimination subgroup and discussed key findings from a Delphi study conducted by these groups on the elements of an optimal genetic discrimination policy.

Key takeaways and action items for April Connect 2024:

- a draft manuscript will be shared with the group in December;
- all contributors can fill out the list of contributors table. A copy of the key findings report has been shared. Comments to the document were collected through 9 October.

GKS: return of the road map

Agenda and minutes • Slides • Recording

This session recapped the other sessions relevant to GKS activities covered at Connect 2023, and included cross-team planning (Variant Representation, Variant Annotation, and Sequence Annotation) for the 2023-2024 Q4/Q1 road map. Some specific VR topics covered included incorporating models from the Structural Variation SIG into VRS v2.0a, advancing VRS v2.0 models to a community beta, and progress on a Study Group for categorical variation. VA and SA plans were discussed, as were plans to align work with external communities (e.g. HL7). Participants included those with an interest in developing and applying the GA4GH genomic knowledge standards.

Neuroscience Community Study Group

Agenda and minutes • Slides • Recording

This was an exploratory meeting to gauge interest in developing a GA4GH Neuroscience Community of Interest. The session consisted of brief introductions to a variety of Canadian neuroscience groups interested in interfacing with the neuroscience data sharing standards community. They provided feedback on potential opportunities and challenges associated with the development of a GA4GH Neuroscience Community of Interest.

The Task Execution Service

Agenda and minutes • Slides • Recording

This session focused on developing the TES API, and working on issues around implementation and engines that use it.

Key takeaways and action items for April Connect 2024:

- goal of TES community development (workflow engines and servers) for the next year;

- goals for TES v1.2;
- plan for developing authentication/authorisation elements for TES.

Tracing ethical provenance for genomic and related health data across the research life cycle: concepts, strategies, and next steps

Agenda and minutes • Slides • Recording

This session:

1. highlighted the pathways by which individuals and communities communicate their preferences for data reuse, and how research institutions and repositories execute these preferences through data access, transfer, and use agreements in a process we refer to as ethical provenance;

2. identified the existing practical and policy challenges for tracing ethical provenance across the research data life cycle from the perspectives of institutional data stewards, including signing officials;

3. demonstrated how software, large language models, and other computational tools can support the synthesis of legal texts and granular development of machine-readable terms for data reuse.

Key takeaways and action items for April Connect 2024

- next steps for evidence-based development of contractual clauses for:
    - data access agreements;
    - data access requests;
- identify inter-group synergies with Data Use and Researcher Identities (DURI) Work Stream, Data Access Review Standards (DACReS) subgroup, and Regulatory and Ethics Work Stream (REWS).

Cohort building from diverse phenotypic data

Agenda and minutes • Slides • Recording

This session identified continuing cross-Work-Stream themes from the Computable Cohort hackathon in 2023, such as:

—DARathon: how to make a successful data application;

—developing shared understanding of GA4GH and other standards, and how they

allow assembly of cohorts based on domain-specific phenotypic and sample data (Data Connect, Beacon, Phenopackets, FHIR);

—continuing with the asthma example as test case, and identifying up to two scientific use cases that would benefit from aggregating data or workflow results from dbGap, EGA, and JGA;

—planning a GIF Project and/or hackathon as future collaborative activities.

Key takeaways:

- clear desire from multiple people/groups to have a proof-of-concept project building a cohort from diverse data sources to answer a real-world scientific question;
- potential scientific areas of interest from those in attendance include neurodegenerative disorders (e.g. Alzheimer's disease, Parkinson's disease), rare diseases, and healthy cohorts followed over longitudinal studies.

Rare Disease Community: challenges in representing, acquiring, and sharing rare disease phenotypes

Agenda and minutes • Slides • Recording

The GA4GH Rare Disease (RD) Community aims to promote global cooperation, data sharing, and collaborative rare disease research through identifying the need for new standards, and/or implementing existing GA4GH standards. This mission will be accomplished by collaboratively defining scientific rare disease use cases that may benefit from global cooperation, by a commitment to advancing those use cases through either the development or implementation of GA4GH standards, and by providing practical examples and guidance for rare disease sharing and analysis.

The RD Community Connect session focused on one of the most critical dimensions of RD diagnosis and care coordination: phenotypes. Deep phenotyping of patients suspected of or diagnosed with a RD — using the Human Phenotype Ontology — has become standard practice in the last ten years. And while the adoption of standardised terminologies and an increasing willingness to share data internationally have laid the foundation for progress in the RD field over the course of the last few years, for many patients the diagnostic odyssey will ultimately be futile: more than 50% of RD patients do not receive a molecular diagnosis.

This session covered some of the challenges underpinning further progress, including

the encoding of longitudinal phenotypes, the communication and exchange of phenotypes in human language, and the definition and sharing of statistical data on phenotypes in diagnosed cohorts.