Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

GIF AMA

GA4GH Implementation Forum: Ask Me Anything

May 21st, 2025

Featured standard

PHENOPACKETS

Developed by the GA4GH Clinical and Phenotypic Data Capture Work Stream, Phenopackets offer a human and machine-readable way to structure clinical and phenotypic data about a patient or individual.

Learn more here.

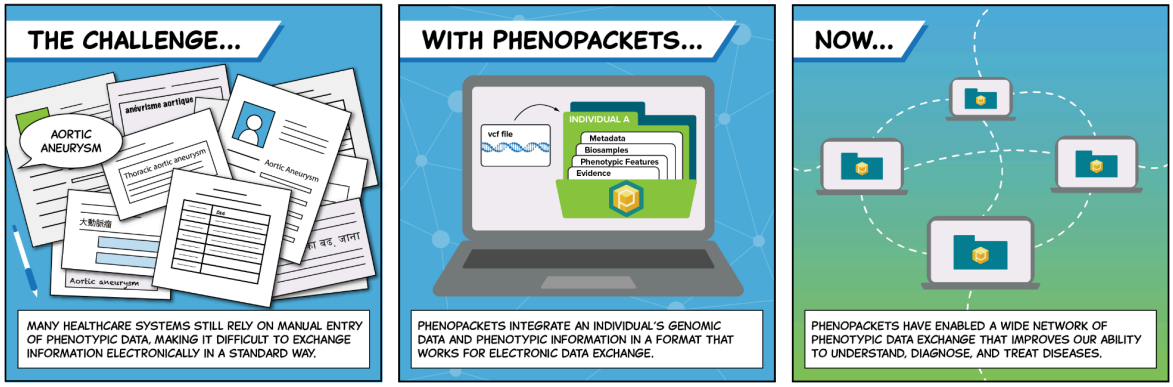


Image summary: *Phenopackets help researchers and clinicians exchange phenotypic data in a standard way to accelerate research and improve patient care.*

Meeting Overview

Title	GIF AMA: Phenopackets
Guests	Open to all, including • Phenopackets Product Team • Pehopackets Implementers • GA4GH Implementation Forum • Phenopacket-curious Implementers
Start time	May 21, 2025 11:00 AM EDT
End time	May 21, 2025 12:00 PM EDT
Location	Zoom
Description	This GIF AMA session will focus on the Phenopackets standard, which provides a human and machine-readable way to structure phenotypic and clinical information about a patient or individual. Designed as an interactive discussion, this session will provide current and prospective implementers with direct access to the Phenopackets product team and experts who have implemented Phenopackets in diverse settings. Attendees will have the opportunity to ask technical and practical questions, learn from real-world implementations, and explore solutions to adoption challenges. Meeting recording

Agenda

	Item	Presenter	Time
1.0	Welcome	Benjamin	2’
1.1	Introduction to session and Phenopackets experts	Mallory	3’
2.0	Latest & Greatest updates on the Phenopackets standard	Jules	5’
3.0	Q&A from attendees	Mallory (chair)	45’
4.0	Wrap-up and next steps	Mallory	4’
4.1	Close	Moni	1’

Info

- Meeting Recording: https://us02web.zoom.us/rec/share/PUVxzdWDOvGIVlRg1aMGRFOiR7Z0H4ixAopVcdxNljhXf8y6cj9PyFJXlFdMV84O.kfhWkdMt6VHnyL44
- Meeting slides

Action items

- [ ] Complete feedback survey on this GIF AMA event
- [ ] Make resources available linked from this document
- [ ] Join GitHub or join our Slack Workspace #clin-pheno_ws to ask additional Phenopackets-related questions

Questions and Answers

Note: This is a community-driven resource, and your contributions help make it better! If you spot a mistake, see an opportunity to add a resource, or have more details to share, please jump in and make changes. We welcome and encourage your input!

Meeting summary:

Intro to Phenopackets:
Jules Jacobsen gave a brief update on the current state of Phenopackets, emphasising stability over frequent changes to avoid breaking implementations. He provided a quick schema overview, explaining how Phenopackets tie together genomic and phenotypic data at the individual level, with support for families and cohorts. Key components include diseases, genomic interpretations, biosamples, phenotypic features, and medical actions. Jules pointed to resources like the formal specification on ReadTheDocs and implementations in Protobuf and JSON Schema for validation. Peter Robinson added that over 8,000 curated example cases from literature across 500+ diseases are now available for public use and analysis.

Question and Answer section:

💡Idea/ suggestion
👉 Response
❓Question
🔗 Link / resource
👋 Implementer story

❓Question (David Bujold): What is the **vision for the next 1-2 years of Phenopackets** development, and is there a real intention to address known gaps and consider a potential update (e.g., version 2.1), especially given the feedback gathered from users like us who have identified areas for improvement?

👉 Moni Muoz Torres: Any future updates to Phenopackets, such as a version 2.1 or 3.0, must be driven by clearly identified community needs. There is always room for improvement, but the evolution of the standard should only happen in response to real-world use cases and user feedback, rather than speculative additions.

💡 David Bujold: Perhaps the bi-monthly Phenopackets meetings should include dedicated time for in-depth discussions on identified gaps or missing features in the standard, rather than focusing solely on project presentations. Detailed feedback often gets lost on Slack - perhaps we need a structured forum to address these issues collaboratively.

👉 Jules: Yes, it's really valuable to get feedback based on real use cases. Please raise specific issues directly on the Phenopacket GitHub repo, ideally with a clear description of the challenge and a possible solution. That way, others can see and contribute to the discussion, and it can be brought to the regular meetings. The process should be guided by actual user needs and collaborative input.

🔗 Issues and feature requests here:
https://github.com/phenopackets/phenopacket-schema/issues

👉 Response (Peter): highlighted two specific issues:

1. *Phenotypic features lack an "age observed" field, which makes it hard to model recurring traits like seizures.*
2. *There's no way to relate measurements directly to phenotypic features, which is currently underspecified.*

💡He suggested these could be added as non-breaking changes and proposed workshops where users bring real examples and discuss what additions would make Phenopackets more useful. He also raised broader questions about whether certain complex or sensitive data types (like environmental exposures or race/ethnicity) should be included in Phenopackets or handled separately.

👋 Anne Deslattes Mays shared that she's working on a project with Richard Sherman (NLM) and Chris Mungall (ontologies), with Lifebit leading federated data analysis. They're **integrating**

environmental exposures and social determinants of health using GIS layers, while ensuring patient data remains local and only aggregated data is shared.

💡 She sees a big opportunity for Phenopackets in this context and is curious about where it might fit in, potentially as part of a demonstrative project.

👉 We'd love to have you join our new Human Exposome Study Group launching on June 6th.
🔗 https://www.ga4gh.org/product/human-exposome-data-standards/

💡 Michael Baudis stressed the **need for clearer definitions and standardised approaches within GA4GH, especially regarding complex concepts like race and population background.** He highlighted ongoing challenges with inconsistent interpretations across databases, such as in Beacon, and emphasised the importance of developing agreed-upon vocabularies and data structures, along with clear explanations, to ensure consistency in epidemiological and population data use.

👉 Response from Jules: There are challenges in standardising the representation of race and ethnicity across different regions and systems, with significant variations between countries like the UK and the US. Existing classifications are often inadequate and inconsistent, making it difficult to create a universal standard.

👉Michael: We have similar race/population discussions at Beacon, and e.g., we use a compositional model in our resources, but this hasn't made it to Beacon. See e.g., issue https://github.com/ga4gh-beacon/beacon-v2/issues/215

🤝They suggested the need for general guidelines on technical representation and clear definitions but acknowledged this could require a dedicated group, possibly involving ethical considerations, to address properly.

👉 Orion Buske (chat): We ended up pointing to several resources, including the HANCESTRO terminology and the ClinGen Ancestry & Diversity Working Group

---

❓Question (Mallory Freeberg): For those who are currently implementing phenopackets: How did you decide that **implementing Phenopackets was right for your project**?

👉 Victor Rocheleau (McGill) (chat):
We have some use cases that we'll discuss in the GitHub issues as soon as possible. We recently extended our schema implementation to include geo-location data for biosamples. We would

also be interested in describing variants of interest outside the context of a disease. Looking forward to discussing if you have similar use cases :)

> 👋 Victor Rocheleau shared that their team has implemented a Django service with a Postgres database to handle Phenopacket data and recently migrated from v1 to v2. They are now supporting new use cases, including agricultural and environmental data, where capturing biosample geolocation is important, so they extended their biosample schema accordingly. Another use case involves expressing 3D variant interpretations without necessarily linking them to a disease, which currently poses challenges. Victor plans to open detailed GitHub issues for these cases and invites others with similar interests to collaborate.

👉 Nicholas Owen (UCL) (chat):
I am exploring Phenopackets and other GA4GH products to improve sharing of human disease information across research silos, without sharing the underlying identifiable data. To improve and strengthen research outputs

👋 Kais Ghedira explained that their project, the African Genomics Data Hub, is focused on genotype-phenotype correlation. Is Phenopackets suitable for supporting genotype-phenotype correlation in projects like the African Genomics Data Hub, and how can it be effectively used for this purpose?

👋 Dave Gordon from Nationwide Children's Hospital described their clinical use case involving the consolidation of manually curated phenotypes from genetic counselors and phenotypes extracted from clinical notes. They are developing a system called Phoebe, which uses a network graph to represent phenotype evidence and genomic data. They plan to use Phenopackets for tool integration and to export identified patient cohorts for clinical and research purposes.

👋 Orion Buske shared that they implemented Phenopackets early within the Phenotips tool, which manages structured phenotype and family history data primarily in clinical genetics workflows. Phenopackets are used as an export format for patient and family records, including phenotype and genetic diagnostic data. This standardised export facilitates moving data to downstream analysis systems, such as variant prioritisation tools, making data transfer more convenient - similar to the use case described by Dave Gordon.

❓ Question (Benjamin Berk) (chat): Has anyone **automated the generation** of Phenopackets from EMR? Has anyone developed Phenopackets **capture or validation UI**?

👉 Gordon Krieger (chat): There's definitely a need for Beacon <-> phenopackets mappings

👉 Liina Nagirnaja (chat): https://github.com/EGA-archive/phenopackets-to-BFF

👉 Mireia Marín Ginestar (chat): At the EGA we have developed a mapping between beacon and phenopackets schema

👉 (chat): Gordon Krieger (chat): Yes, I've seen that, and it's great, although doesn't exactly match my use case (which is query mapping, not record rewriting)

👉 Benjamin Berk:
https://www.sciencedirect.com/science/article/pii/S1532046423002794
https://github.com/CNAG-Biomedical-Informatics/convert-pheno

👉 Dave Gordon (chat): Depending on your definition, we're working to automate EHR -> term extraction -> graph database -> phenopackets. We unfortunately don't have direct EHR access for it at this time, so we're dependent on internal ETL processes rather than FHIR or similar.

👋 Michael Baudis (in chat): we provide phenopackets through our Beacon APIs as additional data model for progenetix.org and other resources. But we don't have a use case; this is just to show as a way for Beacon to provide data.
https://progenetix.org/beacon/phenopackets/?testMode=true

❓Question from David Bujold: Is GA4GH currently aware of **how many Driver Projects are actively using Phenopackets** - beyond just expressing interest - such as making data available as Phenopackets or sharing them online as usable resources?

👉 Angela Page: GA4GH has basic information from Driver Project surveys about whether projects plan to implement Phenopackets, but the data lacks detail on actual usage.

👉 Mallory: GA4GH is exploring a central registry to track real-world implementations, though questions remain about how to scale, maintain, and update it sustainably. Phenopackets could benefit from and potentially lead in this initiative.

💡Victor Rocheleau (McGill) (in chat): Having a web page linking to community implementations would be great!

❓Question from Benjamin: How are people currently **generating PhenoPackets** - are they being created automatically from EMRs, converted from formats like FHIR, or generated in other ways - and is there interest in tools to support this?

👉 Response from Peter: PhenoPackets will become truly valuable once more supporting software is developed. He highlighted existing tools that already use Phenopackets, such as Exomiser, and mentioned ongoing efforts to build tools for genotype-phenotype analysis and data conversion (e.g., from FHIR or OMOP).

💡He called on GA4GH leadership to help coordinate efforts among groups working on similar tools, suggesting collaboration could prevent duplication and accelerate the development of robust, ontology-aware analysis pipelines.

❓How do others handle **converting diverse data formats into Phenopackets**?

❓Is there a desire for FHIR -> Phenopackets or other standard --> Phenopackets **conversion**?

❓Anne Deslattes Mays (in chat): We have everything for our Vicki Shanmugam's Office of Autoimmune Disease Research federated analysis pilot in NLM under Richard Scheuermann and our clinical data is mapped to OMAP for our cohort browsing across nodes. How do we **convert this to phenopackets**? We are using single cell data and I am automating quality control, etc... but there may be an opportunity for us to create nextflow workflows and modularizing at the process level.

👉 Gordon Krieger (chat): also convert-pheno is a bit confused about whether "Exposure" is a phenopackets field (it's not) ... but it's understandable, since the phenopackets documentation is actually incorrect.

👉 Response from Michael: converting data to PhenoPackets is similar to converting data for Beacon, especially with OMOP mappings. He suggested there's strong synergy between these efforts and recommended active alignment between PhenoPackets and Beacon to streamline integration within the GA4GH ecosystem.

👉 Response from Benjamin Berk suggested the Clin/Pheno meeting as the right forum to discuss priorities like FHIR-to-Phenopackets or OMOP-to-Phenopackets conversions, potentially leading to an implementation guide, API, or library. If we have OMOP->Phenopacket, there exists: https://orchestrate.docs.careevolution.com/convert/fhir_to_omop.html

👉 Response from Jules: he shared a basic OMOP-to-PhenoPackets exporter created during a 2020 biohackathon and encouraged community contributions to improve or productionise it. He noted that ontology mapping, especially to HPO, is a challenge but not strictly required for creating Phenopackets.

👋 Victor Rocheleau shared that at McGill, they support diverse projects via the Bento platform, which requires converting various data sources (e.g., databases accessed via VPN, CSV files) into

PhenoPackets. Historically, they've used ad hoc scripts created by a colleague for these conversions. They're now developing a more general ETL application to streamline and standardise the process, with the long-term goal of making it useful beyond their own projects.

❓ Question from Joshua Proulx (in chat): Does anyone have any use cases where they're needing to **convert their data from another format like TOPMed to Phenopackets**? If so I'd love to chat in more detail! We're working on a grant at Geneial to develop tooling to automate the data transformation from TOPMed to Phenopackets and would love to have some beta testers. My contact info is josh@geneial.com

👉 Emily Hughes (in chat):
I work as part of the BioData Catalyst consortium - we work with TOPMed data quite a bit. Wondering what you mean by TOPMed format? (Happy to take this offline as well)

👉 Jules Jacobsen (in chat):
https://github.com/phenopackets/omop-exporter is a partial mapping of OMOP to Phenopackets. If you want to use this or find it useful, please submit pull requests.

❓ Question from Victor Rocheleau (McGill) (in chat): Would love to know more about how others are **leveraging ontologies for hierarchical search** in their implementations

👉 Michael (in chat): Victor - have a look at Progenetix but this is expansion within term trees ... e.g. https://progenetix.org/subsets/NCIT-subsets/ but the expansion then works on sample search etc.

👉 Dave Gordon (in chat): Nothing fancy yet beyond is_a relationships, but we're directly using SPARQL for hierarchical search.

Resources shared in chat:
🔷 https://github.com/monarch-initiative/pyphetools
🔷 https://phenopacket-schema.readthedocs.io/en/latest/
🔷 Genotype-phenotype correlation: https://pubmed.ncbi.nlm.nih.gov/40093222/