Global Alliance for Genomics & Health

ARCHIVED

# GA4GH Discovery Work Stream

Minutes and Actions 2019-2020

These are the minutes for the GA4GH Discovery Work Stream. For further information, please visit the GA4GH Work Stream page (will link to the page on the website when it is ready) at http://ga4gh.org.

Table of Contents:


Meeting Protocols

Links to Sub-Group Minutes and Actions Documents

List of Meetings

- 2020-03-25 Discovery @ GA4GH Virtual Connect
- 2019-06-24: Case Discovery API Creation
- 2019-03-27: March Updates
- 2019-02-27: Schema Blocks Governance
- 2019-01-30: Discovery relaunch
- 2019-MM-DD: Template


## Meeting Protocols

- Please note that by participating in meetings, attendees agree to adhere to the GA4GH Standards of Professional Conduct.
- Meetings may be recorded for note-taking purposes. Recordings will be deleted within three months of the meeting taking place.
- Dates should be specified in the international format yyyy-mm-dd

## Links to Sub-Group Minutes and Actions Documents

This Work Stream has meetings for the following sub groups.

1. Beacon API - 2017 - 2018 - 2019 - 2020 - 2021
2. Discovery Search - 2018 - 2019 - 2020 - 2021
3. Discovery Networks - 2018 - 2019
4. Discovery Work Stream Previous - 2018 - 2021
5. Case Discovery - 2019

Global Alliance
for Genomics & Health

# List of Meetings

## 2020-03-26 Discovery @ GA4GH Virtual Connect

Chair:
Attendees “Name (Affiliation)”:
Slides:
https://docs.google.com/presentation/d/1APTSzhrRXQe2cJ_kBxEyui-pbioOB_Lv9iXQ9tXjTA8/edit?usp=sharing

| | Actions Arising | Assigned To | Deadline |
|---|---|---|---|
| 1 | | | |
| 2 | | | |

| | Agenda Item | Person/Time |
|---|---|---|
| 1. | Welcome & Confirm Agenda | 5 |
| 2 | Discovery Work Stream | Marc, Michael / 5 |
| 3. | Discovery Networks & GA4GH Network: Plans | Jeremy / 10 |
| 4. | Search API - outline + timeline for GA4GH Approval | Jonathan F / 10 |
| 5. | Beacon API v2 - outline + timeline for GA4GH Approval | Jordi / 10 |
| 6. | Schema Blocks | Michael / 10 |
| | A.O.B. | 5 |

## 2019-06-24: Case Discovery API Creation

Chair: Michael Baudis / Marc Fiume
Attendees “Name (Affiliation)”: Nara Sobreira, Heidi Sofia (NHGRI)
Miro Cupak, Michael Miller
Ian Fore (NCI, DCF),
Mike Qu
David Lloyd,
Francis Jeanson,
Tony Brookes
Jacek Lebioda

Global Alliance
for Genomics & Health

| | Actions Arising | Assigned To | Deadline |
|---|---|---|---|
| 1 | Schema block suggestions | Tony, Ian | |
| 2 | | | |

| | Agenda Item | Person/Time |
|---|---|---|
| 1. | Welcome & Confirm Agenda | 5 |
| 2. | Updates and Plans from Sub Groups (Update, next milestone, issues) 5 mins<br>Case Discovery: (Ben Hutton)<br><br>Beacon API: (Jordi/Juha)<br><br>Discovery Networks:(Miro)<br><br>Schema Blocks:(Michael)<br><br>Search API: (Miro/Aaron)<br>● Can Search refer to GA4GH Schemas and Schema Blocks as its schemas? | 30 |
| 3. | | |
| | A.O.B.<br>- identifiers.org | |

MF: Case Discovery is the new “Search” derived MME interface API

- Proposal sent out
- Allows MME and other drivers to proceed with time constraints while others like to develop longer time scale with Search. Can look to align

**Beacon API**

- MB: Jordi has taken lead on API development. Two main areas of development
- Clinical use cases
    - ...
    - Low implementation threshold, building some practical implementations
    -
- On the European side, lots of ELIXIR AAI system is being set up across

- Calling Beacon 2
    - Will have to split this at certain point
    - DL: Gary has discussed form with Juha,

**Discovery Networks**

- MC: Two items - /service-info in the metadata and Service Registry which manages the metadata. Needs more polish before pushing for release so we will submit in a months time.
- MB: Would be good to update the website

**SchemaBlocks**

- Isuru from Melanies group has been looking at this
- …
- Phenopackets format is providing a test of a complex example
- Format will be JSON Schema. A version has been used which is re-writes. The Phenopackets project is being done as correct JSON Schema, and the tooling will be re-worked after taking lessons from this process
- IF: There are probably some extensions t
    - MB: Ben is JSON Schema lead developer, so …
    - IF: There
    - ..
- MB:

MF: Should the endpoint with SchemaBlocks of GA4GH Schemas?

MF: **is this the repository to look into for schema blocks**?
https://github.com/ga4gh-schemablocks/blocks
or this? https://github.com/ga4gh-schemablocks/blocks-json-schema

- The “blocks-json-schema” is the one where the JSON Schema definitions are being tried out.
-
-

## 2019-03-27: March Updates

Chair: Marc Fiume / Michael Baudis
Attendees “Name (Affiliation)”:
Augusto Rendon (Genomics England)
Sergi Beltran (CNAG-CRG, Barcelona; EJP-RD, Solve-RD)
Heinz Stockinger

Global Alliance for Genomics & Health

Tony Brookes
Jordi Rambla
Francis Jeanson
Gary Saunders
David Lloyd
Ian Fore (NCI GDC/DCF)

| | Actions Arising | Assigned To | Deadline |
|---|---|---|---|
| 1 | Discovery Networks update to the Discovery Web Site | Miro | |
| 2 | | | |

| | Agenda Item |
|---|---|
| 1. | Welcome & Confirm Agenda |
| 2. | **Driver Updates**<br>**ELIXIR (Jordi/Gary)**<br>- Finished 1.0 some time ago, updated ELIXIR Beacon to match<br>- Working on ELIXIR Beacon Network, which is guiding Discovery Network<br>- Expanding Beacon spec to have handover, access levels, what items it can support<br>- Filters being added to the Beacon query e.g. age, gender. In line with data and evidence beacons that have been described so far<br>  - MB: Meeting 3 weeks ago had intense implementation discussion e.g phenotypes etc. Settled on model on ontology based model. E.g. prefix with help matching phenotype<br>- GS: Going to restructure calls post GA4GH Connect to clarify the GA4GH and ELIXIR boundaries, with separate Beacon calls<br>- Beacon 2.0 will need new approval focus. Timelines to be clarified.<br>  - RN: Week 3 of June for the process. Product Proposal Form also required.<br>**NCI GDC+DCF (Ian Fore)**<br>- Been looking at NIH NCI levels of engagement, want greater engagement<br>  - Funded by NCI, run by University of Chicago through Bob Grossman<br>  - NCI to show greater activity<br>- Bio-ITWorld panel on diverse data models panel<br>  - MF presented with Bob Grossman, Mike Violo (dbGap)<br>- Setting to look at pilot implementations of Search/Schema Blocks that might be relevant here<br>- https://gen3.org<br>- https://portal.gdc.cancer.gov<br>- Here are two resources NIH / NCI develops, very relevant to Discovery |

| | |
|---|---|
| | **EJP-RD (Tony Brookes)**<br>- Now onboarded - Sergi Beltran from Barcelona and myself are Champions<br>- Number of projects<br>  - ChPap? Data Resource in Spain, working with MME, one sided discovery<br>  - Efforts to find relevant rare disease resources on metadata level, OrphaNet<br>  - Cafe Variome based SW includes RDNexus for Network Exploration (24 european nodes) of the Unseen (don't see data but can Discover it)<br>  - Populated with data from ERMs<br>  - Want to add consent and data use via ADAM model<br>  - Phenotype similarity, not just matching<br>- Basing phenotypes on phenopackets<br>- Model matching - for model organism researchers via characteristics<br>- Beacon in Cafe Variome system being developed |
| 3. | **Sub Group Updates**<br>Beacon API (Juha)<br>- See ELIXIR section again<br>Search API (Ben/Miro/Aaron)<br>- BH: Most of work I've been doing has been on technical side. Schema Blocks tooling<br>  - Issues of next Release Candidate assigned to developers<br>  - DNAStack have been working on interface to Search API that connects to distributed searching platform software<br>  - Aaron going to do more testing showing Search API and its use connected to Google Platform tooling<br>- Appendix<br>- Demo update from Miro<br>  - Reached a milestone in distributed Search Engine<br>  - Had a milestone. Can submit arbitrary SQL query to several separate data sources. Demonstrate with SQL store.<br>  - Still looking for new connectors<br>- Please submit Use Case Collection document if needed<br>- TB: This Milestone demo vs Basel demo - both are federated searches across separate endpoints in a common query language - what is the difference?<br>  - MC: The key difference I think is flexibility.<br>  - The arbitrary SQL query doesn't rely on specific schema<br>  - Can connect several different types of information. Query one resource with variants, and matching with sample and file information from completely separate storage devices<br>  - TB: How do you build that flexibility in?<br>  - MC: For the SQL and NoSQL store we had a way for the store to describe the structure of the data. DRS has its own schema |

        - MB: So need a custom adapter
    - TB: What is the GA4GH standard here? Is it the way in which the back end tells you it's data structure?
        - MC: This is an implementation that is separate from the standard. It doesn't have a Rest API. It doesn't map to a specific standard.
        - MF: We want to standardise both - **how does a repository reflect what can be searched within in**, and **then how do you perform the search**.
        - BH: It demostrates really well how Search can be used, so it is a really viable and powerful reference implementation. But it may not work for those with their own permissions and data sources. But GA4GH really has standards as its product, not software.
    - TB: We know the problems are, makes sense.

**Discovery Networks (Miro)**

- Service-registry / Service-info
- Small incremental changes to both specs

**Schema Blocks (Michael)**

- Mission statement doc has been moved to the website
- Good interaction with phenopackets w/ cross-repository discussions & contributions
- Need to reach out to GKS groups
- E.g. how do you represent 'age' as a standard? The idea is if using a product for GA4GH look at these and use them.
- Community effort across knowledge areas
    - Scoring system needed - how accepted and used something is
- Variants may have multiple types
- TB: Data models? Are there collections for multiple ways to represent similar things? Indices?
    - MB: Index will be implemented in the future. Current representation is the website formated one.
    - MB: Data models - adopted GA4GH data model. Has hierarchy. But with phenopackets there is a different model.
- Versioning needed. Versioned example implementation can be made to demonstrate how things can be used.
- Will probably have a clear restart after Hinxton
- IF: SchemaBlocks addresses the question of models/data elements. Is the intent that SchemaBlocks is the standard by which we share models?
    - MB: The acceptance of this being the place when they want to share concepts is something we hope for us to make this succeed.
- IF: It's a dynamic problem for me to understand something I have seen not before, and I am not going to be an expert in. Dynamic is opposite to the static nature of a standard
    - MB: I don't think the issue is for is to say which item to use exactly. We can say use one of these items. E.g. a good data model is an Ontology Term structure, you have to decide which one to use

| | |
|---|---|
| | - TB: What are the bounds of these entities?<br>- MB: Not concerned with this. |
| 4. | GA4GH Connect Meeting Plans<br>- Is there any pre-work we'd like Drivers to do?<br>- Review of Breakout sessions<br>- Work Stream and Driver Project Updates |
| | A.O.B. |

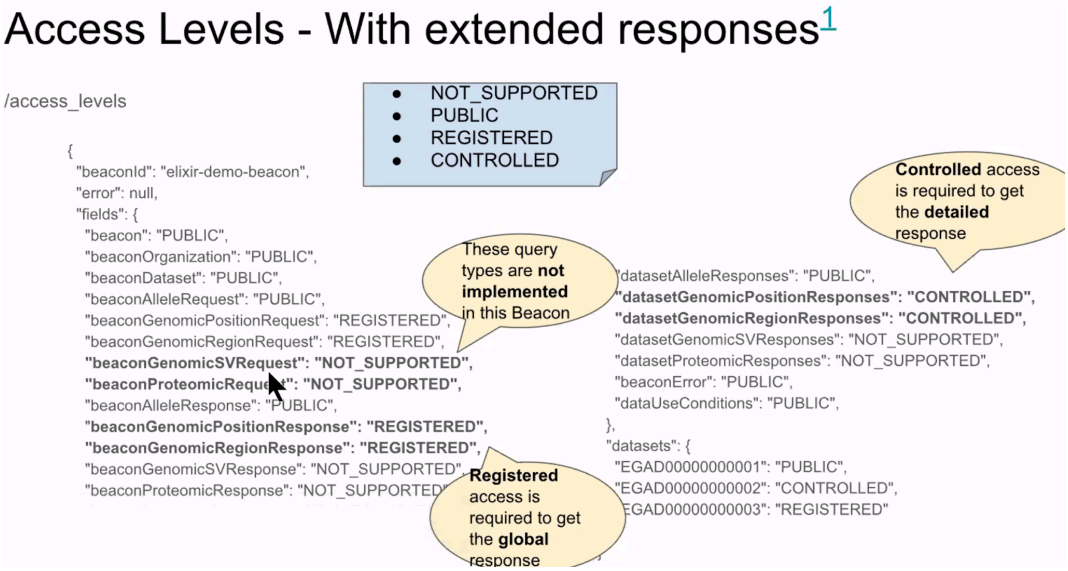
## 2019-02-27: Schema Blocks Governance

Chair: Marc Fiume
Attendees "Name (Affiliation)":
Jordi Rambla (EGA/ELIXIR Beacon/CRG),
Lindsay Smith (GA4GH)
Ben Hutton (Sanger. Search API)
Jonathan Dursi (CanDIG; Hospital for Sick Children)
Steve Laurie (CNAG, Barcelona ; SolveRD)
Miro Cupak (DNAstack)
Alexander Senf (EMBL-EBI)
Ian Fore (National Cancer Institute, NIH, US)
Kris Ganjam (Microsoft)
Orion Buske (Gene42, PhenoTips)
Tony Håndstad (Oslo University Hospital)
Julia Foreman (Sanger, DECIPHER)
Arun Ramani (Hospital for Sick Children, PhenomeCentral)
Melanie Courtot (EMBL-EBI, first half of the call only)
Tony Brookes (University of Leicester)
David Lloyd
Alexander Senf
Andrew P
Junjun
Francis Jeanson
Apporva
Sola

| | Actions Arising | Assigned To | Deadline |
|---|---|---|---|
| 1 | Melanie/Tony/Marc to discuss RI ontology at next DURI call (tonight) | Melanie/Tony/Marc | Discussed |

Global Alliance
for Genomics & Health

| | | | |
|---|---|---|---|
| 2 | Establish a regular Search call | Ben/Rishi/Marc | Setup |

| | Agenda Item |
|---|---|
| 1. | Welcome & Confirm Agenda<br>Product meetings for the sub-groups<br>- Beacon every two weeks strategy<br>- Discovery Search probably ready to resume<br>- Schema blocks is under development. In an incubation phase, calls will be managed by Melissa Konopko when they are ready to do so<br>- Discovery Networks meets once a month |
| 2. | Updates and Plans from Sub Groups (Update, next milestone, issues)<br>**Beacon API - Jordi**<br>- Finishing Beacon 1.1 Going for Beacon 2.0. Want to be backwards compatible<br><br><br>- Idea is to have more diverse request with clearer parameters to define scope of what people can implement<br>- /access_level endpoints<br>- Fields - type of requests beacon can expose, with support level<br>- Datasets - exposure level<br>- Allows spec to grow while being flexible<br>- Access level essential for 2.0<br>- ELIXIR has data outside of human data, e.g. plants, motivating flexible structure<br>Ian Fore: Who would I contact for access control? Can the model hold this information and does it **connect to the DUO model**?<br>- JR: Current Beacon spec uses previous versions of DUO, trying to include it is a work in progress. |

- More than one handover action can be in place,
- TB: DUO is about type about use, not standards relating to classes of access. Researcher Identity does that, but may need its own ontology to complete the goal
    - MC: RI Ontology base can be developed once elements of RI are in place.
        - Will discuss on next RI call
- Francis: Researcher Identity group is working on mapping user info, which will look at the mechanism for terms of access. Should be used in conjunction with Beacon and other projects
- JR: Also have Beacon Network, presenting on Discovery Networks call tomorrow

**Discovery Search API - Ben**

- MF: Continued to solicit use cases. Folder with use cases collected so far.
- Two broad use cases
    - **Search for data** - working with Melanie for Data Use aware search
    - **Search for subjects** - MME as a basis for this
- Writing implementation allowing for arbitrary connectors

Miro: Working on Proof of Concept for the Hinxton Implementation meeting

- Building on a federated data sub-level search on top of DRS, using EGA Biosamples data
- Demo developing using distributed query system tool called Presto
    - Implementing connectors for DRS and EGA/Biosamples
    - MC: Please contact me/rishi to join in building connectors
- TB: Is that a description of what search is now doing?
    - MC: The Search API is in the repo and distributed search being implemented is a superset of what Search does at the moment. WIll create a mapping between Presto and the Search
- TB: Are we getting use cases from Driver Projects
    - MF: Yes, these are being collected here at the folder. We also look at the items generated by other work streams.
    - MF: What we are doing is a framework as a standard for discovering GA4GH data models. It is a lower level API, can be used to drive diverse applications.
- BH: This implementation is a proof of concept of the API, not the 'Search API tool' - it's a standard that people can use in their software.

TB: How is this different from the first version attempted last year?

- MF: The feedback from Basel was the Drivers wanted to be more involved, demonstrated cases was too narrow, so we are aiming to show a >2 use cases from Driver Projects demo

Ian Fore: Can you clarify 'Search would enable Discovery of the model'? Ie. can you formulate a query

- MF: As conceptualised
    - one endpoint like /components allowing reflection of the model
    - One endpoint /search to compose rough boolean logic to mix/match data models in

a search query
- IF: Are both of those implemented at this stage?
- MF: No. This is our call to action. We (DNAStack, Isuru from Melanies team for DUO) are implementing some Presto connectors we are inviting the community to develop more.

AP: Is there work going on about connecting nodes to allow GA4GH to share data without sharing patient info etc.
- MF: Search probably not going to be prescriptive about what you expose at the endpoint
  - There are MME, Beacon groups though looking at how you serve a specific use case by bounding items
  - No implementation will be the de facto one
  - Combine search and networks

IF: Other GA4GH APIs are not inhently going to be queryable. Will they, through their own APIs, be accessible. The situation where Search seems powerful is when the models are more dynamic
- BH: From MME experience there was disparity in all the individual databases, even though trying to do same overarching goals. But it makes sense to accommodate them trying to add something like DUO or other specifications as they are developed by GA4GH
- IF: Whether a model is static or dynamic, I think it is worth defining them the same way, and made queryable.
- BH: That is exactly the end goal in how we are doing this
- TB: By **dynamic do you mean data model or data values are changing**?
  - IF: Data model e.g. NCBI dbGap study has a table of attributes that varies between studies. To my mind if structure of table is different, it is a data model
- BH: Need to be able to federate a query
- TB: Will you need extensive set of components
  - MF: Yes, so API needs to be flexible enough
    - Governance needs to also be flexible - some GA4GH ones, and some NCBI for instance that can be generated as required
- BH: Been working on governance mode

MF: GA4GH Data Models will have their own searchability. As we progress we would hope people move to DRS/Search. We are looking to this as a framework for items such as Clouds DRS project - esserntially a wrapper around a URL. Also allowing for combination on the different data models in the query syntax.
- TB: Worry about too much diversity for components e.g. multiple representations for subjects. How do we help consensus emerge
  - MF: Two components here - the query language and the components
    - GA4GH has a process around schema blocks in development
    - Don't want Search to take too big a scope
  - TB: Start simple, define some limited number of component types first, and see who wants to put a proposal for each of those

| | |
|---|---|
| | - BH: Governance model in development allows Driver Projects to go to a Work Stream which then gets added to a central location<br>- BH: Initial use case was MatchMaker v2 and we've being collecting more Please contribute further use cases.<br><br>BH: Apart from governance been working on **JSON Schema tooling**.<br>- Governance allows people to develop items at their own pace, based on GitHub, makes expectations clear.<br>- Long document: https://docs.google.com/document/d/1cgU7Vp44VyFMxiI_nteh-jLjpoS0RUNfOoBSNN5ueZA/edit?usp=sharing<br>-<br>- Workflow diagram: https://drive.google.com/file/d/1qKILYS2KztNYqE8OEjYfNunnIhvAXgXT/view?usp=sharing<br><br>Reaching a consensus on GA4GH wide data models.<br><br>Please contact me or Marc for feedback on these documents.<br><br>- For schema blocks register issues on schema blocks repository, prefer over Google Doc. High level discussions going on at the moment so can't confirm this at the moment |
| 3. | |
| 4. | |
| | A.O.B.<br>- Future meeting timings / Rishi<br>- ISMB attendees |

Global Alliance
for Genomics & Health

# 2019-01-30: Discovery relaunch

Chair: Marc Fiume / Michael Baudis

Attendees “Name (Affiliation)”:

Ayad (Optum) - Beacon/MatchMaker nodes at United Health
- Personally leading a team with Cloud Workflow, Beacon, DUO/DURI involvement
- Interested in Discovery with related to Data Use

Jordi Rambla (EGA - CRG -ELIXIR Beacon Project ES) >
- Beacon, Networks, Registry, Search
- See us also being heavy users of SchemaBlocks

Apoorva, Kyle Fahr (Illumina) - Caselog - Illumina implementation of Discovery Search API
- end goal is to put data nodes (esp Rare Disease) on Discovery Networks
- Apporva: Also in ClinPheno
- Joerg Hakenberg - worked on Discovery Search API, want to look at clinical use cases

Michael Baudis - co-lead of Work Stream and Beacon Project
- Schema blocks

Rishi Nag - GA4GH Programme Manager

Andrew P (SunBay)

Ben Hutton (Sanger - DECIPHER Project, MME API, Search API co-lead)

Jonathan Dursi (CanDIG, CINECA)

Colin Veal - University of Leicester / Cafe Variome Discovery Project - Search

David Lloyd - ELIXIR - Beacon main point of interest

François Schiettecatte - Johns Hopkins, Discovery Search for VariantMatcher

Heidi Sofia - Programme Officer at NIH - trying to co-ordinate GA4GH activities in NIH

Ian Fore - NIH / National Cancer Institute
- Science Officer on BioCaddie project at UCSD now evolved into NIH Data Commons then Cancer Research Data Commons

Juha Törnroos (ELIXIR-FI) Beacon - also DURI, some Cloud.
- Mainly interested in Beacon and Search API

Melanie Courtot (EMBL-- Schema Blocks / DURI WS

Jules Jacobsen: QMUL / Phenopackets and schema blocks integration interest

Julia Foreman: DECIPHER Project Manager with Ben Hutton / Search API

Miro Cupak - co-lead Search and Network groups

Junjun Zhang - OICR ICGC-Argo Driver Project / Search API interest

Steve: CNAG in Barcelona, representing RD-Connect / SolveRD - European Consortium rare disease project

Tony Handstad (Oslo University Hospital): ELIXIR Beacon - Discovery Search API

Vagelis Ladas - University of Leicester - Rare Diseases, interested in Seach

Global Alliance
for Genomics & Health

| | Actions Arising | Assigned To | Deadline |
|---|---|---|---|
| 1 | Join mailing lists as required | All - contact Rishi to be added | |
| 2 | | | |

| | Agenda Item | Person/Time |
|---|---|---|
| 1. | Welcome and Introductions<br>Work Stream leads update | Marc/Michael 5 |
| 2. | Updates and Plans from Sub Groups (Update, next milestone, issues)<br>- slides from SC meeting<br>- Beacon API<br>- Discovery Search API<br>- Discovery Networks<br>  - spec for service-info endpoint, just started<br>  - service-registry building towards first release, needs syncing with ELIXIR Beacon Network<br>- Schema Blocks<br><br>*Additional Discussion*<br>Q. Is there a plan to join MM/Beacon and Search?<br>BH: The format of MM API is similar to the way Search API is structured<br>- Noone pushing for changes.<br>- Technical changes imposed a cost to update - too high a cost for some nodes. We therefore lost some<br>- MM will probably not change into Search API but could be built upon on the Search API. MM team considering building a new network<br>MF: I imagine Search as a Query language across GA4GH Data Models, and MM could be a constrained version of the Search API. Can use concepts like variant, subject in schema blocks.<br><br>Juha: To clarify - Search API is a REST API introducing a query language with plug-ins to build up the query<br>- BH: Broadly speaking yes. Look to break the MM format down to individual component schemas. Applications adding value on top of Search API will put together schema blocks in this manner<br>- JT: What use cases?<br>- BH: Initial use case was MM based, had a document with what was wrong defined. Beacon and MM have different levels of funding. Search defines a common framework to allow applications to add value at their own pace. | 30 |

| | | |
|---|---|---|
| | MF: Another use case we have is for 'Beacon for data' to identify affiliated with Data Repository Service. E.g. find me VCF file for individual who has this variant<br>- BH: Also can add DURI restrictions into this | |
| | A.O.B. | 5 |

## 2019-MM-DD: Template

Chair:
Attendees "Name (Affiliation)":

| | Actions Arising | Assigned To | Deadline |
|---|---|---|---|
| 1 | | | |
| 2 | | | |

| | Agenda Item | Person/Time |
|---|---|---|
| 1. | Welcome & Confirm Agenda | |
| 2. | Updates and Plans from Sub Groups (Update, next milestone, issues)<br>Beacon API:<br><br>Discovery Networks:<br><br>Schema Blocks:<br><br>Search API:<br><br>Case Discovery: | |
| 3. | | |
| | A.O.B. | |



Global Alliance
for Genomics & Health