HUMAN GENOMES PLATFORM PROJECT

Virtual Cohort Assembly

DISCOVERY PHASE REPORT

National Community Needs & Candidate Solutions

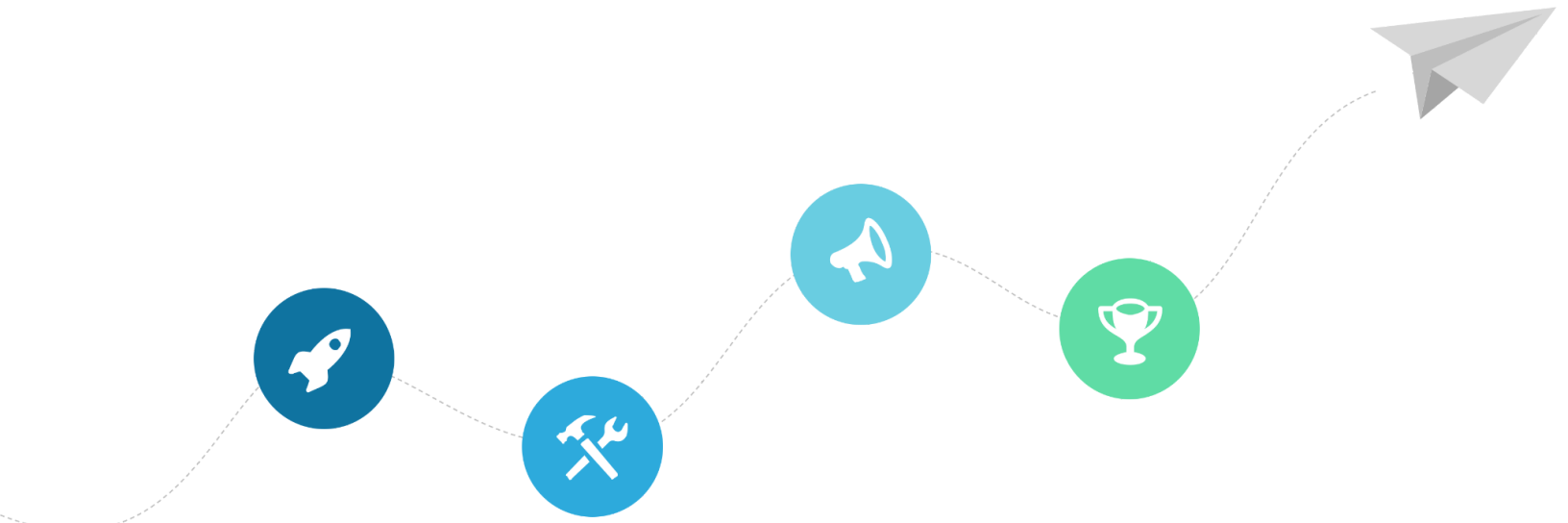

Table of Contents

Glossary

CCIA	Children's Cancer Institute
GA4GH	Global Alliance for Genomics and Health
HGPP	Human Genomes Platform Project
JCSMR	John Curtin School of Medical Research (ANU)
MVP	Minimum Viable Product
NCI	National Computational Infrastructure
QIMR Berghofer MRI	QIMR Berghofer Medical Research Institute
SNV	Single nucleotide variant
UMCCR	University of Melbourne Centre for Cancer Research
WGS	Whole genome sequencing
ZERO	Zero Childhood Cancer Program (led by the Children's Cancer Institute)

Acknowledgements

The HGPP formed part of Australian BioCommons' Human Genome Informatics initiative and was funded by NCRIS via the Australian Research Data Commons (https://doi.org/10.47486/PL032) and Bioplatforms Australia. Contributions were also made by partner organisations: Australian Access Federation, Garvan Institute for Medical Research, National Computational Infrastructure, QIMR Berghofer Medical Research Institute, The University of Melbourne Centre for Cancer Research, the ZERO Childhood Cancer Program and Children's Cancer Institute.

Authors

in alphabetical order by surname

Cowley, Mark - ZERO CCIA
Downton, Matthew - NCI
Holliday, Jessica - BioCommons
Kummerfeld, Sarah - Garvan
Leonard, Conrad- QIMRB
Lin, Angela - ZERO CCIA
Pope, Bernie - BioCommons
San Kho Lin, Victor - UMCCR
Ravishankar, Shyamsunder - Garvan
Shadbolt, Marion - BioCommons
Syed, Mustafa - ZERO CCIA
Taouk, Kamile - ZERO CCIA
Wong-Erasmus, Marie - ZERO CCI

1. Introduction

The Human Genomes Platform Project (HGPP) is a nationally-funded collaborative research project aiming to enhance capability for securely and responsibly sharing human genomics research data. National and international connectivity will maximise the utility of these sensitive and valuable assets. The partners on the project represent many of the largest human genome sequencing and analysis efforts in Australia.

Currently there is no way to identify virtual cohorts of individuals who have had their genomes sequenced nationally as it is not possible to query across the separate assets from each participating genomics repository. This work aims to implement a system that can be used to identify cohorts of individuals and related data assets across the repositories located at each of the partner institutes (i.e., UMCCR/Australian Genomics, QIMRB, ZERO/CCIA, Garvan and NCI; Figure 1).

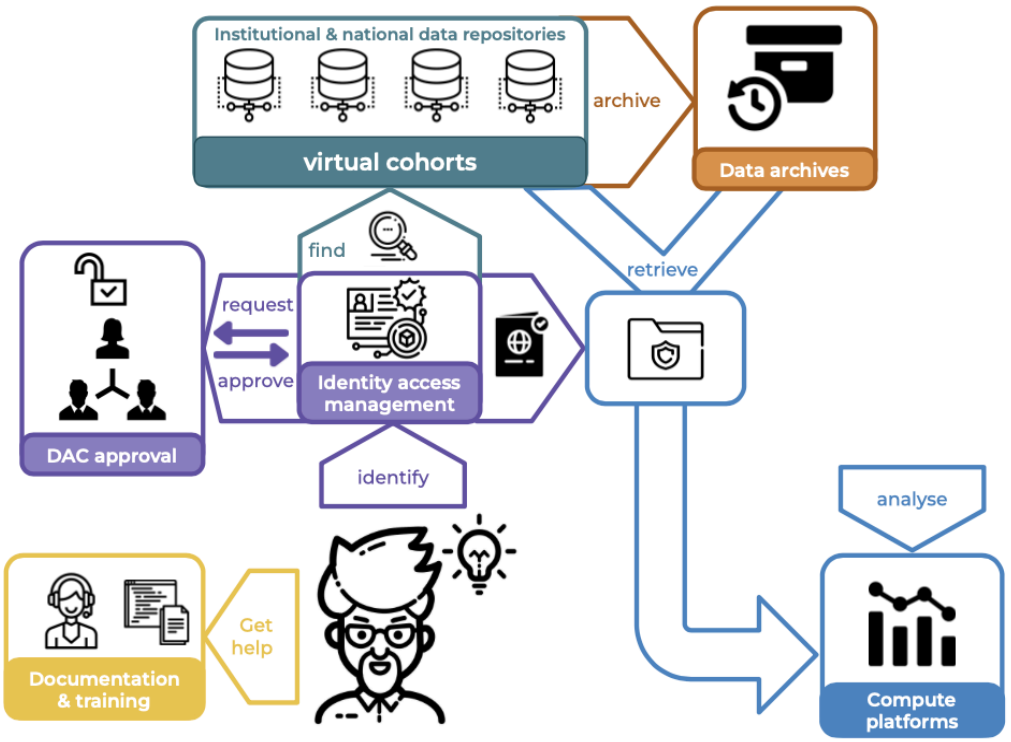

Figure 1. ***The HGPP infrastructure ecosystem from the perspective of the research end user showing the key elements of the human genomics data sharing toolbox and data and information flows.*** **(Drafted by Marion Shadbolt, Human genome data specialist, HGPP.[1])**

Concept. A user (e.g., researcher, clinician) facing interface that enables a national search across participating genomic repository data holdings by relevant demographic and clinically related metadata elements and return of information to the user (i.e., does the data exist and, if so, where the data is housed).

Impact goals: Researchers and clinicians can easily identify virtual cohorts of individuals of relevance to their research studies, across multiple repositories.

Infrastructure goals: A system is established, managed and made available to researchers/clinicians to enable cross-repository search, and identification of where data is held across participating repositories. Specifically, success is expected to look like:

1. A user-facing interface that allows queries to be defined
2. A set of API services at each repository that are responsive, highly available and integrated
3. A user-facing interface that allows return of aggregate query results and further information on how to gain access to the underlying genomic data.
4. An aggregator API that works to deliver aggregated query results in point 3.

Sub-project team members are the subject matter experts and represent multiple organisations:

- Australian BioCommons (BioCommons)
- ZERO Childhood Cancer Program (ZERO) led by Children's Cancer Institute (CCI) and Kids Cancer Centre at Sydney Children's Hospital, Randwick
- University of Melbourne Centre for Cancer Research (UMCCR)
- Garvan Institute of Medical Research (Garvan)
- Queensland Institute of Medical Research Berghofer Medical Research Institute (QIMRB)
- National Computational Infrastructure Australia (NCI).

The initial focus of the virtual cohorts sub-project within the HGPP was a knowledge discovery and recording phase to define:

- the current state of cross-institutional human genomic data querying in Australia
- the set of problems that need to be addressed
- key stakeholders and their (likely) requirements.

As such, this document records:

- the current state of processes and tools for virtual cohort querying
- national community needs
- candidate solutions to enable cross-institutional virtual cohort querying
- recommendations on preferred technology and proposed implementation architecture

This document will be used as a reference to plan the pilot for a system that addresses prioritised requirements to create a Minimum Viable Product (MVP). The primary audiences for this document include the HGPP sub-project team, other HGPP stakeholders, and the project reference group.

2. Current State Findings

Clinical and research genomics data has most value when it is discoverable. Absent is a centralised repository; this requires metadata from different providers to be made findable, searchable and shareable. At all times, of course, the data and metadata available must be restricted to only the level of access for which the user is authorised.

For an Australian genomics federation to be successful, widespread adoption of the new processes and systems are needed. To foster widespread adoption, we consulted experts across all partner institutes directly to comprehensively understand the current state and the future needs of the national human genome research community.

Two essential preconditions for findable, searchable and shareable data from federated sources:

1. Adoption by providers of a common data model (ontology) to annotate data from different sources. This is a prescription for the sharing layer only; the shared model will typically be different to each provider's internal data description and annotations for this purpose will be generated by mapping from internal data models. A key challenge in specifying the common model is to balance generality — allowing future refinements — with sufficiently concrete annotations to be useful even at early proof-of-concept stages.
2. Implementation of distributed search/query protocol or framework. Assuming a minimal common data model to begin with this can be quite lightweight. Key questions here include whether to implement existing technologies or to prototype a new solution; how to implement federated identity management; and how to specify and perform handoff to raw data locations and/or computational

Community Needs Analysis

Specific requirements of the partner organisations involved in the project were collected in the form of user stories, summarised in the next section of this document.

In addition to the specific requirements in the collected user stories, stakeholders described their needs in a more general fashion as follows.

Garvan:

- Establish a shared approach for cohort metadata that allows researchers to find relevant studies
- Develop a common approach to patient-level phenotypic/clinical information
- Implement a working prototype that allows the above metadata and patient-level data to be queried consistently across cohorts managed by different institutions

CCI / ZERO:

- Share ZERO's molecular/genomics data with the research community via a live data commons

- Develop/establish a virtual cohorts platform for discovering data at multiple participating sites
- Harmonise data models to facilitate cross-institute querying

UMCCR / Australian Genomics (AG):

- Share AG cohort data via a Beacon v2 endpoint (locked down with limited information is acceptable)
- Be able to query other Beacon endpoints for related information (again, just 'open' Beacons are acceptable; preference would be to also support user authentication at Beacon V2 endpoints)
- Have (some) cohorts with a subset of Beacon v2 metadata available in a warehouse for interactive exploration, with a preference towards Gen3

QIMR Berghofer:

- Establish common use-cases and user stories for researchers and clinicians wanting to query across datasets.
- Guided by use-cases, establish a common data model across partner organisations for querying across datasets and creating virtual cohorts. It is to be expected that to begin with, this data model may be quite minimal but the underlying framework must be flexible enough to grow and change.
- Establish a common mechanism across partner organisations to perform federated query/virtual cohort discovery
- Host a local instance of federated query/virtual cohort discovery service node, backed by public data from QIMR Berghofer

Summary

All partners expressed a desire to share their data with others, and to be able to query others' data at both a cohort and patient/phenotype level. Implicit in this is an agreement to adopt a minimal shared data model. Project partners agree that the discovery/query service should interface with the Federated IAM sub-project service to provide appropriate access to controlled data. Integration with the DAC sub-project service for data request handoff is desirable.

Project partners agree that a working prototype should be deployed within the life of this project.

There was some early support for establishing data sharing technologies that would both deliver upon the goals of the HGPP project, and also enable some sites to establish a live genomic data commons. However, technology assessment performed early in the project indicated that no existing technologies (in particular Gen3, see below) could achieve both of these goals, in particular the need to establish virtual cohorts, so standing up genomic data commons at each site was deprioritised in favour of delivering the agreed to HGPP objectives.

3. User Stories

Representatives from project partner organisations, with experience working in relevant roles, contributed role-based user stories to describe their virtual cohort assembly needs. A ranked list of the highest priority user stories can be found in Table 1 below.

Common criteria required by project partners to assemble virtual cohorts of interest include:

- Data types (e.g., sequencing, primary, secondary, WGS)
- Sample types (e.g., Blood, Fresh Frozen tumour biopsies and additional sample types)
- Specific phenotypes, such as cancer diagnosis, or a patient's genetic disorder
- Specific mutations, such as SNVs, indels, and gene fusions
- Consent, data location, data access, and data use requirements
- Clinical metadata, such as survival time, age, follow-up, disease status
- Cohort level data, such as number of patients of a certain type

Table 1. Top six user stories ordered by priority for each project partner institute. The full list of user stories can be viewed here.

Rank	QIMR Berghofer MRI	Garvan	NCI (JCSMR)	UMCCR	ZERO
U.S.1	***As a research user:*** I want to know who holds sequencing data for PDAC cases ***So that:*** We can build a virtual cohort of cases for discovery	***As a research user:*** I want to identify all individuals with a particular set of clinical characteristics and obtain primary data ***So that:*** We can build a virtual cohort of cases for discovery	***As a research user:*** I have an interest in research topic X. What datasets have the required consents for me to use to address research topic X? ***So that:*** We can build a virtual cohort of cases for discovery	***As a curator:*** I want to find variant information for cancer samples of a given subtype ***So that:*** We can assess a novel variant	***As a research user:*** I want to find all medulloblastoma samples, get access and download the data ***So that:*** We can utilise them for research
U.S.2	***As a research user:*** I want to know who holds sequencing data for PDAC cases, from fresh-frozen tissue ***So that:*** We can build a virtual cohort of cases for discovery	***As a research user:*** I want to identify all individuals with a particular set of variants and/or clinical characteristics and obtain primary data ***So that:*** We can build a virtual cohort of cases for discovery	***As a research user:*** I want to know what restrictions I have on the use of data? ***So that:*** We can build a virtual cohort of cases for discovery	***As a research user:*** I want to find primary / read level data for published cancer cohorts stored in Australia ***So that:*** So I can re-process / harmonise data	***As a research user:*** I don't have access to large storage - where can I run my analysis on your samples ***So that:*** I can perform my analyses in the virtual cohort
U.S.3	***As a research user:*** I want to know who holds sequencing data for PDAC cases, from fresh-frozen tissue, with survival timepoints ***So that:*** We can build a virtual cohort of cases for discovery	***As a research user:*** I want to run analyses on my virtual cohorts in situ (i.e. bringing compute to the data) ***So that:*** We can analyse the data in the virtual cohort	***As a research user:*** Can I download the data and share it with my collaborators in Australia and/or overseas? ***So that:*** We can establish allowed uses of the data	***As a research user:*** I want to find primary / read level data for published cancer cohorts stored in Australia of a given phenotype / with minimal metadata requirements ***So that:*** So I can re-process / harmonise data	***As a research user:*** How can I find all paediatric samples? (age < 21 yrs) ***So that:*** I can consolidate my data with yours
U.S.4	***As a research user:*** I want to know how frequently a particular germline variant occurs in cases of healthy normal/never diagnosed ***So that:*** We can better understand variant distribution in the Australian population	***As a research user:*** I want to share data and analyses on my virtual cohorts in situ (i.e. bringing compute to the data) ***So that:*** We can analyse the data in the virtual cohort	***As a research user:*** Where can I perform computation on data once I have identified all required samples to comply with DAC requirements? ***So that:*** We can analyse the data in the virtual cohort	***As a research user:*** I want to find primary / read level data for published cancer cohorts stored in Australia of a given phenotype / with minimal metadata requirements and with data access control requirements matching my research plan ***So that:*** So I can re-process / harmonise data	***As a research user:*** I want to be able to access metadata for various cohorts and studies ***So that:*** I know how to normalise my data with the virtual cohort(s)

Rank	QIMR Berghofer MRI	Garvan	NCI (JCSMR)	UMCCR	ZERO
U.S.5	***As a clinician researcher user:*** I want to know who holds clinical data including treatment regime and survival timepoints, for PDAC cases with KRAS G12D mutation ***So that:*** We can build a virtual cohort of cases for analysis	***As a research user:*** *I want to identify samples with a particular set of clinical characteristics and/or variants that have available tissue for follow up studies* ***So that:*** *We can perform follow up research*	***As a research user:*** [How] can I reconnect with participants for follow up sample, additional information, or return results of incidental findings? ***So that:*** We can perform follow-up research and potentially return results	***As a research user:*** I want to *share* primary / read level and secondary / variant level data for our own research cohorts alongside agreed-upon phenotype and minimal metadata annotation restricted by DUO codes ***So that:*** others can use our data	***As a clinician researcher:*** I want to build a cohort where after last follow-up the patient has stable disease ***So that:*** I can build a virtual cohort for survival analysis
U.S.6	***As a data custodian:*** I want to limit which users can view which information — e.g. public access for catalogue type data (what do we hold) plus possibly somatic variants ***So that:*** Access to data is restricted or exposed as appropriate				***As a research user:*** I want to identify all Neuroblastoma patients with ALK fusions, with their disease status at most recent follow up ***So that:*** I can determine the prognostic impact of this driver mutation (SNPs and AMP are well established biomarkers in this disease).

4. Candidate Solutions

The GA4GH Computable Cohorts working group performed a landscape survey of relevant technologies in this space, which the virtual cohorts sub-project team reproduced with minor edits for clarity in Appendix A.

Described below are some of these in more detail, with a focus on Data Models and API/Registry technologies. Although Gen3 is not covered in that survey, it is included here for detailed analysis as several partner organisations have deployed instances and are interested in exploring its native suitability as a Virtual Cohorts solution.

CINECA — Gecko Data Model

The Common Infrastructure for National Cohorts in Europe, Canada and Africa (CINECA) project sought to construct a data model that supports data discovery and analysis across 10 cohorts from three continents. In the CINECA Cohort Representation work package (WP3), a minimal metadata model called the Genomics Cohorts Knowledge Ontology (GECKO)[1] was created.

The method used to create the model was to collect existing cohort models and publicly available data dictionaries. The Maelstrom Research data harmonisation methodology was used to categorise variables and find overlaps between the models. Use cases for querying were compiled from project partners to ensure these could be met by the model.

The minimal metadata set was ontologised with existing terms with the majority from the National Cancer Institute Thesaurus ontology (NCIT). The fully ontologised model is available through EBI OLS and their github.

Separately from the CINECA Cohort Representation work package (WP3), it is noted that the CINECA Federated Data Discovery and Querying work package (WP1) implementation is envisioned as a Beacon network for data discovery.

GA4GH — Data Connect API

The Data Connect API developed within the GA4GH is a standard that seeks to provide a way to describe data and its model, as well as providing a method for querying this data. Its strength lies in the fact it does not prescribe a particular model or structure to the data, though it does recommend the use of GA4GH SchemaBlocks. It can essentially work with any data that can be transformed into an array of JSON objects. A detailed look at the use cases it is aiming to support can be found on github.

This technology is a middleware specification that abstracts common infrastructure for building searchable, federated networks. As such it generalises the approach taken by beacon (see below), and neither defines nor assumes any particular data model. Example implementations are given for various data models and backends, but no reference implementation — which makes sense given its nature as a generic interchange specification.

[1] Jin, Vivian, & Brinkman, Fiona. (2020). CINECA Cohort Level metadata Representation D3.1. Zenodo. https://doi.org/10.5281/zenodo.4575460

GA4GH — Beacon v2

Beacon v2[2] provides a framework and a default data model for enabling discovery of collections, cohorts, datasets, genomic variations, individuals, biosamples and analyses. The framework defines the rules around setting up an API to perform the expected beacon queries. An implementation can be verified by the approval service registry. The default data model is specified as a series of JSON schemas that describe each object type. While the scope of the default model is quite broad, the number of required fields is limited, allowing users to decide how much of the standard to meet. In addition, schemas are extensible with additional properties allowed to be added to the default schemas (Figure 2).

A challenge that is likely to confront users of the default Beacon v2 model is the extensive use of ontologies. While specific ontologies are not generally enforced, suggestions are made for appropriate ontologies and querying relies on terms being curated with consistent ontologies. Depending on the state of the original metadata, this curation process may need to be performed manually or assisted with automated curation tools (e.g., ZOOMA).

Implementing Beacon can be lightweight. It can be a standalone web service hosting metadata in a database backend such as MongoDB or RDBMS, or even flat files. Deploying Beacon instances as cloud-native serverless is another approach. The CSIRO's Transformational Bioinformatics team implemented serverless Beacon v1[3] as a service. They are currently working to develop a version 2 implementation of this service. QIMR Berghofer has also previously implemented and deployed a standalone beacon v1 service in the cloud [private repo, available on request].

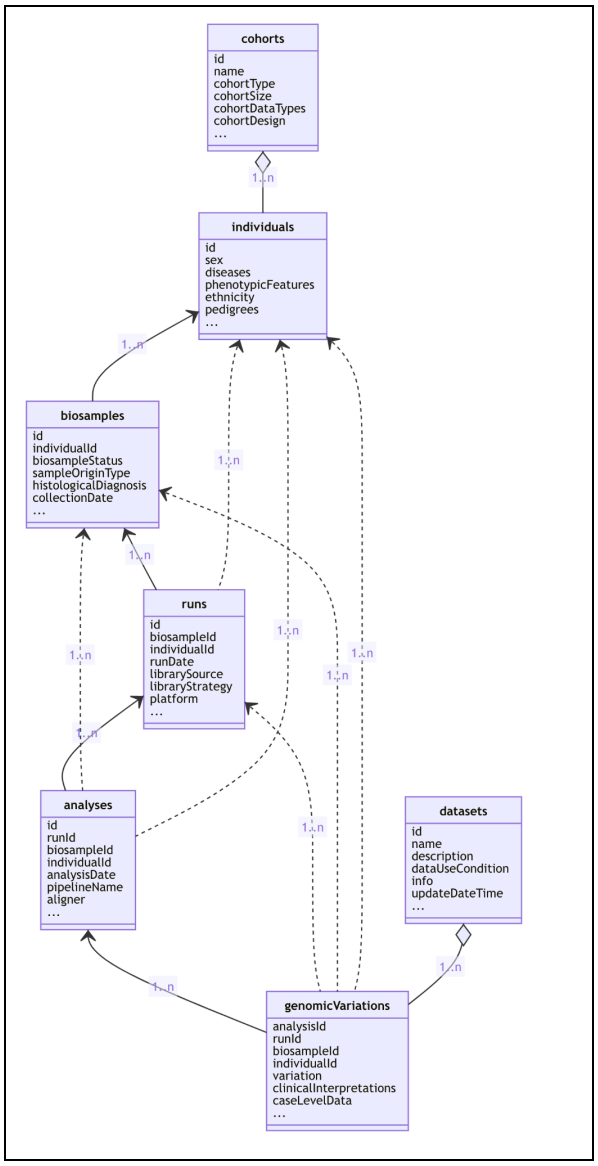

Figure 2. ***Beacon graphical data model from https://docs.genomebeacons.org/models/#introduction***

[2] Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L. A., Navarro, A., Paloots, R., Rueda, M., Saunders, G., Singh, B., Spalding, J. D., Törnroos, J., Vasallo, C., Veal, C. D., & Brookes, A. J. (2022). Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond. Human Mutation, 43(6), 791–799. https://doi.org/10.1002/humu.24369

[3] Serverless beacon – a cloud-native genomic data exchange platform (ABACBS Talk by Yatish Jain)

It is noted that the default Beacon v2 data model does not address all concept domains of interest to all partner organisations; for example structural variants and gene abundance outliers are not currently modelled. The Beacon v2 framework is, however, explicitly designed to be data model agnostic and extensions to the model are expected not to require major changes in the framework (Table 2).

Table 2. Summary of the *pros and cons of Beacon v2 as a solution to virtual cohort querying*

Pros	*Cons*
Allows querying/discovery across multiple data sources without compromising confidential data	No native Data Access Committee tool integration
Lightweight	No native authentication layer
Loosely coupled services	Access control at the level of project rather than individual
GA4GH standard	Beacon Network has rudimentary UI
Flexible implementation options	Does not support direct access to genomic data files
Supports aggregated queries through a Beacon network	Highly nested schema structure, sometimes difficult to understand
Data model is adaptable and extensible	
Supports sharing of SNV and indel variants natively, with extensions to support copy number variants[4]	

GA4GH — Phenopackets

Phenopackets are a GA4GH standard that aims to describe all aspects of clinical metadata for a given study participant to enable standardised 'deep phenotyping'[5]. A domain analysis was conducted to review the requirements and use cases aiming to be met by phenopackets. The v2 schemas are highly normalised describing a data model with approximately 55 classes split across 7 class categories (see Figure 3 below). Fields may be optional, recommended or required with multiplicity of linking between objects also prescribed. Specific ontologies for particular fields are not mandated, so there is flexibility to use ontologies appropriate to particular domains, though there is a set of recommended ontologies for several fields. Phenopackets are generated in Protobuf and can then be exported to a number of formats including JSON.

[4] The Progenetix Beacon + data model: https://progenetix.org/beaconPlus

[5] Jacobsen, J.O.B., Baudis, M., Baynam, G.S. et al. The GA4GH Phenopacket schema defines a computable representation of clinical data. Nat Biotechnol 40, 817–820 (2022). https://doi.org/10.1038/s41587-022-01357-4.

There are java and python libraries to support their creation and validation. Phenopackets aim to be interoperable with the FHIR standard with an implementation guide available.

As a schema/data model specification, Phenopackets could be implemented in parallel with an API/registry data discovery technology e.g. Beacon, with Phenopacket schema serving as the common data model.

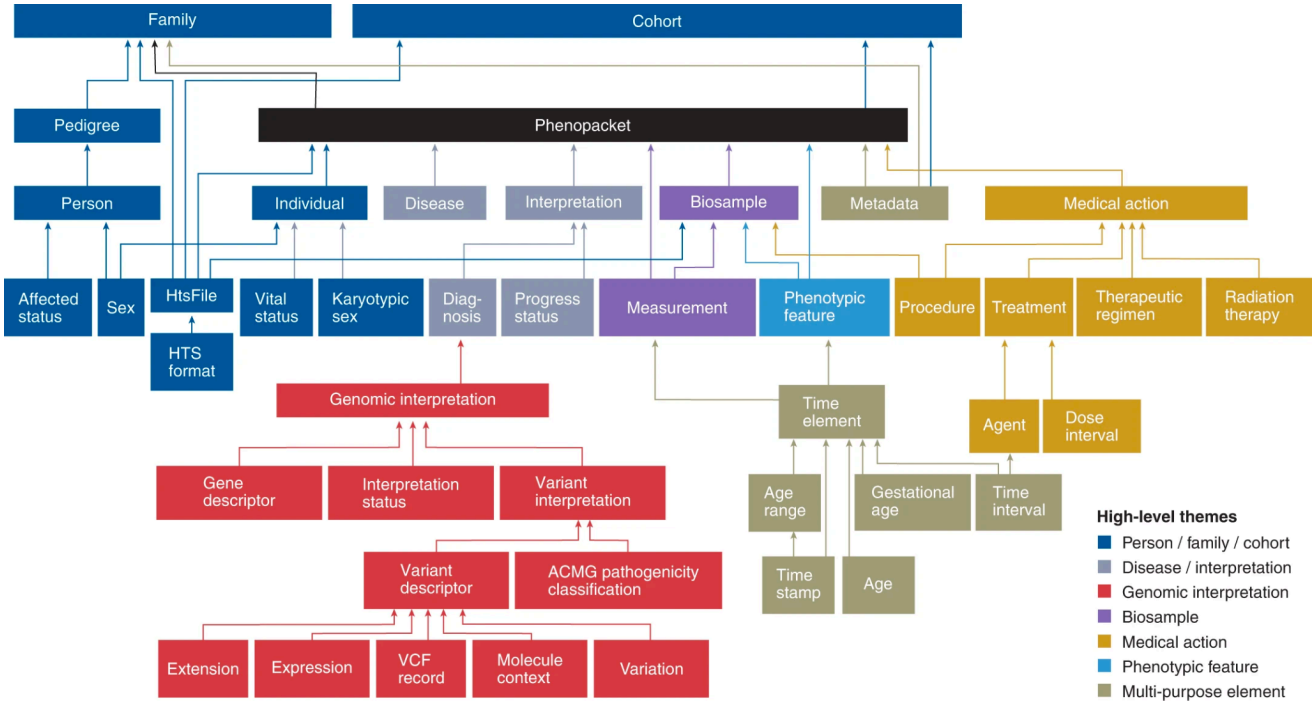

Figure 3: ***Diagrammatic representation of phenopackets schema from Jacobsen et al. 2022.***

Gen3

Gen3 is a data platform system for building data commons and data ecosystems. It comes with a data submission framework for capturing metadata and genomics data files.

For those participating institutes with a Gen3 instance, one option could be to come up with a standard template for a clinical data dictionary that would suit many use cases, but also allow extension to specific use cases. Having a core set of required variables, shared ontologies and common structure would allow greater cross-Gen3 queries, and also allow common tools to be developed and usable across instances, such as EGA submission, Phenopacket creation and a common Beacon mapping abstraction layer. None of the off-the-shelf data models available in Gen3 satisfied the project partners, so we investigated mapping alternative data models onto the Gen3 data dictionary (Table 3). These included beacon-v2 (see below) or Phenopacket model as a template for clinical metadata, plus INSDC model for sequencing/data template.

Table 3. Summary of pros and cons of a Gen3 solution

Pros	*Cons*
Handles authn/z via microservices	Manual entry of groups and users
All microservices centralised in single kubernetes cluster	Components are tightly coupled
Can manage permissions and access to projects at a group + individual level	Complex infrastructure
Provides a comprehensive solution for a live data commons including data portal for UI based exploration and data access	Cannot currently query across multiple instances — this appears to be a blocker for the purposes of this subproject
	Restricted to AWS and requires custom engineering for other cloud platforms
	Overweight solution for the terms of reference of this subproject

Gen3/Beacon Hybrid

Gen3 is flexible with respect to which data dictionary you use, but it is clear that if two sites wanted to share data, that they would need to share the same data dictionary, or undertake a data harmonisation and mapping process which is outside the scope of this sub-project. To this end, we investigated a Gen3/beacon hybrid, potentially using a subset of the Beacon v2 model (Table 4).

Table 4. Summary of pros and cons of a Gen3/Beacon hybrid solution

Pros	*Cons*
Can utilise a subset of Beacon v2 data model in Gen3 data dictionary	Novel proposal, so will require a lot of custom engineering to reach a solution
Will facilitate cohort discovery and cross-instance querying	Required mapping of Beacon v2 model into Gen3 data dictionary, which was difficult to achieve (see below)

Pros	*Cons*
Eliminate maintaining two separate system instances of similar metadata (i.e. as data are submitted into the Data Store (Gen3); the metadata capturing closely coupled and aligned with Data Discovery (Beacon) service requirement)	Still required to implement Beacon API within Gen3 platform
	Basic terminology definition support in Gen3 data modelling; whereas Beacon v2 model require extensive ontology mapping

Mapping between Gen3 data dictionary and Beacon v2 data model

To investigate the feasibility of a Gen3/Beacon hybrid solution, we undertook an extensive mapping exercise to map the Beacon v2 data model onto a Gen3-compatible data dictionary. The result of this exercise is available interactively[6], and shown in Figure 4 (next page). Despite the relatively lightweight Beacon v2 data model, the mapping into Gen3 was complex. The main challenge was the modular nature of the Beacon model and the degree of nesting within their schemas. In contrast, schemas specified for a Gen3 data dictionary cannot be nested, they need to be linked into a graph-like structure, which quickly becomes unwieldy with the number of interconnected nodes involved in the Beacon v2 model. The complexity of this data dictionary would have further complicated the task of ingesting data into a Gen3 system using such a complex data model. Furthermore, difficulties were noted in the way that genomic variants could be represented in Gen3. We conclude that the mapping between Beacon v2 data model and a Gen3 data dictionary is not straightforward, but may be addressed with a more focussed effort.

[6] By cloning and deploying the the code in this github repo https://github.com/umccr/umccr-dictionary and specifying dd=beacon

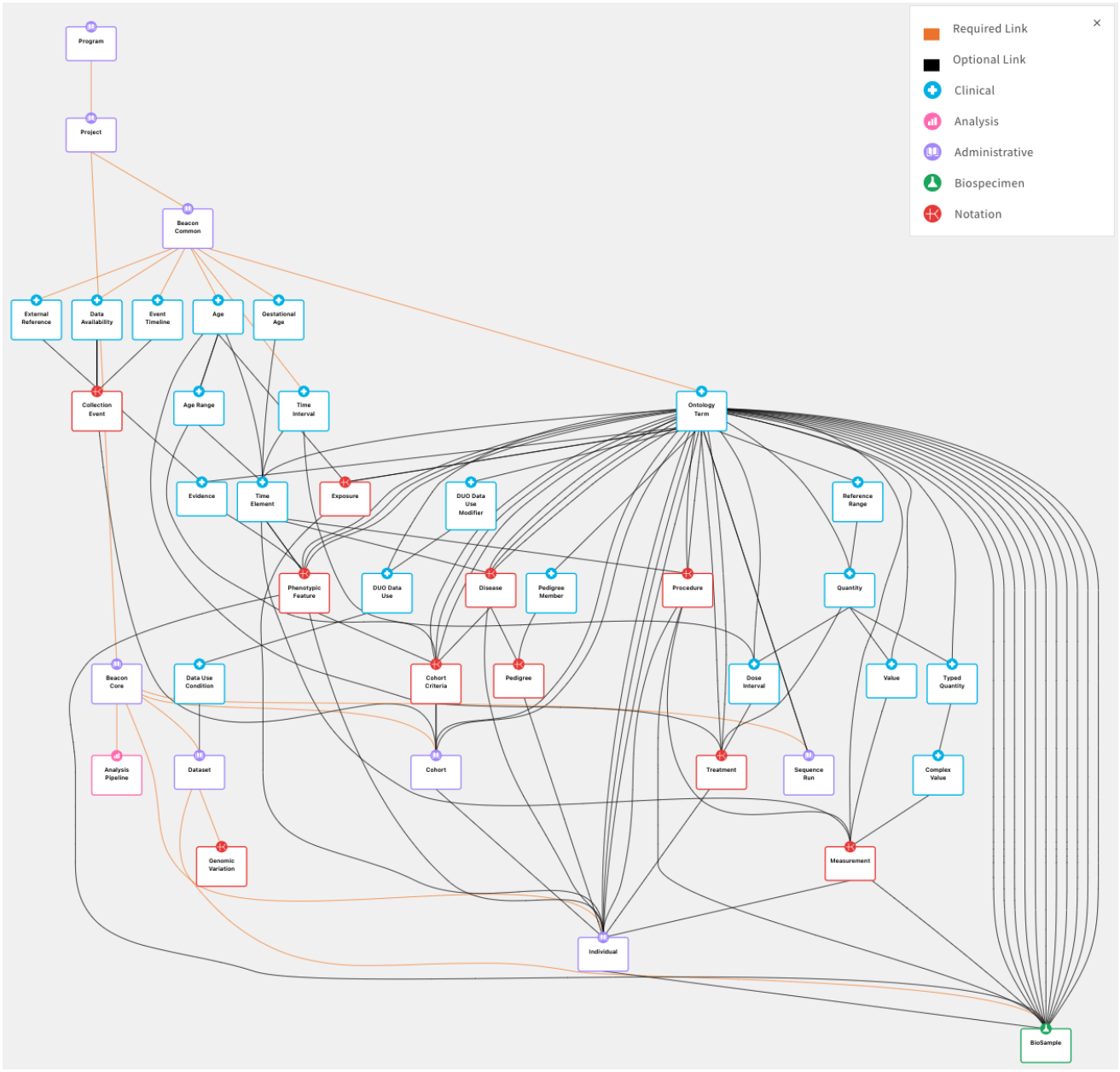

Figure 4. ***Mapping of Beacon v2 phenotype/clinical models into Gen3 data dictionary – source available at https://github.com/umccr/umccr-dictionary***

OMOP CDM

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)[7] is an open community data standard, designed to standardise the structure and content of observational data and to enable efficient analyses that can produce reliable evidence. The OMOP CDM is rich, complex and has been successfully used for initiatives that seek to extract data from medical records across a range of health domains, including a

[7] https://ohdsi.github.io/CommonDataModel/

large-scale pan-USA covid response dataset. Recent extensions to the OMOP CDM have been established for oncology[8]. This is added here for completeness, as this model was largely not known to the implementation group at the outset of the project, who were primarily focussed on technologies for sharing data, rather than modelling the data and as such has not been investigated in depth. Its popularity and large community of support suggest it should be considered an alternative. Efforts to map data from OMOP CDM to the Beacon v2[9] and Phenopackets[10] data models are currently underway with initial proof of principle solutions available[9,10].

A new data model and/or query framework

Rather than adopting an existing data model and/or query framework, one option briefly canvassed was to develop our own. This proposal was rejected as being needless reinvention. Adoption, adaptation or extension of existing technologies was agreed to be a more efficient approach. It is essential that whatever existing technologies we agree upon be fit for purpose and flexible enough to handle extensions and modifications to suit the diverse requirements of the project partners.

5. Recommendations on Preferred Approaches and Technologies

Common Data Model

The virtual cohort team investigated various Data Models (Section 4 — Candidate Solutions) and agreed that Beacon v2 Model fits for Minimal Metadata data harmonisation needs, and for the purposes of proof-of-concept the template data model will be the Beacon v2 default data model. This is comprehensive enough to support real-world use-cases and can be extended as needed to support growth in requirements.

The Beacon v2 Default Model comprises seven base entities including **Individual** and **Cohort** models. Within these entities, it reuses existing ontologies and GA4GH specifications such as GA4GH Phenopacket and GA4GH VRS (Variation Representation Specification) for both phenotypic and genotypic meta information properties that is required to cater for federated genomic data discovery.

Adopting the more comprehensive CINECA GECKO or Phenopackets data models, or elements therefrom may be considered post-proof of concept stage.

Data Sharing/Query Layer

The virtual cohort team recommends the adoption of the Beacon v2 data sharing/query layer. For example, the EGA-Archive implementation.

There are broadly two approaches for deployment:

[8] https://ohdsi.github.io/CommonDataModel/oncology.html
[9] https://github.com/elixir-luxembourg/BH2021-beacon-2.x-omop
[10] https://github.com/phenopackets/omop-exporter

1. Standalone Beacon Instance
2. Implementing inside an aggregated network

Below we describe a beacon network architecture that includes both sorts of nodes.

In what follows we use the following terms:

b_i - Beacon instance data

dm_i - data store data model

dm_c - common data model

ds_i - data store

POC = proof of concept
common data model = the agreed-upon data model used by the virtual cohort discovery/federated query service. This may not be the same data model used internally by partner organisations.

POC common data model = data model used during proof of concept/demonstrator technology deployment phase

Production common data model = the common data model agreed upon by the end of the subproject. It should represent sufficient concepts to satisfy at least the priority use-cases from each partner org.

Theoretical Architecture for Beacon-enabled Virtual Cohort querying

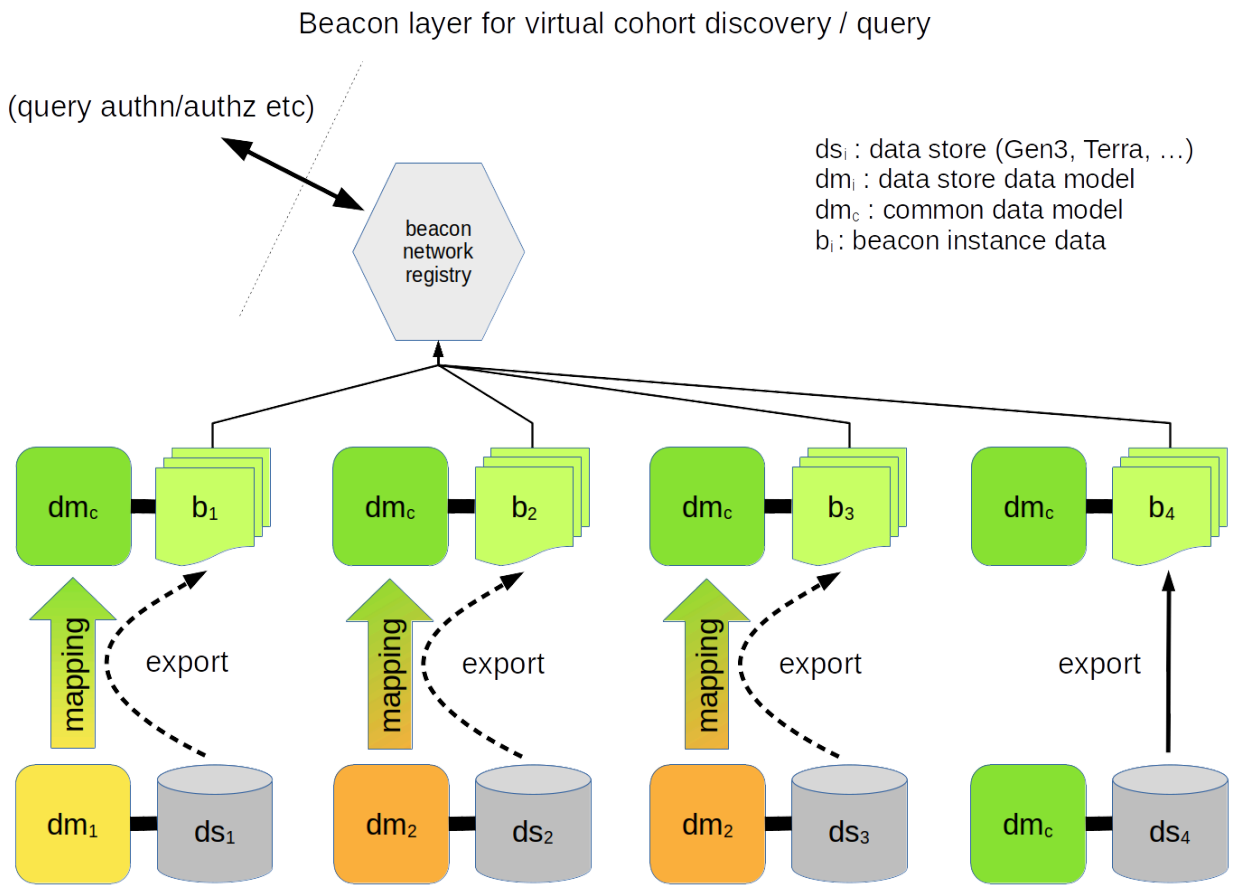

Figure 5. ***Schematic diagram of theoretical architecture of multiple beacon instances, united with a beacon network registry***

In this proposed architecture for example we have four participating partner sites with their own data stores (we make no distinction at this level between data and metadata) — these may be Gen3 instances, some other integrated solution, or a combination of data/metadata stores e.g. object store, relational database, graph database etc.

There exists an agreed-upon common data model for federated query/virtual cohort discovery, designated **dm_c**.

In the example the established data store at site 1 has an internal data model **dm_1** that is strongly coupled to the implementation of the store **ds_1**. The established data stores at sites 2 and 3 share the same data model

dm2 that is different from the common data model $\mathbf{dm_c}$ and also from $\mathbf{dm_1}$. The greenfield data store at site 4 has decided to adopt the common data model $\mathbf{dm_c}$ from the very beginning as its internal data model.

Each partner site hosts a lightweight beacon $\mathbf{b_i}$ backed by data from their data store exported via a scheduled ETL process. In the case of sites 1, 2 and 3 this is by way of a mapping from their internal data model into the common data model. Creation of the ETL process and the mapping from internal to shared data model is a work unit required from each of sites 1, 2 and 3.

Individual beacons are registered and communicate with the beacon network aggregator/ registry service which serves as the point of entry for user queries.
Implied is integration with an IAM service in front of the aggregator/registry to control access to restricted data. A solution to federated identity and authorisation management is being investigated by a separate sub-project[11] of the HGPP and is not within the scope of this sub-project.

Benefits:

- The query/presentation layer software (beacon nodes and the central aggregator/registry) are both lightweight and decoupled from partner sites data store solutions — partner sites are free to implement their own chosen data store/data commons/data processing solutions. Beacon sits on top of these rather than substituting for them.
- The common data model is decoupled from partner sites' data models — shareable harmonised data is generated via local ETL processes and there is no need to re-code or restructure existing data within the context of the store itself, and freedom for sites to implement their own data models internally appropriate to their domain. Conversely, the common data model may be extended without requiring all (or any) partners to update their internal codings. "What do we want to share" becomes a different conversation from "what do we have to capture/represent internally".
- Greenfield sites may at least initially choose to implement a sub- or super-set of the common data model as their internal data model, saving development resources.

Costs:

- Partner sites internally using data models that are not simple subsets or supersets of the common data model are required to implement data mappings and ETL processes.
- Beacon does not provide direct access to genomic scale data, i.e. is not a data commons

Proposed implementation approach

- A key efficiency will be achieved if work on the production common data model can be separated from work on the discovery/query technology proof of concept. The beacon framework allows this naturally because the data model is decoupled from the implementation.

[11] Carnuccio, Patrick, Cowley, Mark, Davies, Kylie, Downton, Matthew, Dumevska, Biljana, Holliday, Jessica, Kummerfeld, Sarah, Lin, Angela, Monro, David, Patterson, Andrew, Pope, Bernie, Ravishankar, Shyamsunder, Robinson, Andrew, Scullen, John, Shadbolt, Marion, Syed, Mustafa, Wood, Scott, & Wong-Erasmus, Marie. (2022). Human Genomes Platform Project: Federated Identity and Access Management (IAM) Discovery Phase Report (2.0). Zenodo. https://doi.org/10.5281/zenodo.6644009

- In the technology POC/demonstrator phase we propose to deploy beacon instances at partner sites adopting the default beacon v2 data model “as-is” as the POC common data model. The data may be dummy data or real public data. Private data should not be used as no authn/authz integration is assumed at this point.
- In parallel with work on the technology demonstrator we propose working towards a first production common data model that represents all concepts required to satisfy the use cases and user stories established in the requirements phase.

References and Links

General References

ELIXIR	https://elixir-europe.org/about-us
GA4GH	https://www.ga4gh.org/
HGPP	https://www.biocommons.org.au/hgpp
OMOP CDM	https://www.ohdsi.org/data-standardization/
ZERO	https://www.zerochildhoodcancer.org.au/

Virtual Cohorts Sub-Project Team Artefacts

User Stories

Endnotes

1. Icons from the Noun Project: search by Flatart, database by Start Up Graphic Design, identified by Tippawan Sookruay, group by Gregor Cresnar, Data File by Blangcon, Unlock by Arthur Shlain, archive by Adrien Coquet, support by Komkrit Noenpoempisut, documentation by lastspark, Scientist by Maxim Kulikov.)

Appendix A

GA4GH Cohort Representation Landscape Analysis

Credit: adapted from document shared at GA4GH computable cohorts meeting 2021-10-13

Resource/ Initiative Name	Resource Type	Description	Comments / Relevance	Link(s)
BeaconAPI v2.0	API	API for discovery data (genomic, cohorts, individuals, biosamples...) and facilitate the next step in the data sharing process.	It suggests a default schema for cohorts as a "meeting point" between services that don't have a direct connection, but foster using alternative or preferred schemas, like the ones suggested in the GA4GH Cohorts group.	https://beacon-project.io/
Data Connect API v1.0	API	Compact API for storing, accessing, listing, and optionally querying Tables of semantically annotated data. A Table row need not be flat; it can be an arbitrary JSON object. Each table is described by a JSON Schema, which constitutes a semantic description of the contents of a Table by referencing externally defined concepts.	- Extensibility where needed: We could agree on a mandatory core set of concepts that are essential to all cohorts, and also take advantage of the extensibility offered by the ability to reference external concepts, even one-off concepts that are specific to a single study. - Storage, transmission, and snapshots: a Data Connect query result is itself a Table, which is a self-describing unit that can be stored, shared, and published as is.	https://ga4gh-discovery.github.io/data-connect/ https://github.com/ga4gh-discovery/data-connect
Phenopackets v2.0 - Cohort Element	Data Model			
Cohort Statistical Profiles	Cohort Summary	A public data source about clinical data that is completely open, transparent and has nothing identifiable. It is de facto a	-Freely shareable, PHI-free	https://clinicalprofiles.org/

		"statistical summary" of real data for a defined phenotype. Includes counts, std devs, distributions, percentages from the usual quantifiable suspects: labs, meds, diagnoses, demographics. Tries to preserve first order correlation among variables.		
CINECA/IHCC Ontology and Cohort Discovery Platform	Cohort Registry	Built a simple model built on the consensus overlap across CINECA cohorts. Looked at overlap between dictionaries. Formalized as an ontology. The usage of that model right now has been for cohort discovery. Built a UI on top of the data to find cohorts of interest	Would be interested to know what others are doing for value harmonization.	
HDR UK Phenotype Library	Data Model	Building a phenotype library, collect phenotype definitions for disease. Making it computable by generating Phenotyping algorithm workflows in Phenoflow. Once we had defined a phenotype with all of these codes, we have built a cohort discovery tool. Not at the level of IHCC but more at the granular query level.		https://portal.caliberresearch.org/ https://kclhi.org/phenoflow/phenotype/all/ https://www.healthdatagateway.org/
National COVID Cohort Collaborative (N3C)	Data Model	N3C, is building a centralized national data resource — the NCATS N3C Data Enclave — that the research community can use to study COVID-19 and identify potential treatments as the pandemic continues to evolve. Specifically, the N3C will enable the rapid collection and analysis of clinical, laboratory and diagnostic data from hospitals and health	Patients are selected with a computable phenotype, there is a need for this phenotype to evolve. Have a local script run to extract those records to then provision those data	https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition/wiki/Latest-Phenotype

		care plans.		
International HundredK+ Cohorts Consortium (IHCC)	Cohort Registry			https://ihccglobal.org/
Answer ALS	Application / Knowledgebase	A public data portal with filtering options allowing extraction of data at the level of participants thus creating cohorts.		https://dataportal.answerals.org/search
Open MRS Cohort Builder	Application / Knowledgebase	"perform ad-hoc queries for patients with defined characteristics, and combines multiple queries into more complex ones"		https://guide.openmrs.org/en/Using%20Data/cohort-builder.html
Deep 6 AI Cohort Builder	Application / Knowledgebase	"Deep 6 AI's Cohort Builder is a software tool researchers can use to find more, better-matching patients for clinical trials. It allows research staff to search structured and unstructured data in the Electronic Medical Record (EMR), looking at millions of documents in a matter of seconds."		https://deep6.ai/update-cohort-builder-3-1/
NCI ISB-CGC Platform Cohort feature	Application / Knowledgebase	"The Cohort Builder/Data Explorer is an ISB-CGC web interface which allows you to build cohorts based on clinical demographics and molecular filters. Compare patient cohorts with various exploration tools including IGV viewer, image viewers, and analytical visualization."		https://isb-cancer-genomics-cloud.readthedocs.io/en/latest/sections/DataExplorer.html
Health Catalyst	Terminology / Identifier System	Describe the nuanced approaches to building cohorts in either a patient-centric ("asthma"),	More healthcare centric than research centric but valuable insights.	https://www.healthcatalyst.com/Defining-Patient-Populations

		episode-centric ("pregnancy") or encounter-centric ("appendectomy") way with factors such as supplemental administrative codes (e.g., ICD9 code for wheezing for the asthma cohort), sentinel medications (e.g., patients taking metformin for the Type 2 diabetes cohort), and clinical observations such as results of imaging studies or lab tests (e.g., cardiac ejection fraction and brain natriuretic peptide [BNP] for the heart failure cohort)."		
HDR UK Cohort Discovery Service	API	A cohort discovery services that federates cohort queries across multiple nodes of healthcare datasets across UK	Supports OMOP / FHIR / i2b2 with a relatively simplistic cohort query API with inclusion and exclusion criteria that uses the standard terminologies and Phenotype definitions	https://www.healthdatagateway.org/about/cohort-discovery
OHDSI APHRODITE	Data Model	OHDSI's Auto- mated PHenotype Routine for Observational Definition, Iden- tification, Training and Evaluation		https://github.com/OHDSI/Aphrodite
PheKB	Application / Knowledgebase	Knowledgebase for discovering phenotypes for EHRs		https://phekb.org/
SAIL Concept Library	Data Model	Next version of the HDR UK Concept Library		https://conceptlibrary.demo.saildatabank.com/phenotypes/
Finngen Risteys	Application / Knowledgebase	Finngen's Risteys platform relates phenotypes temporally, listing those phenotypes that a patient is likely to exhibit either before or after exhibiting an- other (e.g. the onset of depression after exhibiting bipolar dis- order)		https://risteys.finngen.fi/

Phenoflow	Data Model	Database extraction code generated from the HDR UK Phenotype Library		https://kclhi.org/phenoflow/phenotype/all/
OpenSafely Codelists / Study Definitions	Data Model	Pythonic Cohort Definition and extraction system		https://www.opencodelists.org/
Pathling	API	Advanced FHIR Data Analytics Server - Aggregation, Search		https://pathling.csiro.au/