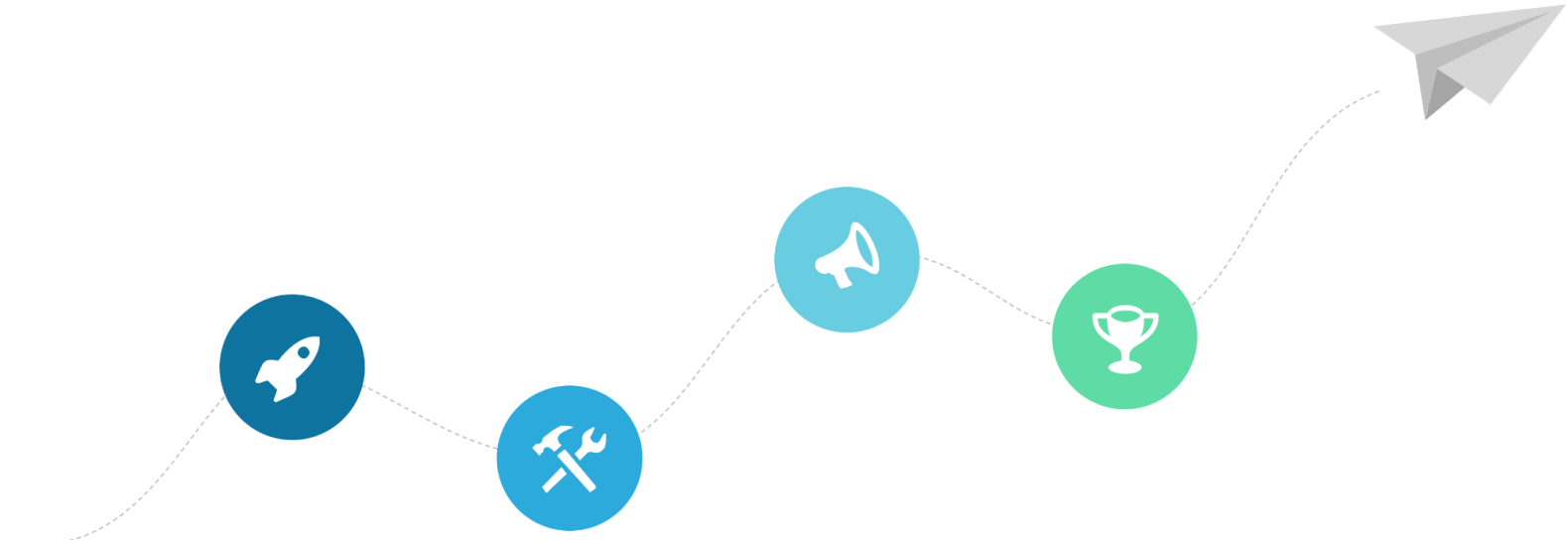
HUMAN GENOMES PLATFORM PROJECT

Virtual Cohort Assembly

PILOT PHASE REPORT

Dec 2023



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, Bernard - BioCommons

San Kho Lin, Victor - UMCCR

Ravishankar, Shyamsunder - Garvan

Shadbolt, Marion - BioCommons

Syed, Mustafa - ZERO CCIA

Taouk, Kamile - ZERO CCIA

Wong-Erasmus, Marie - ZERO CCIA

Acknowledgements

We would like to acknowledge Zoe Kamarinos and Dionne So for their extensive work on creating the Beacon Network User Interface front end described further in the report.

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, Children's Cancer Institute, 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.

Table of Contents

Glossary

CCIA	Children's Cancer Institute
GA4GH	Global Alliance for Genomics and Health
ETL	Extract-Transform-Load: wrangling data from multiple sources into a single consistent store. May involve cleaning, combining, normalising, re-coding and formatting.
HGPP	Human Genomes Platform Project
JCSMR	John Curtin School of Medical Research (ANU)
NCI	National Computational Infrastructure
QIMR Berghofer MRI	QIMR Berghofer Medical Research Institute
UMCCR	University of Melbourne Centre for Cancer Research
ZERO	Zero Childhood Cancer Program (led by the Children's Cancer Institute and Kids Cancer Centre at Sydney Children's Hospital, Randwick)

Introduction

The Human Genomes Platform Project (HGPP) is a collaborative research project aiming to enhance secure and responsible human genomic data sharing for research purposes. National and international connectivity is important to maximise the utility of these sensitive and valuable assets. The project partners represent many of the largest human genome sequencing and analysis organisations in Australia.

The goal of the Virtual Cohorts sub-project within the HGPP is to implement systems to identify cohorts of individuals and related genomic data assets across repositories located at each of the partner institutes. In the preceding Virtual Cohort Discovery Phase report[1] we examined the current landscape of systems for data access requests and data sharing, and documented a set of problems, user stories and requirements for further exploration. For full context, this document should be read in conjunction with the Discovery Phase Report. In that report we described:

- National community needs and stakeholder requirements.
- The current state of processes and tools for virtual cohort querying.
- Candidate solutions to enable cross-institutional virtual cohort querying.
- Recommendations on preferred technology and proposed implementation architecture.

In this Pilot Phase report we describe:

- Pilot implementations of the recommended technologies from the discovery phase (GA4GH Beacon v2[2]) at the partner organisations.
- Assessment of pilot performance against requirements and relative strengths against each other.
- A novel user interface for Beacon Network queries developed by one of the partner organisations (CCIA).
- Work on integration of the Virtual Cohorts and Federated Identity and Access Management (IAM) sub-projects to provide controlled access to data for authenticated users.
- Technical challenges encountered and outstanding unmet needs.
- Recommendations regarding best practices for Beacon ETL processes, data annotation, and software deployment.

The Pilot Phase of the HGPP represents a significant step forward in advancing secure and responsible human genomic data sharing. Partnering with major genome sequencing organisations in Australia, the Virtual Cohorts sub-project successfully implemented systems to identify cohorts of individuals and related genomic data assets across multiple repositories.

One of the key accomplishments in this work is the successful deployment of Beacon instances by three partners—CCIA, UMCCR, and QIMR—each populated with a shard of the CINECA synthetic dataset[3]. The report evaluates different Beacon implementations, emphasising the strengths of the reference implementation, serverless Beacon (sBeacon) by CSIRO, and the Java-based jBeacon. The assessment

[1] https://zenodo.org/record/7439886#.Y5ud9exBzZl
[2] https://www.ga4gh.org/product/beacon-api/
[3] https://www.cineca-project.eu/cineca-synthetic-datasets-old

informs a recommendation favouring the reference implementation for pilot phase deployment, acknowledging its alignment with GA4GH standards.

A pivotal aspect of the pilot phase is the development and deployment of the Beacon Network, enhancing the effectiveness of individual Beacons by enabling network-wide queries. This report assesses the Beacon Network implementation, highlighting CCIA's and Australian Biocommons' contributions. The associated Beacon Network UI, designed for user-friendly query interactions, was developed to make querying the network more intuitive for researchers. The report underscores the interface's success in abstracting complexities but notes limitations in query parameters and ontology choices.

In a technical evaluation we address critical aspects of the pilot such as the choice of data model, ontologies, reference genome, and ETL processes. While the default data model and prescribed ontologies receive positive assessments, challenges arise with aligning ontologies where they are not specified. The requirements evaluation highlights the alignment of Beacon v2 with user stories, but acknowledges limitations in evaluating specific cases due to the CINECA test dataset's narrow scope and missing features in Beacon and Beacon Network implementations. The report underscores the ongoing efforts toward integration with other HGPP sub-projects, notably the Federated IAM and REMS, with successful CILogon integration providing authenticated access to the Beacon Network.

Beacon v2 Pilot Implementations

Beacon instances were deployed by three partners (CCIA, UMCCR, QIMR) and each populated with a shard of the processed CINECA[4] synthetic dataset[5]. CCIA also deployed a Beacon instance populated with original data from Wong et al. (2020). UMCCR additionally deployed an instance of the CSIRO's sBeacon implementation and an instance of BSC's jBeacon implementation.

Beacon v2 Reference Implementation

An implementation of the Beacon v2 specification[6] (Rueda et al., 2022) as developed by EGA may be found at: https://github.com/EGA-archive/beacon2-ri-api. We refer to this henceforth as 'the reference implementation'. It is a project under active development: at time of writing it consists of Python modules providing database interaction and API request handling services, as well as a Javascript frontend. There are also separate repositories for ETL process tools at: https://github.com/EGA-archive/beacon2-ri-tools, and a UI frontend at: https://github.com/EGA-archive/beacon2-ri-ui.

Deployment instructions for the reference implementation may be found at: https://github.com/EGA-archive/beacon2-ri-api/blob/master/deploy/README.md.

ASSESSMENT: The reference implementation is well documented and has recent activity on its GitHub repository. It has strengths in terms of its developer connections with GA4GH members involved with the Beacon specification, and will likely continue to be the most fully aligned with the current version of the specification. When on-premise server resources are available as a deployment target it likely has a cost advantage over sBeacon.

RECOMMENDATION: The reference implementation is a good choice for pilot phase deployment.

sBeacon Implementation

CSIRO has developed a serverless implementation of the Beacon specification using the AWS platform (Wickramarachchi et al., 2023) .

Working closely with UMCCR, the CSIRO sBeacon[7] team added support to sBeacon for the recent GA4GH Beacon v2 specification. The UMCCR team has piloted this — code and documentation are available at: https://github.com/umccr/sbeacon-exploration/tree/main. An instance backed by test data from the CINECA dataset is hosted at: https://beacon.demo.umccr.org.

ASSESSMENT: sBeacon is well documented and has recent activity on its GitHub repository. It has strengths in terms of minimising ETL processes (able to use indexed VCFs directly as a data source), scalability across millions or billions of records, and reduced costs when compared to a cloud-hosted server deployment (e.g., EC2). The sBeacon implementation may sometimes lag the official GA4GH

[4] https://www.cineca-project.eu/
[5] https://ega-archive.org/datasets/EGAD00001006673
[6] https://github.com/ga4gh-beacon/beacon-v2
[7] https://github.com/aehrc/terraform-aws-serverless-beacon

Beacon v2 specification. There can be delays in resolving queries to sBeacon's if the service has gone idle, resulting in slow query times (aka the cold start problem).

RECOMMENDATION: sBeacon is useful for large datasets when funding is available for AWS services. Care must be taken when deploying a non-reference implementation within a Beacon Network to ensure the implementation endpoints are compatible with the version of the specification supported by the aggregator site.

jBeacon Implementation

The UMCCR team explored an alternative Java-based implementation[8] of the Beacon specification, as documented here: https://github.com/umccr/beacon-doc.

ASSESSMENT: The Java-based implementation does not appear to have as much in the way of community activity or documentation compared to the reference implementation or sBeacon.

RECOMMENDATION: Sites should be free to choose whichever Beacon implementation best suits their needs as long as its data model and API endpoints are compliant with those agreed on by partners in the Beacon Network.

[8] https://gitlab.bsc.es/inb/ga4gh/beacon-v2-docker-demo/

Beacon Network Pilot Implementation

Beacons in themselves are effective but they are made significantly more effective with the introduction of the Beacon Network. You can query individual Beacons using the Beacon API or by querying multiple Beacons using the Beacon Network.

Beacon Network API deployment

An implementation of the Beacon Network as developed by CSC - IT Center for Science may be found at: https://github.com/CSCfi/beacon-network. At the time of writing, it consists of Python modules providing three microservices:

1. A database microservice which saves details about all registered Beacons.
2. Registry microservice to register a Beacon and keep track of different Beacons that are participating in the Beacon Network.
3. An aggregator microservice which sends out queries to all Beacons in the network and aggregates results from queries into a single response object.

CCIA deployed a Beacon network/registry instance and developed and deployed the Beacon network UI. Australian Biocommons also deployed an instance of the Beacon network and Beacon network UI. Beacon instances hosted by CCIA and UMCCR were connected to the network as a proof of concept.

Australian Biocommons, as an entity independent of the other network nodes, agreed to host an aggregator service. A version of the above Beacon code base[9] was stripped back, removing all databases and registries, but leverages the existing aggregator microservice. Authentication flows were added to the aggregator along with the ability to host the Beacon Network UI. CILogon was used to implement a basic authentication flow in which the user logs in via the UI and receives a hydrated authentication token in return. Finally, a deployment script was added for deploying a production-grade instance in AWS.

ASSESSMENT: The Beacon Network implementation is well documented and detailed documentation is available at: https://beacon-network.readthedocs.io/en/latest/. The Beacon Network version available at the time of deployment does not support POST API methods. We have branched from the original code and added a POST method to the Beacon Network API. We note that a pull request[10] made after a code update was not merged into the main branch with the owner citing lack of resources to maintain the project.

RECOMMENDATION: We did not find another public Beacon Network repository with more active development and support at the time of deployment. We recommend using a more active fork or implementation of the specification, if available.

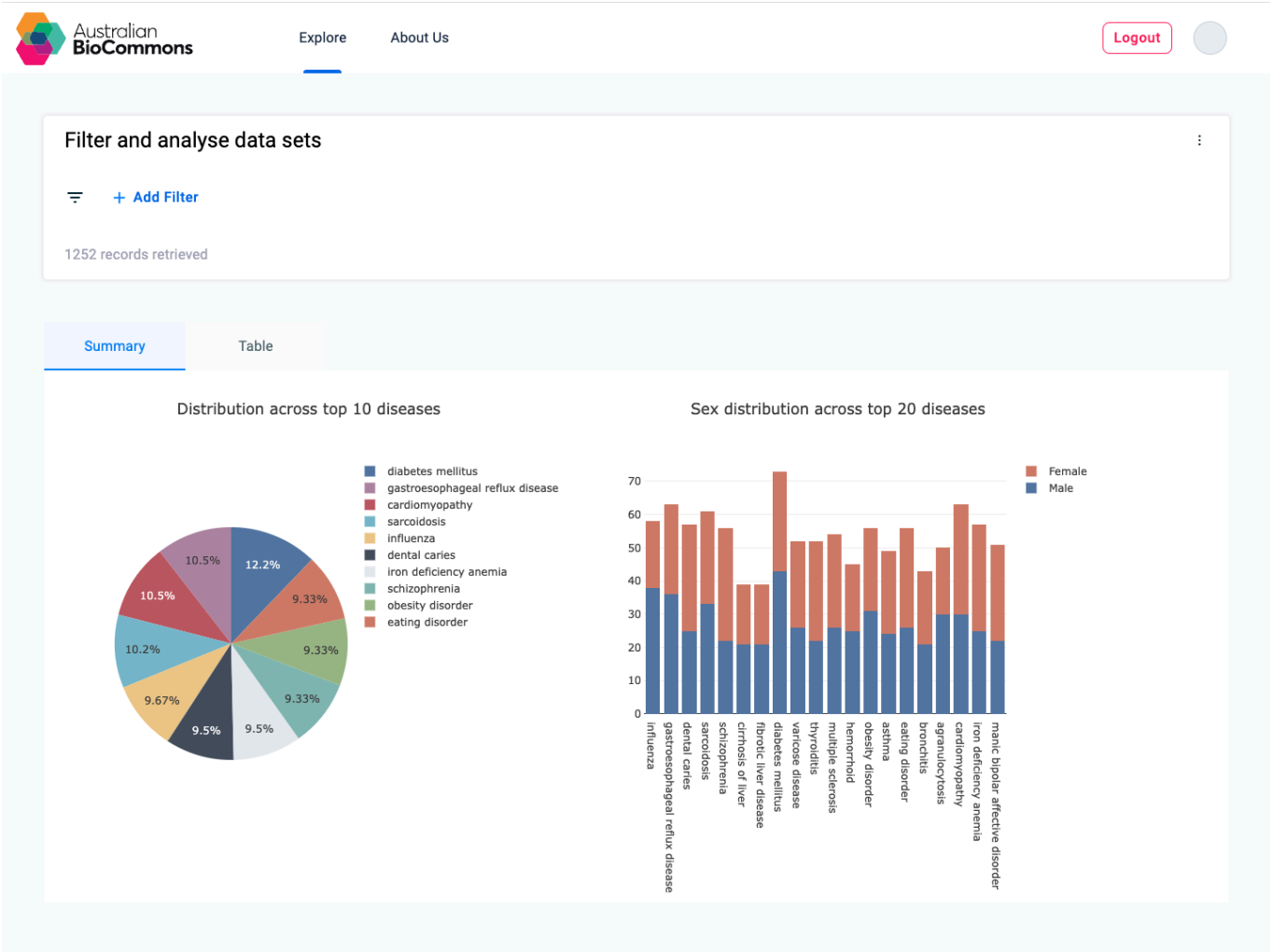
[9] https://github.com/AustralianBioCommons/beacon-network/tree/master
[10] https://github.com/CSCfi/beacon-network/pull/230

Beacon Network UI

The queries required as input to the Beacon Network are not user-friendly and have precise requirements for syntax and formatting. Because of this, we elected to design and develop a user interface that can abstract the complexities of the Network and provide a more accessible and streamlined experience for users.

Code and documentation for the Beacon Network UI may be found here: https://bitbucket.org/cciacb/cci-hgpp-beacon-ui/src/master/.

The interface includes an easy-to-use input module, where users can specify filter parameters with no concern for formatting. As a proof of concept, only the disease, sex, and age of onset filters under the 'Individuals' node were implemented. Also included is a tabular representation of the data returned, as well as a visualisation of the statistical distribution of the data below.



The interface supports conditional rendering based on the user's authorisation, which is made possible by the CILogon integration.

ASSESSMENT:

The Beacon Network UI was designed and implemented with the intention of making the process of querying the network streamlined and more accessible to non-technical users (such as researchers). This overall goal was achieved via a polished, intuitive design. Users are able to authenticate via a CILogon integration, or can navigate directly to the dashboard and peruse statistics about the data within the network. From here, query parameters can be specified in the form of filters via the UI. The submission of these filters will automatically trigger a new request to the aggregator and immediately propagate to the visual elements on the page. For the users that are authenticated via CILogon, there is an additional tab through which more specific tabular data from the Beacon results can be viewed.

While the interface successfully abstracts much of the tedium and technical knowledge to query the network directly, there are still several limitations. By virtue of the scope of the project, the query parameters accessible to the users are restricted to a handful of fields within the 'Individuals' node of the default data model. In addition, while all the Beacon instances registered within the network can theoretically use multiple ontologies within their datasets, we enforced the use of only the MONDO and NCIT ontologies as proof of concept. A noticeable consequence of these limitations is the focus on individual-centric data; the data visualisation, as well as the table structure, heavily cater to data of a specific format.

RECOMMENDATION:

The Beacon Network UI remains a useful starting point for teams wanting to utilise an interface for a network without starting from scratch. Given that the code has been designed and implemented in a way that facilitates easy customisation and extension, teams forking the repository can refine the design and implementation with minimal effort. Almost all of the customised logic surrounding the parsing of aggregated results was written with the default data model in mind, and so is readily extensible to more complex or nuanced use cases.

Technical Evaluation

Beacon instances were deployed by three partners (CCIA**,** UMCCR**,** QIMR**)** and each populated with a shard of the processed CINECA synthetic dataset. CCIA also deployed a Beacon network/registry instance, an instance of the Beacon network UI, and a Beacon instance populated with original data from Wong et al. (2020). UMCCR deployed an instance of the CSIRO sBeacon implementation and an instance of the jBeacon implementation. While we did not conduct a structured evaluation of all technical aspects of the pilot, this section presents our impressions and findings in significant areas.

Choice of Data Model

The Beacon v2 reference implementation provides a default data model[11] with seven primary entities:

- genomic variation
- analysis
- dataset
- run
- biosample
- individual
- cohort

Within the Beacon framework, this data model may be customised or replaced entirely according to need, but for the purposes of the pilot implementation stage, we used the default model without alteration.

ASSESSMENT: The Beacon default data model satisfactorily models the entities required to represent query-able genotypic and phenotypic data across several clinical domains.

RECOMMENDATION: The Beacon default data model is a good starting point from which to establish a network of Beacons. All node sites should use the same Beacon data model (schema) version so that the aggregator site may distribute a query across Beacons without modification. In the pilot phase, we are currently targeting beacon-v2-default-model 519aa57[12]. Possible customisations are of two kinds: structural alterations to the model which require all nodes to adopt them for compatibility, and extensions which might be implemented by only some nodes (e.g., additional annotation fields). A Beacon Network should have a policy about if and how such changes may be made.

Choice of Ontologies

The Beacon v2 default data model makes three sorts of statements around coding and ontologies to be used for entity fields:

[11] https://github.com/ga4gh-beacon/beacon-v2/tree/main/models

[12] https://github.com/ga4gh-beacon/beacon-v2/tree/bc8f02a6a388140c854f3e17da3ccb779f2cbc0b/models/json/beacon-v2-default-model

- prescriptions (e.g., GAZ for geographicOrigin[13])
- recommendations (e.g., NCIT for genders[14])
- list of examples (e.g., ICD10CM, OMIM, HP for diseaseCode[15])

Where ontologies were prescribed or recommended in the Beacon default data model they were adopted.

ASSESSMENT: Where the default data model prescribed or recommended ontologies these functioned well. Where the default data model did not prescribe or recommend an ontology there were often differences in coding requirements and existing usage between the partner organisations. For example, for disease codes, CCIA uses MONDO, UMCCR uses HANCESTRO, SNOMED and other ontologies, and QIMR uses ICD-O.

RECOMMENDATION: Where an ontology is prescribed or recommended by the Beacon v2 default data model, that should be used. These are (mostly) "industry standards" like GAZ country codes. For more domain-specific fields relating to phenotype and clinical data, organisations with different focuses (e.g., adult cancer, paediatric cancer, or rare disease), will have different requirements. It may not be desirable or even possible to require re-coding of all organisations' data onto a single ontology — likely why the Beacon data model makes neither prescription nor recommendation in these cases. Our recommendation is that, where possible, organisations should adopt existing ontologies when establishing a new dataset, and that the central Beacon query point provides an ontology mapping service to allow for queries across differently annotated datasets.

Choice of Reference Genome and Coordinate System

Reference genome and coordinate system for genomic variations entities are not prescribed by the Beacon v2 default data model.

ASSESSMENT: Queries by genomic location are only useful if all variants use the same reference and coordinate system.

RECOMMENDATION: Use GRCh38 (GCA_000001405.15) and a 1-based (SAM/MAF/VCF-like) coordinate system. Note that patches to a given reference assembly do not change chromosome coordinates[16].

ETL process

An essential step in lighting a Beacon is to map the organisation's data from its internal data sources into the Beacon data model. Choices must be made around what fields to populate within Beacon and what ontologies to use. Beacon makes recommendations for certain fields about certain ontologies but is not prescriptive so that partners may use their own preferred standards and acceptable ontologies. The choices made around fields and ontologies for each Beacon can limit the compatibility and query ability across Beacons that are joined in a network. As a group, we aimed to find alignment on which

[13] https://github.com/ga4gh-beacon/beacon-v2/blob/main/docs/schemas-md/obj/geographicOrigin.md
[14] https://github.com/ga4gh-beacon/beacon-v2/blob/main/docs/schemas-md/obj/genders.md
[15] https://github.com/ga4gh-beacon/beacon-v2/blob/main/docs/schemas-md/obj/diseaseCode.md
[16] https://www.ncbi.nlm.nih.gov/grc/help/patches/

ontologies and fields we populated to maximise the query ability across the Beacons that we joined in a network.

The ETL process for the CINECA synthetic dataset used in the pilot implementation involved transforming the CINECA metadata into the JSON files required for the default data model of the Beacon reference implementation. Two approaches were used for this. CCIA transformed a published dataset from their landscape paper (Wong et al. 2020). A bespoke ETL process was used for CCIA's data since the data is stored in an in-house database.

The ETL process for transforming this data can be found here:
https://bitbucket.org/cciacb/cci-beacon-etl/src/master/
https://github.com/EGA-archive/beacon2-ri-tools

UMCCR used an alternate ETL process as they operated a different implementation of Beacon V2. This also involved curation of MONDO disease codes rather than the non-hierarchical ICD codes provided in the reference implementation. The script for this may be found in the following repositories:

https://github.com/umccr/sbeacon-exploration/blob/main/scripts/CINECA_UK1C-MS.ipynb
https://github.com/umccr/sbeacon-exploration/blob/main/scripts/Re-curate%20ontologies.ipynb

QIMR Berghofer developed proof-of-principle ETL code to transform data from their internal triplestore into Beacon-compatible JSON files. This may be found at the following repository:

https://github.com/delocalizer/beacon_v2_test_data

ASSESSMENT: Each organisation hosting a Beacon instance must develop ETL processes to transform their existing data into a format compatible with the data model, codings and ontologies defined for the network. Significant variation is to be expected between the internal data sources and data representations of organisations in the network so these processes may not have many details in common.

RECOMMENDATION: Organisations should share their Beacon ETL code in public repositories — the details may not be directly usable by others but they can demonstrate broadly useful approaches to the problem (e.g., examples of batch data processing, de/serialisation or schema validation).

Input Data Validation

ASSESSMENT: By default, input data is not validated against the Beacon data model schema when following the data loading instructions from the reference implementation.

RECOMMENDATION: Each node site must validate all Beacon entity data against the agreed version data model (schema) *prior to loading* into the node backend (database). This is a relatively straightforward JSON schema validation task.

Example validation tools include:

https://github.com/EGA-archive/beacon2-ri-tools/tree/main/utils/bff_validator (reference implementation validator)

https://github.com/delocalizer/beacon2_data_validator (simple validator)

Deployment and Resources

ASSESSMENT: The reference implementation deployment notes[17] are sufficient. Issue #92[18] was encountered which was resolved as per the workaround detailed in the issue discussion.

RECOMMENDATION: With the reference implementation deployed as per the documentation to AWS EC2 as a stack of docker services, the following resources were found to be the bare minimum system requirements to have the service come up after "docker-compose up -d beacon":

- 8G disk
- 2G mem
- 1 vcpu

Corresponding to a t2.small @ $AU1/day in the ap-southeast-2 region (Sydney). In a production setting, a larger instance type with more disk would be required, as well as provision for load-balancing and high-availability.

Security

ASSESSMENT: Using the default Beacon Network configuration, CCIA experienced cyberattacks against some of the deployed services. This required action by their IT security team and reconfiguration of the deployment.

RECOMMENDATION: With the reference implementation default configuration, all public access to the deployment host should be denied except to port 5050 (or whichever port is actually serving the API). If deploying to AWS, port access should be controlled using an EC2 security group. It is worth considering restricting Beacon instance public access to specific IP addresses or ranges (e.g., an allow-list containing only the Beacon aggregator and trusted organisations). There is always a trade-off between security and convenience.

[17] https://github.com/EGA-archive/beacon2-ri-api/blob/master/deploy/README.md
[18] https://github.com/EGA-archive/beacon2-ri-api/issues/92

Requirements Evaluation

In this Pilot Phase, Beacon instances were deployed by three partners (CCIA, UMCCR, QIMR) and each populated with a shard of the CINECA synthetic dataset. CCIA deployed a Beacon network/registry instance, and an additional Beacon instance populated with original data from Wong et al. (2020). Demos and adhoc testing of these deployments were used to assess the capabilities of Beacon and the Beacon network against our user requirements as established in the Discovery Phase report. Following is a table of the top six user stories ordered by priority for each project partner institute, with an assessment of requirement capability at the Pilot Implementation phase.

ASSESSMENT: Although Beacon v2 satisfies a significant number of these requirements in principle, many of the specific cases cannot be evaluated at this stage. This is partly due to the narrow scope of the CINECA test dataset and partly due to specific features currently missing from Beacon and Beacon Network implementations.

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 **ASSESSMENT:** This can be achieved via the Beacon Network API as well as the Beacon UI. A user may list all individuals with PDAC and the results will show which Beacon instance is listing the data.	***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 **ASSESSMENT:** The Beacon Network API can be used to query individuals with the appropriate filters. Currently the merge/join to obtain the intersection must be performed client-side.	***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 **ASSESSMENT:** The Beacon Network API can be used to query datasets.	***As a curator:*** I want to find variant information for cancer samples of a given subtype ***So that:*** We can assess a novel variant and its reporting status **ASSESSMENT:** This can be partially addressed with the Beacon Network but relies on harmonised metadata for the cancer type and an agreement on how to describe reporting status.	***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 **ASSESSMENT:** This can be achieved via the network API. While a user may successfully list all the medulloblastoma samples in the network, the ability to request access and download via the UI is not supported by the beacon data model, nor is the UI connected to a data access system.
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 **ASSESSMENT:** The Beacon Network API can be used to query individuals, biosamples, and genomic variants entities with the appropriate filters. Currently the merge/join to obtain the intersection must be performed client-side.	***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 **ASSESSMENT:** The Beacon Network API can be used to query individuals, biosamples, and genomic variants entities with the appropriate filters. Currently the joins and aggregation to obtain the count must be performed client-side.	***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 **ASSESSMENT:** Not assessed	***As a research user:*** I want to find primary / read level data for published cancer cohorts stored in Australia ***So that:*** So I can request relevant read or variant information and re-process / harmonise data **ASSESSMENT:** The Beacon Network API can be used to query individuals, biosamples, and genomic variants entities with the appropriate filters. Data requests for the primary information are out of scope for Beacon	***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 **ASSESSMENT:** The current implementation does not support storage and analysis.

Rank	QIMR Berghofer MRI	Garvan	NCI (JCSMR)	UMCCR	ZERO
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 **ASSESSMENT:** The Beacon Network API can be used to query individuals, biosamples, and genomic variants entities with the appropriate filters. Currently the merge/join to obtain the intersection must be performed client-side.	***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 **ASSESSMENT:** The current implementation does not support data analysis.	***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 **ASSESSMENT:** Access to data is unavailable in the current implementation of the Beacon UI and Beacon Network API.	***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 **ASSESSMENT:** The Beacon Network API can be used to query individuals, biosamples, and genomic variants entities with the appropriate filters. Data requests for the primary information are out of scope for Beacon.	***As a research user:*** How can I find all paediatric samples? (age < 21 yrs) ***So that:*** I can consolidate my data with yours **ASSESSMENT:** A query can be made to the Beacon Network API to find samples of any specific age range. The Beacon UI however, does not yet support this specific query. Though the implementation of this feature would be very similar to ZERO.U.S.1's.

Rank	QIMR Berghofer MRI	Garvan	NCI (JCSMR)	UMCCR	ZERO
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 **ASSESSMENT:** The Beacon Network API can be used to query individuals, biosamples, and genomic variants entities with the appropriate filters. Currently the joins and aggregation to obtain the count must be performed client-side.	***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 **ASSESSMENT:** The current implementation does not support storage and analysis.	***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 **ASSESSMENT:** The current implementation does not supply information for how and where to perform computation on 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 and with data access control requirements matching my research plan ***So that:*** So I can re-process / harmonise data **ASSESSMENT:** The Beacon Network API can be used to query individuals, biosamples, and genomic variants entities with the appropriate filters. Data requests for the primary information are out of scope for Beacon.	***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) **ASSESSMENT:** Not available via the Beacon UI but queryable using the API.

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 **ASSESSMENT:** The Beacon Network API can be used to query individuals, biosamples, and genomic variants entities with the appropriate filters. Currently the joins and aggregation to obtain the count must be performed client-side.	***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* **ASSESSMENT:** The Beacon Network API can be used to query individuals, biosamples, and genomic variants entities with the appropriate filters. Currently the joins and aggregation to obtain the count must be performed client-side.	***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 **ASSESSMENT:** Not available.	***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 **ASSESSMENT:** External UI and search tools can make use of the Beacon V2 API to provide relevant sample and variant information; read level data would have to be obtained through other data access approaches.	***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 **ASSESSMENT:** Not available via the UI but users can query for the “pathologicalStage” property of the biosamples and use the individual’s object. Currently, the joins and aggregation to obtain the count must be performed client-side.

Rank	QIMR Berghofer MRI	Garvan	NCI (JCSMR)	UMCCR	ZERO
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 **ASSESSMENT:** Beacon by default supports public (anonymous) requests. Integration work with CILogon (IAM subproject) has enabled authenticated access. Further integration work is required to support specifically authorised access.				***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). **ASSESSMENT:** A user can technically retrieve this data by using the Beacon Network API to query the individuals, biosamples, and genomic variants objects and then getting the intersection of these results. The Beacon UI currently does not support join/merge functions between the objects, however, and does not support gene fusions.

Table 1. Top six user stories ordered by priority for each project partner institute with assessment of requirement capability at Pilot Implementation phase.

Integrations

Integration with services provided by the other HGPP sub-projects is stated as a desirable outcome in the Discovery Phase report. Here we describe work done towards this goal.

Integration with CILogon

This work was done in collaboration with the HGPP Federated IAM sub-project. See the section **Custom JWT for Beacon Network** in the IAM Pilot Phase Report for technical details. The objective here is to enable three levels of access to Beacon data:

1. Anonymous (public) access to non-sensitive data — for example, non-identifying or aggregate data; boolean responses for data existence/availability.
2. Authenticated access to more detailed data (not requiring a data access agreement).
3. Authorised access to controlled data (requires a data access agreement).

A proof-of-concept implementation of the CSCfi Beacon Network with a minimal CILogon integration was developed. The code may be found here:

https://github.com/AustralianBioCommons/beacon-network

Most of the code was stripped out from the reference implementation. A login endpoint was created where the user could pass in a request and receive a basic JWT that is hydrated with the relevant credentials. See Appendix A Figure A1: Data flows of JWT bearer tokens throughout the Beacon Network.

ASSESSMENT: This Pilot Phase CILogon integration provides authenticated access to the Beacon Network. CILogon can also provide authorisation information in the tokens, e.g. group membership, although this is not currently used by the Beacon network to control access to data. The simplest level of authorisation to implement will be at the level of datasets and pre-defined cohorts, and this might be achieved quite quickly. Authorisation at the level of user-defined cohorts will require more work.

Integration with REMS

Resource Entitlement Management System (REMS) is the platform piloted by the HGPP DAC Automation sub-project for managing access rights to resources such as research datasets. A goal of Beacon integration in this context is to hand off to the REMS system when a Beacon user requests access to raw data supporting Beacon records (e.g., vcf or bam files, as described in a Beacon dataset that support *genomicVariant* records).

ASSESSMENT: Details of integration with REMS have not been worked out at this stage.

RECOMMENDATION: Use the REMS API to make the handoff process as smooth as possible — the data access request can be pre-populated with information known at the point of handoff (e.g., the user id

and data authorisation level of the Beacon user, as well as the identifier(s) of the requested dataset(s)). Note that this is dependent on the integration with the Federated IAM solution being in place.

Conclusions

The goal of the Virtual Cohorts sub-project was to evaluate technologies that can identify groups of individuals and their related genomic data assets across various repositories located at each of the partner institutes. Initially, we analysed the existing systems for data access requests and sharing, and documented a range of issues, requirements, and user stories for further exploration. After the examination, we decided to use the Beacon platform because of its lightweight framework, which would be easier to implement across multiple institutions.

We successfully piloted Beacon v2 as a platform for creating virtual cohorts across several organisations within the HGPP. Different implementations of Beacon were used, with CCI adopting the Beacon reference implementation, UMCCR choosing to run with sBeacon, and QIMR using a different approach. Each Beacon joined the Beacon Network, allowing the catalogue of data from each site to be queried simultaneously. We also worked with the Federated IAM sub-project to incorporate a level of controlled data access for authenticated users through CILogon as a proof of concept.

Beacon and the Beacon network have the potential to be a solution for cross-organisational data sharing, provided that the underlying data is of the same type, such as all cancers. We highlight the pros and cons of the reference and serverless implementations of Beacon, what they can and cannot do, and their limitations.

The following items may be considered for future work:

- Currently, the user interface provides three basic data filtering options which can be expanded to other data filtering options using Biosample, Run, Analysis, Dataset, and Cohort Beacon entities.
- Tabular view of the user interface can be updated to provide more information and data from all above entities besides “Individual entity”.
- Filtering options using ontology terms from a specific ontology that can be replaced by a free text search. Free text must be mapped to one or more ontologies used by different Beacons in the network. This ontology mapping service can be part of Beacon Network service.
- Beacon specifications mentions three different types of data access: registered, controlled and anonymous. Controlled data access is not working as expected in the Beacon implementation tested in this project. At the time of writing this report, we are working with the reference Beacon implementation team to fix this issue.
- The current Beacon implementation does not support structural variant data which can be added in future specification and data model.

This Pilot Phase report demonstrates significant progress in implementing a robust system for querying virtual cohorts of individuals with genomic data across distributed sites. The recommendations provided, based on thorough assessments and lessons learned, pave the way for future phases of the HGPP, emphasising collaborative, secure, and responsible genomic data sharing for research purposes.

References

1. Cowley, M., Downton, M., Holliday, J., Kummerfeld, S., Leonard, C., Lin, A., Pope, B., San Kho Lin, V., Ravishankar, S., Shadbolt, M., Syed, M., Taouk, K., & Wong-Erasmus, M. (2022). Virtual Cohort Assembly Discovery Phase Report: National Community Needs & Candidate Solutions. Zenodo. https://doi.org/10.5281/zenodo.7439886
2. Rueda, M., Ariosa, R., Moldes, M., & Rambla, J. (2022). Beacon v2 Reference Implementation: A toolkit to enable federated sharing of genomic and phenotypic data. Bioinformatics, 38(19), 4656-4657. https://doi.org/10.1093/bioinformatics/btac568
3. Wickramarachchi, A., Hosking, B., Jain, Y., Grimes, J., O'Brien, M. J., Wright, T., Burgess, M. A., Lin, V. S. K., Reisinger, F., Hofmann, O., Lawley, M., Wilson, L. O. W., Twine, N. A., & Bauer, D. C. (2023). Scalable genomic data exchange and analytics with sBeacon. Nature Biotechnology. https://doi.org/10.1038/s41587-023-01972-9
4. Wong, M., Mayoh, C., Lau, L. M., Khuong-Quang, D.-A., Pinese, M., Kumar, A., Barahona, P., Wilkie, E. E., Sullivan, P., Bowen-James, R., Syed, M., Martincorena, I., Abascal, F., Sherstyuk, A., Bolanos, N. A., Baber, J., Priestley, P., Dolman, M. E., Fleuren, E. D., … Cowley, M. J. (2020). Whole genome, transcriptome and methylome profiling enhances actionable target discovery in high-risk pediatric cancer. *Nature Medicine*, *26*(11), 1742–1753. https://doi.org/10.1038/s41591-020-1072-4

Virtual Cohorts Pilot Phase Artefacts

Artefact	Owner	Location
sBeacon *documentation and source code*	UMCCR	https://github.com/umccr/sbeacon-exploration/tree/main
sBeacon *test deployment*	UMCCR	https://beacon.demo.umccr.org
jBeacon *documentation*	UMCCR	https://github.com/umccr/beacon-doc
Beacon network *test deployment*	CCIA/Australian Biocommons	https://beacon-network.test.biocommons.org.au/
Beacon deployment *documentation*	CCIA	https://docs.google.com/document/d/1250Tj5PZmeEVMPdup7dO8nmKADAAMqsYgSVO4Vxvygo
Beacon network UI *source code*	CCIA	https://bitbucket.org/cciacb/cci-hgpp-beacon-ui/src/master/
Beacon ETL process *source code*	CCIA	https://bitbucket.org/cciacb/cci-beacon-etl/src/master/
Beacon ETL process *source code*	QIMR Berghofer	https://github.com/delocalizer/beacon_v2_test_data
Beacon data validation *source code*	QIMR Berghofer	https://github.com/delocalizer/beacon2_data_validator
Beacon network with CILogon integration *source code*	Australian Biocommons	https://github.com/AustralianBioCommons/beacon-network

Appendix A

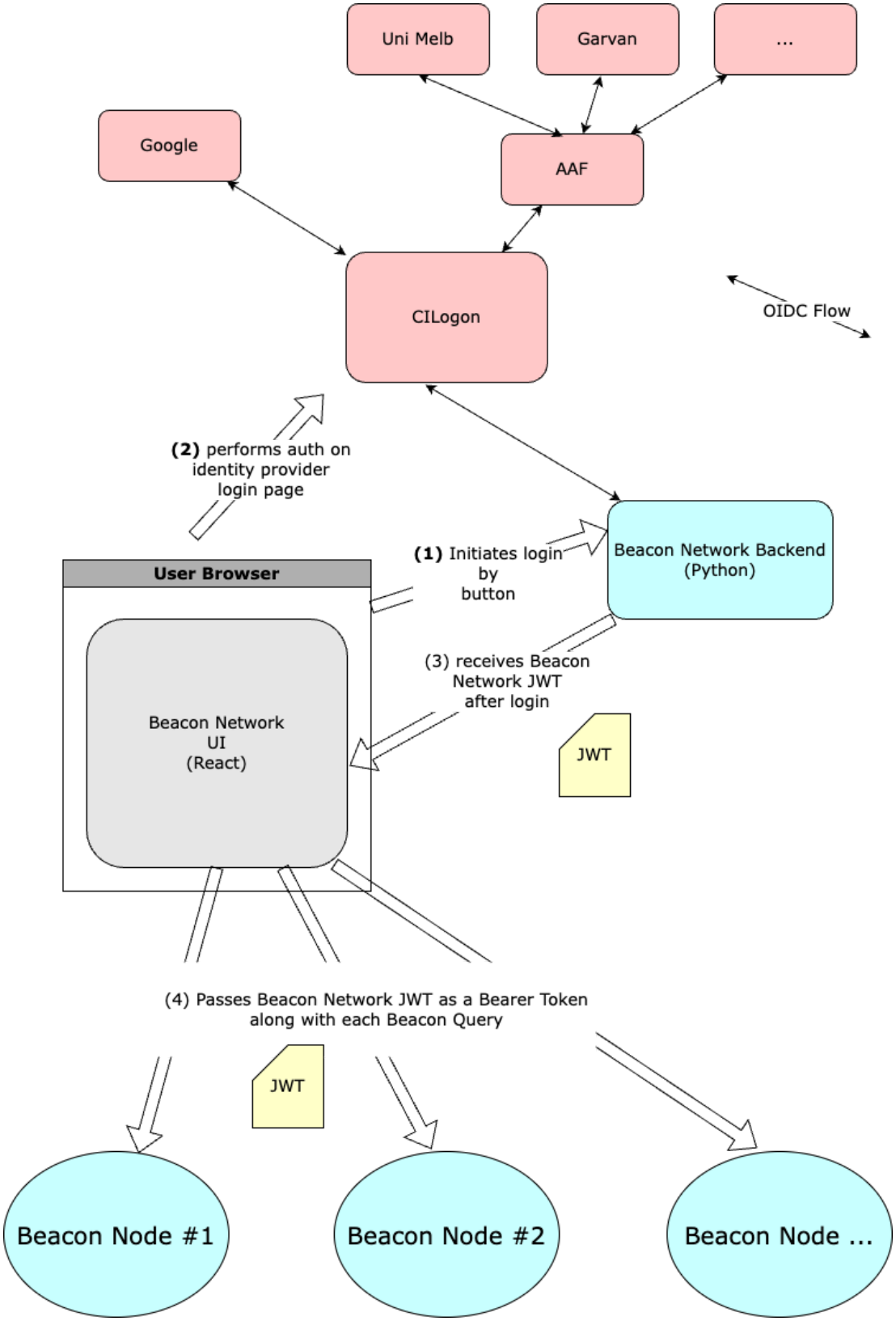


Figure A1. Data flows of JWT bearer tokens throughout the Beacon Network