Beacon v2
Migration Workshop

Discovery

Date: 2022-11-16
Time: 14:00-15:30 UTC (15:00 CET | 06:00 PST | 09:00 EST | 14:00 GMT | 01:00 AEDT)
Meeting Chair(s): **Michael Baudis,**
Abstract: The goal is to map out the upgrade path for migrating a version 1 Beacon to version 2, preferably using real world examples. As an alternative solution, we can discuss the technical feasibility of and interest in a “translation” middleware that would allow v1 Beacons to join a v2 Beacon network, without direct upgrade.

	Agenda Item	**Speaker**	**Time**
1	Introduction	Michael Baudis	
2	Beacon reference implementation w/ documentation	Manuel Rueda	
3	Migration paths - Existing data (Y/N by coord) - Extended data (gender, age etc.)	Open Discussion	
4	Survey of Beacon v1 implementers	Zoom Poll	
	Short Break before second session		

Attendees

Fabio Liberante (GA4GH), David Salgado (INSERM, IFB, ELIXIR-FR), Babita Singh (EGA-CRG), Manuel Rueda (CNAG-CRG), Alex Tsai (GA4GH), Gordon Krieger (McGill/C3G), Tshikala Eddie Lulamba, Dmitry Repchevsky (BSC), Sergi Aguiló Castillo (BSC) , ,

Notes/Links

Zoom recording

https://us02web.zoom.us/rec/share/7KEYeMRMwAsevt5Qt3nFjGO-hAqj9lS5UjyWrl4ozvw-mLtpfy-Cn4CTrQeZuPGO.2rJdHQ7QkOPVALYR
Passcode: #Bb^0ECd

Michael Baudis slides are here https://drive.switch.ch/index.php/s/9rN9xmuAn8SQFd8

Key takeaways

- Beacons need a way to share or harmonize their query/response types
- Should gather documentation, example data, implementations, use cases, filters, mappings and tooling centrally
-

Summary

- Queries to a Beacon rely on IDs - therefore prescriptive recommendations are important e.g. “Use RefSeq for Gene IDs”
- An API specification is a “plan for building a house”, you specify which materials you use to build it
- The Beacon Friendly Format is not an official or even standard concept, it reflects an intermediary step for facilitating Beaconization of data
- Differences in approach - some implementations convert all data to BFF once, others have built tooling to do it on the fly
- Harmonizing labelling takes considerable time in Beaconizing data, especially clinical data - people are building tooling to convert/map it into Phenopackets/BFF see https://convert-pheno.readthedocs.io & https://github.com/phenopackets/omop-exporter
- Some members of the community, clinicians especially, feel that providing a specification without a supporting implementation severely limits uptake of Beacon
- While Beacon is very capable, it was envisioned as a *Discovery* tool, not for research. That should usually come later/separately.
- Encouraging a set of standard/minimal terms for data representation would help the growing network harmonize as it grows
- Collecting example data sets and usage would facilitate further uptake
- There already exists a number of backend data to BFF mappings and tooling, but this is not being shared in a centralized location next to Beacon
- The case is similar for documentation, CRG B2RI and Progenetix has some too, but this does not sit with Beacon standard or Starter Kit resources
-

Minutes

Michael Baudis - Intro on what was there before for Beacon. Slides here.
Asking for resources which have been Beaconized before - or could be.

Comparison of Beacon v1 and v2.
Can translate a v1 Beacon without providing access to extra data.
Additional variant parameters were present in v1, but can be similarly used or not in v2.
Beacon v1 queries represented in v2 format. Some ids are kept as legacy from v1, so some queries will still work.
Beacon v2 offers 4 response types. Boolean fallback. More “chatty” than v1. Count also includes count.
Beacon v2 implementation could be very simple - not unlike v1.

DS: Seq ref Q. I find it difficult as a querier to know exactly which reference they use. Is that something that should be listed for a particular query? A catalog of the sequence references somewhere?
MB: RefSeq ID vs chromosome ID etc.? It is neat to use standard IDs. If you use transcript IDs - nightmare. Possibility for implementers to do they own, but then you need translation in a network. Need to be soft “enforcing” or verbose in what we really recommend people do. “For humans you should use RefSeq version XYZ” for example. Capitalisation of chromosomes.
Patrick Ruch conversation. Should be more prescriptive?

Manuel Rueda - Reference Implementation Overview - Slides here
Some lessons I learned as a user/implementer. Happy for technical questions.
People need to understand that Beacon is an API specification. Nothing to do with programming languages. Always need examples. Database is not important. Floor plans for a house - you need to choose building materials yourself. 3 options to create a Beacon

> *MB (in chat): +1 Manu; I frequently get this "data stored in the Beacon format" which doesn't make sense (although we do store data in Progenetix according to the Beacon model this is NOT the common case)*

A - Internal Database and API -> Beacon
B - Internal Database and ready-made/custom API -> Beacon
C - Start with data (unstructured, excel etc.) - want to Beaconize my data
B2RI is free software - CLI interface and published recently

> *DR (in chat): Well in many systems ideally the data should be just transformed to "Beacon" on the fly. otherwise one will fall to the problema to sync databases between "beacon" and "non beacon" formats.*

4 components

1. loading data into DB - this is the first and least glamorous
2. DB (switch to MongoDB)
3. API
4. Example CINECA dataset - very important

At which point do you transform your data into Beacon v2 models? On the fly or in the DB? In B2RI - we store data in MongoDB in Beacon v2 format.

MB (in chat): Metadata - "Everything but the Sequence" (I got this on an early GA4GH plenary slide presented by David Haussler ....)
We store data in Beacon v2. Some re-writes (e.g. Phenopackets exports)....
and we re-write variants, since we store each variant instance & Beacon aggregates "identical" variants).

Hierarchical data store. We provide a tool to convert their data, but no concrete requirements for labelling/IDs etc.
Excel to Beacon Friendly Format tool. BFF
VCF conversion tool. Re-Annotating VCF with SnpEff to have a homogenous nomenclature for all your variables. JSON output. This goes into MongoDB. Then Beacon API on top.
Key issue - transformation of clinical data into Beacon. https://convert-pheno.readthedocs.io - common clinical formats into Beacon Friendly and Phenopackets.

BS: We use your Ref Implementation. We say Beacon is a specification, but common researchers cannot set this up. This is not optional. People want to Beaconize, but we cannot say that it is just a specification. It should come with the implementation.
MR: We have a polarized community - experts - then we have the rest of the world. They don't know if this thing is going to pay off. We provide only a basic solution. Need work from the community.
MB: It's tricky. I am not using Ref Implementation. Obviously a push for data rich Beacons makes sense; however, one of our goals is the worldwide Discovery of data, before sharing the data itself. Simple / common phenotype parameters (diagnosis, genotypic sex …) etc. as common "onboarding" parameters. "Tweets" or "books" depends on resources, environments, permissions. Any resource environment with some development can stand up a website. For the all important "onboarding" of v2 Beacons: Keep it simple at the beginning. Explain what Beacon is. Then demonstrate / help with the implementation of a Beacon with filter for phenotype or disease code. No real need to recapitulate the complete data model, just a value responding to a certain query type.
KR (in chat): Great presentation. Love the clin/pheno transformation. I agree with your assessment that this is the bottleneck.
Gentle encouragement to align to standards by making recommendations on how to represent the data.
DS (in chat): Be careful with re-annotation of data. Davis McCarthy and colleagues shown in Genome Medicine 2014 that only 87% of exonic SNP were concordant using Ensembl transcipts between ANNOVAR and VEP (two only system tested at that time) and much more discrepancies with intronic variants
JJ (in chat):Something to help with transforming OMOP into Phenopackets, if people need to do this:
https://github.com/phenopackets/omop-exporter

MB (in chat): I think we should provide some very simple examples & document them, such as this "VCFs from patients w/ a phenotype".

DR: I did not find Beacon that complicated to implement. The hard part is the data model. My Java implementation separates interface and backend. If someone developed the same in Python this would benefit the community. Ref Implementation doesn't use Python model for data. If data is already in MongoDB - it doesn't help people understand the model. OMOP, BioPortal etc. Some Python package?
MB: We have this in Progenetix. Import templates and mappings. Would be good to have others helping. Most use cases are simpler. Onboarding is the most important. Models live in their own environment. Can benefit from sharing a subset of their data that is easily shareable. NCBI GEO resources have been remapped by us, but this is not easy. Do the things you can easily do correctly.

Michael - second slide deck slides 15+
Progenetix in 2022
Beacon sits under the UI
Our stack aggregates over all the data - intersection and individual annotations
Aligned with VRS and Phenopackets - experimental Phenopackets endpoint.
Bycon - python Beacon stack

BS: With B2RI - providing training. Could you do the same with bycon? CINECA training session/documentation.
MB: Getting data from some input format into something you can work with. Example input file - transformed by some magic into BFF by bycon. How to get from one to the other is less important than which data can we transform and what standards formats we use. How do you Beaconize "this" data? How do we make the data Beacon ready?
BS: Prepare use cases.
MB: VCFs from patients from a certain study cohort - attributes already there by being part of that cohort. Simple table.

KR (in chat): Just wanted to confirm that the training materials we've been discussing will become part of a GA4GH Starter Kit (if they're not already), right?
MB: BS - where are you sharing your documentation?
BS: We have this on the Beacon docs webpage. This is easier as it is mapping a table. We would need another training document for more complex use cases.
KR (in chat): Those 2 use cases would be a great starter kit. :)
MB: We have a multi sample - proprietary format. Observations and tagging them in one film without a DB. No standard for variant plus annotation. Don't want to go ahead with a use case that is "this would be good" without a real world example.

BS: **How to make sensitive data discoverable - Online workshop**
https://docs.google.com/document/d/1SIe0bwFz_vnmpluB0Xd2m5EEt4Qi6TubX3aL9vVtGzg/edit

Beaconize your data type

Discovery

Date: 2022-11-16
Time: 15:30-17:00 UTC (16:30 CET | 07:30 PST | 10:30 EST | 15:30 GMT | 02:30 AEDT)
Meeting Chair(s): Michael Baudis,
Abstract: Beacon version 2 is a much more powerful platform for data discovery due to its modular design, which allows almost any data type to be “beaconized”, including Structural Variant information and newer VCF formats. This is a working session where attendees can bring their own data types and discuss methods for “beaconizing” their data. The goal is to boost uptake of the Beacon v2 and demonstrate its potential to the community.

	Agenda Item	Speaker	Time
1.0	New query types - ranges & brackets	Michael Baudis	
2.0	Supported & unsupported variant types ● Progenetix ● reflecting on VCF 4.4	Michael Baudis	
3.0	Copy Number Variant Resource	David Salgado	
4.0	Filters for metadata and individual data - adapting the Reference Implementation	Vatsalya Maddi & Umar Riaz	

Attendees

Fabio Liberante (GA4GH), Heidi Sofia (NIH), Vatsalya Maddi (Leicester), David Salgado (INSERM, IFB, ELIXIR-FR), Melanie Courtot (OICR), Tshikala Eddie Lulamba, Dmitry Repchevsky (BSC), Sergi Aguiló Castillo (BSC), Manuel Rueda (CNAG-CRG)

Notes/Links

Zoom recording
https://us02web.zoom.us/rec/share/X6D3Pn65UVJib54m2FaebwW7GZvL6U7e3wvSapcBYP3dPc8xRyJNwk1IRE9TK38H.RLMpm9q7DuzJKmsR
Passcode: 2X..+u%A

MB - Intro - Slide 22+ here.

Ranges

https://docs.genomebeacons.org/variant-queries/#beacon-range-queries
Brackets, e.g. CNVs of a given location & size
https://docs.genomebeacons.org/variant-queries/#beacon-bracket-queries

Cytoband translation? Trisomy?
=> e.g. https://progenetix.org/services/cytomapper/?cytoBands=21&assemblyId=GRCh38
for getting coordinates (21:0-46709983) & use them in a query (no cytoband support in Beacon v2 although HGVS is in principle supported but no known implementers…)

Key takeaways

- Beacon does not report on missing filtering terms, only "0" results, this needs to be addressed somehow
- Community would benefit from a "safe" Beacon concept with input from REWS

Summary

- Beacon can interpret queries in different ways, this means that each query type does not necessarily need its own data model
- We need greater feedback from (would-be) implementers on their use cases
- Beacon can be built in different ways - e.g. PHP on PostgreSQL or Python/Django on MongoDB
- Some groups have begun implementing Beacon's filter capabilities
    - This helps standardize query-response as pairs, but is considerable work
    - Solve-RD working to create "standard" filter set for consortium
- Docker instances have helped implementers understand the specification
- Again Discovery focus of Beacon - it doesn't allow chained Boolean response, so research questions are difficult e.g. cannot ask x<y<z
- No standard way to share "What queries/IDs/filters does this Beacon support?"
    - Info endpoint supports information, as in "Patient age", but not type, e.g. integer
    - Is there a risk of data leak with open sharing of Beacon query structure?
- Variant types is not prescribed, and no way to share this
- The Beacon documentation could be better and resources aggregated
- If a filtering term doesn't exist - a Beacon does not report missing term, only "0" results, so querying across multiple Beacons is impossible if the presence of filters varies
    - Again if you openly report which terms do not exist, this could be a data leak
    - Tony Brookes is collaborating on a potential solution to this
- A "safe" Beacon concept could be created, with input from Legal/Ethics teams
    - Limited queries etc. as *implementation* considerations
-

Minutes

MB - Intro - Slide 22+ here (here as backup).

How to move from a simple SNP query into something more substantial.
V1 to V2 query comparison.
“variantType” focus now. CNV - copy number variants.
CNVs come in many different formats
Confusing “INFO” field. VCF v4.4 tries to disambiguate this. Allows imprecision. (Timothe Cezard)
How is all this interpreted in Beacon queries?
SVCLAIM field - allows qualification as to how the CNV is supported by evidence.
https://docs.genomebeacons.org/variant-queries/#term-use-comparison
Query types vs. data models - interpretation of query does not mean a new model for each.

Want feedback by implementers on their use cases.

David Salgado - BANCCO - Beacon Implementation - **slides**
BANCCO - French national DB for CNV. 10 participating centres. Array CGH data.
http://bancco.fr
BANCCO+ - increase interoperability and to collect NGS data. 100k patients by 2028
Beacon in BANCCO - Beacon4CNV project - used Beacon v2 beta specification
API Laravel 6 PHP on top of PostgreSQL
Starting on range and bracket queries and gene ID queries
Implementation of some filters
Docker instance created by Jordi’s group was very important during our implementation - test data helped us play with the system

Vatsalya “Rinni” Maddi & Umar Riaz - University of Leicester experience
Vatsalya “Rinni” Maddi - How we used Beacon as a solution.
What, why and how?
Formats numerous and complex.
Beacon uses filters to help us standardize output.
Beacon framework - Requests & Filters
Cafe Variome
Specify a standard set of filters - requests and responses are paired.
LeHMR & Solve-RD - Umar Riaz
LeHMR
A dataset query UI with query builder.
We have only metadata. Mapped this into BFF.
Wrote my own script to generate json
Solve-RD
Individual level filters - security key to make these queries
Collaborating with others to create a standard filter set.
www.cafevariome.org

Discussion

MB: Example - Beacon does not allow you to chain Boolean response, but you could have things like a negated filter. Could be possible, but not easily implemented by most. We use only filters for our databases. Where do people see gaps or conceptual problems?
GK (in chat): One issue I've had is that querying is so open that it's not clear, on encoutering a new beacon, what kinds of filters I can use, (ontology term query, "alphanumeric" query, etc) and what the exact format is
MB: Every Beacon can provide terms via Info endpoint. Can be tricky - security concerns. You may gain information via open filters. But the technology is there to share these. When do you do the translation?
GK: Doesn't the endpoint tell me the information, but not the terms themselves, e.g. alphanumeric.
MB: We don't use alphanumeric.
VM: filtering terms - depends on data type
GK : Discovery issue.
MB: Alphanumeric. What type of query would be supported?
TB: This is version 1 of Beacon v2. This is great feedback. Probably need to be more precise in our definitions of the filtering terms.
MB: We don't have a variant query types - as it was seen as too prescriptive. But this is a problem now. Will it support this range query, but the protocol doesn't inform if it supports this or not.
TB: You've all taken the Beacon standard and implemented it. Some project specific things, some more general. What was the biggest bottleneck to having a live service?
DS: Identify the resources - where the information was? Little documentation about v1 - GitHub - various repositories and lots of documentation about the new specification. We have discussed with Jordi and Michael - which will be the next one.
MB: It is better now. We merged into one repository.
TB: Do you agree?
DS: It is clearer now.
MB: One big problem with our documentation. Developers did the documentation. Not necessarily the best information for newcomers.
TB: Anyone else?
VM: In final stages of implementation - Beacon is flexible. We keep getting back 0 terms. Making a very standardized set of filters - so you get something back as a researcher.
MB: Excellent issue - if you use a filtering term - it is not defined if the term is checked if it exists at all. If filtering term query in this domain - then provide this response.
TB: Big issue for across sites. 12 elements agreed, but won't be present across all of them. Possible solution to that we will present at next Beacon meeting

DS: Another bottleneck. Sometimes when you make a choice as an implementation - you don't know if you are doing something that the Beacon network can address on your behalf. Should I be allowing more complex queries?
TB: This is at a higher level than the spec. Could have some suggestions in future documentation.
MB (in chat): i.e. partial matches w/ indication which elements did not exist...
MR: I have always heard "We should not be prescriptive" - people can't fly without a recipe.
MB: We see this also in Phenopackets. We don't like your ontology etc. then eventually more prescriptive. Social process and documentation process. Beacon being a very open framework is not always the best solution. Could it be better to scope a bit more narrow?

TB: We've made it flexible. Like a wheel. GA4GH invented a wheel. Car, trolley, train. I think it was correct to leave them open. If you put Beacon on top of your dataset - we could have put a security layer, but we deliberately didn't. We'll get a myriad of ways of how to do this. As a community - whether we shouldn't make a more limited version Beacon - someone can't do serial queries, or age in ranges etc. If you follow this recipe it's a "safer version".
MB: Problem not just Beacon, but also data types. Many would present problem if patient information is there. Important to delineate things as a potential security risk. While Beacon v1 was well understood, considering this otherwise it's more secure etc. Beacon v1 already breaks an envelope.
TB: It would be nice to have a suggestion for a safe Beacon v2 design. Anyone else wants to bring their challenges? Please let us know. Offer this to the world.
MB: Some use case scenarios - what could go wrong? But you cannot make it secure in all cases. HIV database etc. Depends always on the data how you break it. Useful to describe these scenarios. Perspectives from Policy and legal people too.

GK (in chat): some things like rate limiting seem like implementation issues rather than belonging explicitly in the spec
TB (in chat): @Gordon - yes, it spans safe/obfuscating data, relative risk of different filters, security wrappers (e.g., AAI), rate limiters, tracking users queries, and other implementation options