# Proposed beacon v2 schema changes discussed today and maybe a couple other suggestions

color code>

**Endpoint name either old or new**

endpoint field added today in discussion

endpoint field that is kept exactly the same as in pdf

my comments

new endpoint field that I am suggesting now but we didn’t discussed today, or I’m suggesting to change name

**Organism**

taxonId: taxon ID of species from where variants come from, for example, virus and not human in the case of viral beacons) (where it goes is pending, maybe in new Organism endpoint-and a new table in database where all taxons their refseq ids and stuff are stored or in Run endpoint?)

**Individual**

individualId: alphanumeric ID

datasetId: alphanumeric ID

taxon\_id: alphanumeric ID (reference taxon ID for this individual human, animal or plant)

sex: categorical value (ontology ID)

ethnicity: categorical value (ontology ID)

geographicOrigin: categorical value (ontology ID)

phenotypicFeature: list of

phenotype: categorical value (ontology ID)

level/severity: categorical value (ontology ID): (see explanation below in same field for diseases)

diseases: list of

disease: categorical value (ontology ID) > to include ontologies for diseases of other species as well

dateOfOnset: alphanumeric value (ISO8601 duration format)

ageOfOnset:

age: alphanumeric value (ISO8601 duration format)

ageGroup: categorical value (ontology ID)

stage: categorical value (ontology ID) > stage should be kept general, disease and species agnostic: “acute”, “sporadic”, “chronic” and “recurrent”….and “remission” , “solved” maybe .see what applies to plants and animal diseases as well in order to suggest other ontology(s)

level/severity: categorical value (ontology ID)> Now that we decided not to remove relevant stuff I would add this to general beacause can be useful in other cases and it’s useful in covid this category of mild vs severe disease. Also phenopacket has it I think, al least for phenotypic feature

outcome categorical value (ontology ID) Outcome of disease e.g fatal or non-fatal

familyHistory: boolean

pedigrees: list of

pedigreeID: alphanumeric ID

disease: disease format (list)

pedigreeRole: categorical value (ontology ID)

numberOfIndividualsTested: numeric

info

General Comments from pdf on Individual endpoint:

1. Keep Individual endpoint like this,.. only for Plant beacons or others with no human metadata this endpoint is not used; to make distinction when this is or not the sequenced species but Host or other interactor species…the info comes from taxon\_id in the Database or other dedicated endpoint ,…if taxon id is other species, then this individual data is individual host / interactor of the main species of interest

2. Alternatively, one Individual can be an individual Plant or an individual Animal…only some fields like ethnicity will not be used in those cases > we could rename the endpoint as Subject instead ….

Note that alternatively, this human Individual endpoint implies that endpoints for Plant, Fungus, etc should be created with their own relevant fields.> so I advocate for just one general one for all Organism be it the sequenced ones or hosts of sequenced pathogens.

Note that non-human studies where also no host organisms animals or plants or fungi…but only pathogens such as viruses or bacteria are studied independently of any host data…for example, viral fitness in cell culture…this Individual or Subject endpoints will be empty. Or shall we deal with cell lines or rather bacteria or viruses themselves as Individuals/Subject? Cell lines certainly are not dead matter but have their own genomes&transcriptomes and thus phenotypic features that can influence and being influenced by infecting entities e.g on viruses some cell lines have and other lack IFN-I antiviral cascade, or other permisive/limiting factors for viral entry, replication, etc.

**Biosample**

biosampleId: alphanumeric ID

individualId: categorical value (ontology ID)

description: free text

biosampleStatus: categorical value (ontology ID)

collectionDate: alphanumeric value (ISO8601 duration format)

IndividualAgeAtCollection: alphanumeric value (ISO8601 duration format)

sampleOriginType: categorical value (ontology ID)> This is to specify categories from which sample comes from: “organism primary tissue”, “organism xenograft”, “organism-derived fluid”, “cell culture”, “environmental sample”, “mixed” , “unknown” (check Ontology values for those) and then matching ontologies to those adding specific instances: eg, if cell culture, ontology specifying cell line. Let’s say here it is set to “Cell culture”

sampleOriginDetail: categorical value (ontology ID) > This is where the specific matching instances go e.g HEK293T (or specifically, CL ontology ID)

obtentionProcedure: categorical value (ontology ID) > just make broader now, to include other obtention procedures unrelated to human samples such as “field collection”, “culture cells sorting” or “microorganism isolation/purification” etc> I am to look for good ontologies capturing all this to suggest in spec

cancerFeatures: list of > I understood this was to be kept, right?

tumorProgression

tumorGrade

info

**Variant Annotation**

variantId: alphanumeric ID

genomicHGVSId: alphanumeric ID (HGVSId descriptor at genomic level) > these Im not sure are used for all species

transcriptHGVSId: alphanumeric ID (HGVSId descriptor at transcript level) > Not sure why we didn’t have it, databases usually expose the three levels

proteinHGVSId: alphanumeric ID (HGVSId descriptor at protein level)

genomicRegion list of

class: categorical value (ontology ID) eg: protein coding, intergenic, untranslated region

featureID: > id either genes, genomic regions, subgenomic regions, transcripts, and proteins that are affected by the variant: genomic region ref seq accessions (NC, NM, YP )and map to their aliases or names such as ORF1, 3UTR, S1

annotationToolVersion: alphanumeric value e.g SnpEffVersion=4.3t (build 2017-11-24 10:18)

molecularEffect: categorical value (ontology ID) (here will come predicted effect at nucleotide level eg: “STOP\_GAINED” as opposed to the description at protein level for protein affecting variants eg. “Nonsense” that goes into molecularConsequence)

molecularConsequence(or change name to functionalClass): categorical value (ontology ID)

aminoacidChange: string (this comes direct in vcf and needs transformation to be shown as proteinHGVSId, and also maybe this endpoint it’s valuable to show like this rather than the other one. The one could be an ‘alias’ on top on this one plus refseq for user to make queries)

phenotypicEffect: list of

phenotypeId categorical value (ontology ID) everything that is not a disease

phenitypeEffect categorical value (ontology ID). Phenotypic effect classification

references list of PMIDs

clinicalRelevance: list of

disieaseId: categorical value (ontology ID)

clinicalEffect: categorical value (are there ontologies?, I am to search this)

references: list of PMIDs

allelleOrigin: categorical value (ontology ID)

info

General comments on Variant annotation endpoint:

1. Still doubts on where to put effect of variants from pathogens’ point of view, i.e in the fitness of pathogens themselves (these would to a different endpoints> eg SARSCoV2 variants can be annotated as clinicalRelevance: “COVID19 pneumonia” and its own phenotpic feature: increased patogenicity or transmisibility or increased host range)

**Run**

runId: alphanumeric ID (external accession) e.g "SRR10903401"

experimentId: alphanumeric ID (external accession) e.g "SRX7571571"

librarySource: categorical value e.g “Metagenomic”, “Viral RNA”

libraryStrategy: categorical value e.g “WGS”

librarySelection: categorical value e.g “RANDOM”, “RT-PCR”

libraryLayout: categorical value e.g “PAIRED” “SINGLE”

platform: categorical value e.g “Illumina”, “Nanopore”

platformModel: categorical value e.g “Illumina MiSeq” , ”GridION"

info (or handover maybe):

experiment\_info:

experimentId: alphanumeric ID (external accession) e.g "SRX7571571"

experimentTitle: string e.g ”Total RNA sequencing of BALF (human reads removed)”

study\_info: alphanumeric ID (external reference)

studyId: e.g "SRP242226"

studyRef: PMIDs

General comments on Run endpoint:

1. Note that all but rundId belongs naturally to experiment and there will be lot of duplications for all runs coming from same experiment, we might include Experiment endpoint or at least this might be a different table in database (well, this is for database experts to decide)

**Variant in Sample**

variantId: alphanumeric ID

runId: alphanumeric ID

variantCaller:

biosampleId: alphanumeric ID

individualId: categorical value (ontology ID)

variantFrequency: numeric value

zygosity:

alleleOrigin:

clinicalRelevance: list of

disieaseId: categorical value (ontology ID)

clinicalEffect: categorical value (are there ontologies?, I am to search this)

info

**Encounter (phenopacket)**

encounterID: alphanumeric ID

encounterDate: alphanumeric value (ISO8601 duration format)

ageAtEncounter:

age: alphanumeric value (ISO8601 duration format)

ageGroup: categorical value (ontology ID)

clinicalFindings: (non quantifiable or non quantified findings)

id: categorical value (ontology ID)

level/severity

measurements:

id: categorical value (ontology ID)

value: numerical value

units: categorical value (ontology ID)

info

General comments on Encounter endpoint:

1. This is a suggested new endpoint that is the equivalent to one phenopacket (on encounter or medical visit, it’s lacking other phenopackets stuff like everything that is outside like Biosample, Diseases, but they are mappable with this encounter through date or age values. I suggest for other things such as clinicalFindings and measurements (clinical or not) that have observation timelines associated which can come in handy for hospital data, say, for filtering by clinicalFindings of “arrithmia” at age X or at time Y after treatment Z (well, treatments is still missing from here and from phenopackets but we could include as well in Individual endpoint or a dedicated endpoint something like this—it’s something pending for EGA and needed for CGAT )

treatment:

dateAtOnset: alphanumeric value (ISO8601 duration format)

ageAtOnset:

age: alphanumeric value (ISO8601 duration format)

id: categorical value (ontology ID) eg. chemotherapy

dose: numerical value

units: categorical value (ontology ID)

schedule: free text for now eg. 3/week

duration: alphanumeric value (ISO8601 duration format)

intervention:

inverventionDate: alphanumeric value (ISO8601 duration format)

age: alphanumeric value (ISO8601 duration format)

id: categorical value (ontology ID) eg. vasectomy