

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

Note: whatever is in automatic/black font is kept the same as in current version, things in blue font means additions or changes. Things marked with gray boxes (■) are fields currently in viral beacon (the mapping of those into the corresponding fields here is in https://github.com/clauw87/virusbeacon/blob/raw_ideas/virus_beacon_schema_v1_to_generic.md)

Variant basic ■

refAssemblyId

startPos

endPos

ref

alt

variantType Variant classification e.g SNV, indel, CNV, structural variant

Organism ■

taxonId categorical value (ontology ID). Taxon ID of species from where variants come from, for example, the SARSCoV 2 taxon id and not human's in the case of viral beacons)

Individual

individualId ■

datasetId

taxon_id categorical value (ontology ID) (reference taxon ID for this individual human, animal or plant)

sex ■

ethnicity




geographicOrigin ■

phenotypicFeatures list of

phenotypId categorical value (ontology ID) Phenotypic feature observed (not disease)

level/severity categorical value (ontology ID) Level/severity when and as applicable to phenotype observed e.g “mild”, “severe”




diseases list of

diseaseId 
dateOfOnset alphanumeric value (ISO8601 duration format) Date of onset/diagnosis of disease
ageOfOnset
 age alphanumeric value (ISO8601 duration format)
 ageGroup categorical value (ontology ID)
stage  categorical value (ontology ID)
outcome  categorical value (ontology ID) Outcome of disease e.g fatal or non-fatal
level/severity categorical value (ontology ID) Level/severity when and as applicable to disease observed e.g “mild”, “severe”
familyHistory
treatments list of
 treatmentId categorical value (ontology ID) eg. “chemotherapy X”
 dateAtOnset alphanumeric value (ISO8601 duration format)
 ageOfOnset
 age alphanumeric value (ISO8601 duration format)
 dose numerical value
 units categorical value (ontology ID)
 schedule free text for now eg. “/week”
 duration alphanumeric value (ISO8601 duration format)
interventions list of
 InterventionId categorical value (ontology ID) eg. “vasectomy”
 date alphanumeric value (ISO8601 duration format)
 ageAtIntervention alphanumeric value (ISO8601 duration format)
 age alphanumeric value (ISO8601 duration format)

pedigrees list of
 pedigreeId
 disease disease format
 pedigreeRole
 numberOfIndividualsTested

info

Biosample

biosampleId 
individualId 
description
biosampleStatus
collectionDate  alphanumeric value (ISO8601 duration format). Date at which sample

is collected.

IndividualAgeAtCollection

sampleOriginType categorical value (ontology ID) Category of sample origin e.g “organism primary tissue”, “organism xenograft”, “organism-derived fluid”, “cell culture”, “environmental sample”

sampleOriginDetail categorical value (ontology ID) Specific instance of sample origin matching the category set in sampleOriginType e.g “HEK293T”, “nasopharyngeal swab”
obtentionProcedure categorical value (ontology ID)

cancerFeatures list of

tumorProgression

tumorGrade

info

Variant Annotation

variantId

genomicHGVSId

transcriptHGVSId alphanumeric ID (HGVSId descriptor at transcript level)

proteinHGVSId

genomicRegion list of

class categorical value(s) (ontology ID) Class of genomic regions altered by the variant eg “protein coding”, “intergenic”, “untranslated region”, “transcript”

featureID categorical value(s) IDs matching class (of genes, genomic regions, subgenomic regions, transcripts, other RNA species and proteins that are affected by the variant names or genomic region ref seq accessions (NC, NM, YP))

annotationToolVersion alphanumeric value. Tool used for annotation and prediction of variant effects e.g “SnpEffVersion=4.3t (build 2017-11-24 1018)”

molecularEffect categorical value (ontology ID) 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

aminoacidChange string. Change at aminoacid level for for protein affecting variants eg. “V304*”

phenotypicEffect categorical value (ontology ID) Annotated effect on any phenotypic feature other than a disease

phenotypId Phenotype associated

phenotypicEffect categorical value (ontology ID). Phenotypic effect classification

references list of PMIDs

clinicalRelevance list of

disieaseId

clinicalEffect previously variantClassification

references

alleleOrigin list of
info

Run

runId ■ alphanumeric ID (external accession) e.g "SRR10903401"
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 Sequencing platform group e.g "Illumina", "Nanopore"
platformModel ■ categorical value Sequencing platform model e.g "Illumina
MiSeq", "GridION"
info (or handover maybe)
 experiment_info ■
 experimentId ■ alphanumeric ID External experiment accession e.g
"SRX7571571"
 study_info
 studyId alphanumeric ID External study reference/accession e.g
"SRP242226"
 studyTitle ■ string e.g "Total RNA sequencing of BALF (human reads
removed)"
 studyRef list of PMIDs

Variant in Sample

variantId ■ alphanumeric ID
runId ■ alphanumeric ID
variantCaller ■ categorical value e.g GATK vxx
biosampleId ■ alphanumeric ID
individualId ■ categorical value (ontology ID)
variantFrequency ■ numeric value
zygosity
alleleOrigin
clinicalRelevance list of
 diseaseId categorical value (ontology ID)
 clinicalEffect categorical value (ontology ID)
info

Encounter

[encounterID](#) alphanumeric ID

[encounterDate](#) alphanumeric value (ISO8601 duration format) Date of encounter/medical visit

[ageAtEncounter](#)

age alphanumeric value (ISO8601 duration format)

ageGroup categorical value (ontology ID)

[clinicalFindings](#) Non quantifiable or not quantified clinical findings

[finding](#) categorical value (ontology ID) eg: "arrhythmia"

[level/severity](#) categorical value (ontology ID) e.g "mild"

[measurements](#) (list of) measurements taken during encounter

[id](#) categorical value (ontology ID)

[value](#) numerical value

[units](#) categorical value (ontology ID)

[info](#)