The comprehensive virus genome database based in Jena

Shahram Saghaei[1,2], Noriko Cassman[1,2], Hamed Ziraksaz[1], Manja Marz[1,2,3,4]

Introduction

Bioinformatics research in virology requires the support of prior knowledge on viruses. Databases allow for efficient access to multidimensional, structured data due to a large volume capacity, simultaneous access for multiple users, and search and sort functionality on the data. Currently, virus databases vary in terms of FAIRness and can contain errors propagated from user-submitted data in primary sequencing repositories. The goal of the VirJenDB (VJDB) group is to develop an umbrella database covering all viruses in a single platform which will provide user-friendly access to curated virus sequences and metadata data for downstream analyses.

~12.5M viruses

35 metadata fields

~10K Species

Methods & Results

The VJDB runs on the de.NBI[5] cloud as an OpenStack service consisting of multiple instances. Sequences and metadata are hosted on the Aruna Object Storage system. We ingested data from several sources including the ICTV[6], NCBI Virus[7] and the BV-BRC[8] and integrated these into a virus data model using a MySQL database in the backend. The backend connects to the front end through a RestAPI and the code is developed via GitHub. The resulting merged dataset is available through the web interface. The VJDB beta version v0.1 is currently available (see below) for users to search, browse, select, and download genome sequences and metadata.

Discussion

Figure1- Data insertion

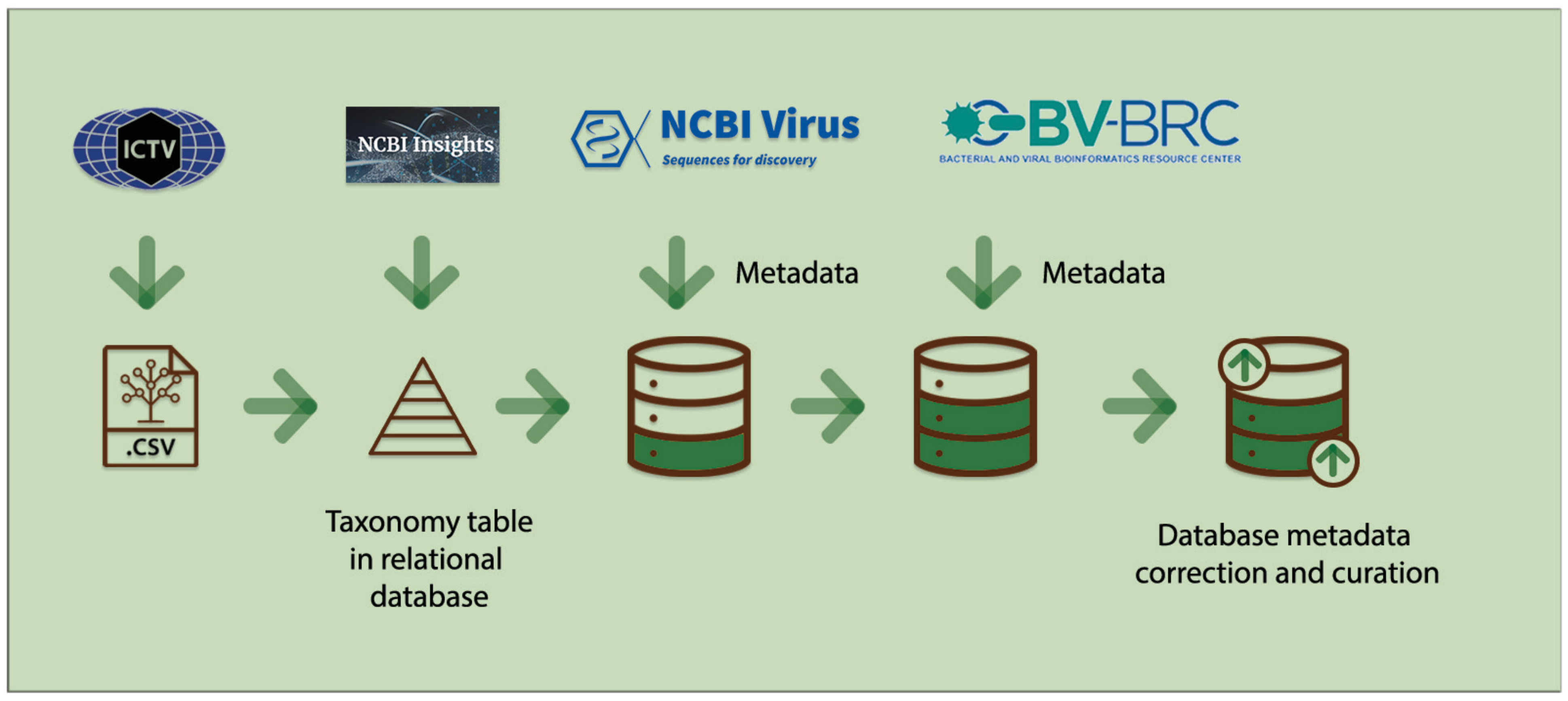

ICTV
NCBI Insights
NCBI Virus
Sequences for discovery
BV-BRC
BACTERIAL AND VIRAL BIOINFORMATICS RESOURCE CENTER
Metadata
Metadata
.CSV
Taxonomy table in relational database
Database metadata correction and curation

Figure2 - Network structure of the service

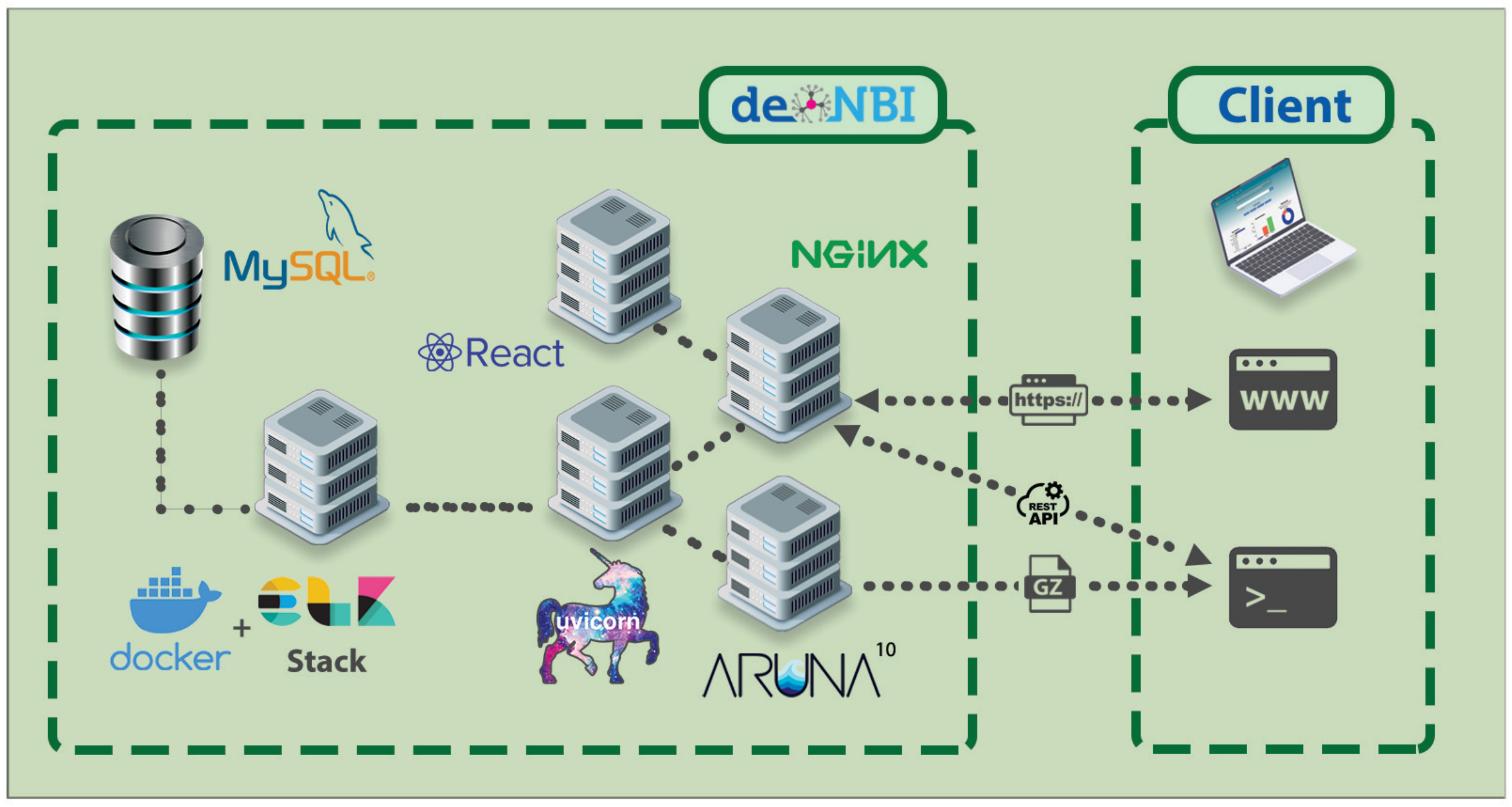
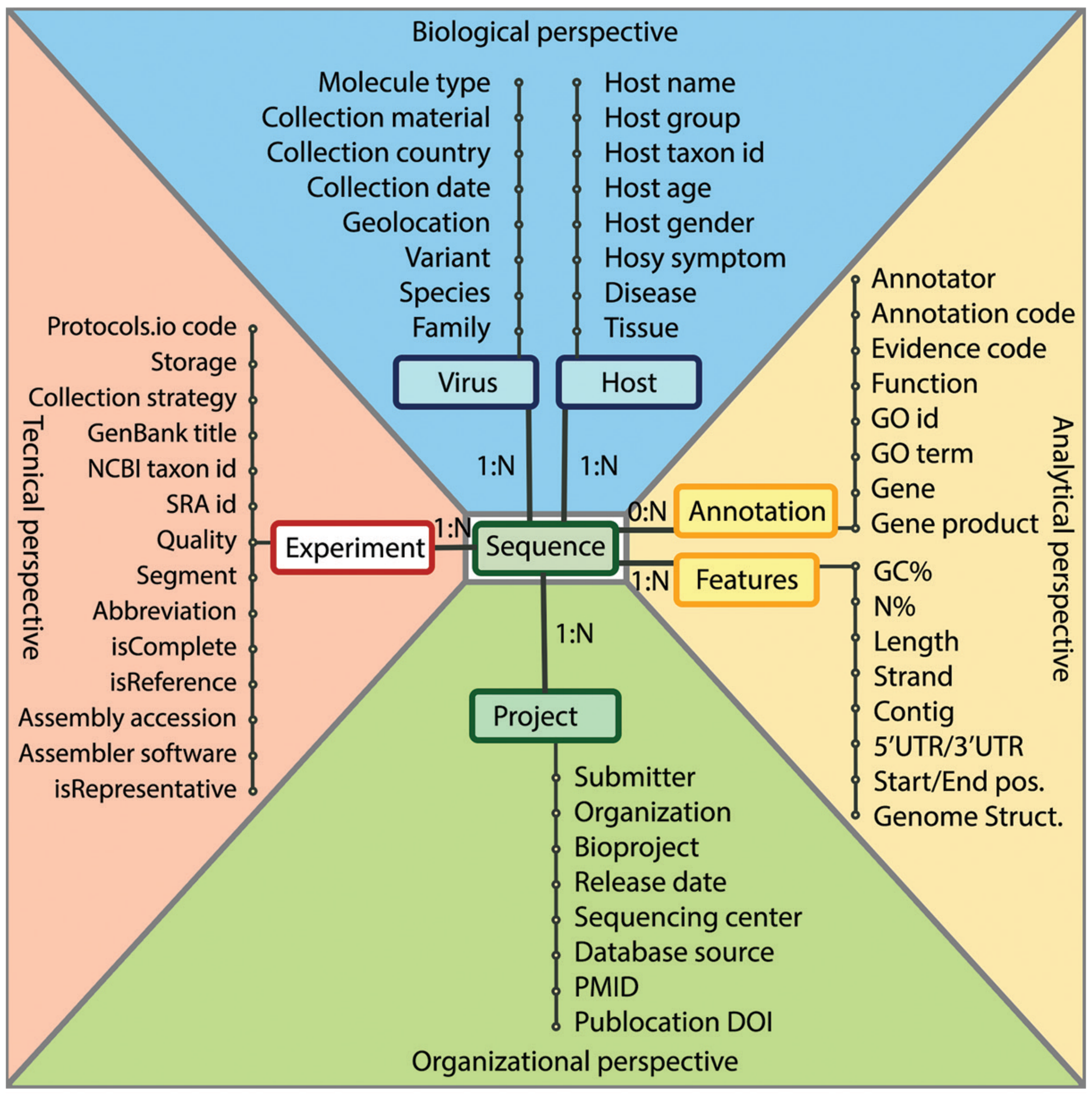


Figure3 - Metadata schema



Biological perspective

Molecule type
Collection material
Collection country
Collection date
Geolocation
Variant
Species
Family

Host name
Host group
Host taxon id
Host age
Host gender
Hosy symptom
Disease
Tissue

Virus
Host

Tecnical perspective

Protocols.io code
Storage
Collection strategy
GenBank title
NCBI taxon id
SRA id
Quality
Segment
Abbreviation
isComplete
isReference
Assembly accession
Assembler software
isRepresentative

Experiment
1:N
Sequence
1:N
1:N
0:N
Annotation
1:N
Features
1:N
Project

Analytical perspective

Annotator
Annotation code
Evidence code
Function
GO id
GO term
Gene
Gene product

GC%
N%
Length
Strand
Contig
5'UTR/3'UTR
Start/End pos.
Genome Struct.

Submitter
Organization
Bioproject
Release date
Sequencing center
Database source
PMID
Publocation DOI

Organizational perspective

The data collected from the different sources must be merged, standardized and harmonized with existing standards (e.g. GSC MIxS Vi and UViG). Manual metadata curation and correction on the VJDB v0.1 dataset are in progress while community and automatic curation are planned. Further developments planned for the upcoming year include automating data ingestion and adding data sources. With the help of the infrastructure provided by the de.NCBI and the support of the NFDI4Microbiota[9], various services are being prepared to establish security.

Outlook

The VJDB is the comprehensive virus database developed in Jena, Germany as a service of the NFDI4Microbiota in line with the FAIR and Open Science principles. The VJDB team aims to provide an analysis platform for researchers to access, analyze, visualize and download curated sequences and metadata from all viruses. Further, we aim to help educate and help train researchers build data management skills using virus data resources. The target researcher groups are non-bioinformatician virologists and bioinformatician phage ecologists. The beta version v0.1 is available at https://virjendb.uni-jena.de/.

Contacts: shahram.saghaei@uni-jena.de, noriko.cassman@uni-jena.de.

References

[1] Bioinformatics/High-Throughput Analysis, Faculty of Mathematics and Computer Science, Friedrich Schiller University Jena, 07743 Jena, Germany
[2] European Virus Bioinformatics Center, 07743 Jena, Germany.
[3] FLI Leibniz Institute for Age Research, Jena, Germany.
[4] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany.
[5] de.NBI infrastructure, https://www.denbi.de/literature
[6] Virus taxonomy, https://ictv.global/ PMID: 29040670.
[7] NCBI Virus, Brister JR *et al.,* 2015 Jan, DOI: 10.1093/nar/gku1207
[8] BVBRC, Pickett BE *et al.,* 2012 Nov, DOI: 10.3390/v4113209.
[9] DFG project #460129525 in NFDI4Microbiota https://www.dfg.de/nfdi/
[10] Aruna Object Storage (AOS), https://aruna-storage.org/

VirJenDB Ver: 0.1 · NFDI4 MICROBIOTA · FRIEDRICH-SCHILLER-UNIVERSITÄT JENA · de.NBI GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE · DFG · RNA BIOINFORMATICS & HIGH-THROUGHPUT ANALYSIS