bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io


Creative Commons
This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu
creative commons
Attribution-Share Alike 2.5 Canada
You are free:
to Share — to copy, distribute and transmit the work
to Remix — to adapt the work
Free Cultural
APPROVED FOR
Works
Under the following conditions:
Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.
For any reuse or distribution, you must make clear to others the licence terms of this work.
Any of the above conditions can be waived if you get permission from the copyright holder.
The author's moral rights are retained in this licence.
Disclaimer
Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French
Learn how to distribute your work using this licence

Module 3: Tools and processes for infectious disease genomic epidemiology data curation and sharing

bioinformatics.ca

Emma Griffiths
Infectious Disease Genomic Epidemiology
May 14, 2024

CENTRE FOR INFECTIOUS DISEASE GENOMICS AND ONE HEALTH

SFU

Centre for Infectious Disease Genomics and One Health

Faculty of Health Sciences,
Simon Fraser University

Learning Objectives

By the end of this lecture, you will:

1. Understand the challenges of using genomics contextual data for public health analyses

2. Be able to describe the importance of data curation in public health

3. Know how ontologies, data standards and tools can be used as solutions for streamlining data flow

4. Be able to describe real-world examples of how ontology-based specifications are used

5. Be aware of data sharing principles, considerations (practical, ethical, privacy)

6. Know about different public repositories (GISAID, INSDC) and their submission requirements

Contextual data is critical for interpreting the sequence data.

Sequence data

Contextual data

Sample metadata

Lab results

Clinical/Epi data

Methods

Contextual data (metadata) used for **surveillance** and **outbreak investigations**:

- **characterize** lineages, sequence types, clusters
- identify variants with **clinical significance**
- correlate genomics trends with **outcomes, risk factors**
- **Monitoring and quality control**
- **Comparing results** between laboratories
- **Generating hypotheses** about sources of infection/transmission etc

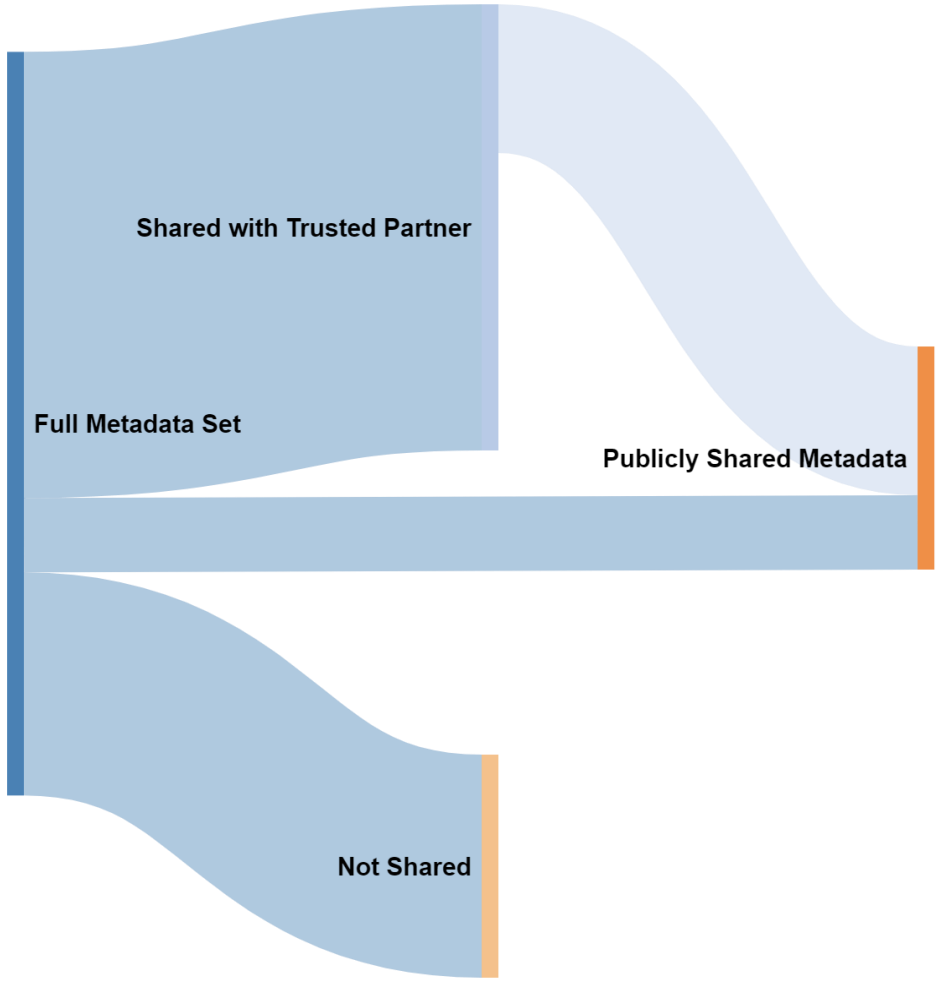
- **inform decision making** for public health responses and **monitor effects of interventions**

Why are we learning about data curation and standards in this course?

- **Establishes a quality framework for contextual data**
  - -auditability, chains of custody
  - -future-proofing assets
- **Part of best practices**
  - -WHO (Guiding Principles for Genomic Data Sharing, 2022)
  - -FAIR (Findable, Accessible, Interoperable, Reusable)
- **Improves efficiency**
  - -save time/money (80:20 data science rule)
  - -human/machine-readability
  - -streamlines data exchange
- **Empower investments contributing to local/national/global surveillance**
  - -breaking down barriers (coordination, communication, data flow)

There are different kinds of data sharing.

- Data comes from **different sources** (labs, departments, databases)
- Data needs to be shared within **organizations**, with **trusted partners**, with **public repositories**, with **international agencies**
- **Everyone uses different systems, processes**



Heterogeneity of values within a field also complicates using the data.

Free text = 

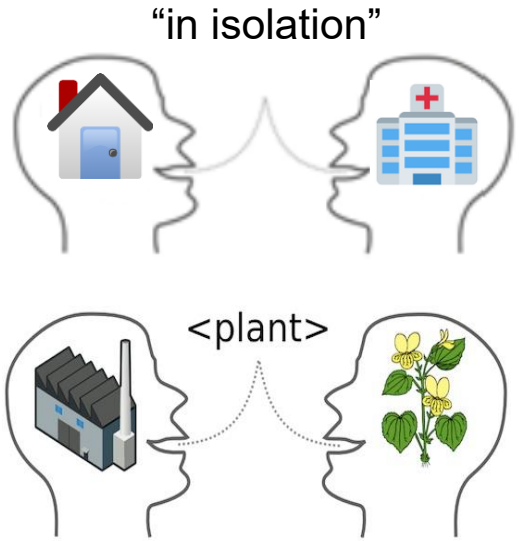
Errors, Jargon, Short hand

In-pateint
Np swad
UTM
NPS

Granularity

Cough
vs
Dry cough
vs
Productive cough
vs
Cough with green phlegm

Semantic ambiguity



Formats

Date:
2021-04-26
April 26, 2021
26-Apr-2021

Different Classifications/ Concepts

RISK FACTORS:
Exposures vs
Pre-existing conditions

Variability in private databases propagates out to public repositories, complicating data integration/analyses.

isolate	SARS-CoV-2/186197/human/2020/Malaysia
collected by	Universiti Malaya COVID Research group
collection date	14-Mar-2020
geographic location	Malaysia
host	Homo sapiens
host disease	COVID-19
isolation source	Nasopharyngeal/throat swab
latitude and longitude	3.1390 N 101.6869 E

source name	Lung sample from postmortem COVID-19 patient
cell type	Lung Biopsy
strain	NA
subject status	No treatment; >60 years old male COVID-19 deceased patient

Data structure impacts function.

It’s difficult to fit it all together. Data clean up takes time, resources.

Solutions to common contextual data challenges

1. Ontologies
   universal language for humans and computers
2. Data standards
   prescribed sets of fields, terms, formats
3. Tools
   software and supporting materials to implement standards
4. Consensus, Coordination, Education
   awareness, uptake & implementation, utility, ease-of-use

Ontologies: Built for harmonization and data linkage

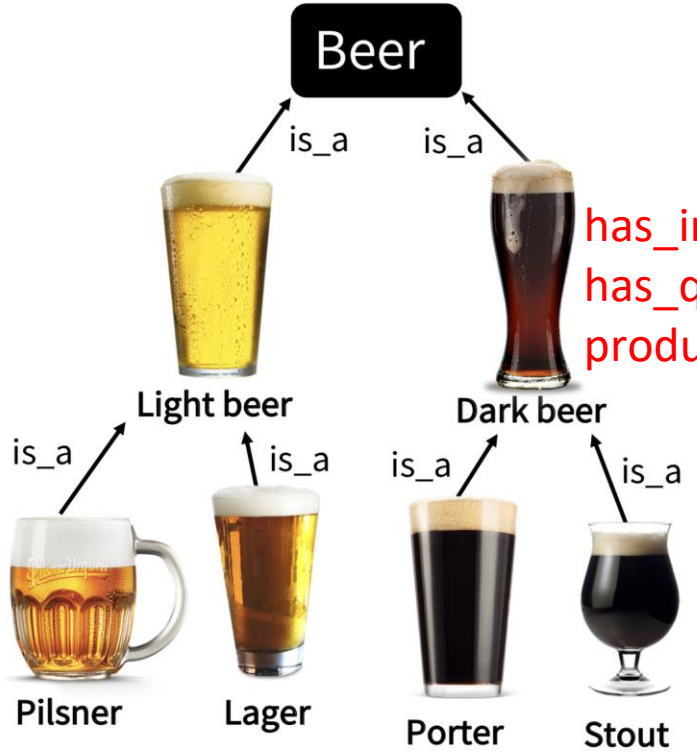
Controlled (standardized) vocabulary
Hierarchy + logic (linked data, enable classification for analyses)

Universality

- Meanings disambiguated with URIs
- Labels/Synonyms (organization-specific/interoperability)
- Principles and practices to enable reuse (BFO, RO)

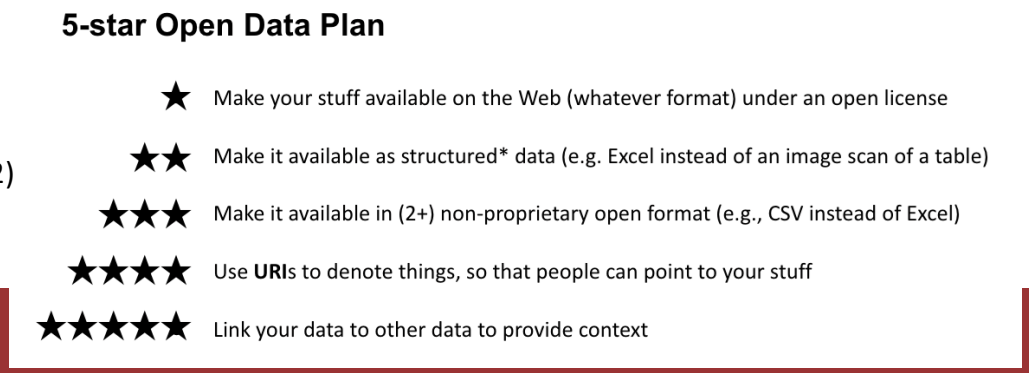
Community

- Community of practice (OBO Foundry, >200 interop ontologies)
- **Registries/Portals** (EBI OLS, Ontobee, BioPortal)
- **Languages/Tools** (Protégé, LinkML, Robot, OntoFox)



Lager beer [BeerO:1234]

FAIR



Hausenblas & Kim (2012)
Berners-Lee (2009)

Standards: ISO 23418:2022

Microbiology of the Food Chain — Whole genome sequencing for typing and genomic characterization of foodborne bacteria — General requirements and guidance

Contextual Data Fields
Sample Collection Lab Contact Information
Geographic Location of Sample Collection
Collection Date
Sample Type
Food Product
Food Processing
Environmental Material
Environmental Location
Collection Device
Collection Method
Microbiology Lab Contact Information
Organism
Strain
Isolate
Serotype
Isolation Media
Isolate Passage History

ISO standard provides tables and annexes to describe…

1. Information about the sample
2. Information about the isolate
3. Information about the sequence

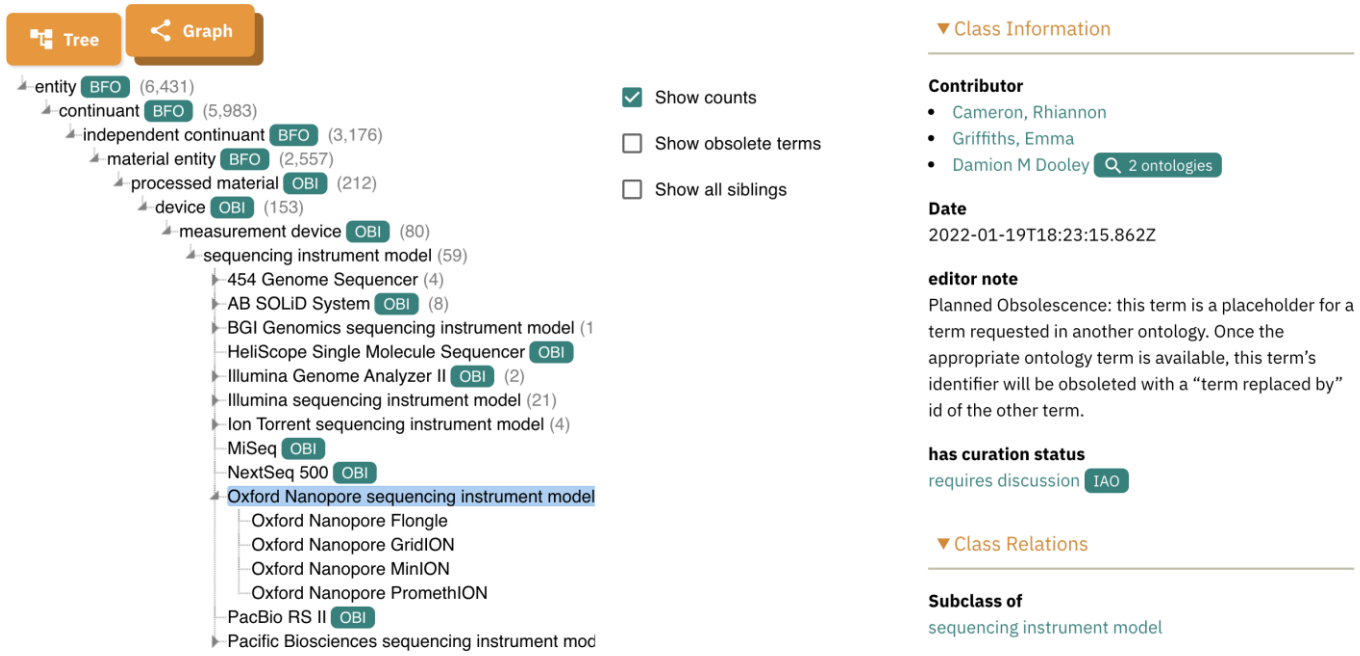
Fields and terms sourced and adapted from:

- Agency documentation
- Public repository submission forms
- Domain expert consultations
- Existing standards and ontologies

ISO slim (package of fields and terms) available:
https://github.com/GenEpiO/iso2017

How data curation & management resources interact

Ontologies



Hierarchies, logic, synonyms, annotations, (OWL)
EBI-OLS, Protege

Specifications

Data types, rules, patterns, max/mins, formats (LinkML: tsv, JSON, YAML)

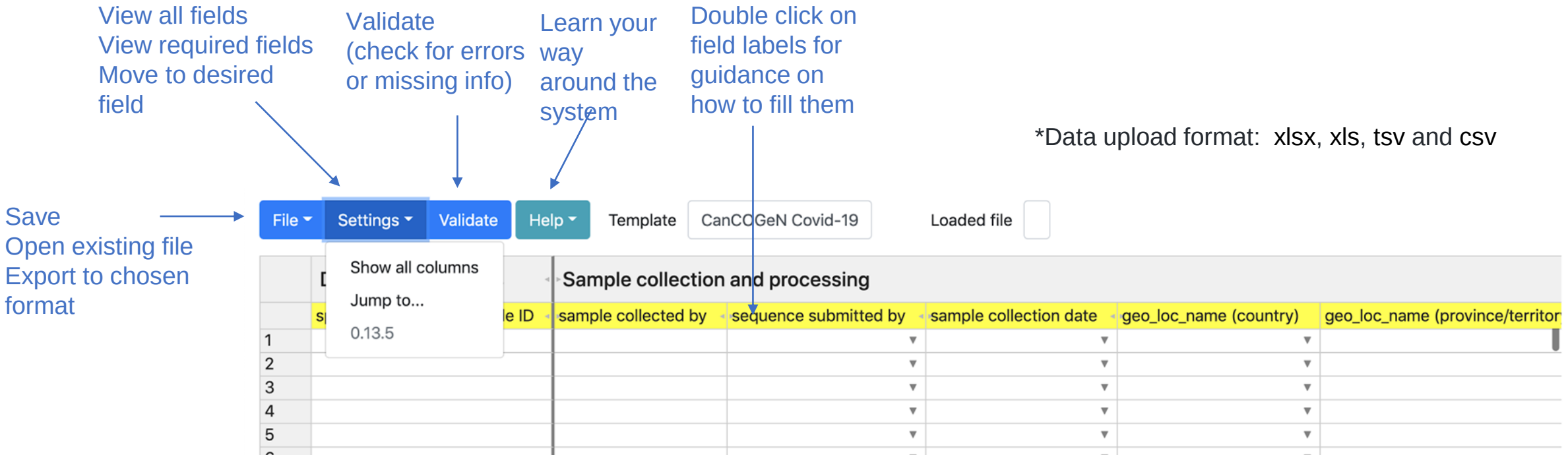
How data curation & management resources interact

Ontologies → Specifications → Tools & Data Management Systems

- Spreadsheets
- LIMS
- Databases
- Online/local tools
    - Look-up services
    - Curation tools
    - Validation tools
    - Transformation Tools
    - Mapping tools

The DataHarmonizer enables standardized data entry and validation.

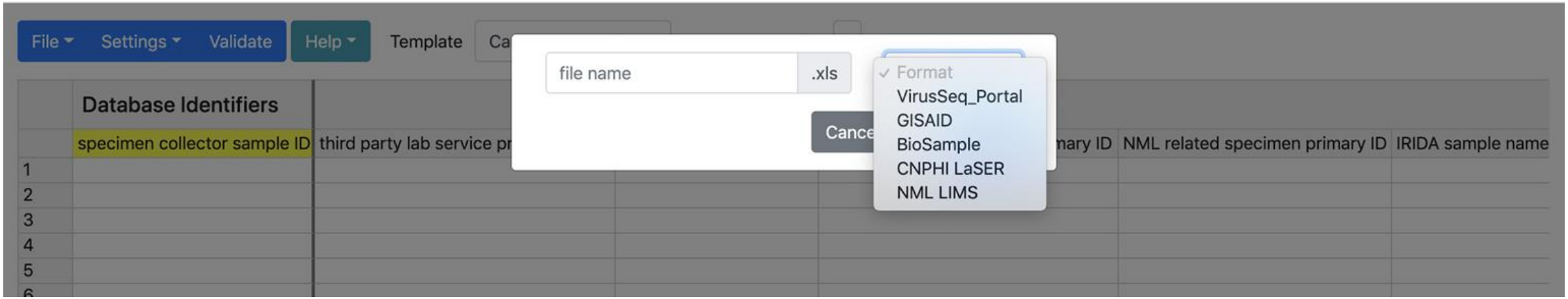
- Tool for data entry and validation developed for pandemic data harmonization
- Spreadsheet-style text editor application
- Colour-coding, picklists, curation features, validation
- Guidance, curation SOP, training



Find the fields you need, learn what to put in them, fill the ones that apply to your sample, check the info is right.

Data transformation is required for different downstream destinations of data.

We will use this later in the lab!

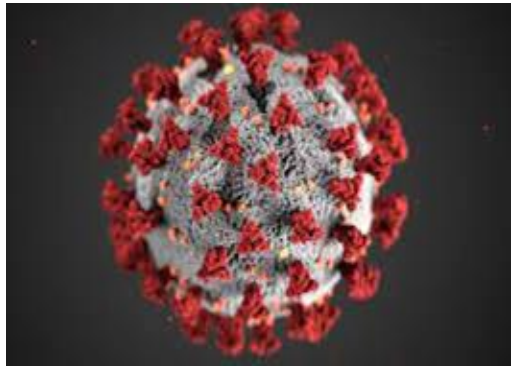


- Enter data once, export in different submission formats i.e. GISAID, VirusSeq Data Portal, NCBI BioSample as well as the national database (NML-LIMS).

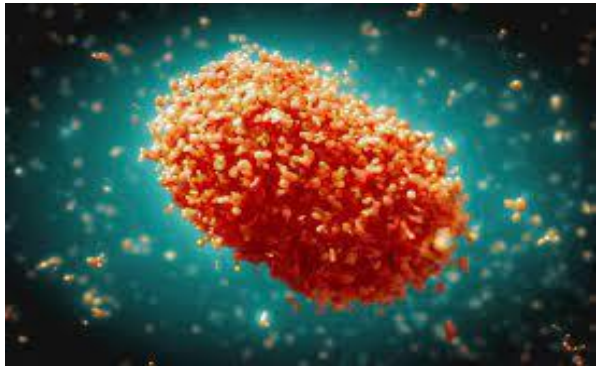
Enter data once,
export for
different uses!

Get the latest version here: https://github.com/cidgoh/pathogen-genomics-package
MGen, 2022: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000908

Real world examples of implementing ontology-based specifications in public health pathogen surveillance



SARS-CoV-2 (CanCOGeN & PHA4GE)



MPXV

One Health AMR

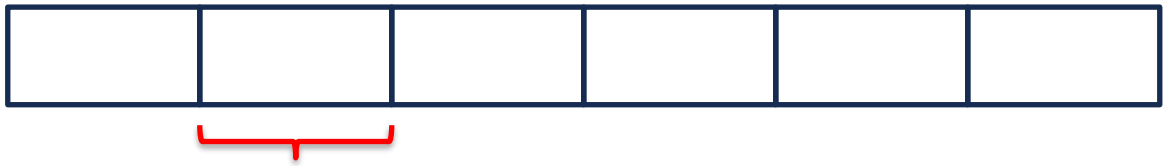
Wastewater

SFU

CENTRE FOR INFECTIOUS DISEASE GENOMICS AND ONE HEALTH

Interoperable specification design

Modular framework and core content (ISO 23418:22)



Modules populated with fields/terms from community-driven ontologies

Putting specs into practice:

- Consultation
- Implementation (tooling)
- Testing & consensus

Thematic Modules

- Database identifiers
- Sample collection and processing
- Host information
- Host exposure information
- Host reinfection information
- Host vaccination information
- Environmental conditions & measurements
- Sequencing methods
- Bioinformatics and quality control metrics
- Lineage/clade information
- AMR profiling
- Pathogen diagnostic testing details
- Provenance and attribution

Benefits of using interoperable data standards

REDUCE

- time
- workload
- uncertainty

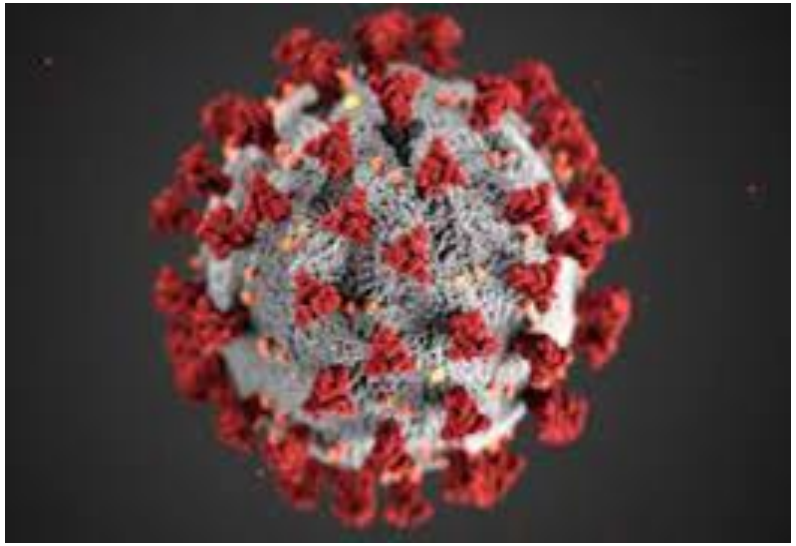
REUSE

- fields/terms
- tools
- training
- protocols

RECYCLE

- expectations
- agreements
- skills

Lesson: Curation/standards enable harmonization across systems



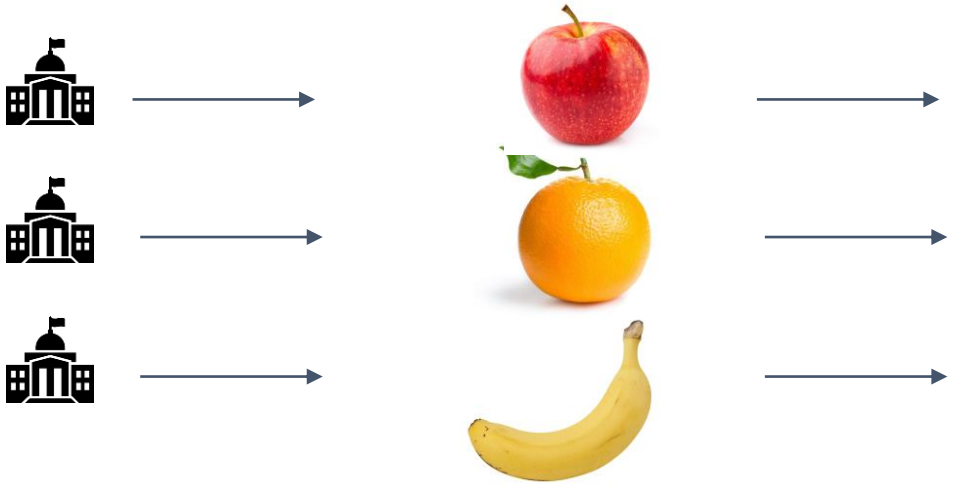
Use case 1: SARS-CoV-2 (CanCOGeN & PHA4GE)

Microbinfie podcast: episodes 26 & 53

Data flow for Canadian SARS-CoV-2 genomic surveillance would be complicated by data heterogeneity.

Colle ct

Data heterogeneity: slows down analysis



- Samples & contextual data collected at frontlines
- Sequenced by different labs/services
- Different provincial systems/priorities
- Submitted to national DB (PHAC)

Integrate

National database (NML PHAC)

National surveillance priorities (coordinated response)

Disseminate

GISAID

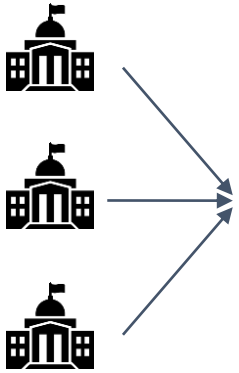
NCBI

VirusSeq Data Portal

CanCOGeN

Data standard development and tools to operationalize standards were critical to enabling harmonization and analyses.

Collect



Harmonize

Standardize

CanCOGeN SARS-CoV-2 Contextual Data Standard

+

The DataHarmonizer

Upload

Routine and analysis-specific curation

Real-time response to changing data needs

Integrate

National database

National surveillance priorities

Disseminate

GISAID

NCBI

VirusSeq Data Portal

CanCOGeN

Public Health Alliance for Genomic Epidemiology

- Global, volunteer organization
- >200 members, 90 organizations, 30 countries

Scope:

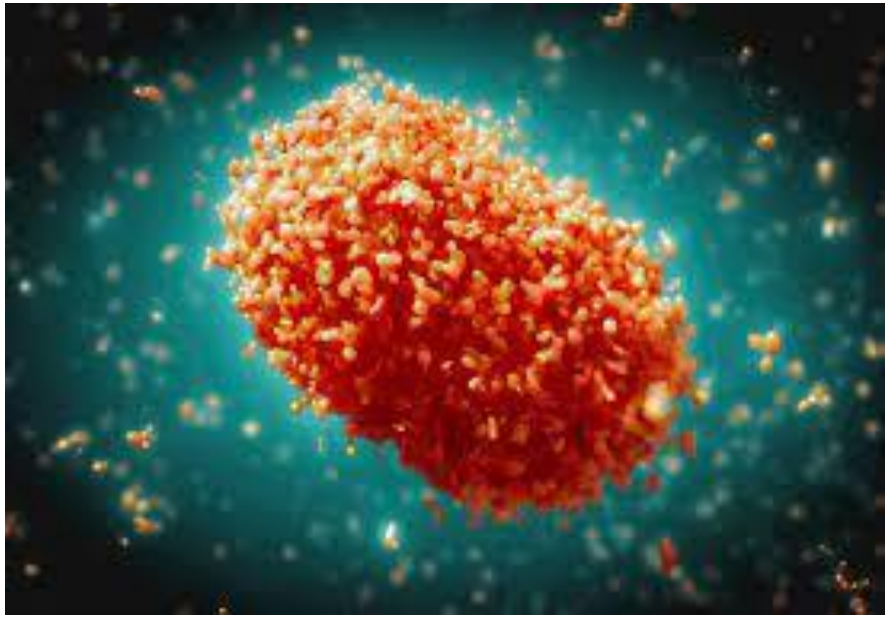
- Reproducibility, interoperability, portability, capacity for public health bioinformatics

Working Groups:

1. **Data Structures**
2. Infrastructure
3. Pipelines & Visualization
4. Training & Workforce
5. Ethics & Data Sharing

- CanCOGeN specification internationalized via PHA4GE
- Implemented in international systems (US CDC, Austrakka, Baobab LIMS, Nigeria CDC, ANLIS-Argentina)

Lesson: Harmonization solutions for specific data types



Use case 2: MPOX

Benefits of a common pathogen surveillance framework:
Repurposing the SC2 specification for the Monkeypox Epidemic

- On July 23, 2022, the World Health Organization (WHO) declares MPXV a **Public Health Emergency of International Concern**
- **As MPXV was growing concern, need for genomic surveillance**
- **Developed a template in the DH**
- **Reused most fields, customize some pick lists** (e.g. anatomical sites – Nasopharyngeal swabs (SC2) vs Groin (MPXV) ; anatomical material - Lesion (Pustule), Lesion (Vesicle))

Avoid proliferation of incompatible specifications!

Solutions for common data problems:

1) variable clinical sample descriptions

original sample description	anatomical material	anatomical part	body product	collection device	collection method	biomaterial extracted
anal dry swab	Not Applicable	Anus	Not Applicable	Dry swab	Not Applicable	Not Applicable
anal ulcere swab	Ulcer	Anus	Not Applicable	Swab	Not Applicable	Not Applicable
Arm Legion-pustule	Lesion (Pustule)	Arm	Not Applicable	Not Applicable	Not Applicable	Not Applicable
Base of RT Arm back lesions	Lesion	Arm; Back	Not Applicable	Not Applicable	Not Applicable	Not Applicable
Crusted skin lesion perineal	Lesion	Perineum	Not Applicable	Not Applicable	Not Applicable	Not Applicable
CSF	Fluid (cerebrospinal (CSF))	Not Applicable	Not Applicable	Not Applicable	Not Applicable	Not Applicable
DNA	Not Applicable	Not Applicable	Not Applicable	Not Applicable	Not Applicable	DNA

2) **specimen processing** (pooling samples) and **sequential sampling** from the same individuals (subject host ID)

scenario	anatomical material	anatomical site	collection device	specimen processing	specimen processing details	host subject ID
Multiple swabs pooled from different penis lesions (5 swabs per sample)	Lesion	Penis	Swab	Specimens pooled	5 swabs per sample	Not Applicable
Swabs from patient (#ABC12345), groin lesions	Lesion	Genital area	Swab	Not Applicable	Not Applicable	ABC12345
Swabs from patient (#ABC12345), groin lesions, 2 weeks later	Lesion	Genital area	Swab	Not Applicable	Not Applicable	ABC12345

3) **One Health** MPXV samples (wastewater, fomites, other hosts)

original sample description	host (common name)	anatomical material	anatomical part	body product	environmental material	environmental site	collection device
wastewater	Not Applicable	Not Applicable	Not Applicable	Not Applicable	Wastewater	Not Provided	Not Provided
Groundhog feces	Groundhog	Not Applicable	Not Applicable	Feces	Not Applicable	Not Applicable	Not Applicable
Human: lesion swab, arm and back	Human	Lesion	Arm; Back	Not Applicable	Not Applicable	Not Applicable	Swab
Bed linen	Not Applicable	Not Applicable	Not Applicable	Not Applicable	Bed linen	Not Applicable	Not Applicable

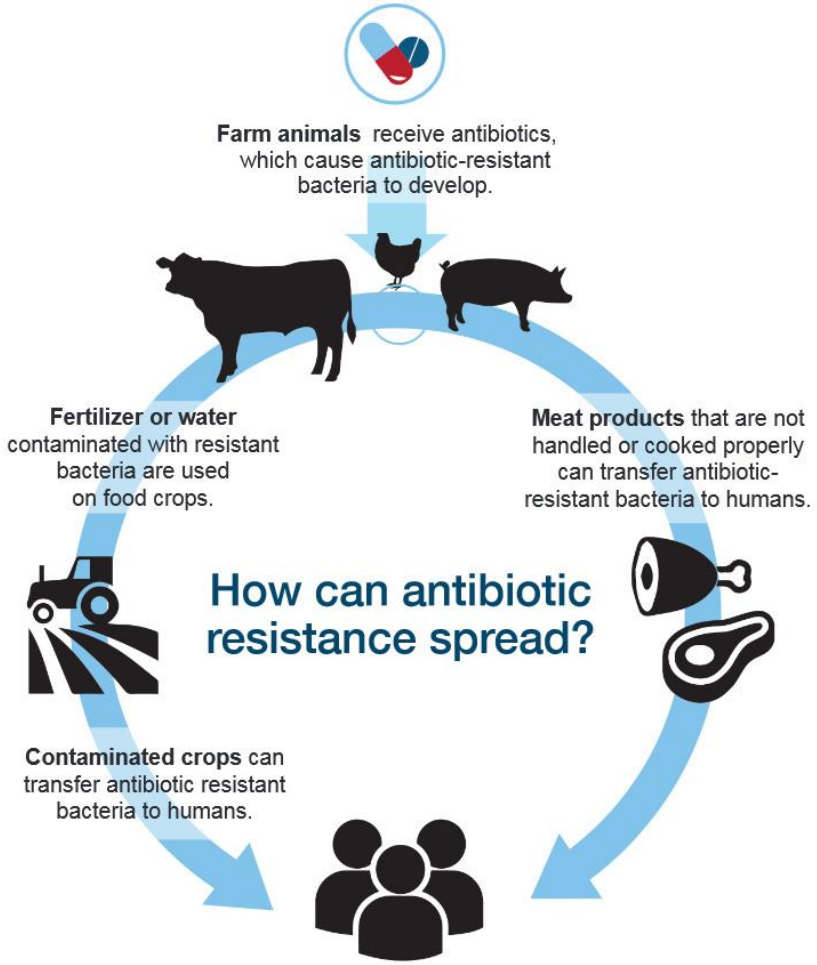
Harmonized contextual data in NCBI: **BioProject PRJNA846794**

Lessons:

1. Implementation Options
2. Mapping & exchange formats
3. Need for Formalization of Curation/Coordination Role

Use case 3: One Health AMR

GRDI-AMR standard: Genomics and Research Development Initiative to support Canada's federal AMR action plan



https://github.com/cidgoh/GRDI_AMR_One_Health

- Scope: Bacteria. **WGS across sectors, commodities, environments, hosts**
- Goal: use genomics and harmonized contextual data to understand foodborne **AMR** in **food supply and environment,** identify interventions
- **Canadian implementation: Federal Interagency** (PHAC, CFIA, AAFC, ECCC, DFO, HC etc)
- **international data exchange**

Technical Implementations – Tools & Databases

Different ways to implement the standard for data management.

1. Spreadsheet-based templates and tools

2. Existing Systems

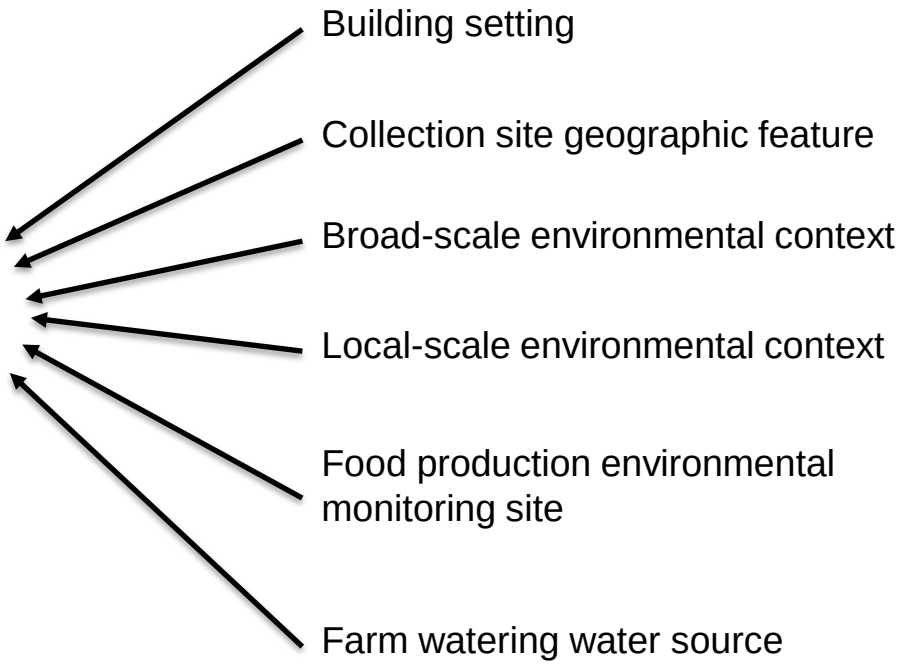
- Mapping
- Interchange formats
- Automated transformations

3. New Systems

Different standards for different needs

No one standard to rule them all

environmental site [GENEPIO:0001232]



Canada: integrative approach

USA: disaggregative approach

Both ontology-based. Mapping relates concepts; creates interoperability

Formalizing curation roles in pathogen genomics

What we learned/confirmed in the GRDI:

- Often different data generators/partners
- Different data needs
- Lack of awareness about appropriate standards, how to customize, how to implement
- Lack of knowledge about semantic best practices
- Need for troubleshooting

Solution:

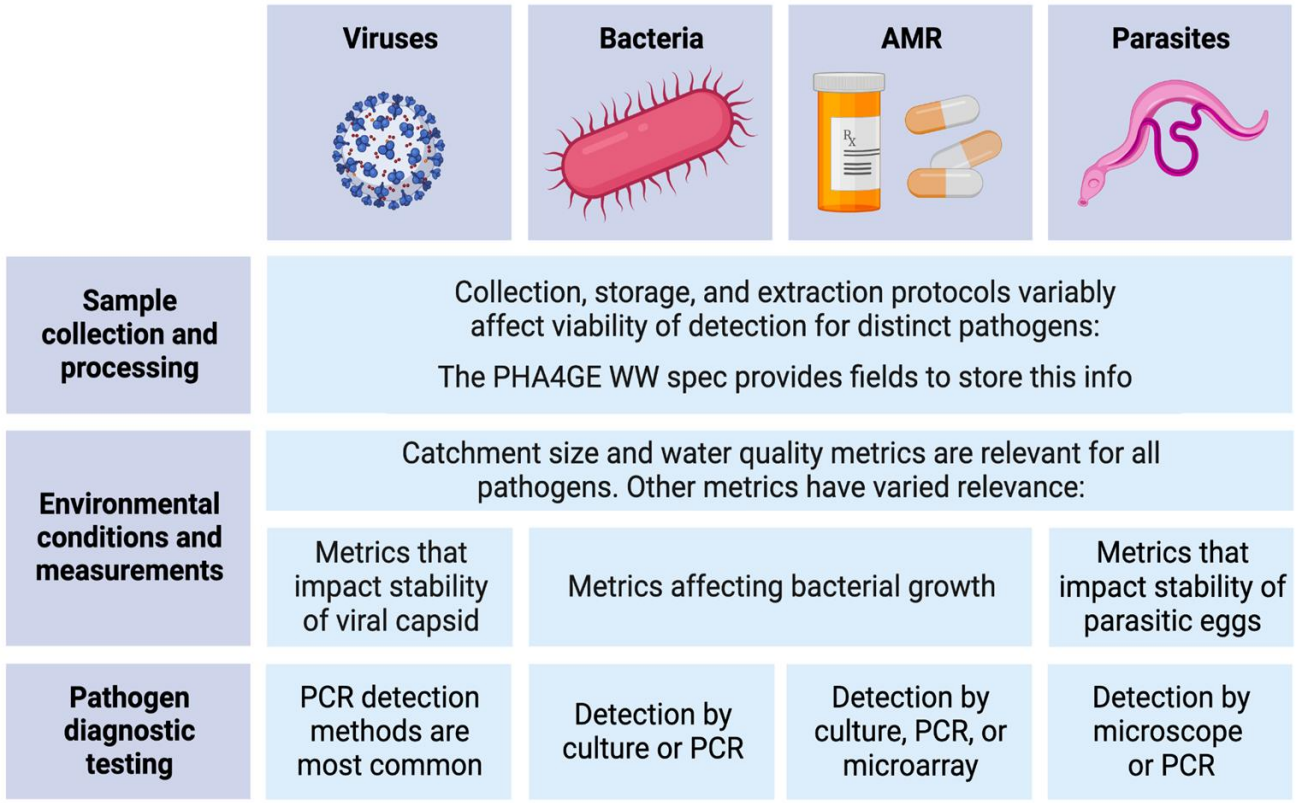
- Dedicated personnel for identifying/developing standards should be embedded in every project/program
- Resources should be provided for curation (so not last minute/second class citizen)

Lessons:

1. Curation/standards reduce “data Wild Wests”
2. Consensus is key in developing and implementing standards

Use case 4: Wastewater

WW genomics has a lot of applications in public health surveillance.



Utility of WW data compromised by "data Wild Wests"

Identify many threats at once

Technically challenging, different (fewer) privacy complications for data sharing compared to clinical surveillance

Many academic projects, PH initiatives (Canada, globally)

Lack of consensus about vocabulary (many environments, geographical locations & differently resourced settings, sample types, pathogen-based communities)

Consensus is key in developing and implementing standards

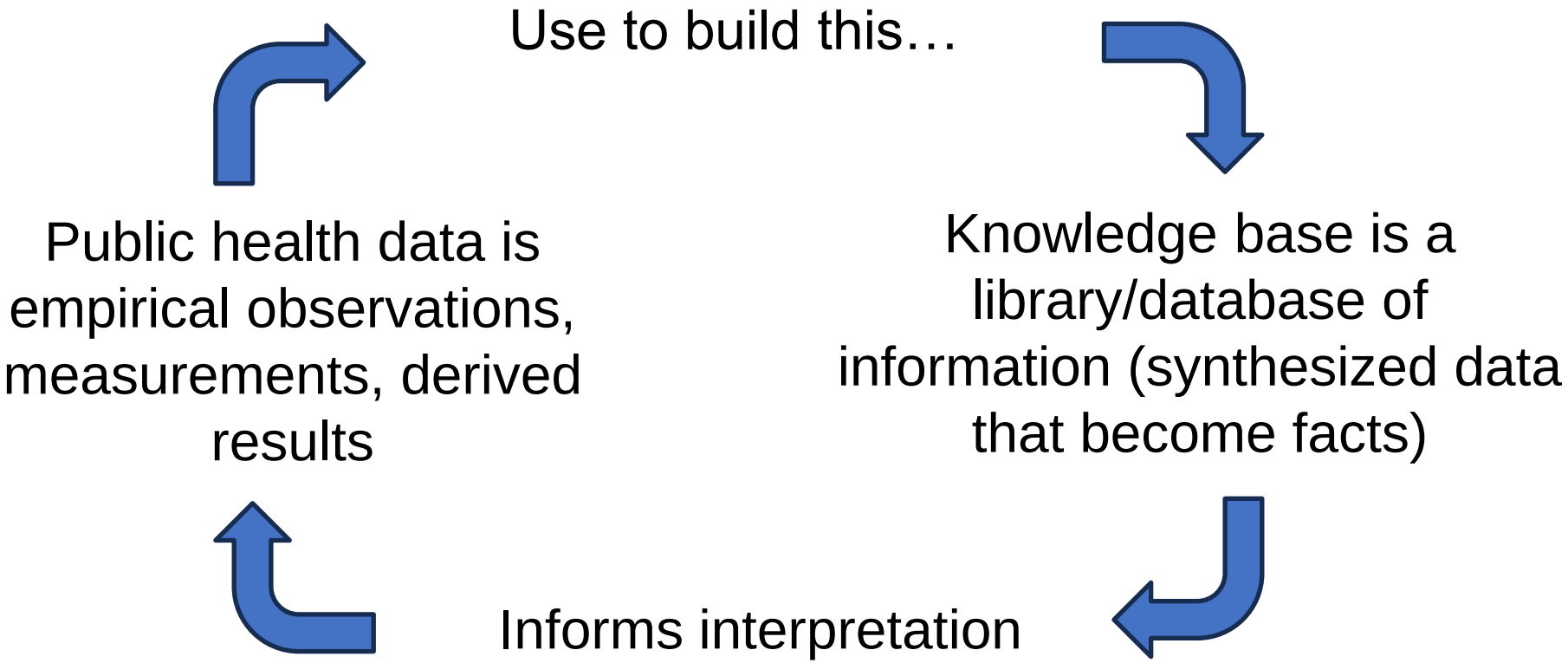
Building relationships:

- **consultation, co-creation** (direct communication – sometimes people say one thing but they really need another)
- People are usually not bad actors, need a **clear path to implementation** (hopefully an easy one)
- You will need to **compromise**
- **Education** is needed

Public Health Contextual Data Curation Best Practices

- Be aware of FAIR, interoperable data standards
- Use data standards (implement curation practices and tools in routine practice)
- Advocate for formalized curation role in your organization (include in experimental design & analysis, systems, funding)
- Communicate with standards/tool developers (CIDGOH, OBO Foundry, PHA4GE)

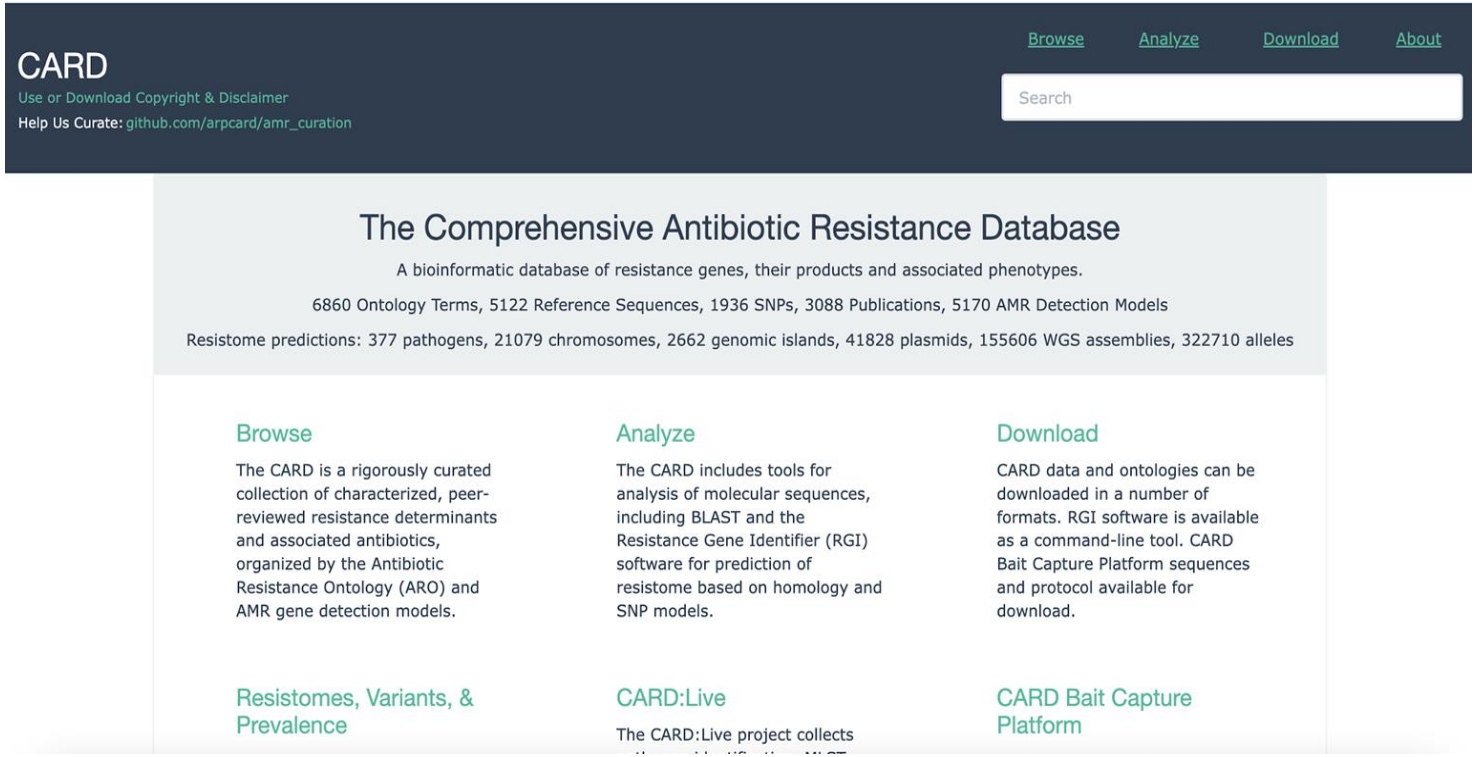
Curation for **public health data** vs **knowledge bases**



Real world examples of knowledge base curation to support public health

Comprehensive Antimicrobial Resistance (CARD) Database

Comprehensive Antimicrobial Resistance Database (CARD)



https://card.mcmaster.ca/

- **Database** that links **genes/mutations** with **drugs**, **mechanisms of action**, and more
- Knowledgebase used by the **Resistance Gene Identifier (RGI)**

- **EVIDENCE based**
- curators identify data and proof
- use data model (structured using **Antibiotic Resistance Ontology**)
- need **curation system** (process, rules)

- More in **Module 7**!

Knowledgebase curation best practices

- Criteria
  - scope
  - rules/cut-offs
  - examples
- Trusted sources & evidence
- Consensus
  - automation preferable to personal interpretation
  - consistency
- Documentation
- Versioning

High quality content
+
Standardized structure
+
Transparent methods
+
FAIR principles
=
Open, interoperable, useful knowledge base (can build tools to turn data into facts & actionable knowledge)

Data sharing and public repositories

Why should I share data?

1. Situational awareness: lack of data sharing creates blind spots

2. Diagnostics/therapeutics (make sure your viruses covered)

3. Having a voice in global decision making (data creates leverage)

4. Data sharing in a human rights framework (GA4GH)

Article 27 of the 1948 Universal Declaration of Human Rights

- *right of all citizens in all countries to the benefits of the advancements of science (duty to share)*
- *right of attribution of scientists*
- *reinforces the right of scientific freedom*

https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/framework-for-responsible-sharing-of-genomic-and-health-related-data/

5. Being a good data citizen

Data stewardship: oversight and practices to ensure data is **accessible, usable, safe, trusted.**

Privacy protection (sharing):

- **Public trust essential**, loss of trust has consequences (protection, transparency)
- De-identified data (**no names/addresses**)
- Be careful of 1) geographical granularity, 2) small case numbers in defined geo_loc/time, 3) combinations of fields
- **Track identifiers** (chain of custody), but personal health IDs/sample IDs may be considered PHII
- **Consult privacy officer** (jurisdictional policies, national legislation)

Security & Quality:

- Provenance, methods (rich details): **attribution, auditability, reproducibility (track methods), accountability**
- Contextual data may require storage with **higher security** than seq data
- Errors **corrected**, **update** as required

Types of contextual data critical for surveillance/ genomic epidemiology (what you can most likely share)

- **Geo_loc (at least country, preferably state/province) – sample collection**
- **Sample collection date (to the day)**
- **Attribution: Who collected sample, who sequenced it**
- **Methods: instrument (platform & model), consensus sequence software, coverage**

- Sampling strategy (random sampling, targeted sampling, outbreak, research)
- Demographics: age/sex (gender)
- Sample type
- Host
- Quality indicators (e.g. Ct values)

- Vaccination
- Exposures
- Travel history
- Hospitalization
- Outcomes

Submitting to public repositories is crucial for global surveillance

- Main public repositories **GISAID** and INSDC (**NCBI, ENA, DDBJ**)
- Different but overlapping requirements
- Submitting to different repos is encouraged, transformation will be necessary
- What you collected, how you structured it (in your spreadsheet or LIMS) may be different than the submission requirements
- What you share depends on your data sharing policies

No matter which repository you choose, you will need to do the following:

Stage 1: Set up an account

Stage 2: Prepare your contextual and sequence data (fastq/fasta) files

Stage 3: Submit

Data Sharing: Public Repositories

Open Access

International Nucleotide Sequence Database Collaboration (**INSDC**)
NCBI (USA)
EBI-ENA (UK)
DDBJ (Japan)

- No restrictions

- Relies on behavioural norms
- Benefits are not equitable distributed
- Data generators have no control

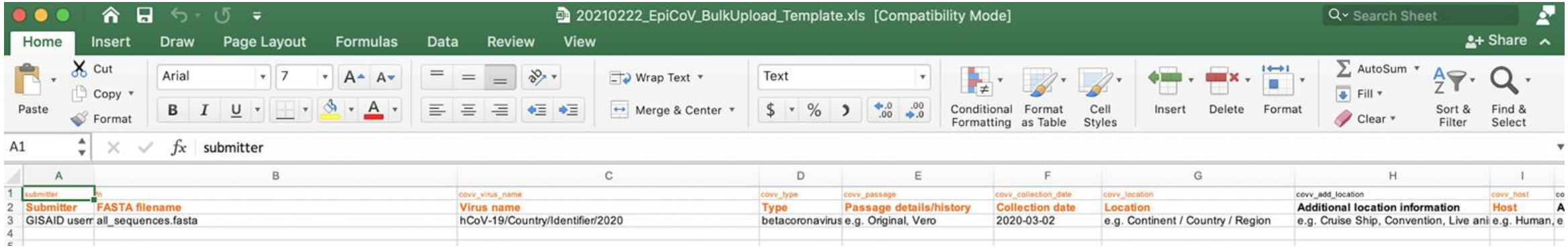
Controlled Access

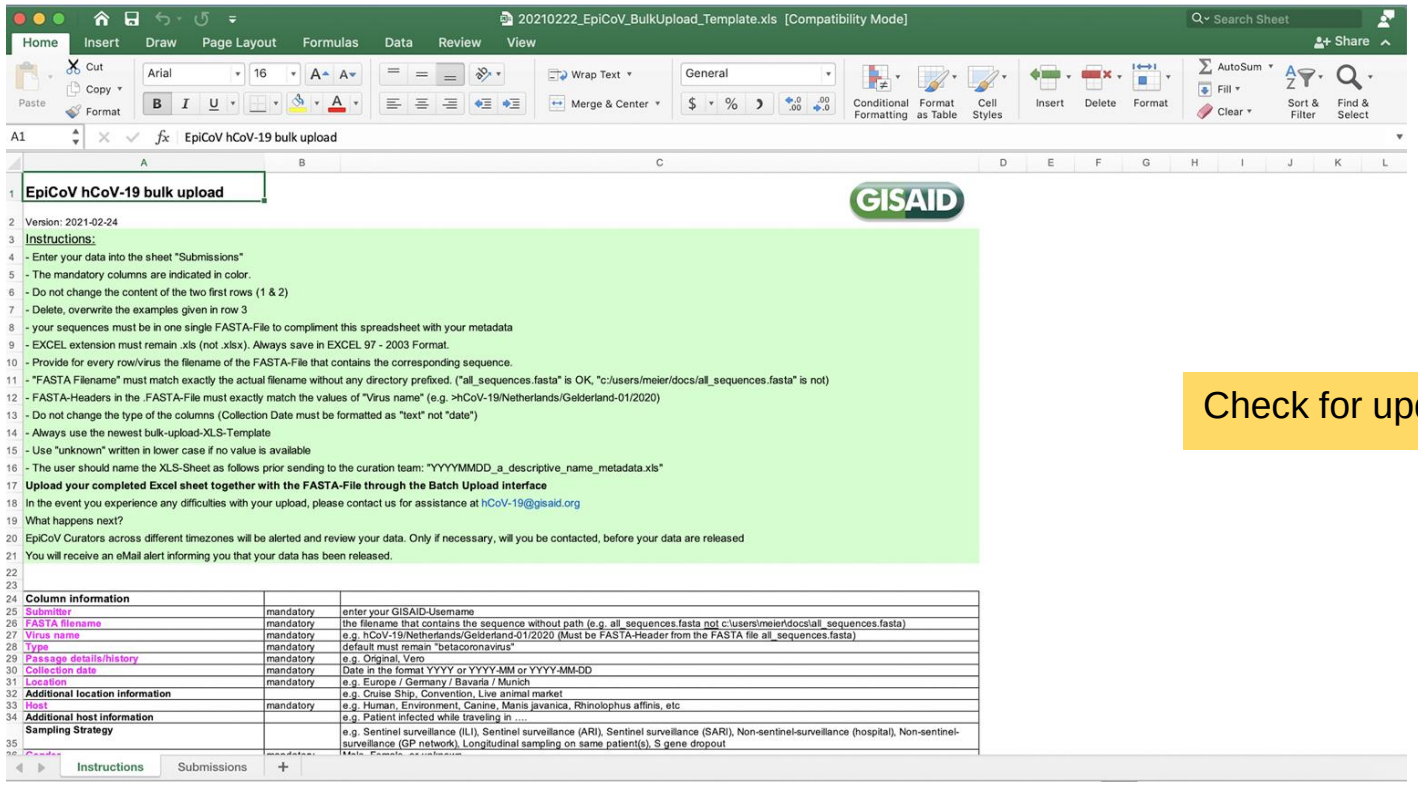
Global Initiative on Sharing Avian Influenza Data (**GISAID**)

- Some protections
- Data use restrictions (Terms of Service specify limitations on reuse, prescribe collaboration & attribution)

- Few solutions to practical issues (how to attribute thousands of labs?)
- Interpretation of infringement & Enforcement
- Impacts on innovation
- Benefits are still not equitably distributed

GISAID





Check for updates!

- Register and sign agreement to access templates (**attribution, collaboration, data use restrictions**)
- Template is what you will submit, includes instructions (different templates for different organisms)

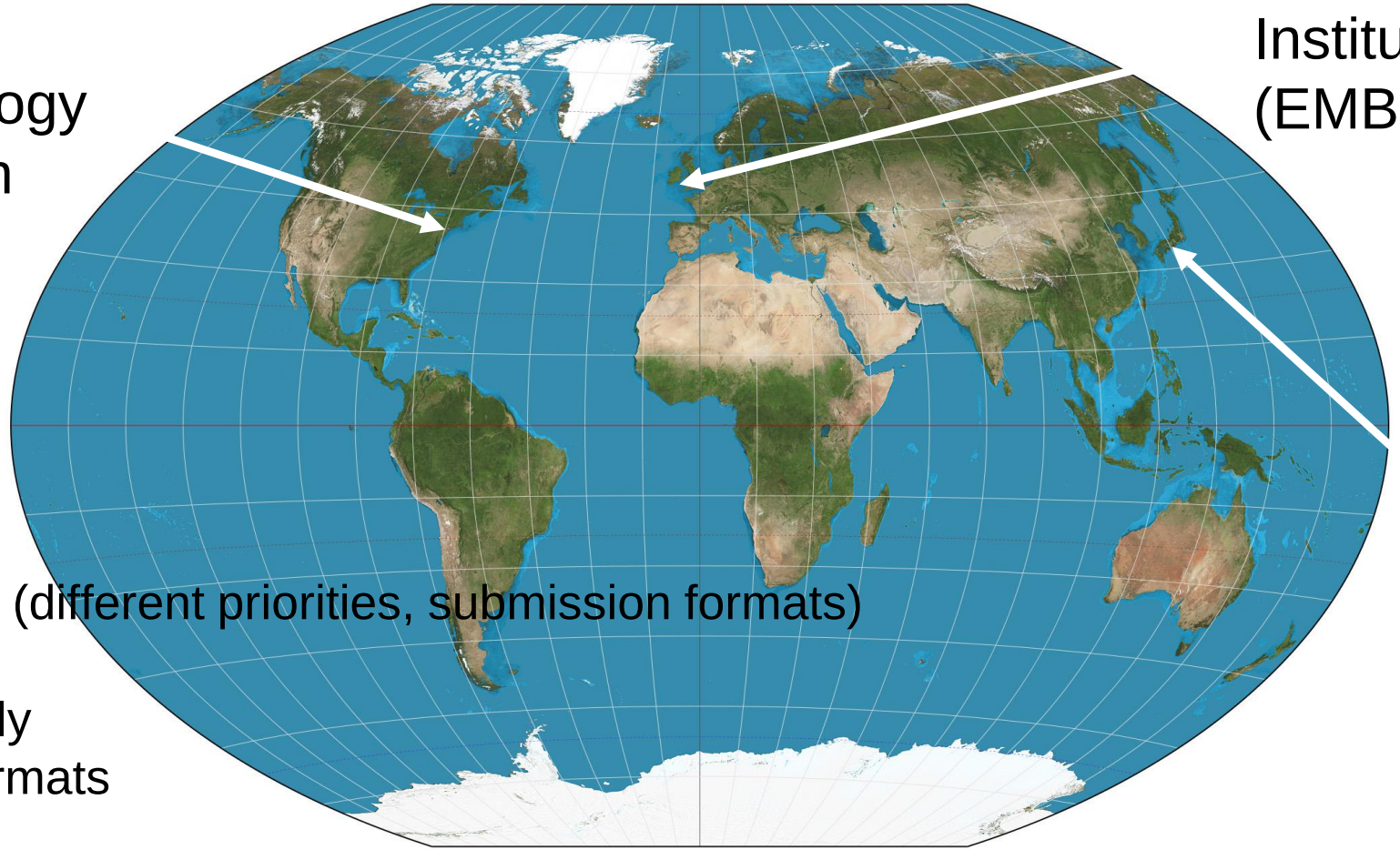
GISAID Contextual data requirements

GISAID Fields (as of 2020-06-19)	GISAID Definition
Submitter	enter your GISAID-Username
FASTA filename	the filename that contains the sequence without path (e.g. all_sequences.fasta not c:\users\meier\docs\all_sequences.fasta)
Virus name	e.g. **hCoV-19/Canada/BC-prov123/2020** (Must be FASTA-Header from the FASTA file all_sequences.fasta)
Type	**default must remain "betacoronavirus"**
Passage details/history	e.g. Original, Vero
Collection date	Date in the format YYYY or YYYY-MM or YYYY-MM-DD
Location	e.g. Europe / Germany / Bavaria / Munich
Additional location information	e.g. Cruise Ship, Convention, Live animal market
Host	e.g. Human, Environment, Canine, Manis javanica, Rhinolophus affinis, etc
Additional host information	e.g. Patient infected while traveling in ….
Sampling Strategy	e.g. Sentinel surveillance (ILI), Sentinel surveillance (ARI), Sentinel surveillance (SARI), Non-sentinel-surveillance (hospital), Non-sentinel-surveillance (GP network), Longitudinal sampling on same patient(s), S gene dropout
Gender	Male, Female, or unknown
Patient age	e.g. 65 or 7 months, or unknown
Patient status	e.g. Hospitalized, Released, Live, Deceased, or unknown
Specimen source	e.g. Sputum, Alveolar lavage fluid, Oro-pharyngeal swab, Blood, Tracheal swab, Urine, Stool, Cloakal swab, Organ, Feces, Other
Outbreak	Date, Location e.g. type of gathering, Family cluster, etc.
Last vaccinated	provide details if applicable
Treatment	Include drug name, dosage
Sequencing technology	e.g. Illumina Miseq, Sanger, Nanopore MinION, Ion Torrent, etc.
Assembly method	e.g. CLC Genomics Workbench 12, Geneious 10.2.4, SPAdes/MEGAHIT v1.2.9, UGENE v. 33, etc.
Coverage	e.g. 70x, 1,000x, 10,000x (average)
Originating lab	Where the clinical specimen or virus isolate was first obtained
Address	
Sample ID given by the sample provider	
Submitting lab	Where sequence data have been generated and submitted to GISAID
Address	
Sample ID given by the submitting laboratory	
Authors	a comma separated list of Authors with complete First followed by Last Name

International Nucleotide Sequence Database Collaboration

National Center for Biotechnology Information (NCBI)

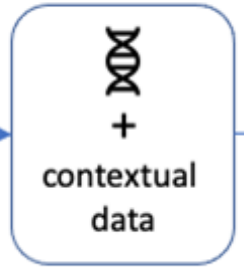
European Bioinformatics Institute (EMBL-EBI)

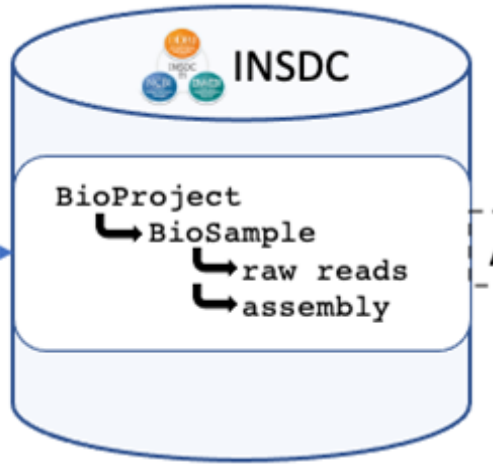


DNA Database of Japan (DDBJ)

Different centres (different priorities, submission formats)
Collaborate
Mirror data nightly
Map between formats

Shared data is useful for many, many applications



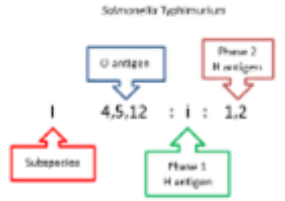


API

BioSample metadata
+
Experiment metadata

Genotyping assays

- Antibiotic resistance
- Serotyping / sequence type
- Virulence predictions
- stress tolerance predictions
- Typing/Sub-typing



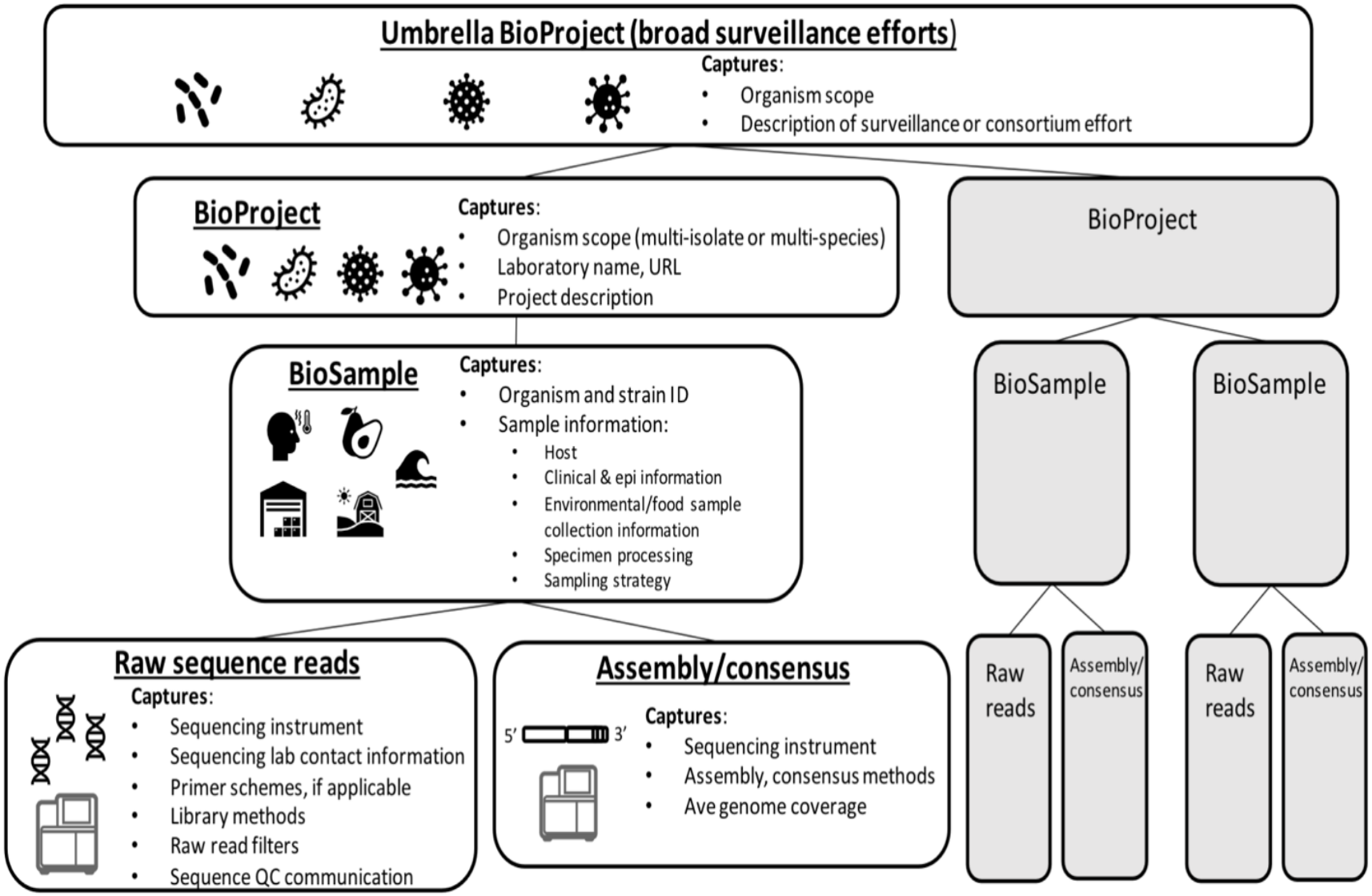
Phylogeny / phylodynamics

- wgMLST Clustering
- Phylogenetic tree inference
- Molecular evolution / selection
- New variant/lineage identification

Public Health/regulatory/clinical applications:

- Outbreak investigations
- Foodborne contamination events
- Pathogen harborage in food processing facilities
- Wastewater based epidemiology
- Evolution of virulence, stress tolerance, etc.
- Monitoring geographic distribution of pathogens
- Baseline surveillance vs targeted sampling
- Tracking antibiotic resistance
- Therapeutics + diagnostics development/validation
- New vaccine formulation

INSDC Data Model (Best Practices)



BioSample Metadata Packages

- sets of specifications (fields) interoperability, standardization
- authoritative source (**Genomics Standards Consortium**)
- describing samples from **different contexts** (e.g. clinical, environment, food) or for different kinds of **sequences** (e.g. single genomes, metagenomes, marker genes)
- user selects most appropriate package **Pathogen packages** (also different **SARS-CoV-2 package**)

Standard Packages

SARS-CoV-2: clinical or host-associated; version 1.0

SARS-CoV-2: wastewater surveillance; version 1.0

One Health Enteric; version 1.0

Microbe; version 1.0

Model organism or animal; version 1.0

Metagenome or environmental; version 1.0

Invertebrate; version 1.0

Human; version 1.0

Plant; version 1.0

Virus; version 1.0

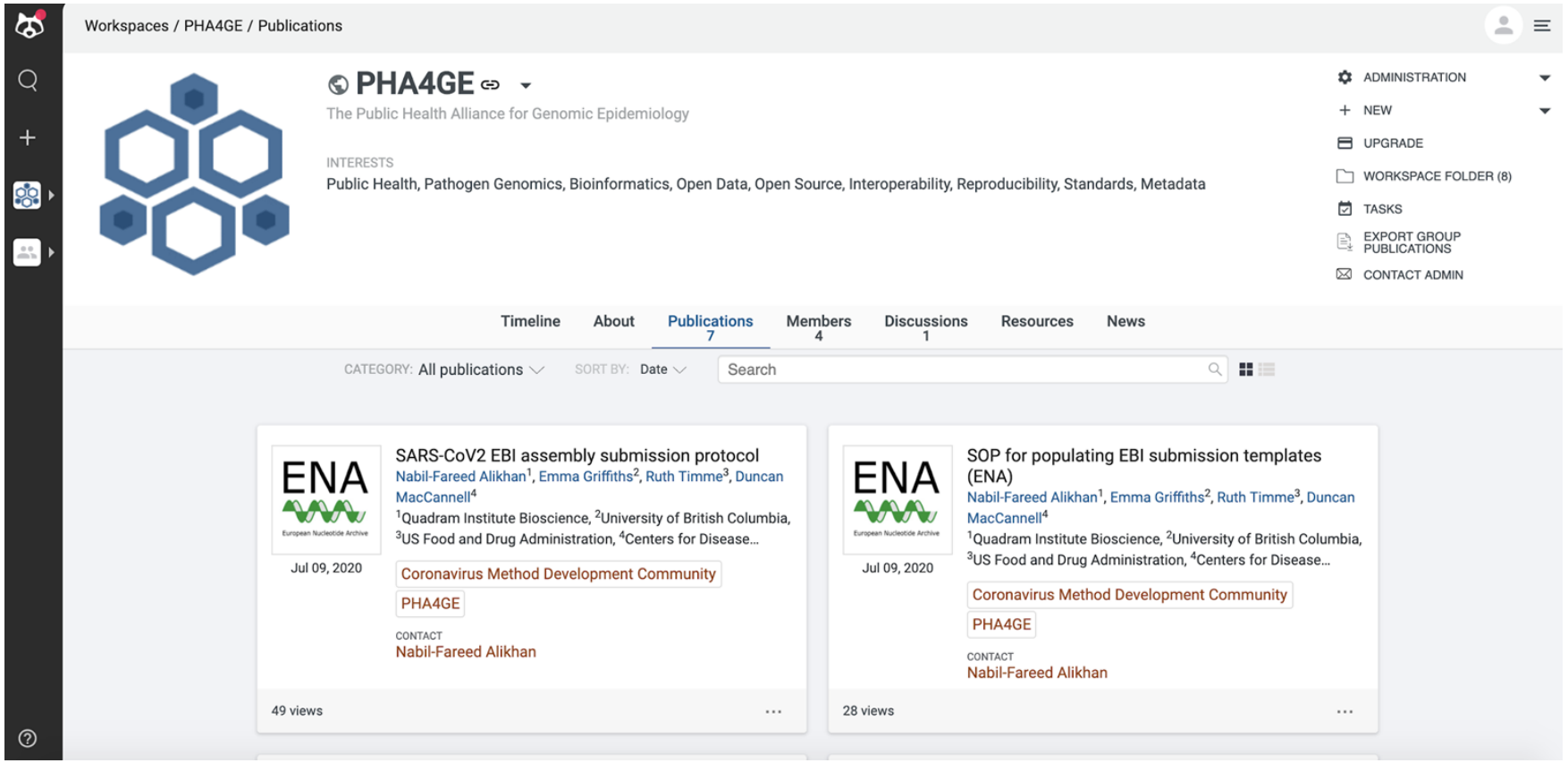
Beta-lactamase; version 1.0

Pathogen

Pathogen: clinical or host-associated; version 1.0

Pathogen: environmental/food/other; version 1.0

Protocols to mobilize harmonized data



https://www.protocols.io/workspaces/pha4ge

Public Health Alliance for Genomic Epidemiology

Data Curation & Standards: Career Opportunities

We live in a data-centric world.

Jobs requiring ontology skills (tech companies, public health organizations: databases, knowledge graphs, AI)

Jobs requiring curation skills (academia, industry, health, etc)

Data science (extract or extrapolate knowledge and insights from potentially noisy, structured, or unstructured data)

Developing or implementing tools

Take Home Messages: Summary

- Data curation and standards are part of quality frameworks (best practices; PHA4GE, GSC)
- Ontology-based data specifications increase interoperability of datasets and systems
- Data management tools (e.g. DataHarmonizer) help operationalize data standards
- Data standardization tools and best practices help build knowledge bases useful for research
- Data sharing is important for situational awareness, decision-making, innovation
- Data stewardship is important considerations important for data sharing (privacy, security, trust)
- Open and access-controlled public repositories have different advantages (GISAID, INSDC), different submission formats

Thank you!

Centre for Infectious Disease Genomics and One Health - CIDGOH

https://cidgoh.ca/
https://github.com/cidgoh/

ega12@sfu.ca
@griffiemma

We are on a Coffee Break & Networking Session

Workshop Sponsors: