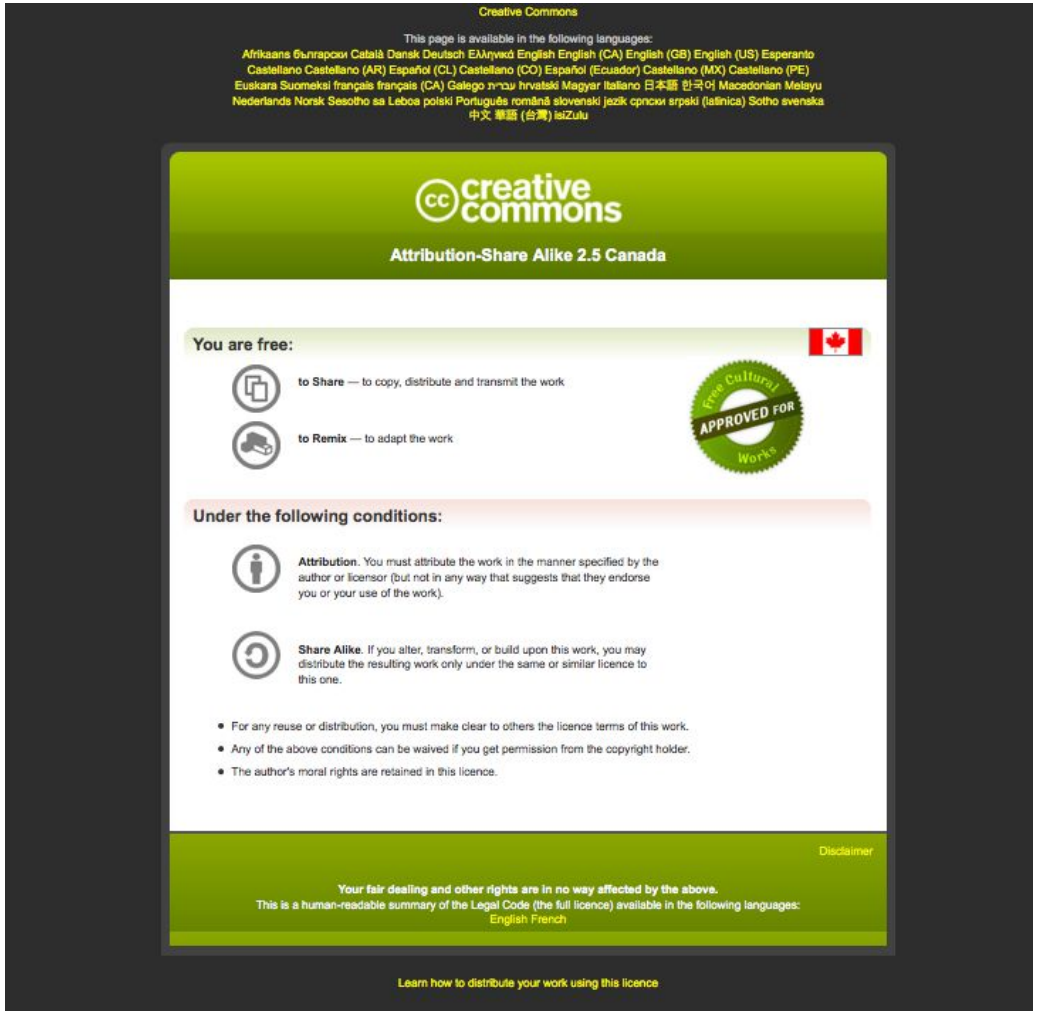
bioinformatics.ca

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

Emma Griffiths

Infectious Disease Genomic Epidemiology

April 18-21, 2023

bioinformatics.ca

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 challenges of using genomics contextual data for public health analyses

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

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

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

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

6. Be aware of data curation best practices

Contextual data is critical for interpreting the sequence data

Sequence data

Contextual data

Sample metadata

Lab results

Clinical/Epi data

Method
s

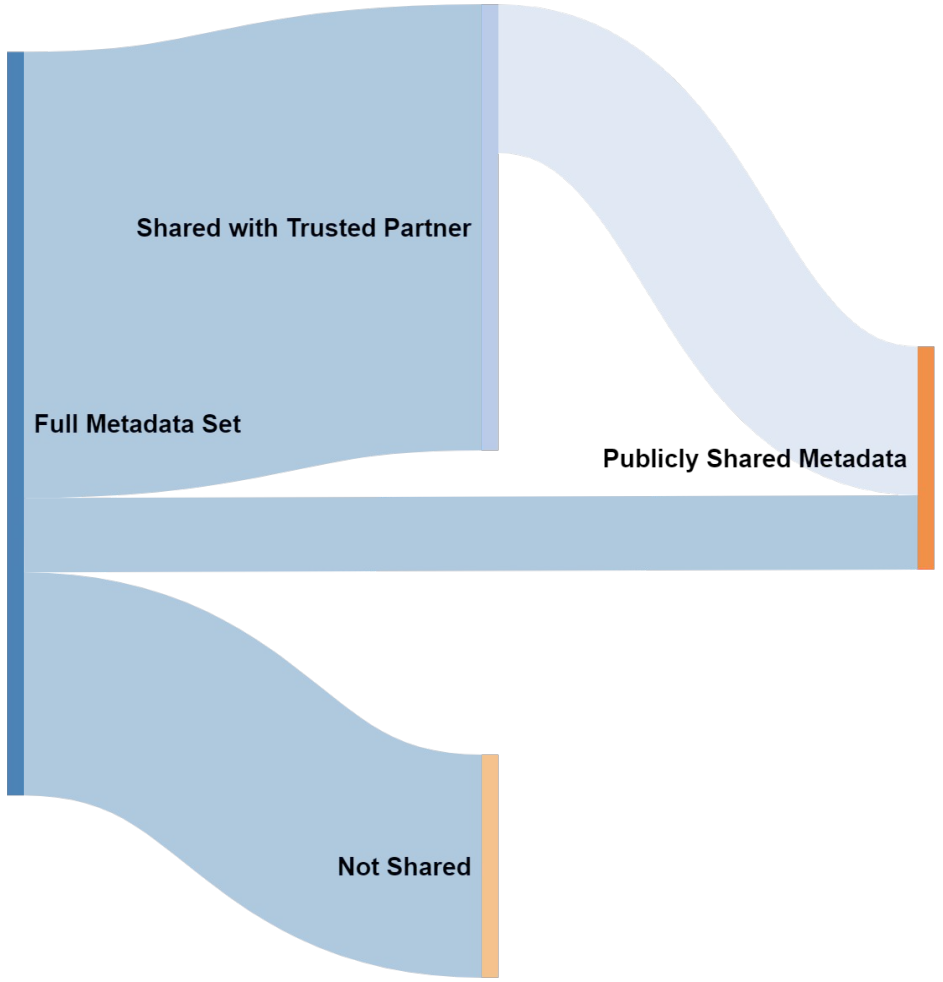
Contextual data (metadata) used for **surveillance** and **outbreak investigations**:

- **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**

There are different kinds of data sharing.

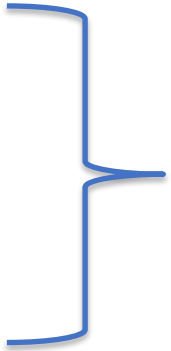
- Data comes from **different sources** (labs, departments, databases)
- Data needs to be shared within **organizations**, with **trusted partners**, with **public repositories**, with **international agencies**
- **Everyone uses different systems, processes**



What does data heterogeneity look like?

Harmonizing fields of data is challenging.

SPECIMENSOURCE_1
Isolation
host_tissue_sampled
Source

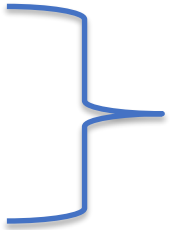


The labs mean "**sample type**"

Differences in labels, Same meaning

Computer doesn't recognize these as the same thing

Source



The lab means "**submitting lab**"

Same label, Different meaning

Computer doesn't recognize these as different

A field by any other name does NOT smell as sweet…

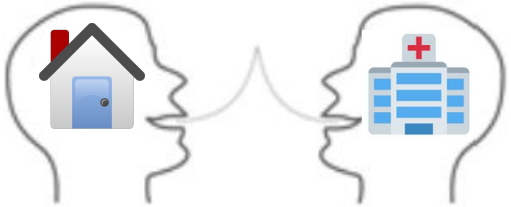
Heterogeneity of values within a field also complicates using the data.

Free text =

In-pateint
Np swad
UTM
NPS

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

“in isolation”



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

RISK FACTORS:
Do you work with animals?
(exposures)
vs
Do you have diabetes?
(pre-existing conditions)

Errors & Short hand

Granularity

Semantic ambiguity

Formats

Different Classifications/ Content

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

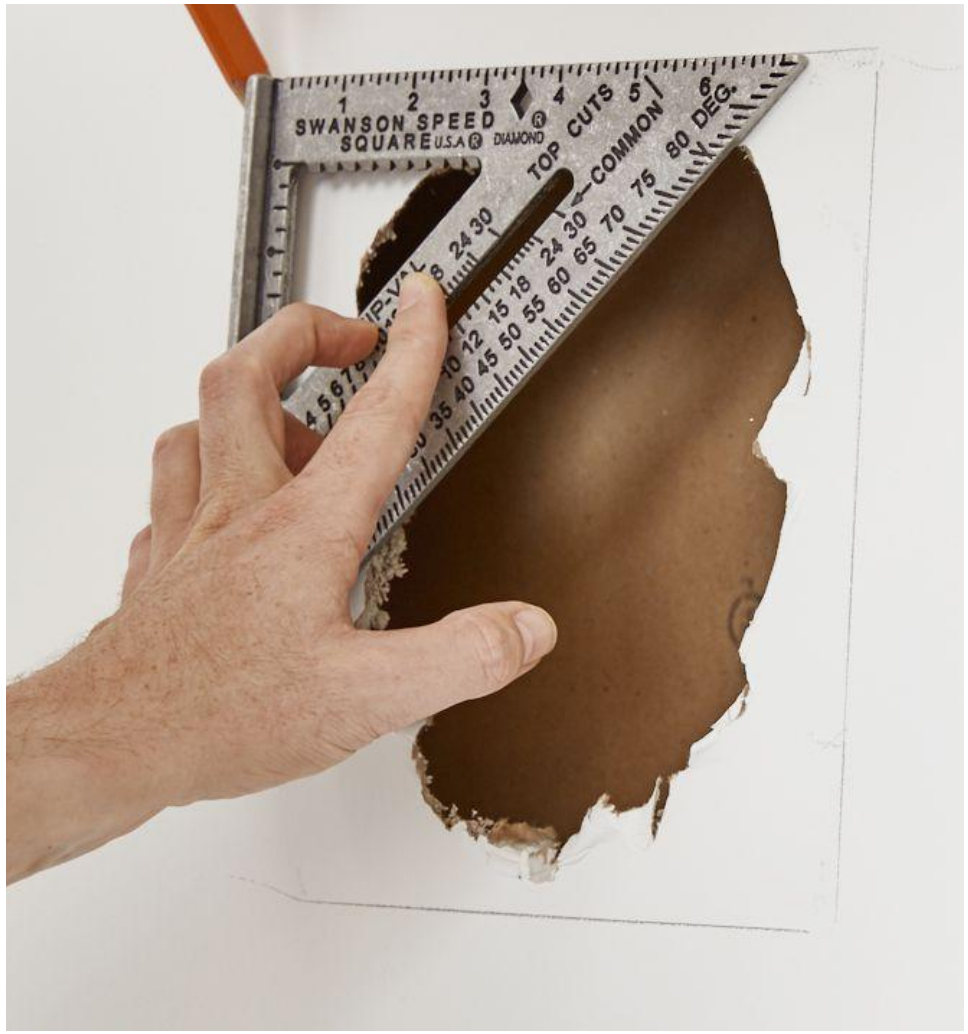
source name	Lung sample from postmortem COVID-19 patient
cell type	Lung Biopsy
strain	NA
subject status	No treatment; >60 years old male COVID-19 deceased patient

Data structure impacts function.

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

How do we fix it?

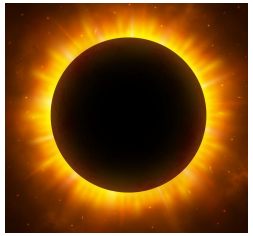
1. Ontologies
   - universal language for humans and computers
2. Data standards
   - prescribed sets of fields, terms, formats
3. Tools
   - software and supporting materials to implement standards



Ontology, A Way of Structuring Information

Ontologies aim to represent truth. *Is this universal?*

- Controlled (standardized) vocabulary
- Hierarchy (granularity)
- Logic, machine actionable
- Definitions and unique IDs (disambiguation)
  - e.g. Lager beer [BeerO:1234]
    A type of beer that uses a process of cool fermentation, followed by maturation in cold storage.
- Synonyms (facilitates mapping/interoperability)
  - e.g. Suds, a cold one, pint, wobblypop
- Deprecation

e.g. Corona  vs 

Corona (sun) [ENVO:5555]

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

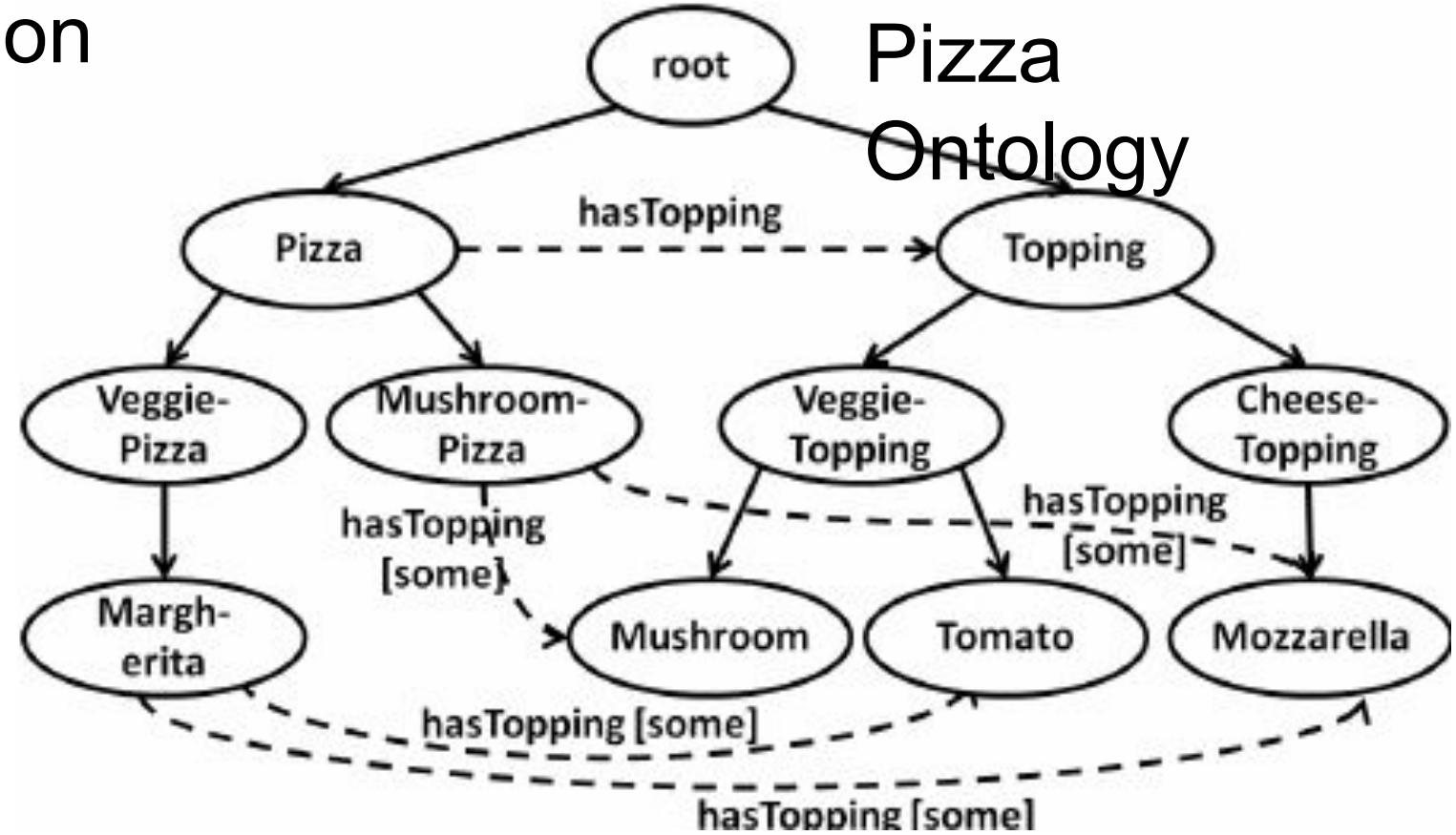
ISO slim (package of fields and terms) available:
https://github.com/GenEpiO/iso2017

Ontologies: not just lists of terms, but how the terms relate to each

Pizza Data Specification

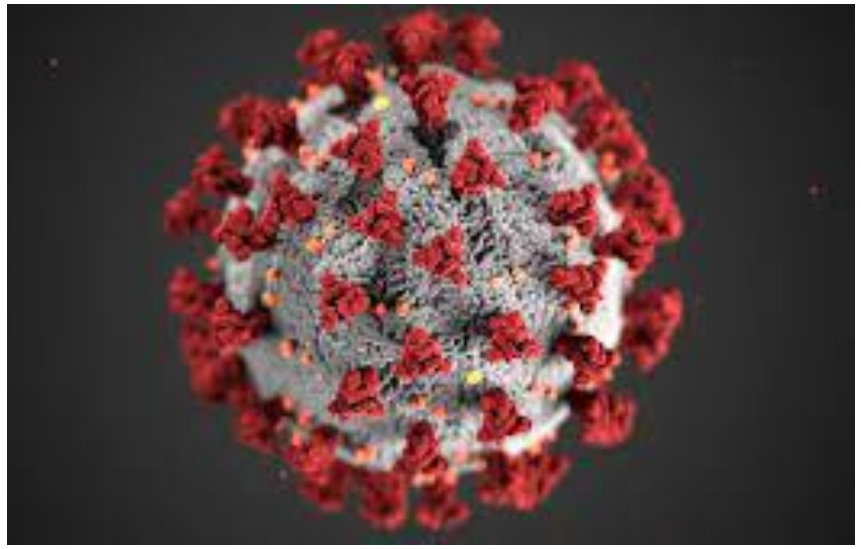
Pizza types
- Veggie pizza
- Mushroom pizza
- Margherita pizza

Toppings
- Mushroom
- Tomato
- Mozarella

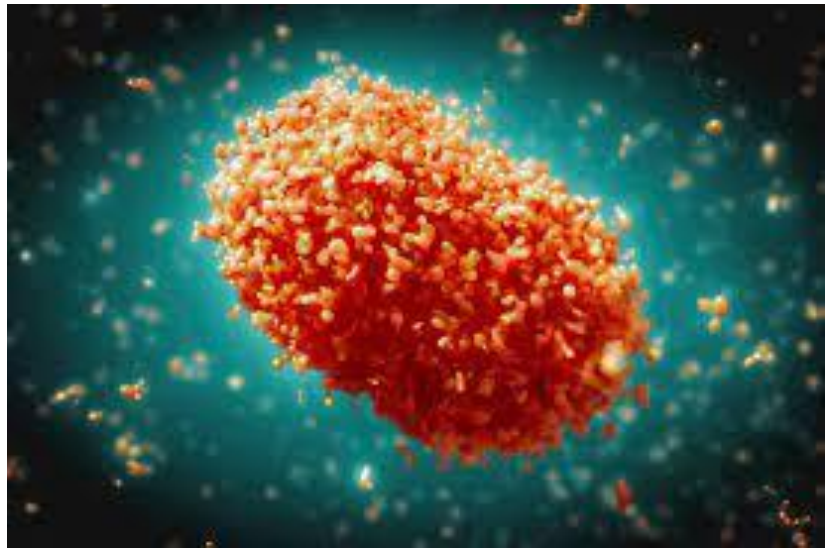


Links particular toppings to particular pizza types

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



SARS-CoV-2
(CanCOGeN)

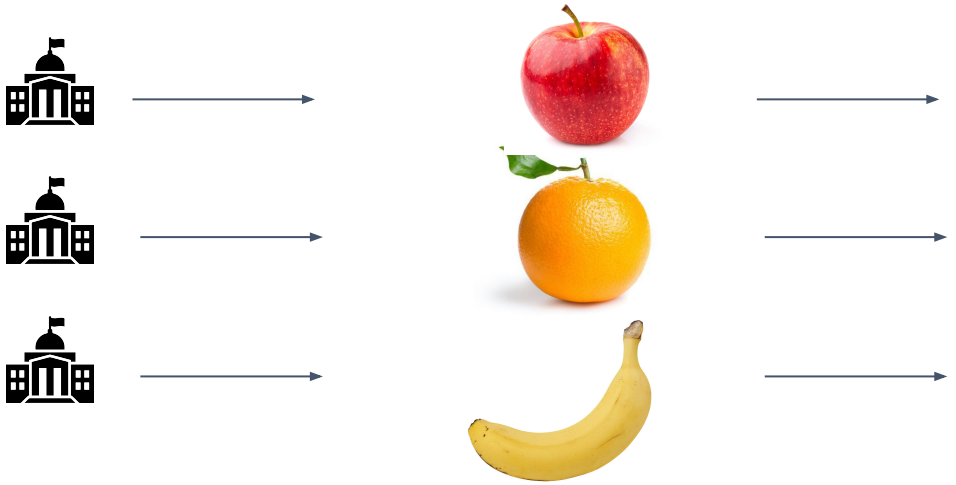


MPX
V

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

Colle ct

Data heterogeneity: slows down analysis



Integrat e

National database (NML PHAC)

National surveillance priorities (coordinated response)

Dissemina te

GISAID

NCBI

VirusSeq Data Portal

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

CanCOGeN

The CanCOGeN SARS-CoV-2 Contextual Data Standard

SARS-CoV-2 Domain Content

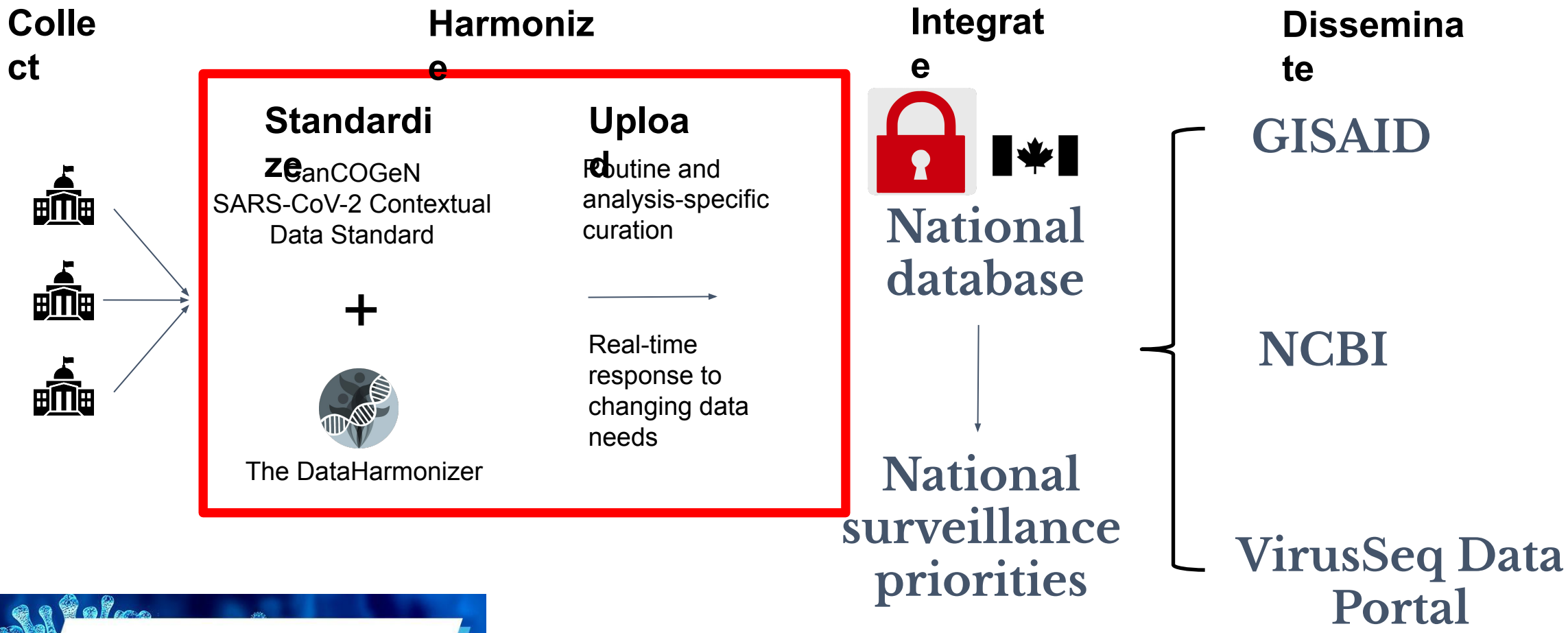
- Repository accession numbers and identifiers
- Sample collection and processing
- Host information
- Host exposure information
- Host reinfection information
- Host vaccination information
- Sequencing methods
- Bioinformatics and quality control metrics
- Lineage and variant information
- Pathogen diagnostic testing details
- Provenance and attribution

Used ~24 ontologies, >1000 ontologized terms

Standardized null values

33 required fields (who sequenced it, when/where/what was sampled, how sequenced, binfx)

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



The DataHarmonizer enables standardized data entry and validation.

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

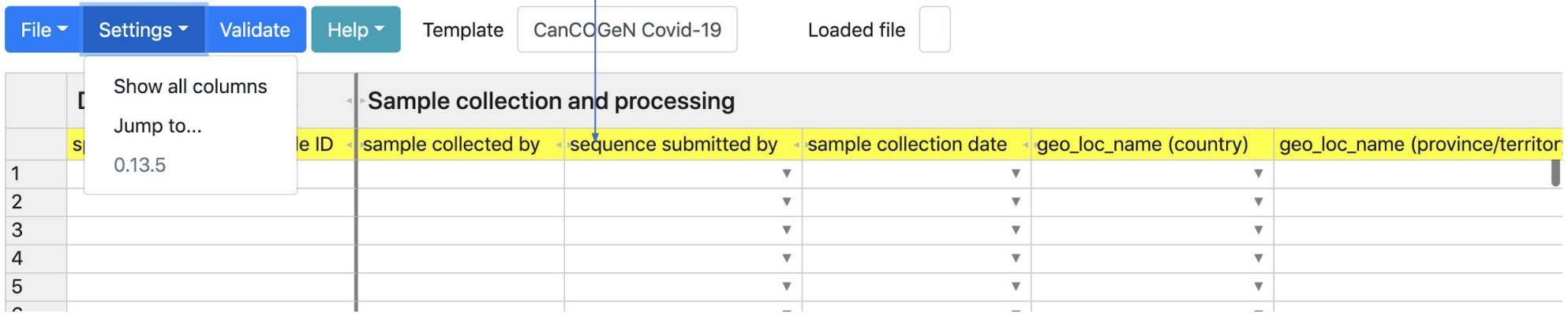
View all fields
View required fields
Move to desired field

Validate (check for errors or missing info)

Learn your way around the system

Double click on field labels for guidance on how to fill them

Save
Open existing file
Export to chosen format



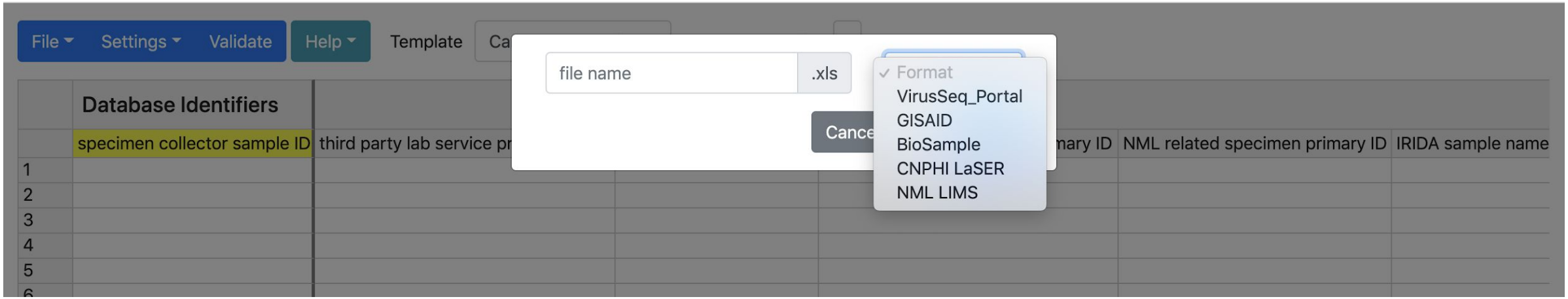
Every submitting provincial public health lab has an instance

*Data upload format: xlsx, xls, tsv and csv

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

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!

Harmonization of:

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**

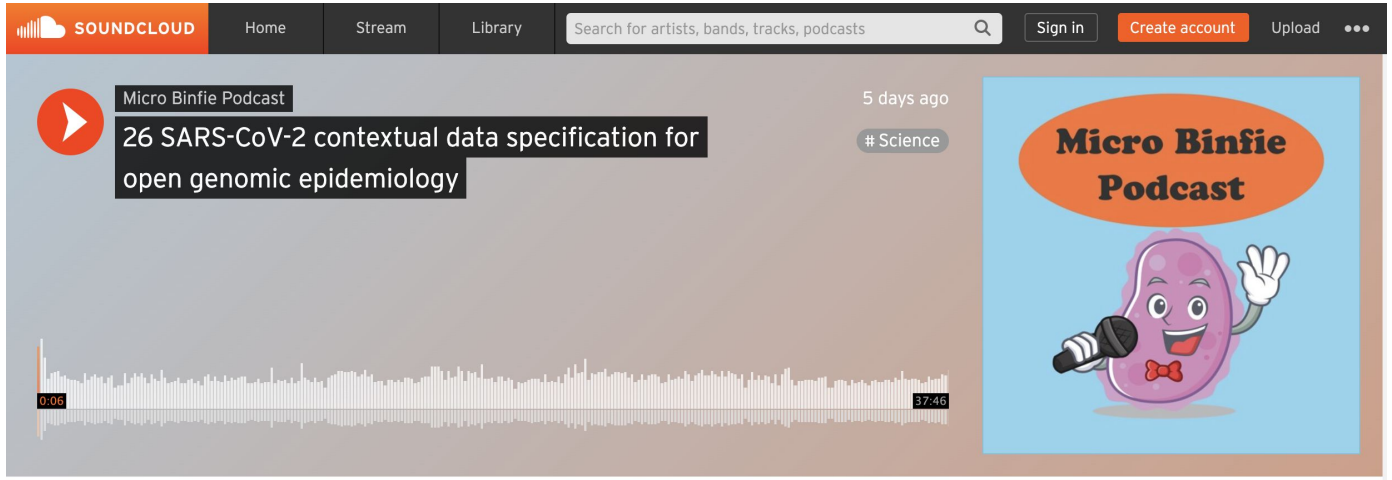
Benefits of using interoperable data standards

- time
- workload
- uncertainty

- fields/terms
- increased interoperability/ standardization

- expectations
- agreements
- skills
- tools/processes
- training, protocols

How can I learn more?



Read the manuscript:
GigaScience, Volume 11, 2022, giac003, https://doi.org/10.1093/gigascience/giac003

Podcast:
https://soundcloud.com/microbinfie/26-sars-cov-2-metadata#t=0:00

Podcast:
https://soundcloud.com/microbinfie/sars-cov-2-surveillance-in-canada-with-cancogen

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)

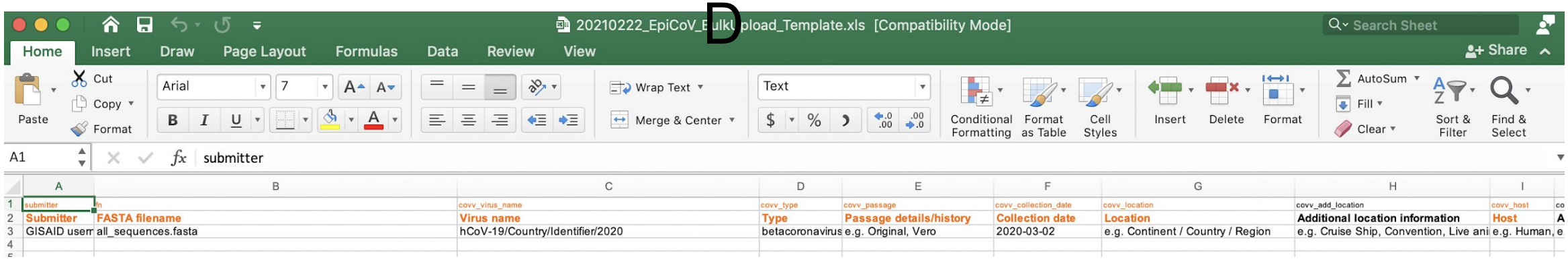
- Nodes that mirror data
- No restrictions
- Everything (all organisms, lots of data types, research & public health)
- Specific databases (assemblies, raw data, Pathogens, AMR, metagenomics, RNA-Seq, etc)

Controlled Access

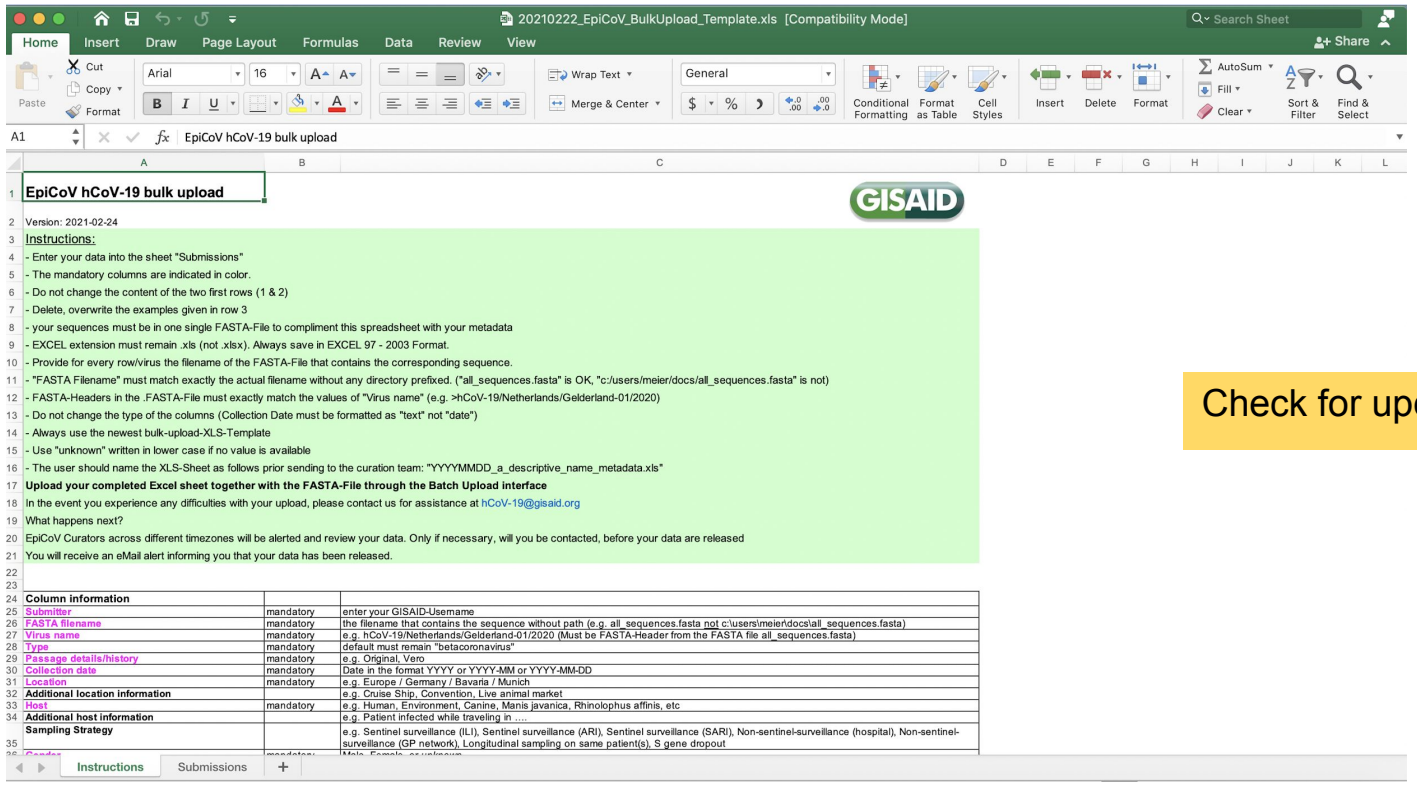
Global Initiative on Sharing Avian Influenza Data (**GISAID**)
Influenza
SARS-CoV-2
MPXV
RSV

- Data use restrictions (Terms of Service)
- Geared towards surveillance and public health
- Assembly/consensus sequence focused
- Dashboards

GISAID



	A	B	C	D	E	F	G	H	I
1	submitter	fn	covv_virus_name	covv_type	covv_passage	covv_collection_date	covv_location	covv_add_location	covv_host
2	Submitter	FASTA filename	Virus name	Type	Passage details/history	Collection date	Location	Additional location information	Host
3	GISAID usern	all_sequences.fasta	hCoV-19/Country/Identifier/2020	betacoronavirus	e.g. Original, Vero	2020-03-02	e.g. Continent / Country / Region	e.g. Cruise Ship, Convention, Live ani	e.g. Human,
4									



Column information		
Submitter	mandatory	enter your GISAID-Username
FASTA filename	mandatory	the filename that contains the sequence without path (e.g. all_sequences.fasta not c:\users\meier\docs\all_sequences.fasta)
Virus name	mandatory	e.g. hCoV-19/Netherlands/Gelderland-01/2020 (Must be FASTA-Header from the FASTA file all_sequences.fasta)
Type	mandatory	default must remain "betacoronavirus"
Passage details/history	mandatory	e.g. Original, Vero
Collection date	mandatory	Date in the format YYYY or YYYY-MM or YYYY-MM-DD
Location	mandatory	e.g. Europe / Germany / Bavaria / Munich
Additional location information		e.g. Cruise Ship, Convention, Live animal market
Host	mandatory	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

Check for updates!

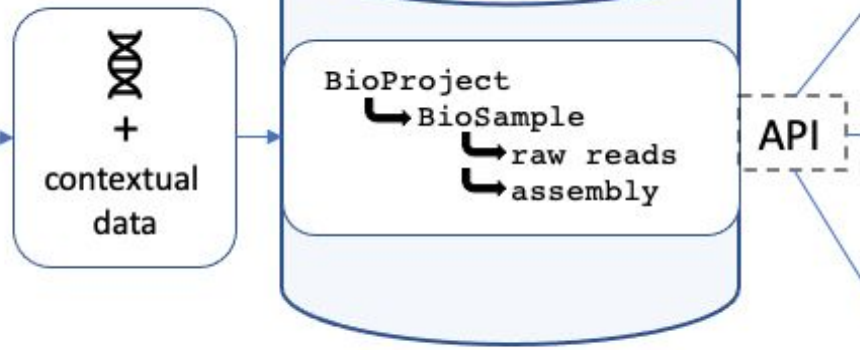
- Register and sign agreement to access templates (**attribution, collaboration, data use restrictions**)
- Template is what you will submit, includes instructions (different templates for different organisms)

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

Submission: programmatically or via web client

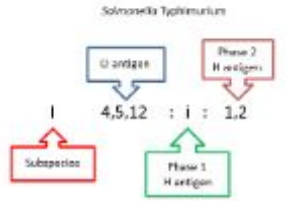


BioSample metadata
+
Experiment metadata

*both accept user-defined fields

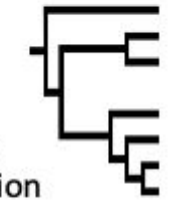
Genotyping assays

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



Phylogeny / phylodynamics

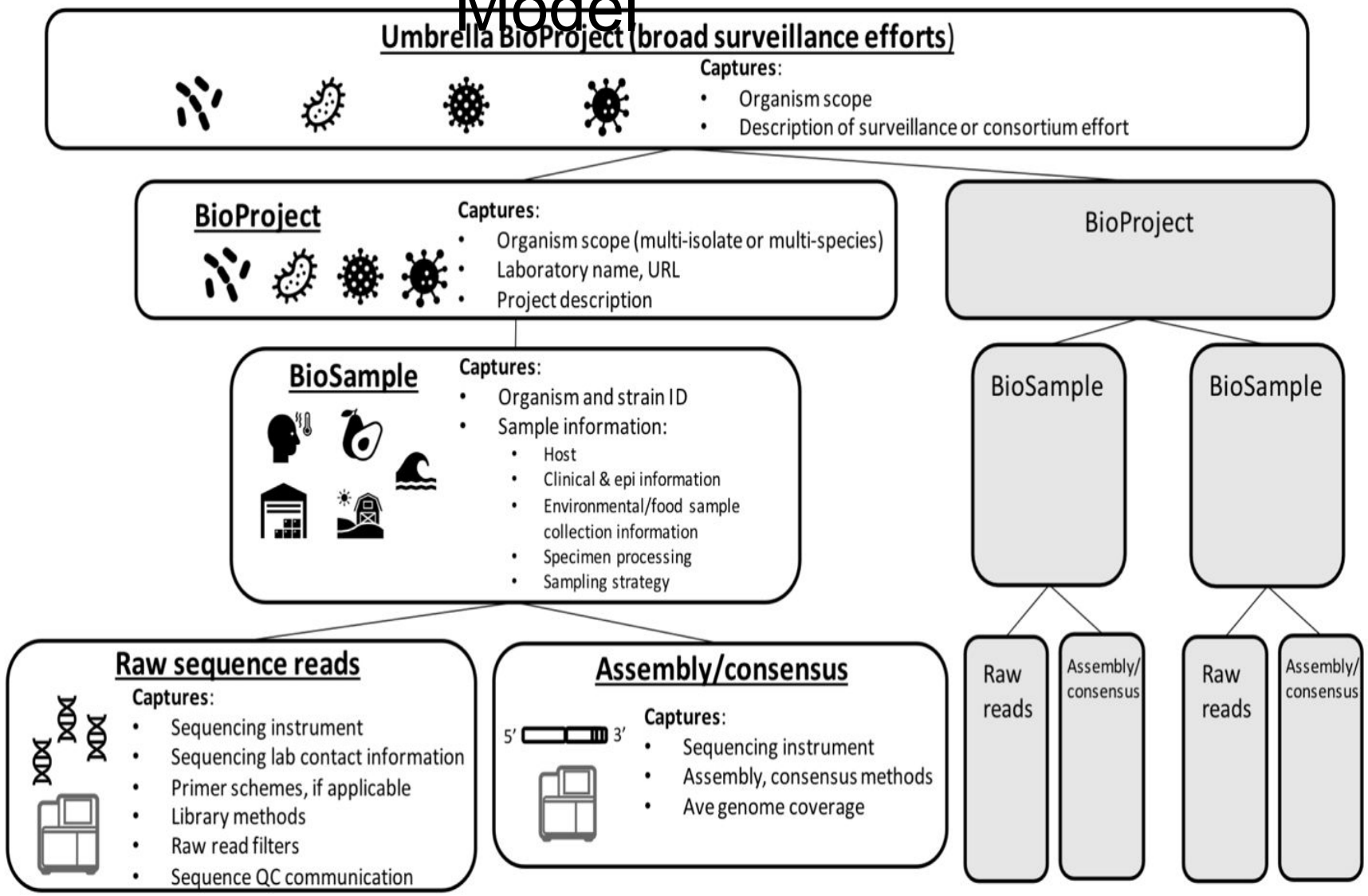
- wgMLST Clustering
- Phylogenetic tree inference
- Molecular evolution / selection
- New variant/lineage identification



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



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

Mapping Between Formats

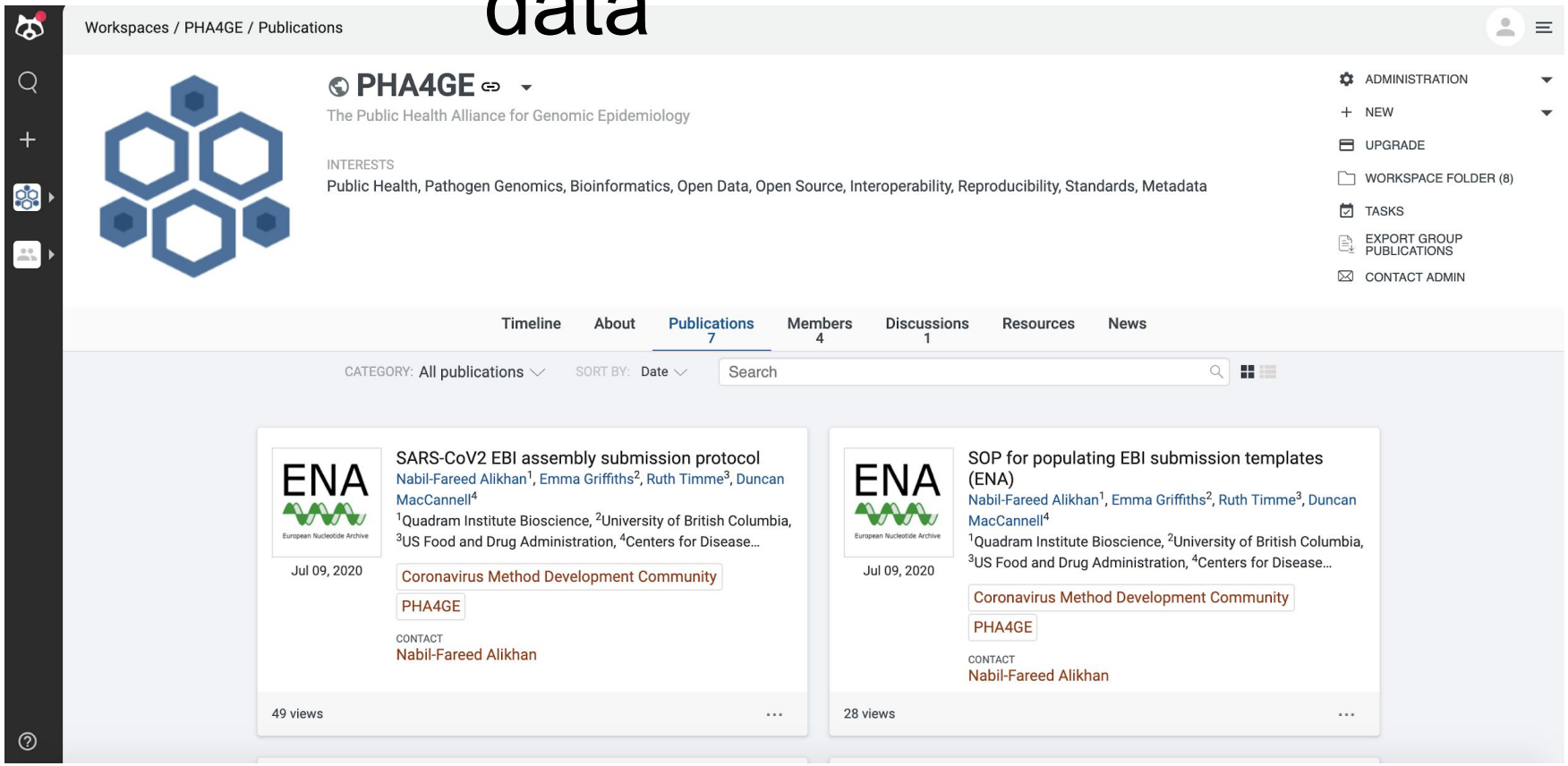
GISAID	ENA Virus Package
Submitter	No equivalent (submit from your account)
FASTA filename	**file_name** (See Experiment metadata)
Virus name	**isolate**
Type	**tax_id** (See Experiment metadata)
Passage details/history	No equivalent
Collection date	**collection date**
Location	**geographic location (country and/or sea)**
Host	**host common name/host scientific name**
Gender	**host sex**
Patient age	**host age**
Patient status	**host health state**
Sequencing technology	**instrument_model** (See Experiment metadata)
Originating lab	**collecting institution**
Address	**collecting institution**
Submitting lab	No equivalent (submit from your account)
Address	No equivalent (submit from your account)
No equivalent	**host subject id**

GISAID Virus name:
hCoV-19/Country/ISO regional code-Identifier/year
hCoV-19/Country/**un**-Identifier/year
e.g. hCoV-19/CANADA/BC-ABCD1234/2021

NCBI Isolate:
SARS-CoV-2/host/country(short)/sampleID/date
e.g. SARS-CoV-2/human/CAN/ABCD1234/2021

*remember, even if something is "required", you can always provide a null value if you need to
e.g. Missing, Not Applicable, Not Collected

Protocols to mobilize harmonized data



- **7 public repository submission protocols** (GISAID, NCBI, EMBL-EBI) on **Protocols.io**
- **PHA4GE-adapted submission forms**

- **Curation SOP**
- **instructional videos**

Different repositories have different fields, but PHA4GE helps standardize what goes into those fields

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

Public Health Alliance for Genomic Epidemiology

Other kinds of data useful for genomic epidemiology:

Best curation practices for associating genomic data with assays, phenotypes, epidemiological data etc

Preparing FAIR data to maximize reuse and reproducibility

FAIR Guiding Principles (Nature, 2015)

Findable - other users can discover your data (humans AND computers)
Accessible - should be easy to retrieve
Interoperable - computer systems or software should be able to exchange and make use of information
Reusable - digital information can have many uses, should be stored in a way that enables reuse (e.g. sufficient context, provenance)

How do we make data FAIR?

Everyone: Use ontologies and data standards (datasets and systems more interoperable)
Standards developers: Make standards/ontologies available in registries/centralized resources (OBO Foundry)
Data providers: Share data in databases and repositories
Provide information on provenance, data limitations
Repositories: Make data searchable with standardized filters/tags

How you build an ontology (architecture) affects how you can use it.

Open Biomedical Ontologies e.g. **Gene Ontology (GO)**
http://www.obofoundry.org/

- *integrate information across sources*

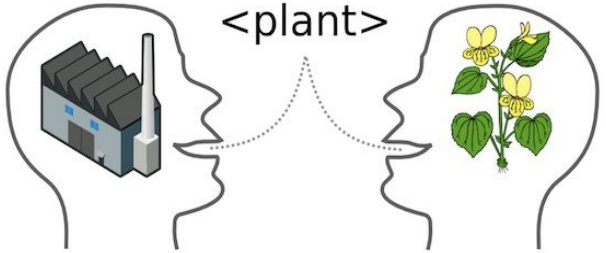
Interoperability depends on:

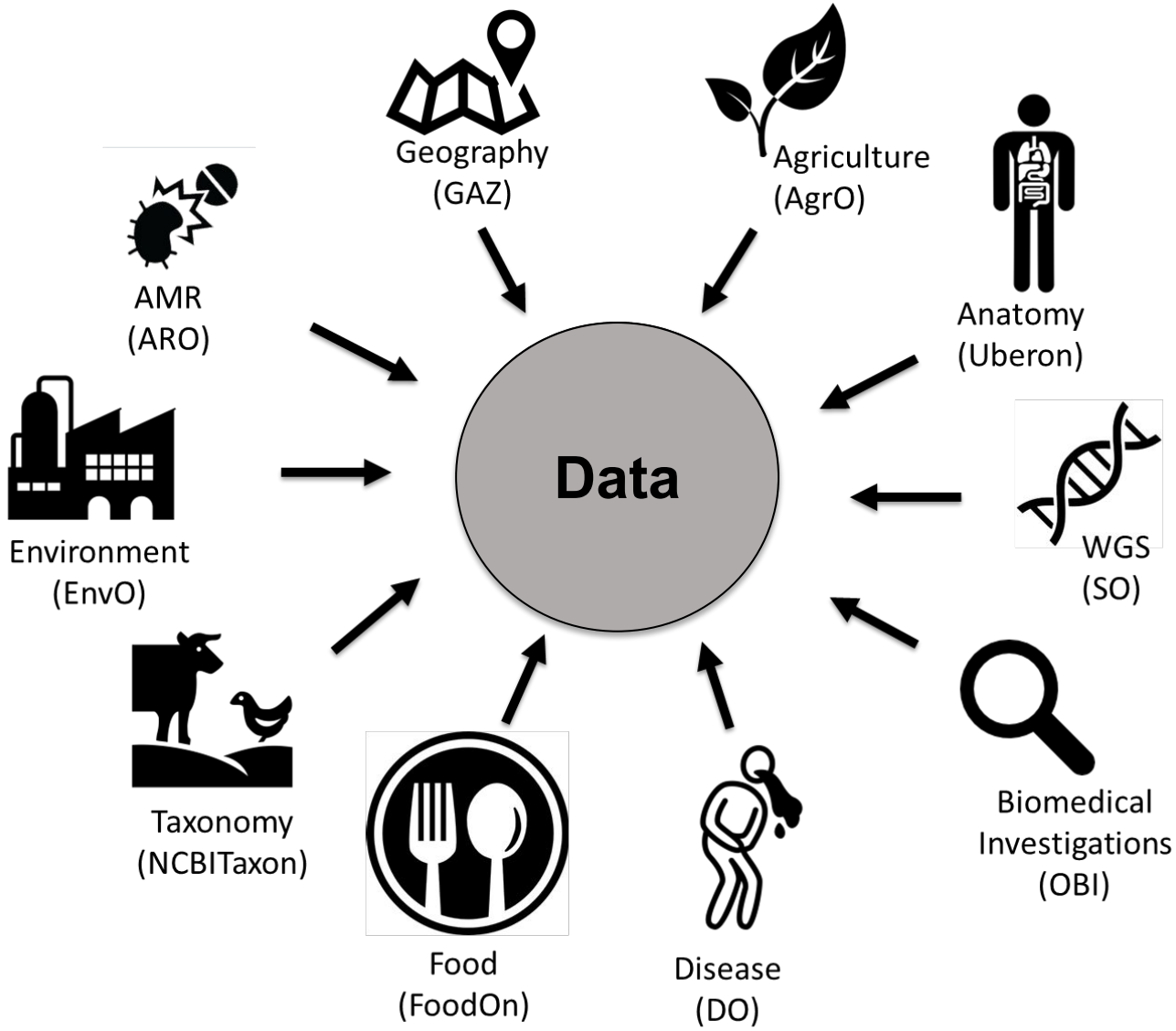
- common architecture

Basic Formal Ontology: how to group things into classes e.g. things, processes, qualities

Relation Ontology: prescribed relations e.g. *is_a*, *adheres_in*, *part_of*, *has_role*

- different ontologies, reuse of terms
- oversight (centralized)
- open source





OBO Foundry library (>200 ontologies)

How to find ontology terms

We will use this later in the lab!

Ontology Lookup Service

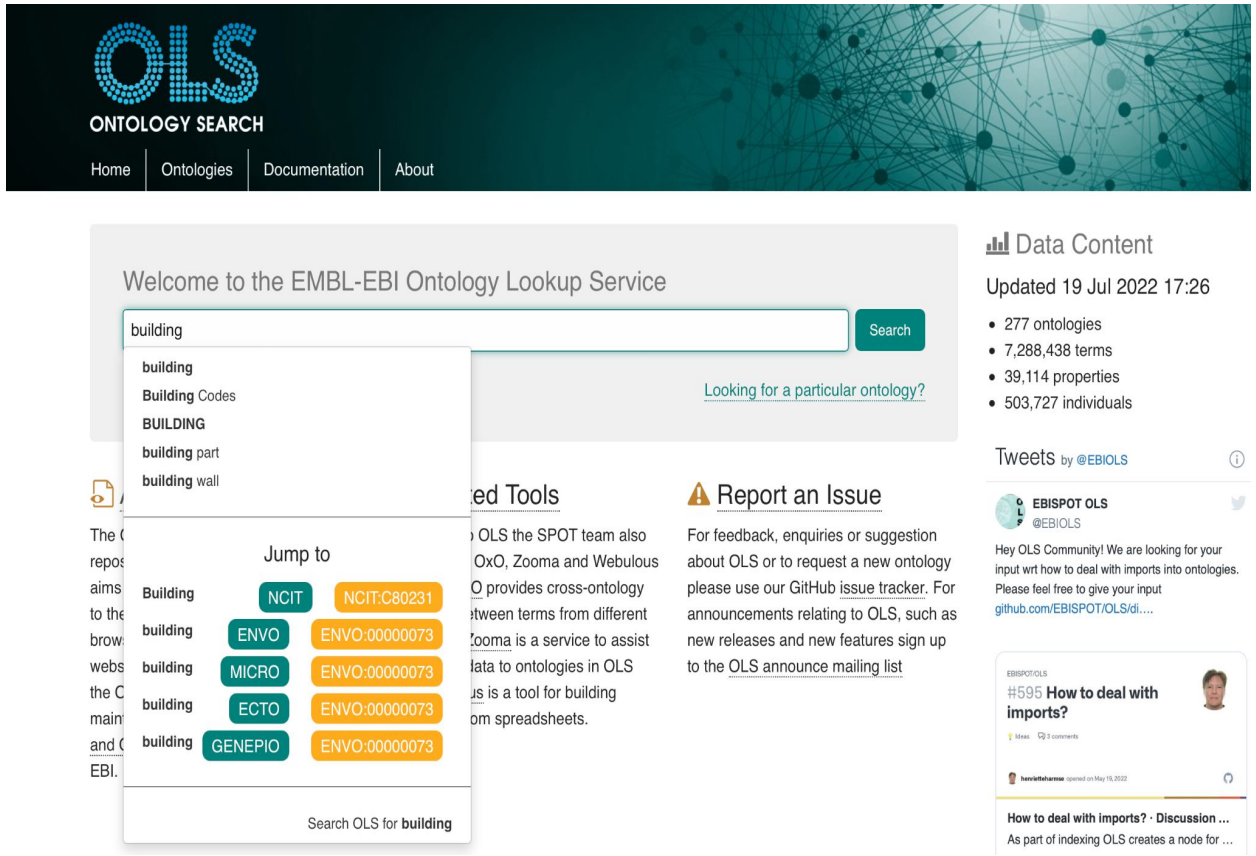
Service to make ontology accessible for human (data curators) and machines (through workflows).

Many routes to the same term

EMBL-EBI OLS

A repository for all Open Biomedical (OBO) Ontologies

- OLS hosts over 250 ontologies in a single place.
- Over 7 million terms.



Website: https://www.ebi.ac.uk/ols

Standardization using ontology terms: medical terms (drugs)

*table contains mock data

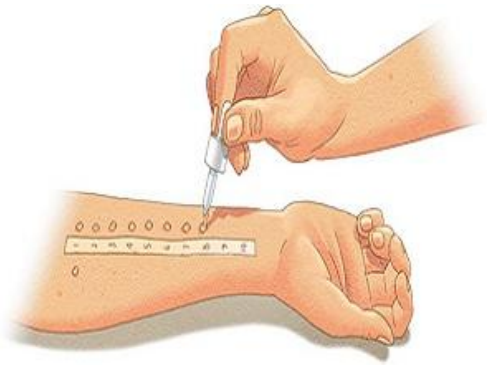
Child ID	Mother ID	Medication_1yr	Medication_1yr (onto_label)	Medication_1yr (onto_ID)
C12345	M23456	Tylenol, antibiotic	acetaminophen antibiotic	CHEBI:22582 CHEBI:46195
C12346	M23457	gentamicin	gentamicin	CHEBI:17833
C12347	M23458	Children's tylenal, ritalyn	acetaminophen methylphenidate	CHEBI:46195 CHEBI:6887
C12348	M23459	Paracetamol neomycin	acetaminophen neomycin	CHEBI:46195 CHEBI:7507

- Chemical Entities of Biological Interest (ChEBI) ontology
- Active ingredients standardized
- Identify trends more easily
- Use "universal language"

Comparing variables with implicit interpretation criteria

Experimental variables:

- implicit assumptions
- methods and interpretation criteria



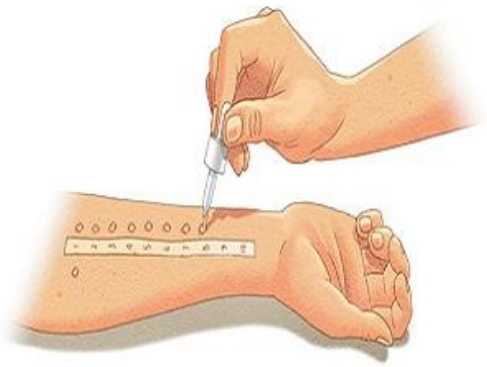
e.g. “positive skin prick test result”

Different thresholds
2mm (Cohort A) vs 5mm (Cohort B)

= different interpretations, different results

Ontologies can resolve different usage/meaning of the same terms

- differentiate variables → make methods and interpretation criteria more explicit



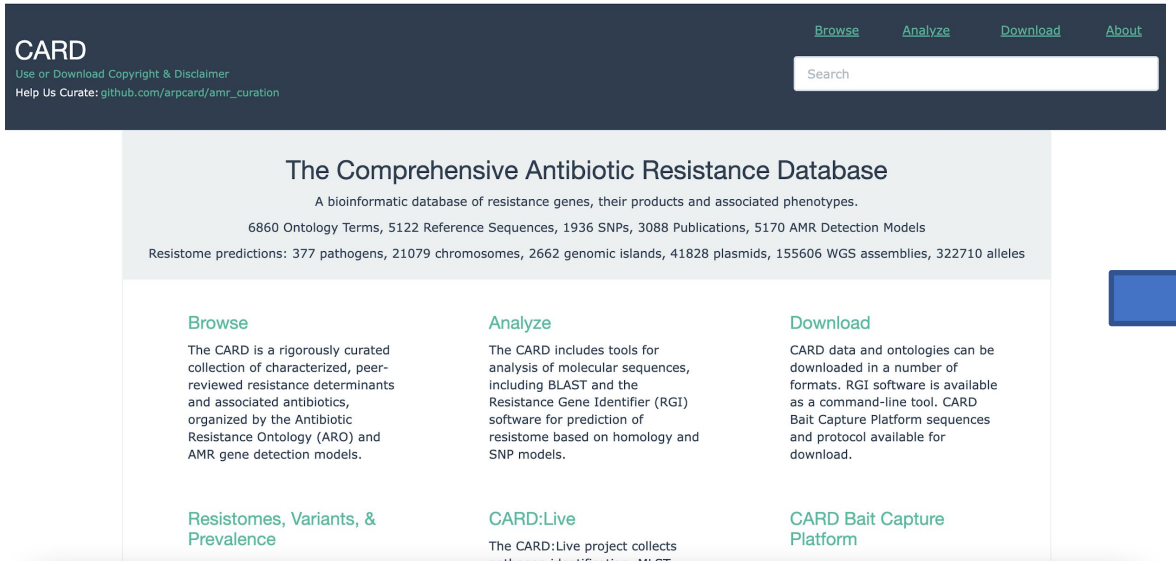
AllergyOnto:1234*
positive allergy skin prick test result (2mm)

AllergyOnto:5678*
positive allergy skin prick test result (5mm)

*these terms do not yet exist, for demonstration purposes only

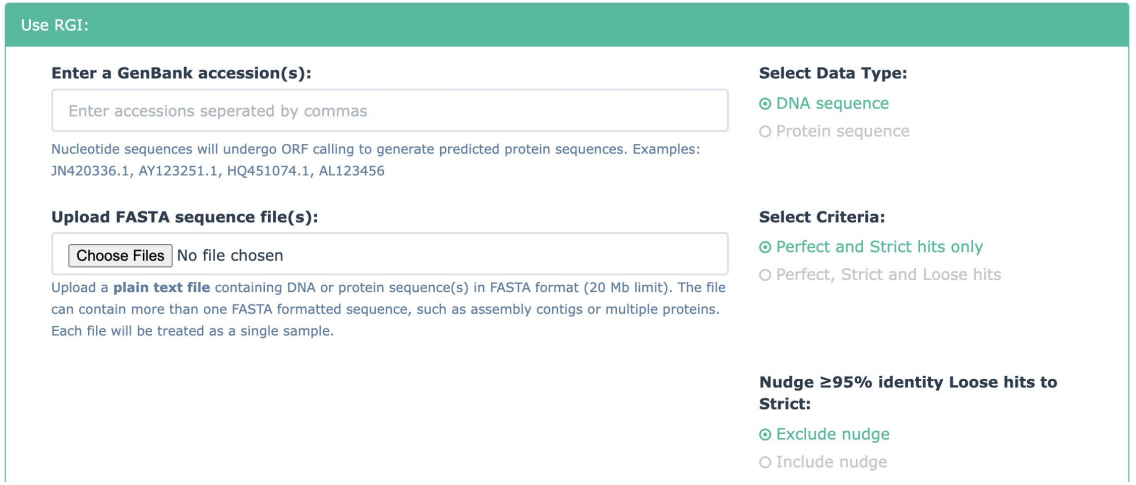
Comprehensive Antimicrobial Resistance Database (CARD)

- associates genomic data with resistance genes, mechanisms of action, drug classes and more
- EVIDENCE based
- curators identify data and evidence
- create data model (structured using ARO)
- need system to ensure facts are correct!



https://card.mcmaster.ca/

- Underpins the Resistance Gene Identifier (RGI)
- Tool for analysis, identifying genomic resistance markers



https://card.mcmaster.ca/analyze/rgi

Data 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 facts into actionable knowledge)

Summary

- Ontology-based data specifications increase interoperability of datasets and systems
- Data management tools (e.g. DataHarmonizer) help operationalize data standards
- 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
- Data standardization tools and best practices help build knowledge bases useful for research

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:

Canadian Centre for Computational Genomics