



CENTRE FOR
INFECTIOUS DISEASE
GENOMICS AND
ONE HEALTH



Putting standards into practice: pathogen genomics contextual data (“metadata”) standards in public health and food safety

Emma Griffiths, PhD

Centre for Infectious Disease Genomics and One Health

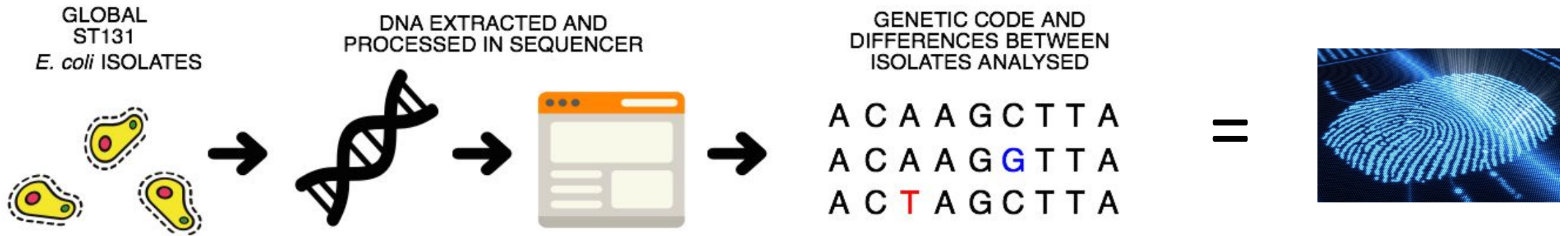
Faculty of Health Sciences, Simon Fraser University

DCMI 2021 – Oct 7

Outline

1. Contextual data is critical for genomic surveillance.
2. Challenges of sharing/using contextual data.
3. Ontologies and data standards as solutions.
 - Tools, implementations
4. Importance of contextual data in COVID-19 response

Microbial genomic sequences can be used as a molecular fingerprint to trace the source of infectious disease.



- Public health/food safety agencies exchange information about these fingerprints



(Dramatic representation from the movie *Outbreak*)

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

Challenges in public health/food safety data sharing



Lack of a mechanism for data sharing
between partners (technical, governance)

Security/privacy

Control over data (attribution, IP, politics)

Semantic interoperability

Manual curation

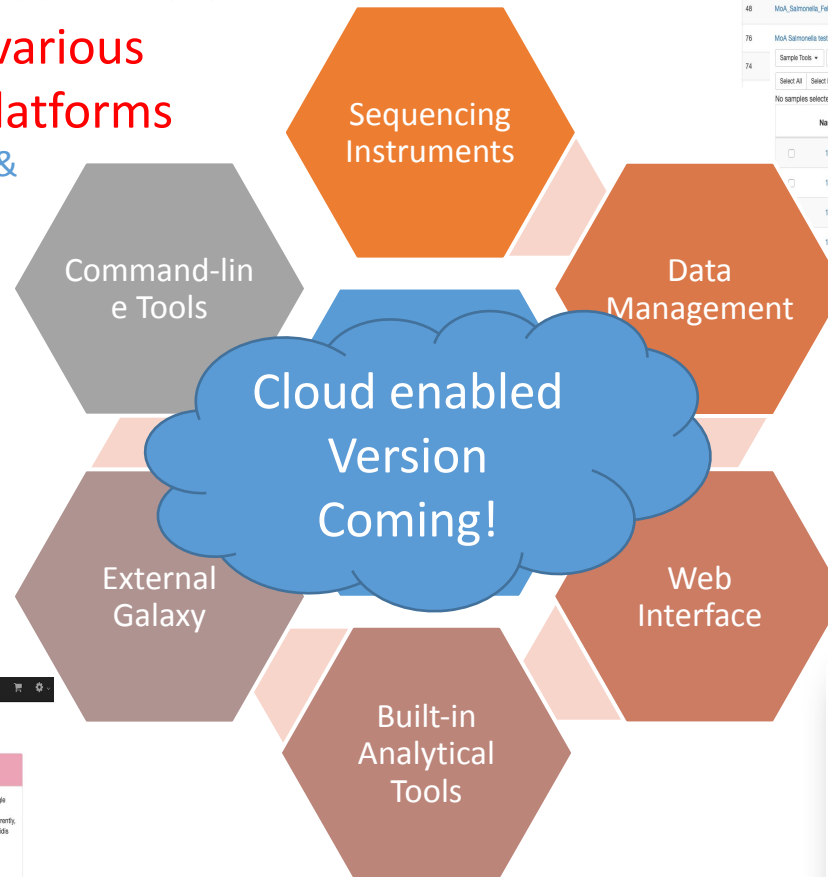


IRIDA

A comprehensive and distributed platform for pathogen genomics analyses and data management

Easy, automated transfer
of data from various
sequencing platforms
(GitHub: SeqUDAS &
IRIDA Uploader)

Reproducible and
versioned workflows
(Assembly, annotation,
SNP, MLST, AMR,
serotyping, and more.)



IRIDA Projects

ID	Name	Organism	Samples	Created	Modified
48	Mock_Salmonella_Feb2018	Salmonella enterica subsp. enterica serovar Enteritidis	176	Feb 9, 2018 9:45 PM	May 2, 2019 2:31 PM
76	Mock_Salmonella_Test	Salmonella	176	May 2, 2019 10:57 AM	May 2, 2019 11:52 AM

Sample Tools Export Add to Cart

Select All Select None

No samples selected

Name	Organism	Project	Created On	Modified On
17-945		Mock_Salmonella_Feb2018	Feb 9, 2018 10:46 AM	Feb 13, 2018 1:51 PM
17-768		Mock_Salmonella_Feb2018	Feb 9, 2018 10:46 AM	Feb 13, 2018 1:51 PM
17-4221		Mock_Salmonella_Feb2018	Feb 9, 2018 10:46 AM	Feb 13, 2018 1:51 PM
17-4155		Mock_Salmonella_Feb2018	Feb 9, 2018 10:46 AM	Feb 13, 2018 1:51 PM
133		Mock_Salmonella_Feb2018	Feb 9, 2018 10:46 AM	Feb 13, 2018 1:50 PM
29		Mock_Salmonella_Feb2018	Feb 9, 2018 10:46 AM	Feb 13, 2018 1:50 PM

Project and
Sample Management

Simple User Interface

IRIDA Projects Analysis Search Help

Select a Pipeline

Your cart has no local samples selected. Select some samples from one of your Projects to run a pipeline.

Assembly and Annotation Pipeline

Generate an assembled and annotated genome from the reads within a sample using PLINK, SPARK, and Prokka. Outputs analyzed and produced separately for each sample include: log files, assembly statistics, the complete set of contigs, filtered contigs with repeats, filtered contigs without repeats, and annotations from Prokka.

Assembly and Annotation Collection Pipeline

The assembly and annotation collection pipeline provides the same results as the assembly and annotation pipeline, but all samples are analyzed together which allows you to download a single package for all samples submitted.

bio_harvest Pipeline

Rapidly subtype microbial genomes using single nucleotide variant (SNV) subtyping schemes (https://github.com/bio-harvest/bio-harvest). Currently, only available for subtyping Salmonella Enteritidis and Salmonella Heidelberg genomes.

IslandViewer Genomic Island Prediction

IslandViewer is a computational tool that integrates four different genomic island prediction methods: IslandPick, IslandPath-DMCB, SIG-HMM, and Islander.

Mentalist MLST Pipeline

Genotype bacterial samples directly from reads, using an efficient k-mer based algorithm.

SNVPhyl Phylogenomics Pipeline

Generate a Whole Genome Phylogeny from a set of samples and a reference genome based on Single Nucleotide Polymorphisms (SNPs) using the SNVPhyl pipeline. This will provide a dendrogram as well as a table of all SNPs used and a SNV distance matrix between each sample.

Expanding Analytical Workflows
with plug-in architecture

SNVPhyl Phylogenomics Pipeline

Ready to Launch?

Pipeline Name: SNVPhyl_20170914

Reference File: 08-5578.fasta (Project 1) Upload New

Parameters: Default Parameters Customize

Description: Enter description of the analysis (Optional)

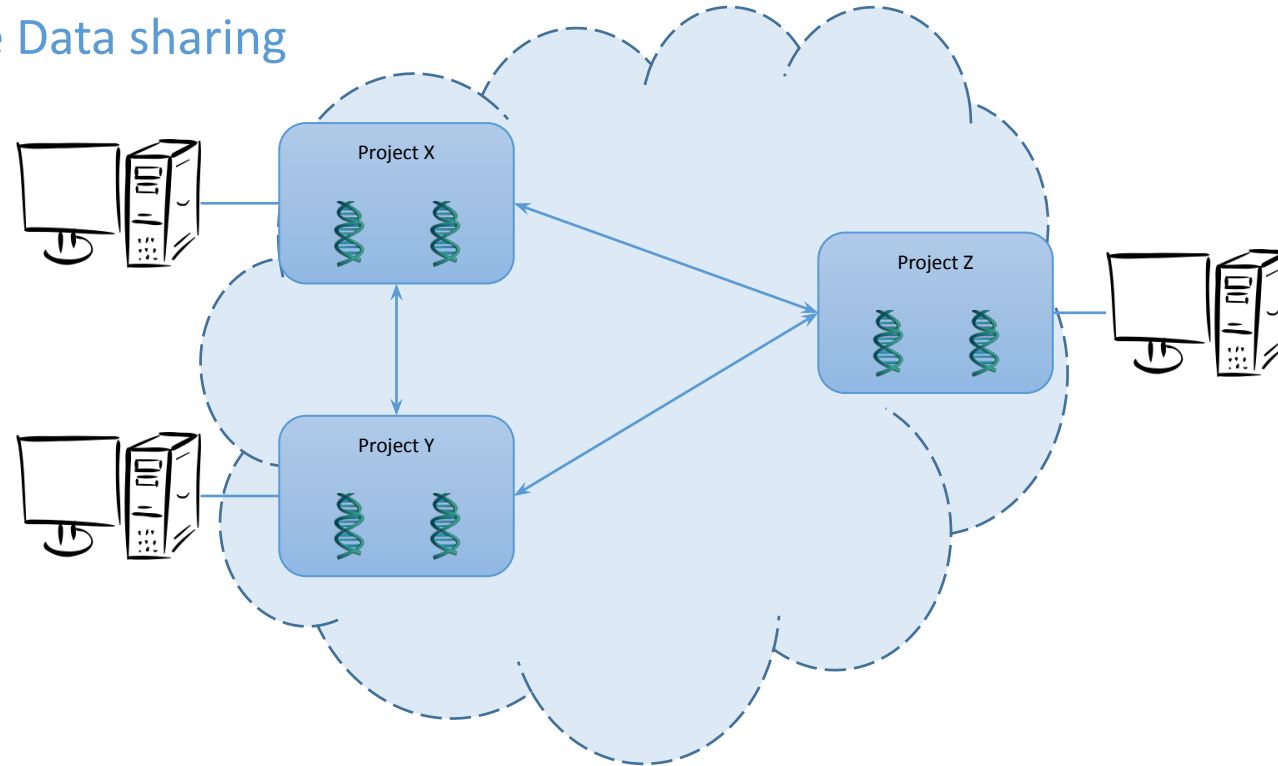
Share Results Project 1

Files: Project 1 / sample1 Remove

08-5578_S1_L001_R1_001.fastq	65.9 MB	14 Sep 2017
08-5578_S1_L001_R2_001.fastq	65.9 MB	14 Sep 2017

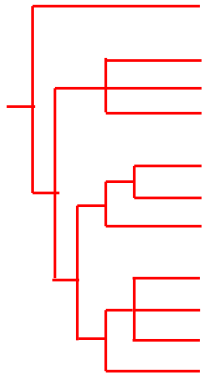
Project 1 / 03-3333 Remove

Decentralized Architecture:
Allow flexible Data sharing



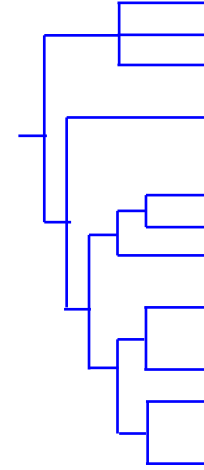
- Goal: fine grained access controls – assign permissions to standardized fields
- users specify fields of contextual data to share

Instance A



ID	SampleType	Commodity	Farm	Geo	VEpi

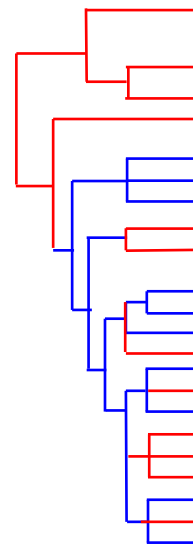
Instance B



ID	SampleType	Epi Associations	Case#	Geo



Shared Data Between Instances



ID	SampleType	Geo	Commodity	Farm	VEpi	Epi Associations	Case#

Both sites
can access


Site A
access only

Site B
access only

Harmonizing fields of data is challenging.

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

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*

...so, you can't just combine fields of data.

Harmonizing data values is challenging.

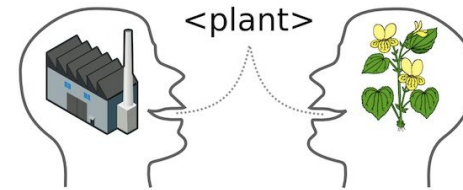
Free text = 

Arugala
Arugila
Arrugula

Errors

Frz chk brst

Short hand



Semantic ambiguity

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

Formats



**Inconsistently
collected**

Swab
Swab of walls inside egg incubator #123

Granularity

Data clean up takes time.
Ultimately impacts time to response.

Variability impacts how you understand/use data.

Variability in data structures between projects, labs and databases due to:

- Different **fields**
- Different **terms**
- Different **meanings**
- Different **granularity of information**
- Different **information types collected**

BEFORE



AFTER



Contextual data standards
improve data harmonization
and integration.


Data standards: Solutions for standardizing contextual data

Minimum Information Checklists:

- **Prescribed fields to describe something (in a particular context)**
- created by authoritative source (**Genomics Standards Consortium**)
- core fields common between checklists (e.g. collected by, collection date)
- specific packages (e.g. air, human gut)
- commonly used by international sequence repositories

e.g. **MIxS** (Min Info About and Seq)

- Field Label
- Definition
- Data type & Syntax
- Examples of use

Specification projects	MIGS	MIMS	MIMARKS	New checklists
Checklists		metagenomics	survey	specimen
Shared descriptors	collection date, environmental package, environment (biome), environment (feature), environment (material), geographic location (country and/or sea, region), geographic location (latitude and longitude), investigation type, project name, sequencing method, submitted to INSDC			
Checklist specific descriptors	assembly, estimated size, finishing strategy, isolation and growth condition, number of replicons, ploidy, propagation, reference for biomaterial		target gene	
Applicable environmental packages (measurements and observations)	<div><div>Air</div><div>Host-associated</div><div>Human-associated</div><div>Human-oral</div><div>Human-gut</div><div>Human-skin</div><div>Human-vaginal</div></div> <div><div>Microbial mat/biofilm</div><div>Miscellaneous natural or artificial environment</div><div>Plant-associated</div><div>Sediment</div><div>Soil</div><div>Wastewater/sludge</div><div>Water</div></div>			

Ontology, A Way of Structuring Information

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

- Controlled (standardized) vocabulary
- Hierarchy (granularity)
- Logic
- 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)

e.g. Food Ontology mapping to **SIREN food database**

Chickpea (whole): FOODON_03306811 *broad_synonym* **CHICK PEA, RAW (F6118)**



How you build an ontology affects how you can use it.



Open Biological and Biomedical Ontology Foundry

e.g. **Gene Ontology (GO)** <http://www.obofoundry.org/>

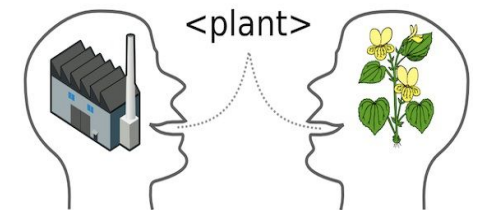
Interoperability depends on:

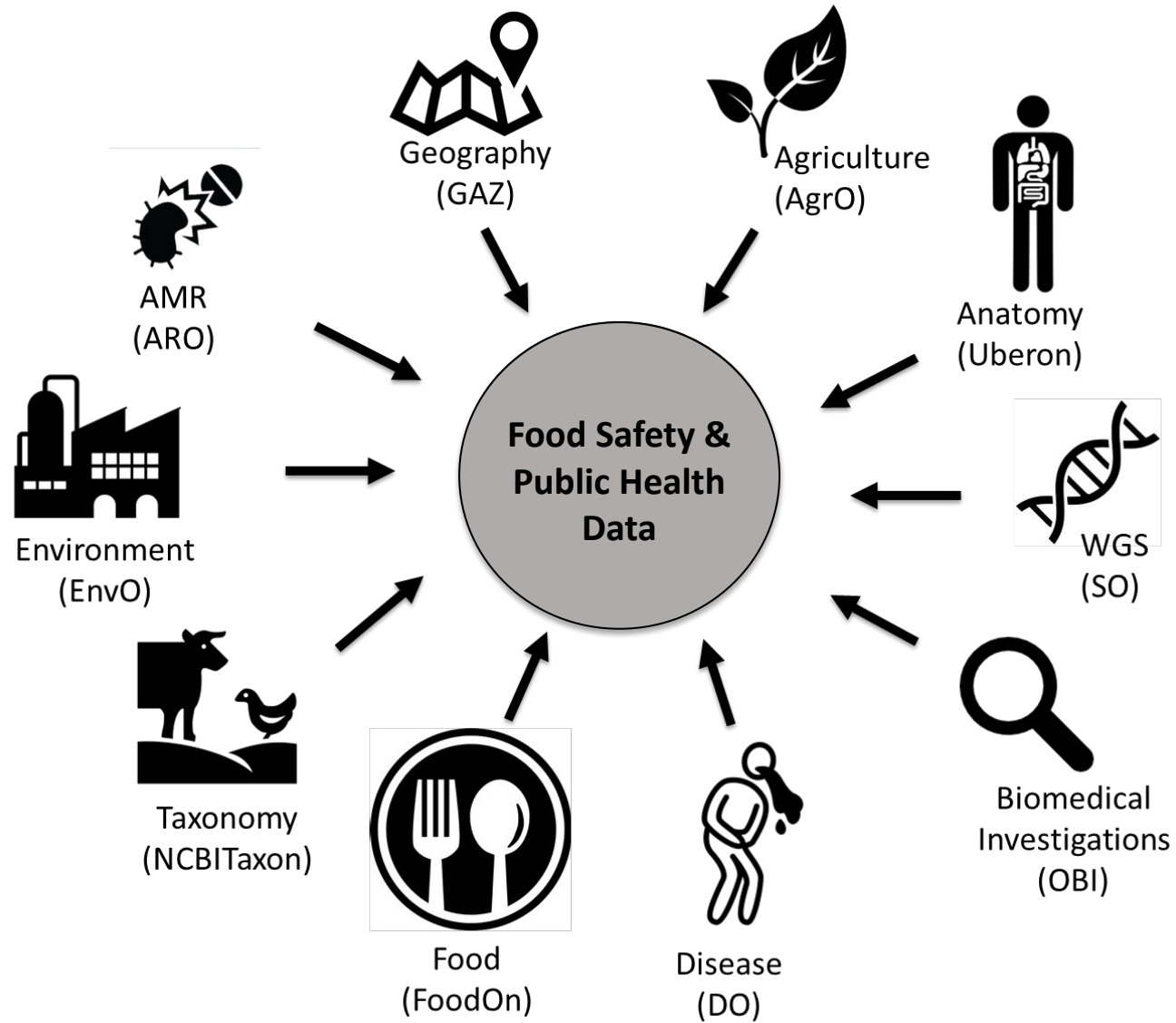
- common architecture

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*

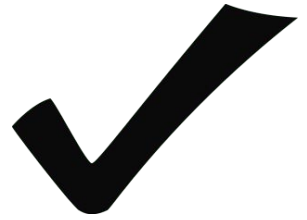
- different ontologies, reuse of terms
- oversight (centralized)
- open source



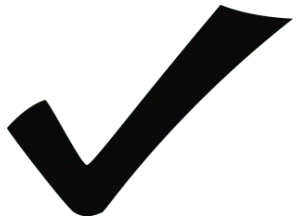


OBO Foundry library (>150 ontologies)

What can data standards do for genomics data?



1. Data is more interpretable by **humans and computers**



2. **Future-proof** contextual data



3. Harmonize and integrate data across labs/databases (**interoperability**)



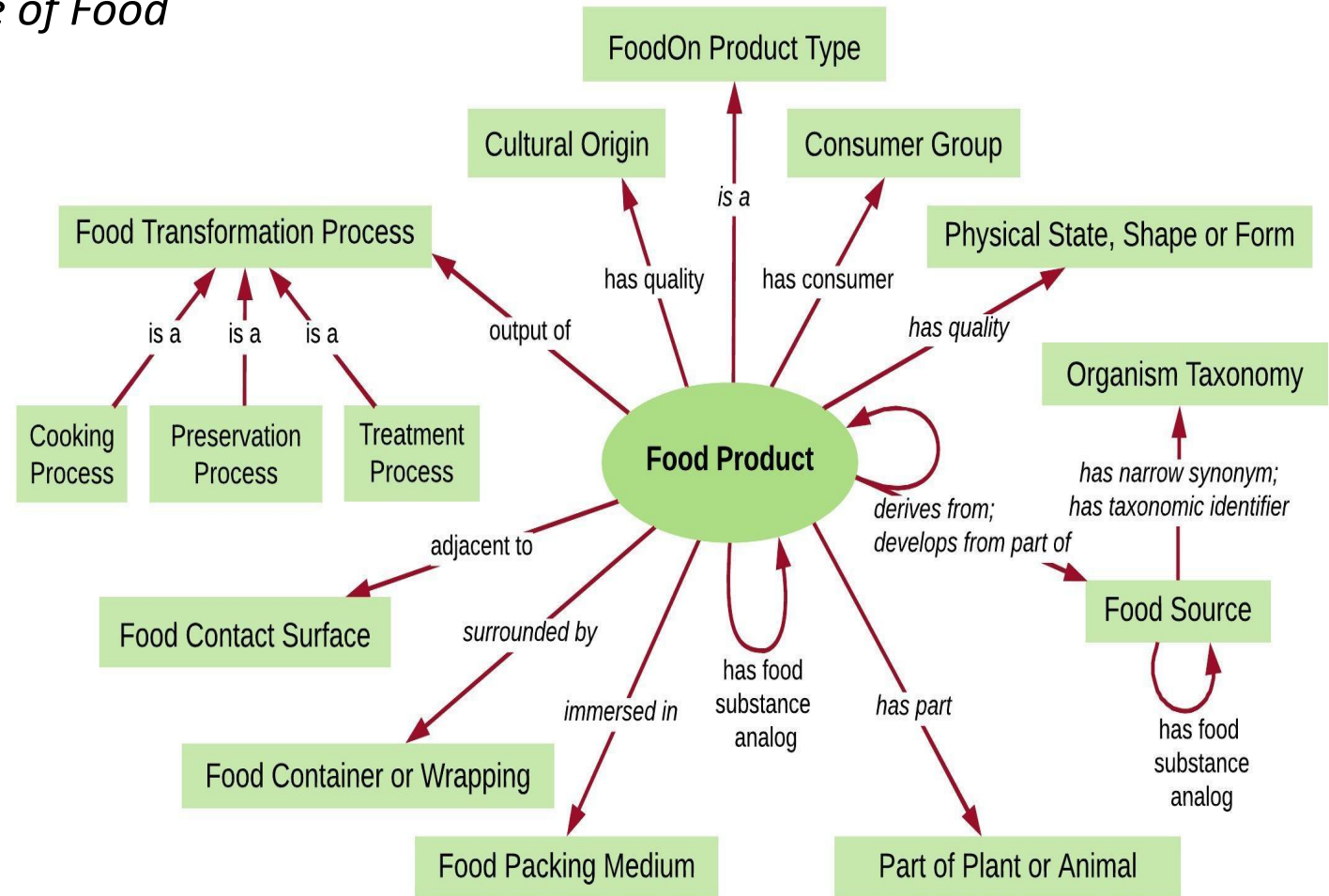
4. **Tools**

Ontologies: The Food Ontology (FoodOn)



Dooley, Griffiths et al (2018), *Nature: Science of Food*

- >28K terms for **food products, feed, sources and processes**
- **interoperable architecture**
- characterizes products by **facets**
- enables **mapping** between international food schemes (e.g. EFSA's FoodEx2)
- FoodOn consortia



Ontologies: Genomic Epidemiology Ontology (GenEpiO)

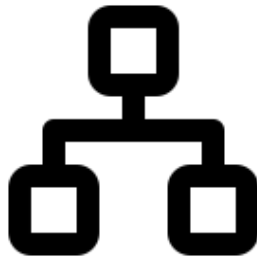
Aim: integrating genomics, lab, clinical and epidemiological data critical for WGS-based microbial pathogen investigations



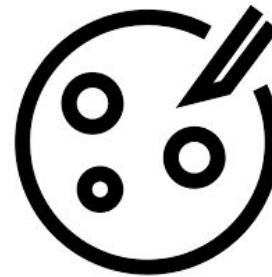
Environments



Anatomy & Body
Products



Taxonomy



Assays, Methods,
Devices



Information
Entities



Genomics

- >5200 terms for describing samples, contexts, instruments, analyses



Implementation: Tracking foodborne pathogens in real-time around the world.

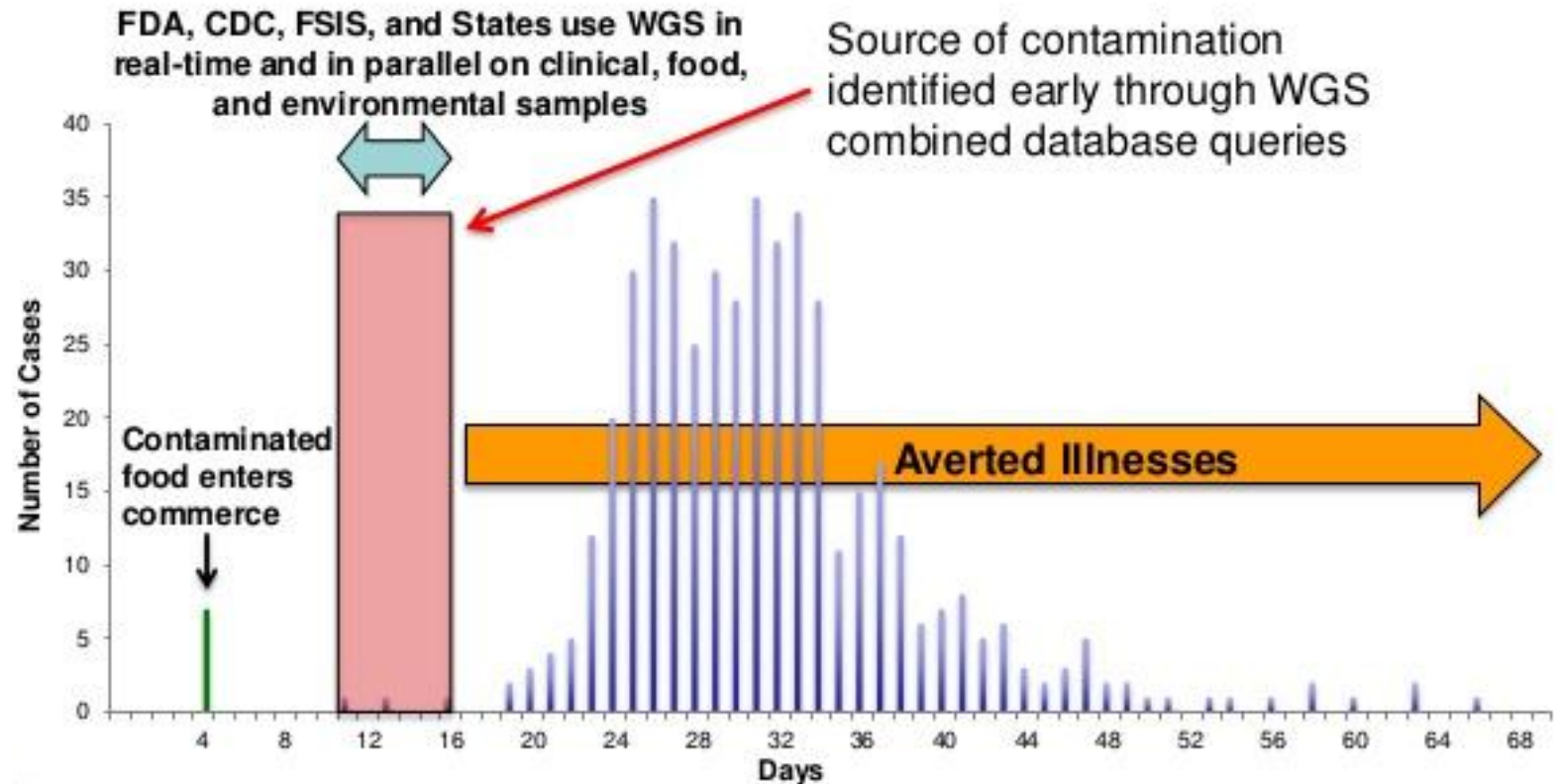
"Tinder for genomics data"

- Global network
- WGS-based surveillance and source tracking of foodborne pathogens
- brokered by FDA, db housed by NCBI

Contains >630k genomes

Free text descriptions of food sources

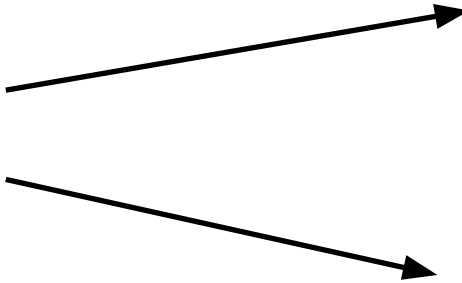
Representative* Timeline for Foodborne Illness Investigation Using Whole Genome Sequencing



Food descriptions vary depending on culture, geographic regions, institutional conventions, how the descriptions are being used.

“Biscuit”

(What does this mean?
Depends who you are talking
to...)



Scone



Cookie

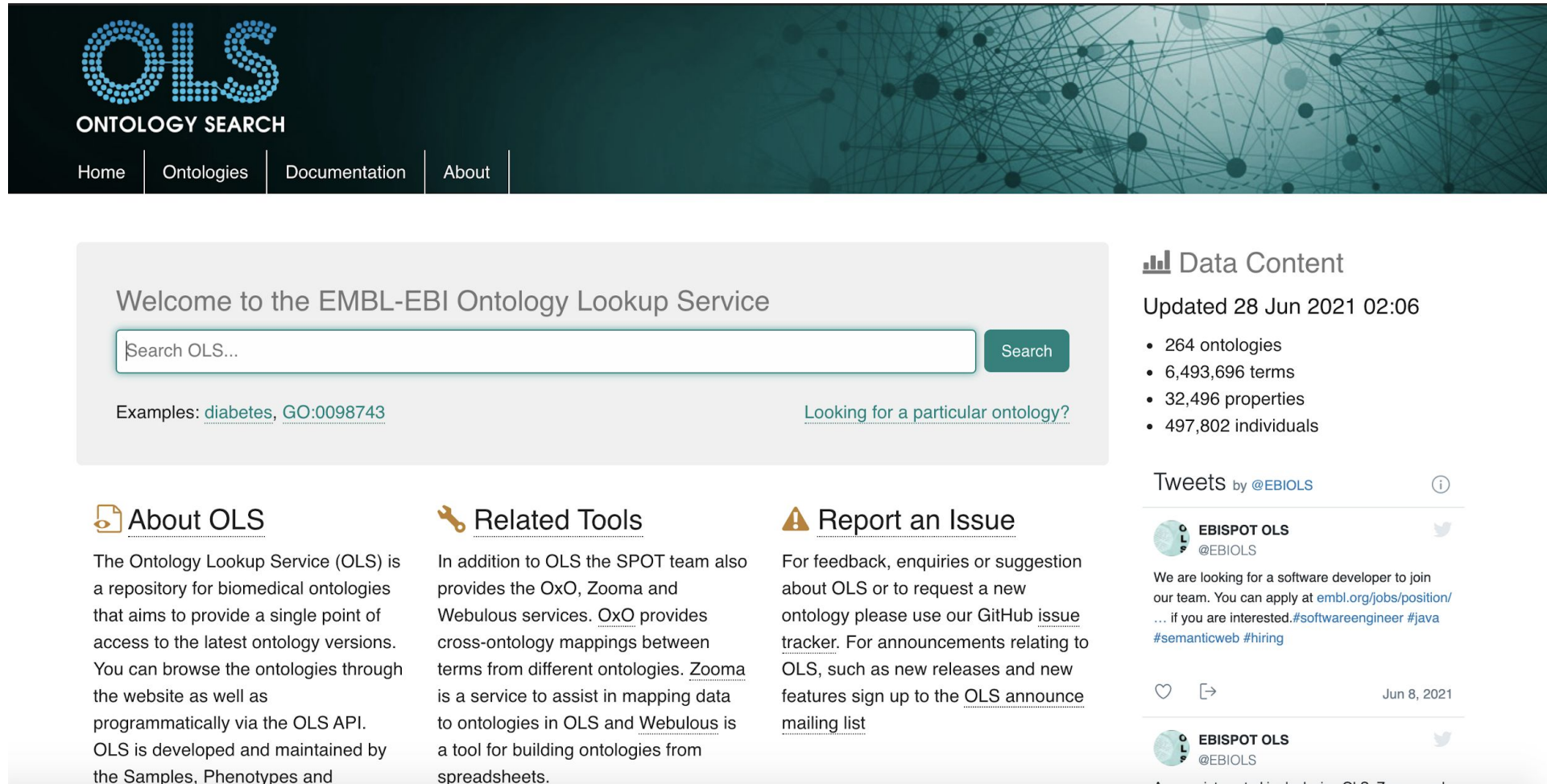


(What are these called?
Depends who you are talking to...)



Chick pea
Chickpea
Garbanzo bean
Cicer arietinum

Tools: Explore terms in ontologies using look-up services



The screenshot shows the EMBL-EBI Ontology Lookup Service (OLS) homepage. At the top, there is a dark blue header with the OLS logo (a stylized 'OLS' made of dots) and the text 'ONTOLOGY SEARCH'. Below the logo is a navigation bar with links: Home, Ontologies, Documentation, and About. The main content area is light gray and features a large search bar with the placeholder text 'Search OLS...' and a green 'Search' button. Below the search bar, there are examples: 'Examples: [diabetes](#), [GO:0098743](#)' and a link '[Looking for a particular ontology?](#)'. To the right of the search bar, there is a section titled 'Data Content' with a bar chart icon, showing 'Updated 28 Jun 2021 02:06' and a list of statistics: 264 ontologies, 6,493,696 terms, 32,496 properties, and 497,802 individuals. Below the search bar, there are three columns of content. The first column is titled 'About OLS' with a document icon and describes the OLS as a repository for biomedical ontologies. The second column is titled 'Related Tools' with a wrench icon and lists other services like OXO, Zooma, and Webulous. The third column is titled 'Report an Issue' with a warning icon and provides information on how to provide feedback. On the right side of the page, there is a section for tweets by @EBIOLS, showing a tweet from June 8, 2021, about a software developer position.

OLS ONTOLOGY SEARCH

Home | Ontologies | Documentation | About

Welcome to the EMBL-EBI Ontology Lookup Service

Search OLS... Search

Examples: [diabetes](#), [GO:0098743](#) [Looking for a particular ontology?](#)

About OLS

The Ontology Lookup Service (OLS) is a repository for biomedical ontologies that aims to provide a single point of access to the latest ontology versions. You can browse the ontologies through the website as well as programmatically via the OLS API. OLS is developed and maintained by the Samples, Phenotypes and

Related Tools

In addition to OLS the SPOT team also provides the OXO, Zooma and Webulous services. OXO provides cross-ontology mappings between terms from different ontologies. Zooma is a service to assist in mapping data to ontologies in OLS and Webulous is a tool for building ontologies from spreadsheets.

Report an Issue

For feedback, enquiries or suggestion about OLS or to request a new ontology please use our GitHub issue tracker. For announcements relating to OLS, such as new releases and new features sign up to the OLS announce mailing list

Data Content

Updated 28 Jun 2021 02:06

- 264 ontologies
- 6,493,696 terms
- 32,496 properties
- 497,802 individuals

Tweets by @EBIOLS

EBISPOT OLS @EBIOLS

We are looking for a software developer to join our team. You can apply at [embl.org/jobs/position/](#) ... if you are interested. [#softwareengineer](#) [#java](#) [#semanticweb](#) [#hiring](#)

Jun 8, 2021

EBISPOT OLS @EBIOLS

Are you interested in developing OLS, Zooma and

<https://www.ebi.ac.uk/ols/index>

Tools: Automating standardization and classification

Free text

Ontologized

3rd Party Scheme

Frz
hamburger
pattie

Hamburger Patty (frozen)

FOODON:03309571



Beef
(IFSAC+)

Data processing,
mapping to ontologies

Map to 3rd party
classification scheme

LexMapr Django is still in the development phase, and is currently catered towards food and environmental samples.

Input file*

Choose File No file chosen

Submit

processing time: minutes

GenomeTrakr is implementing FoodOn and LexMapr as part of its metadata curation system.



Pathogen: environmental/food/other sample from *Listeria monocytogenes*

Identifiers BioSample: SAMN17176170; SRA: SRS7939055; CFSAN: CFSAN109577

Organism [Listeria monocytogenes](#)
cellular organisms; Bacteria; Terrabacteria group; Firmicutes; Bacilli; Bacillales; Listeriaceae; Listeria

Package [Pathogen: environmental/food/other; version 1.0](#)

Attributes	strain	FDA1152605-C001-001
	collection date	2020-12-08
	collected by	FDA
	geographic location	Indonesia
	isolate name alias	CFSAN109577
	latitude and longitude	missing
	isolation source	frozen raw shrimp
	PublicAccession	CFSAN109577
	ProjectAccession	PRJNA215355
	Genus	Listeria
	FDA_Lab_Id	1152605-C001-001
	Species	monocytogenes
	attribute_package	environmental/food/other
	source type	Food
	LexMapr Version	LexMapr-0.7.1
	IFSAC+ Category	crustaceans
	ontological term	shrimp (frozen):FOODON_03301169 shrimp (raw):FOODON_03301837

Standardizing free text
descriptions of sample
sources

More easily queried

BioProject [PRJNA215355](#) *Listeria monocytogenes*
Retrieve [all samples](#) from this project

Submission [CFSAN](#); 2020-12-29



Standards: ISO

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:

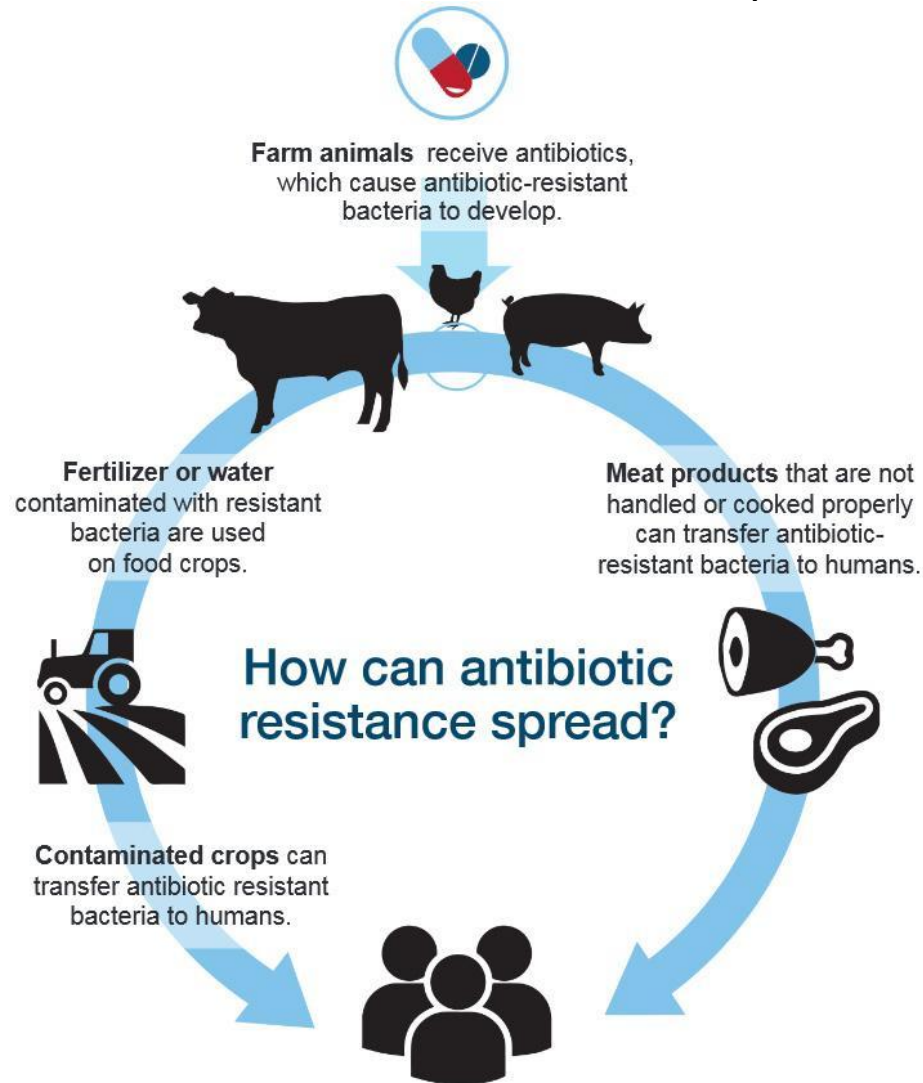
- Agency documentation
- Public repository submission forms
- Domain expert consultations
- Existing standards and ontologies

ISO slim (package of fields and terms) available:

<https://github.com/GenEpiO/iso2017>

Implementation: Canada's One Health Antimicrobial Resistance Genomics Research and Development Initiative (GRDI)

<https://info.grdi-amr.com/en/latest/>



- **Interagency** (PHAC, CFIA, AAFC, EC, HC etc)
- Use **genomics** tools and **harmonized contextual data** to understand foodborne AMR in Canadian food supply
- use **ontologies/standards** to prepare/integrate **contextual data** for **surveillance** and **risk assessment**
- **automate transformations, classification**

Structuring information using the **ISO standard**, **FoodOn** & **GenEpiO**

1. Food sample:
Frozen skinless, boneless chicken breasts from Greg's Groceries



GRDI Standard Field	Ontology Term
Food Product	chicken breast (skinless, boneless)
Food Processing	preserved by freezing
Environmental Location	grocery store
GRDI Standard Field	Ontology Term
geo_loc name (country)	Canada
host_origin geo_loc_name (country)	Country X
Host (scientific name)	Bos taurus
Host disease	listeriosis
Anatomical material	liver

2. Animal clinical sample:
Liver from cow imported from country X with suspected listeriosis, sampled in Canada



GRDI Standard Field	Ontology Term
Environmental Location	contact surface; Abattoir
Collection Device	swab

3. Environmental sample:
Swab of a contact surface in abattoir



- Label trees, machine learning & inferencing, feed into downstream analytical tools

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

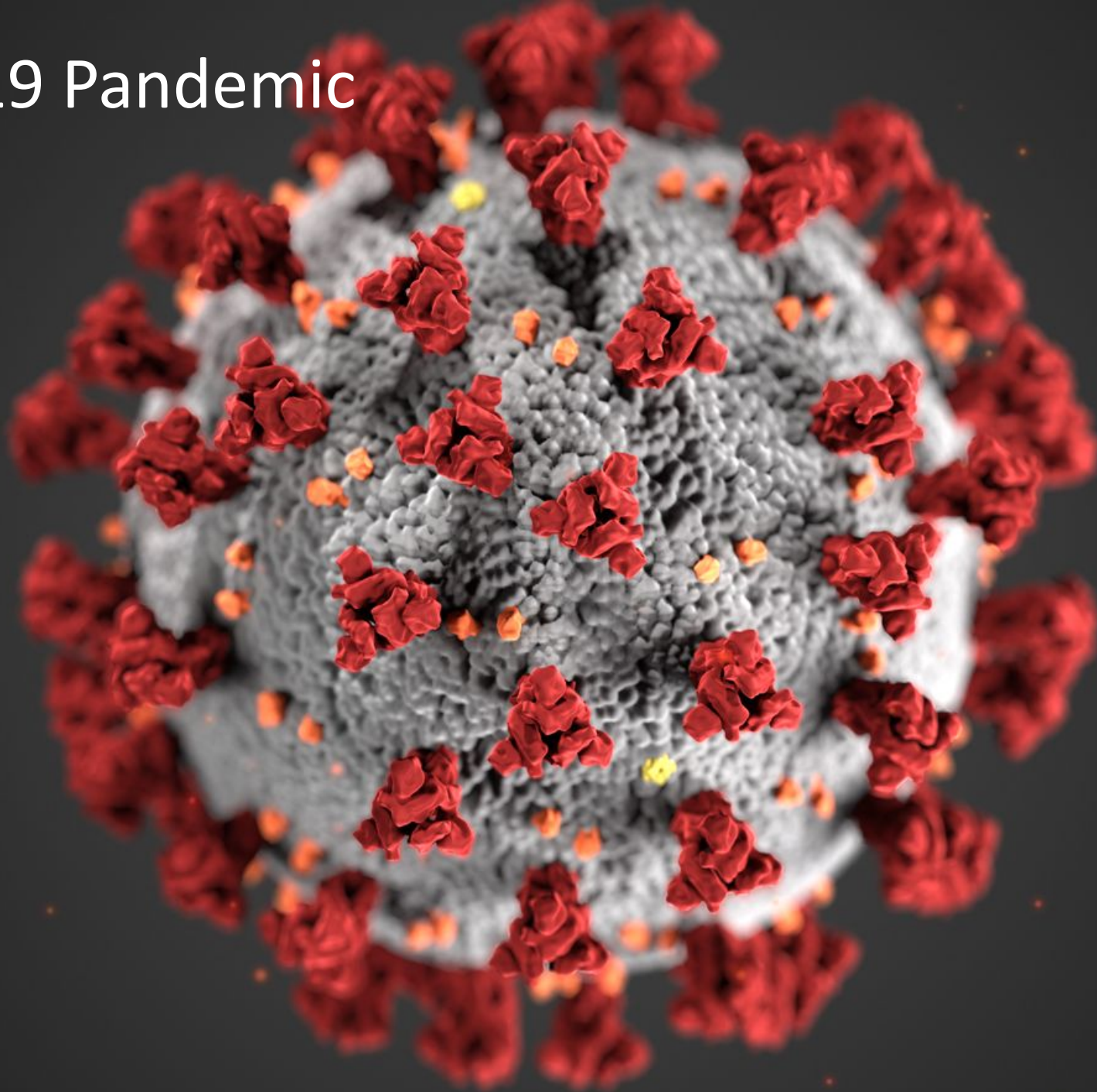
Privacy protection (sharing):

- 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 may be considered PHII
- Consult privacy officer (jurisdictional policies)

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

The COVID-19 Pandemic



Genomics has been a hero of the COVID-19 pandemic.

A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants

 Bethany Dearlove,  Eric Lewitus,  Hongjun Bai,  Yifan Li,  Daniel B. Reeves,  M. Gordon Joyce, Paul T. Scott,  Mihret F. Amare,  Sandhya Vasan,  Nelson L. Michael,  Kayvon Modjarrad, and  Morgane Rolland

PNAS September 22, 2020 117 (38) 23652-23662; first published August 31, 2020;
<https://doi.org/10.1073/pnas.2008281117>

The proximal origin of SARS-CoV-2

Kristian G. Andersen , Andrew Rambaut, W. Ian Lipkin, Edward C. Holmes & Robert F. Garry

Nature Medicine 26, 450–452(2020) | [Cite this article](#)

5.03m Accesses | 706 Citations | 35003 Altmetric | [Metrics](#)

To the Editor – Since the first reports of novel pneumonia (COVID-19) in Wuhan, Hubei province, China^{1,2}, there has been considerable discussion on the origin of the causative virus, SARS-CoV-2³ (also referred to as HCoV-19)⁴. Infections with SARS-CoV-2 are now widespread, and as of 11 March 2020, 121,564 cases have been confirmed in more than 110 countries, with 4,373 deaths⁵.

SARS-CoV-2 is the seventh coronavirus known to infect humans; SARS-CoV, MERS-CoV and SARS-CoV-2 can cause severe disease, whereas HKU1, NL63, OC43 and 229E are associated with mild symptoms⁶. Here we review what can be deduced about the origin of SARS-CoV-2 from comparative analysis of genomic data. We offer a perspective on the notable features of the SARS-CoV-2 genome and discuss scenarios by which they could have arisen. Our analyses clearly show that SARS-CoV-2 is not a laboratory construct or a purposefully manipulated virus.

Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California

 Xianding Deng^{1,2*},  Wei Gu^{1,2*},  Scot Federman^{1,2*},  Louis du Plessis^{3*},  Oliver G. Pybus³,  Nuno Faria³,  Candace Wang^{1,2},  Guixia Yu^{1,2},  Brian Bushnell⁴,  Chao-Yang Pan⁵,  Hugo Guevara⁵,  Alicia Sotomayor-Gonzalez^{1,2},  Kelsey Zorn⁶,  Allan Gopez¹,  Venice Servellita¹,  Elaine Hsu¹,  Steve Miller¹,  Trevor Bedford^{7,8},  Alexander L. Greninger^{7,9},  Pavitra Roychoudhury^{7,9},  Lea M. Starita^{8,10},  Michael Famulare¹¹,  Helen Y. Chu^{8,12},  Jay Shendure^{8,9,13},  Keith R. Jerome^{7,9},  Catie Anderson¹⁴,  Karthik Gangavarapu¹⁴,  Mark Zeller¹⁴,  Emily Spencer¹⁴,  Kristian G. Andersen¹⁴,  Duncan MacCannell¹⁵,  Clinton R. Paden¹⁵,  Yan Li¹⁵,  Jing Zhang¹⁵,  Suxiang Tong¹⁵,  Gregory Armstrong¹⁵,  Scott Morrow¹⁶,  Matthew Willis¹⁷,  Bela T. Matyas¹⁸,  Sundari Mase¹⁹,  Olivia Kasirye²⁰,  Maggie Park²¹,  Godfred Masinde²²,  Curtis Chan²²,  Alexander T. Yu²,  Shua J. Chai^{5,15},  Elsa Villarino²³,  Brandon Bonin²³,  Debra A. Wadford⁵,  Charles Y. Chiu^{1,2,24†}

 [Comment on this paper](#)

Large scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management

 Andrew J Page,  Alison E Mather,  Thanh Le Viet,  Emma J Meader,  Nabil-Fareed J Alikhan,  Gemma L Kay,  Leonardo de Oliveira Martins,  Alp Aydin,  David J Baker,  Alexander J. Trotter,  Steven Rudder,  Ana P Tedim,  Anastasia Kolyva,  Rachael Stanley,  Maria Diaz,  Will Potter,  Claire Stuart,  Lizzie Meadows,  Andrew Bell,  Ana Victoria Gutierrez,  Nicholas M Thomson,  Evelien M Adriaenssens,  Tracey Swingle,  Rachel AJ Gilroy,  Luke Griffith,  Dheeraj K Sethi,  Rose K Davidson,  Robert A Kingsley,  Luke Bedford,  Lindsay J Coupland,  Ian G Charles,  Ngozi Elumogo,  John Wain,  Reenesh Prakash,  Mark A Webber,  SJ Louise Smith,  Meera Chand,  Samir Dervisevic,  Justin O'Grady,

The COVID-19 Genomics UK (COG-UK) consortium

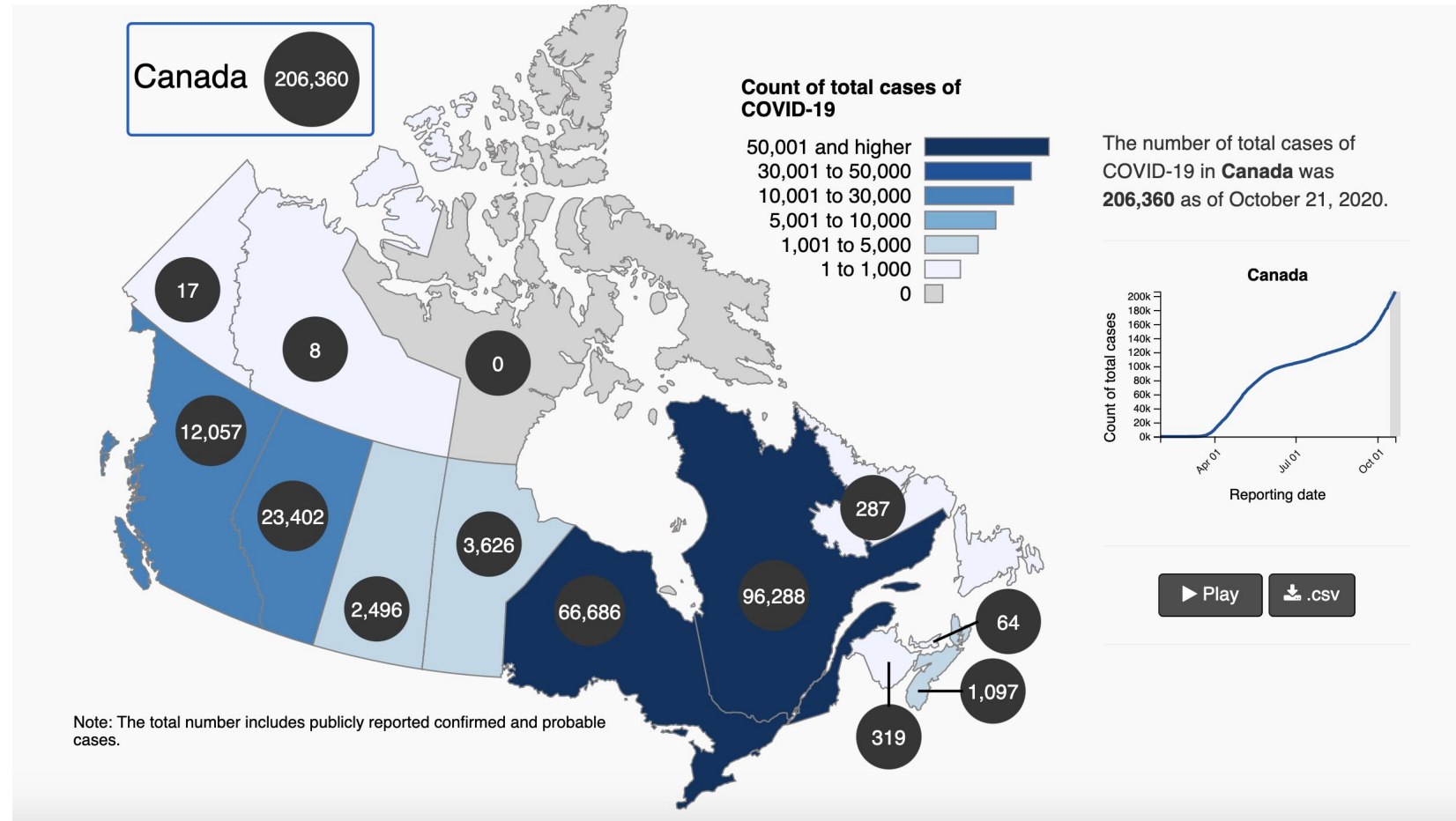
doi: <https://doi.org/10.1101/2020.09.28.20201475>

- Tracking transmission and outbreaks locally and globally, identifying variants, used for developing clinical tests and vaccines, understanding viral origins and evolution



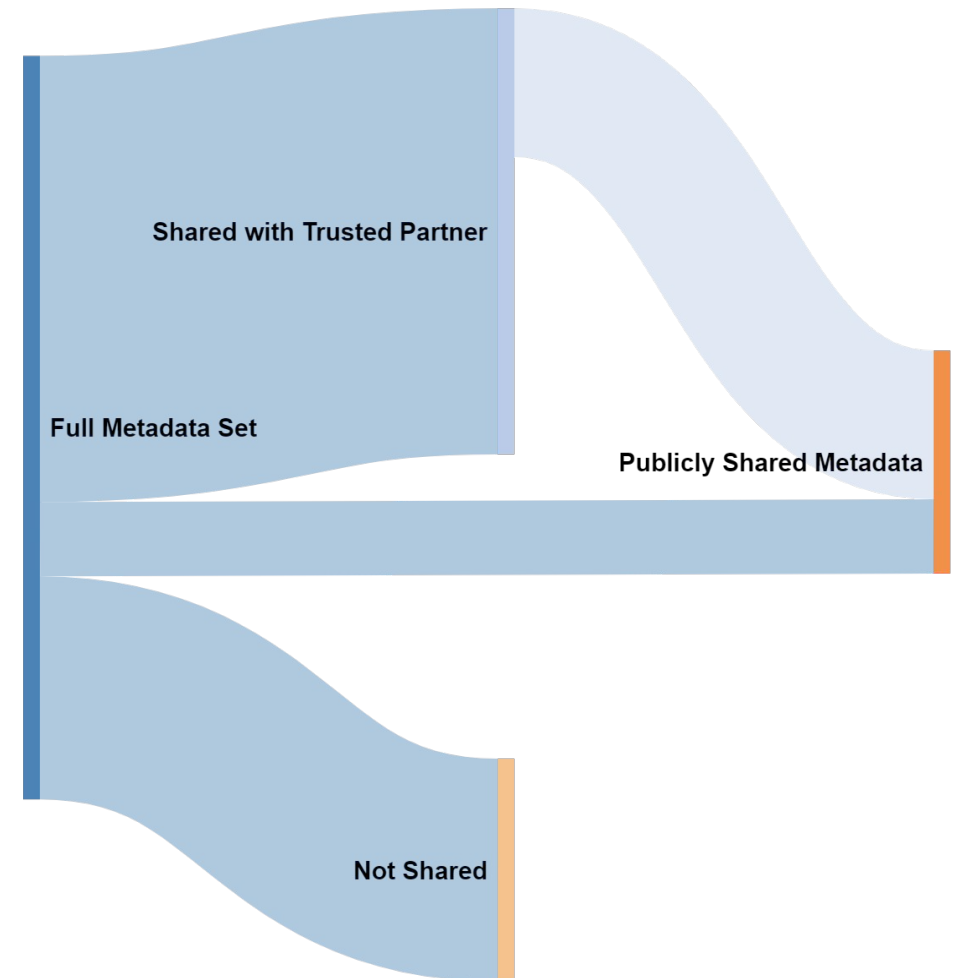
The Canadian COVID-19 Genomics Network

- cross-Canada (10 provs, 3 territories), inter-agency sequencing initiative
- 150K viral genomes, 10K matched human genomes
- public health **surveillance** & genomic determinants contributing to **outcomes**, and **risk**



In public health emergencies, you need to get the right information to to the right people quickly.

- Data comes from **different sources (epi vs clinical vs lab)**
- Data needs to be shared within **organizations, with trusted partners, with public repositories, with international agencies**
- **Structure metadata consistently across data management systems**



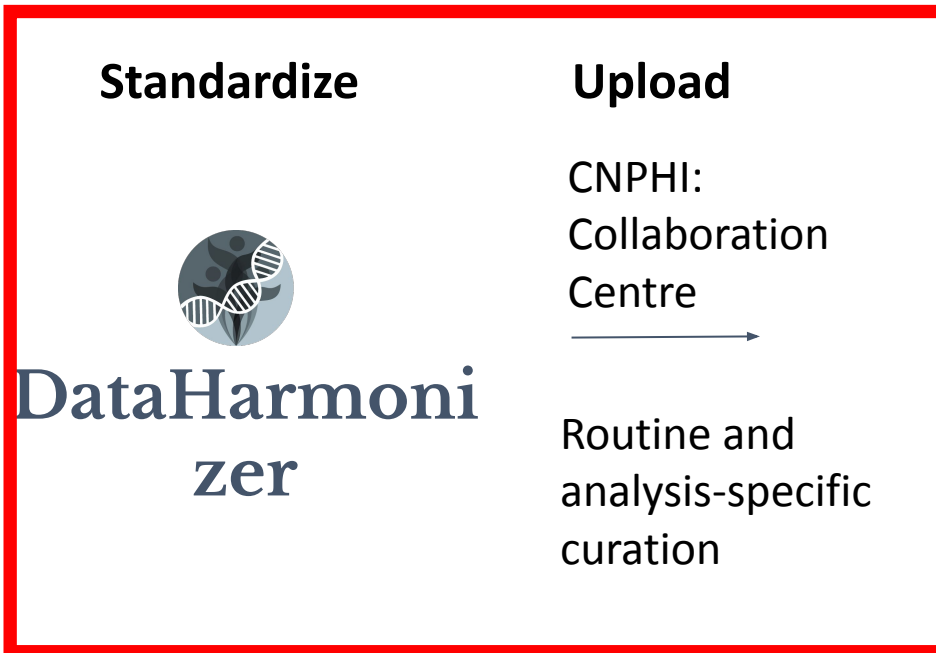
Data Flow for SARS-CoV-2 Viral Genome Contextual Data

Collect

Harmonize

Integrate

Disseminate




**National
database
(IRIDA)**

Public health
analyses

GISAID

NCBI

**VirusSeq Data
Portal**



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

Data Sources

- Case report forms
- Public repository requirements
- Existing metadata standards
- Literature

Mapping to Standards

- MIxS 5.0
- MIGS Virus, Host-Associated
- Project/Sample Application Standard
- WHO/ISARIC case collection form
- **OBO Foundry Ontologies**



Tools: The DataHarmonizer enables data entry, validation, transformation

- Tool for data entry and validation developed for CanCOGeN
- Spreadsheet-style text editor application
- Picklists, data structure, validation, export
- 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

Repurposed for foodborne pathogens and National Microbiome Data Collective

Save
Open existing file
Export to chosen format

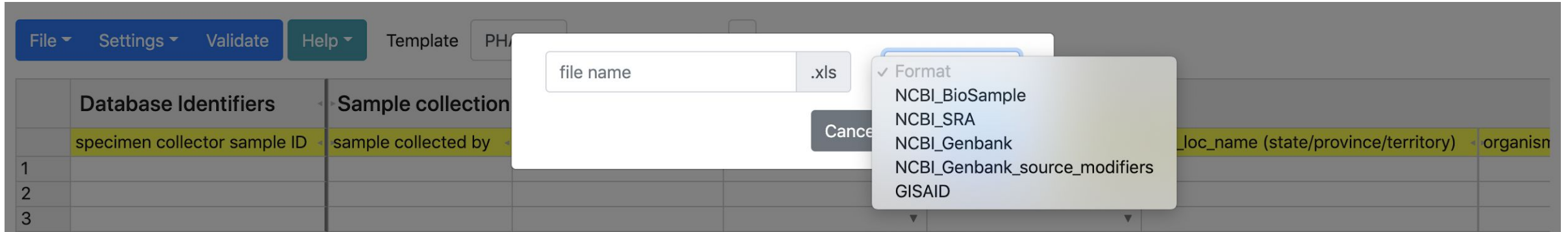
File ▾ Settings ▾ Validate Help ▾ Template CanCOGeN Covid-19 Loaded file

Show all columns
Jump to...
0.13.5

Sample collection and processing						
	Sample ID	sample collected by	sequence submitted by	sample collection date	geo_loc_name (country)	geo_loc_name (province/territory)
1						
2						
3						
4						
5						

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

Tools: The DataHarmonizer enables data entry, validation, transformation



- Enter data once, export in different submission formats i.e. GISAID and NCBI (BioSample, SRA, GenBank)





Public Health Alliance for Genomic Epidemiology

Data Structures | Bioinformatic Pipelines and Visualizations | Infrastructure
Public Repositories | Reference, QC and Validation | Workforce Development
Data Sharing and Ethics | Users and Applications

<https://www.pha4ge.org> | <https://www.github.com/pha4ge> |  @pha4ge

- Promoting implementation of data standards in public health settings, capacity building
- **Goal: Improved PH response and surveillance**

- **Standardized collection template** (colour-coded, yellow=required, purple=recommended, white=optional)
- **Pick lists:** standardized terms
- **Structured formats** e.g. for dates
- **JSON schema**

Guidance documentation

PHA4GE SARS-CoV-2 Contextual Data Template_demo

	A	B	C	D
1	Database Identifiers	Definition	Guidance	Examples
2	specimen collector sample ID	The user-defined name for the sample.	Every Sample ID from a single submitter must be unique.	prov_rona_99
3	bioproject umbrella accession	The INSDC umbrella accession number of the BioProject	Required if submission is linked to an umbrella	PRJNA623807
4	bioproject accession	The INSDC accession number of the BioProject(s)	Required if submission is linked to a BioProject.	PRJNA12345
5	biosample accession	The identifier assigned to a BioSample in INSDC arch	Store the accession returned from the BioSample	SAMN14180202
6	SRA accession	The Sequence Read Archive (SRA), European Nucleo	Store the accession assigned to the submitted "run".	SRR11177792
7	GenBank/ENA/DDBJ accession	The GenBank/ENA/DDBJ identifier assigned to the se	Store the accession returned from a GenBank/ENA/DDBJ	MN908947.3
8	GISAID accession	The GISAID accession number assigned to the sequ	Store the accession returned from the GISAID	EPI_ISL_123456
9	GISAID virus name	The user-defined GISAID virus name assigned to the	GISAID virus names should be in the format "hCoV-	hCoV-19/Canada/prov_rona_99/2020
10	host specimen voucher	Identifier for the physical specimen.	Include a URI (Uniform Resource Identifier) in the form of	URI example:
11				
12	Sample collection and processing	Definition	Guidance	Examples
13	sample collected by	The name of the agency that collected the original sam	The name of the agency should be written out in full, (with	Public Health Agency of Canada
14	sample collector contact email	The email address of the contact responsible for follow	The email address can represent a specific individual or	johnnyblogs@lab.ca
15	sample collector contact address	The mailing address of the agency submitting the sam	The mailing address should be in the format: Street	655 Lab St, Vancouver, British Columbia,
16	sequence submitted by	The name of the agency that generated the sequence.	The name of the agency should be written out in full, (with	Centers for Disease Control and Prevention
17	sequence submitter contact email	The email address of the contact responsible for follow	The email address can represent a specific individual or	Resplab@lab.ca
18	sequence submitter contact address	The mailing address of the agency submitting the seq	The mailing address should be in the format: Street	123 Sunnybrooke St, Toronto, Ontario, M4P
19	sample collection date	The date on which the sample was collected.	Record the collection date accurately in the template.	2020-03-19
20	sample received date	The date on which the sample was received.	The date the sample was received by a lab that was not	2020-03-20
21	geo_loc name (country)	The country of origin of the sample.	Provide the country name from the pick list in the	South Africa
22	geo_loc name (state/province/territory)	The state/province/territory of origin of the sample.	Provide the state/province/territory name from the GAZ	Western Cape
23	geo_loc name (county/region)	The county/region of origin of the sample.	Provide the county/region name from the GAZ geography	Derbyshire
24	geo_loc name (city)	The city of origin of the sample.	Provide the city name from the GAZ geography ontology.	Vancouver
25	geo_loc latitude	The latitude coordinates of the geographical location o	Provide latitude coordinates if available. Do not use the	38.98 N
26	geo_loc longitude	The longitude coordinates of the geographical location	Provide longitude coordinates if available. Do not use the	77.11 W
27	organism	Taxonomic name of the organism.	Select "Severe acute respiratory syndrome coronavirus	Severe acute respiratory syndrome
28	isolate	Identifier of the specific isolate.	This identifier should be an unique, indexed, alpha-	SARS-CoV-2/human/USA/CA-CDPH-
29	culture collection	The name of the source collection and unique culture	Format: "<institution-code>:[<collection-	/culture_collection="ATCC:26370"
30	purpose of sampling	The reason that the sample was collected.	Select a value from the pick list in the template.	Diagnostic testing
31	purpose of sampling details	Further details pertaining to the reason the sample wa	Provide a free text description of the sampling strategy or	Screening of bat specimens in museum

PHA4GE – SARS-CoV-2 Contextual Data Template User Guide and SOP 2.0

introduced to capture different kinds of anatomical and environmental samples, as well as collection devices and methods. These fields include "anatomical material", "anatomical part", "body product", "environmental material", "environmental site", "collection device", and "collection method". **Populate only the fields that pertain to your sample.** Provide the most granular information allowable according to your organization's data sharing policies.

e.g. nasal swab should be recorded:

host (scientific name)	host (common name)	host disease	anatomical part	collection device
Homo sapiens	Human	COVID-19	Nasopharynx	Swab

e.g. saliva should be recorded:

host (scientific name)	host (common name)	host disease	anatomical material
Homo sapiens	Human	COVID-19	Saliva

e.g. human feces should be recorded:

host (scientific name)	host (common name)	host disease	body product
Homo sapiens	Human	COVID-19	Feces

e.g. sewage from treatment plant should be recorded:

environmental site	environmental material
Sewage Plant	Sewage

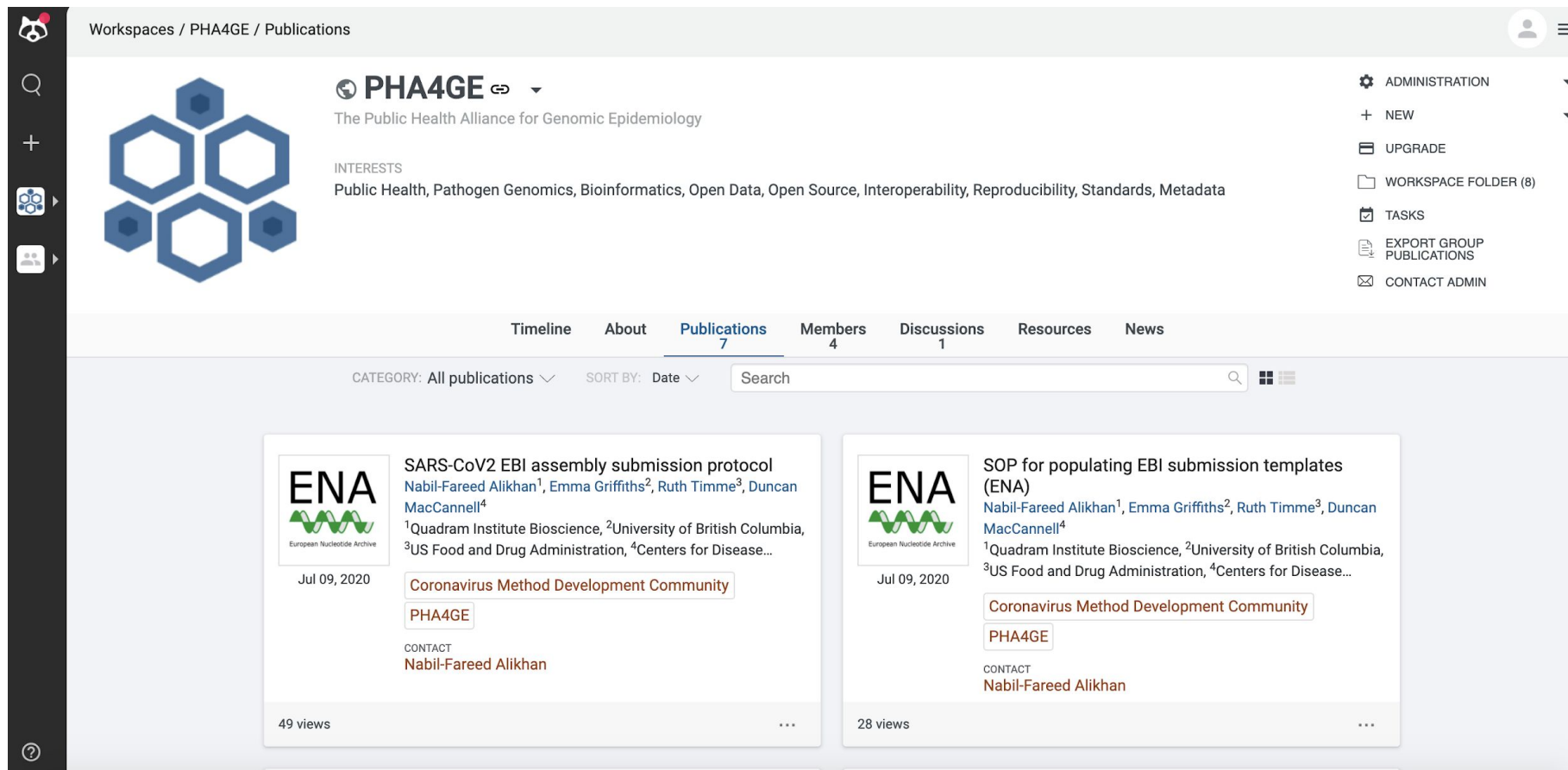
e.g. swab of a hospital bed rail should be recorded:

environmental site	environmental material	collection device
Hospital	Bed Rail	Swab

- **Reference guide:** field labels, definitions, guidance, expected values

- **SOP:** how to curate contextual data

Protocols to mobilize harmonized data



The screenshot shows the PHA4GE workspace on Protocols.io. The header includes the PHA4GE logo and name, a description 'The Public Health Alliance for Genomic Epidemiology', and a list of interests: Public Health, Pathogen Genomics, Bioinformatics, Open Data, Open Source, Interoperability, Reproducibility, Standards, Metadata. A sidebar on the right contains navigation links: ADMINISTRATION, NEW, UPGRADE, WORKSPACE FOLDER (8), TASKS, EXPORT GROUP PUBLICATIONS, and CONTACT ADMIN. The main content area displays a list of publications under the 'Publications' tab. Two publications are visible, both dated Jul 09, 2020, and both associated with the 'Coronavirus Method Development Community' and 'PHA4GE' tags. The first publication is 'SARS-CoV2 EBI assembly submission protocol' by Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, and Duncan MacCannell, with 49 views. The second publication is 'SOP for populating EBI submission templates (ENA)' by the same authors, with 28 views. Both publications list affiliations: ¹Quadram Institute Bioscience, ²University of British Columbia, ³US Food and Drug Administration, ⁴Centers for Disease...

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

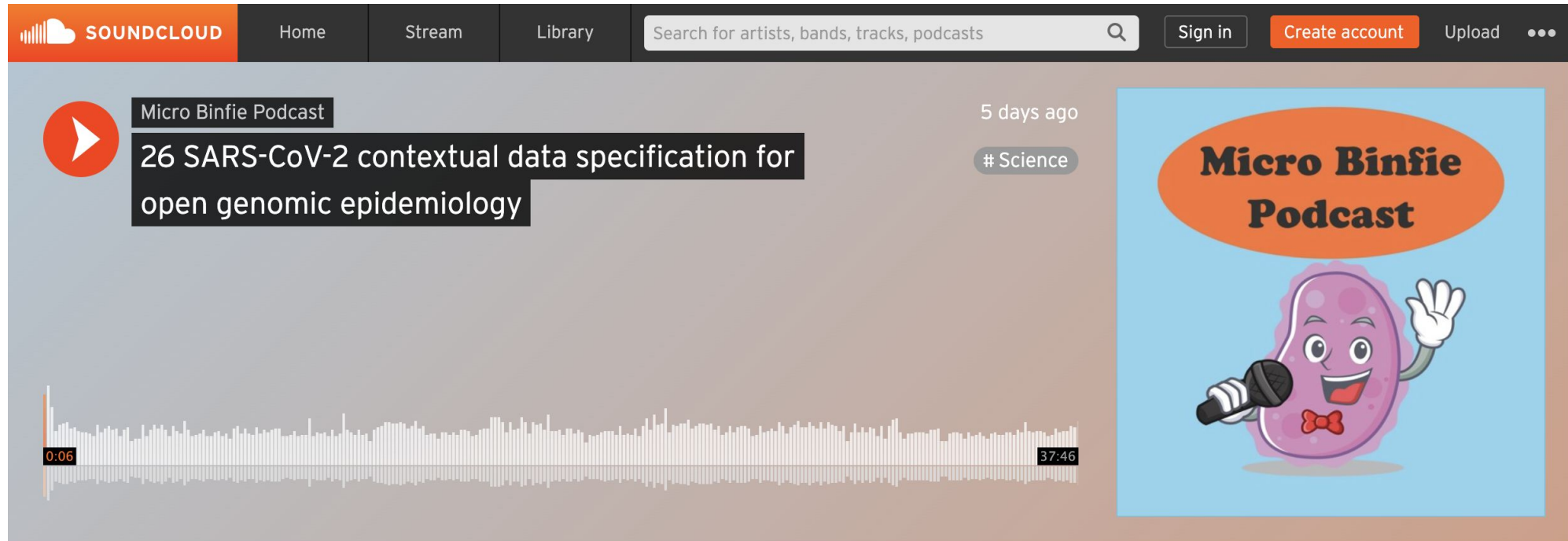
- **Different repositories have different fields, but PHA4GE helps standardize what goes into those fields.**

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

Data standard implementation and expert engagement/feedback is crucial.

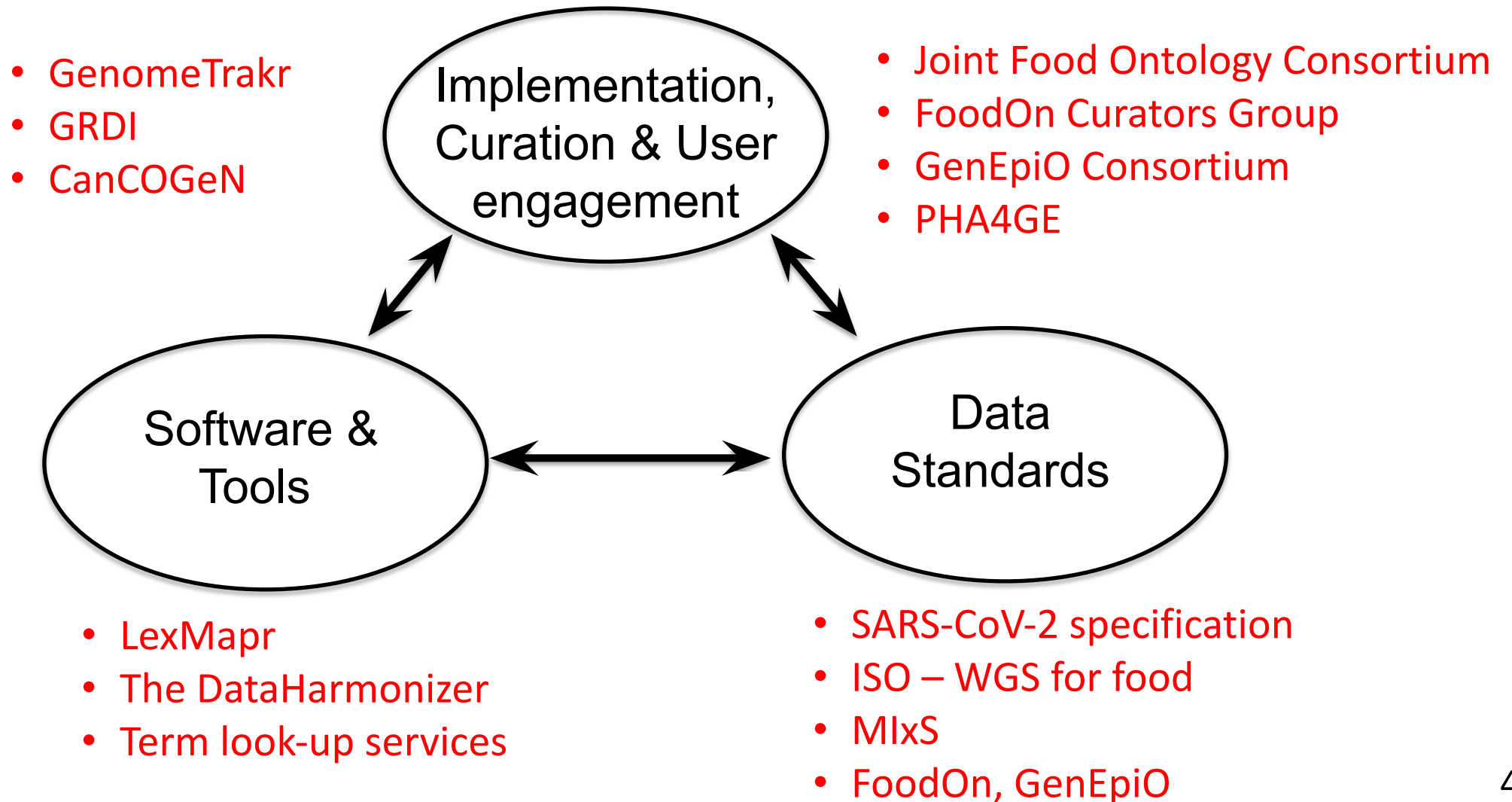


Communicating about standards is part of putting them into practice.



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

Summary: strengthening the contextual data ecosystem increases the value and usability of data



Thank you!

Centre for Infectious Disease Genomics and One Health - CIDGOH



<https://github.com/cidgoh/>

[@griffiemma](mailto:ega12@sfu.ca)