New international standards for Whole Genome Sequencing (WGS) contextual data sharing

Emma Griffiths, PhD

Centre for Infectious Disease Genomics and One Health

Faculty of Health Sciences, Simon Fraser University

Vancouver, Canada

EFSA Sept 5 2023

Outline

1. Challenges of integrating/harmonizing contextual data across labs and databases.

2. Interoperable contextual data standards (for private vs public data) as solutions & benefits.

3. Example of implementation

4. Summary.

Foodborne pathogen genomics contextual data is critical for interpreting the sequence data.

Sequence data

Contextual data

Sample metadata

Lab testing results

Methods & Provenance

Contextual data (metadata) used for documenting **sampling and sample processing**, **isolation methods** and **isolate characteristics** (e.g. AMR and virulence factors), **sequencing methods** and **quality metrics** used for:

- **Monitoring and quality control**
- **Comparing results** between laboratories
- **Characterizing** sequence types/clusters
- **Generating hypotheses** about sources of contamination etc
- **Informing decision making** and **monitoring effects of interventions**

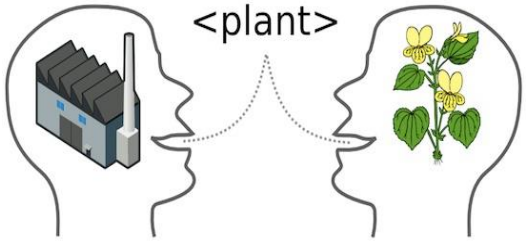
Lack of standardization complicates using the data.

Free text = ☹️

Errors, Jargon & Short hand

Arrugila
Slamonella
Layer crumb
Pizza pocket
Frz Chick Brst

Semantic ambiguity



"Sample source"
Lab A=Lab name

"Sample source"
Lab B=Sample type

Same words, Different meaning

Different Data Structures

Formats

Date:
2021-04-26
April 26, 2021

Concepts/entities/granularity

Poultry
Chicken
Skinless, boneless chicken breast
RTE Cajun style nuggets
Chicken offal excluding liver

Data structure impacts function.

It’s difficult to fit it all together.

Data clean up takes time, resources.

- Complicates reuse of your own data (across time, across organization)
- Variability in private databases propagates out to public repositories, complicating data integration/analyses.

Contextual data standards improve data harmonization and integration.

BEFORE AFTER

Data standards provide a quality framework for your contextual data

- Improves **auditability** (e.g. chain of custody)
- **Provenance** and **acknowledgement**
- Streamlines **re-use** and **data sharing**
- **Reduces uncertainty**
- Creates **expectations** for structure, **requirements**, and **completeness**
- Can **reuse** curation **training/skills, tools, also agreements**
- **Future-proofs data**

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
AMR & Virulence phenotypes

ISO standard provides tables and annexes to describe…

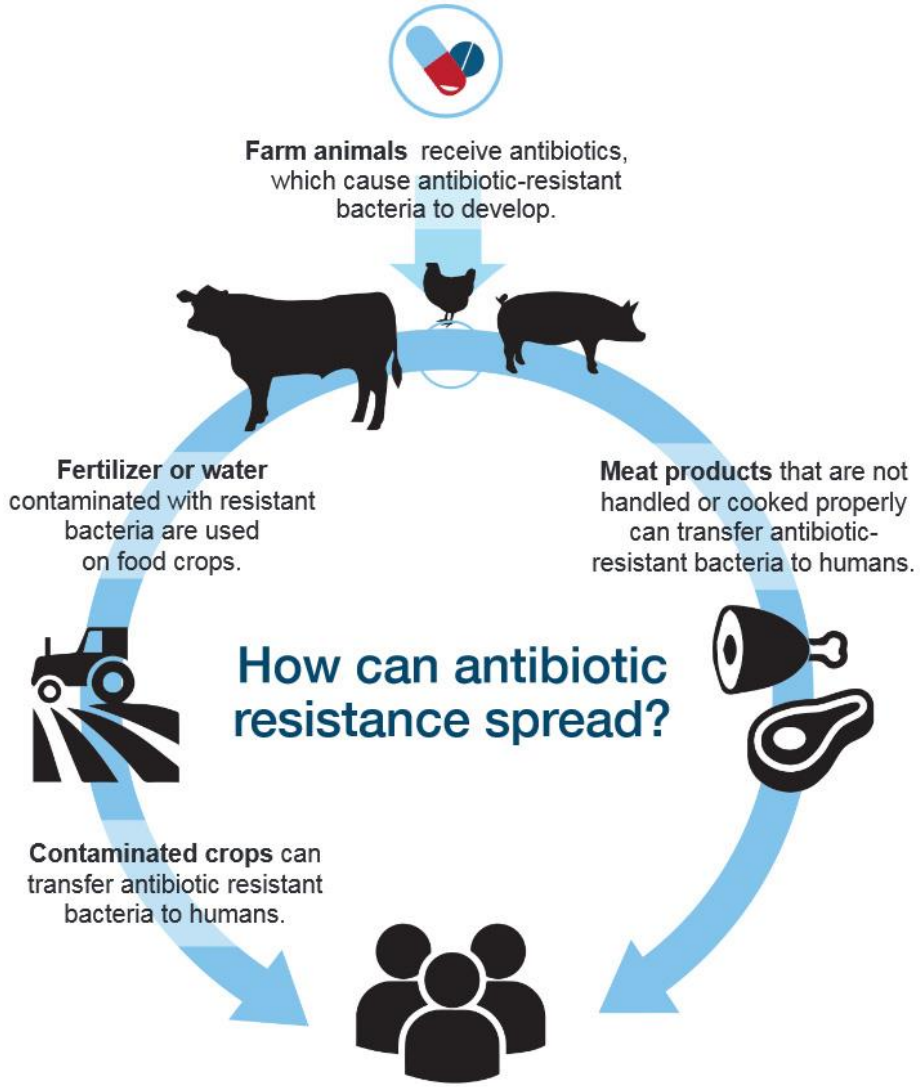
1. Information about the sample
2. Information about the isolate
3. Information about the sequence

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

One Health Antimicrobial Resistance (AMR) Standard for collection & private/collaborative sharing



- Based on ISO framework
- Scope: WGS across sectors, commodities, environments, hosts
- Goal: use **genomics** and **harmonized contextual data** to understand foodborne AMR in food supply and environment
- **Canadian implementation: Interagency** (PHAC, CFIA, AAFC, ECCC, DFO, HC etc), **Next Uganda**

https://github.com/cidgoh/GRDI_AMR_One_Health

What's in it and how do you use it?

Domain Content

- Repository accession numbers and **identifiers**
- **Sample collection and processing**
  - Food **products**
  - Food **processing**
  - Host/food **geo-loc origin** vs **sampling location**
  - **Environments** (abattoir, farm, natural enviros, fisheries)
  - Environmental **materials** (chicken litter, sediment, water, soil)
  - **Anatomical** parts/sites (feces, organ contents)
  - **Presampling activities** (fertilizer, vaccination, decontamination)
  - **Sampling/sequencing strategies** (bias/limitations)
- **Host information (animals, plants, humans)**
- **Sequencing methods**
- **Bioinformatics and quality control metrics**
- **AMR phenotype testing**
- **Risk assessment**
- **Provenance and attribution**

Standardized null values

Standardized fields & Picklists (can be updated)

Only small subset are required! (colour-coded)

Support docs (ref guide/SOP)

Operationalized in spreadsheet and data curation app (DataHarmonizer)

INSDC Attribute (Metadata) Packages for Food & One Health for public sharing

NCBI

- Food – animal and animal feed
- Food – farm environment
- Food – food production facility
- Food – human foods

- Pathogen: environmental/food/other
- One Health Enteric

https://www.ncbi.nlm.nih.gov/biosample/docs/packages

ENA

- ERC000045: COMPARE-ECDC-EFSA pilot food-associated reporting standard

- ERC000044: COMPARE-ECDC-EFSA pilot human-associated reporting standard

https://www.ebi.ac.uk/ena/browser/checklists

How are these different?

- **Overlapping content/concepts** but **different scopes/needs**
- Attribute packages much **smaller** than ISO & One Health AMR (public vs private data management)

Ontologies: Built for harmonization and data linkage

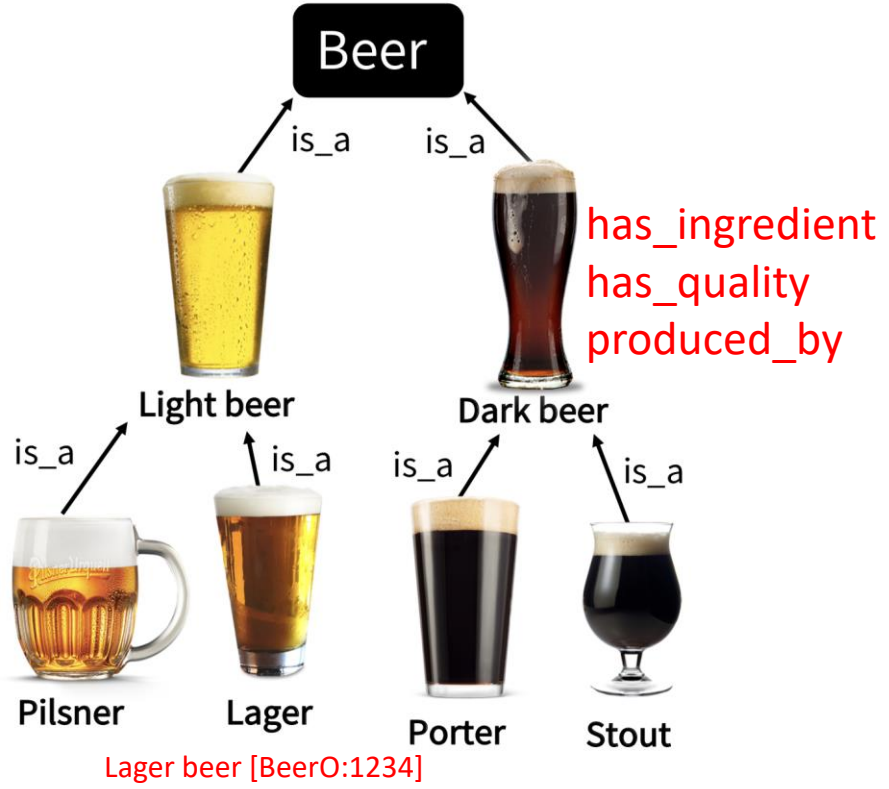
Controlled (standardized) vocabulary
Hierarchy + logic (linked data, enable classification for analyses)

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)

5-star Open Data Plan

★ Make your stuff available on the Web (whatever format) under an open license

★★ Make it available as structured* data (e.g. Excel instead of an image scan of a table)

★★★ Make it available in (2+) non-proprietary open format (e.g., CSV instead of Excel)

★★★★ Use **URI**s to denote things, so that people can point to your stuff

★★★★★ Link your data to other data to provide context

FoodOn:455678

VS

ENVO:009747

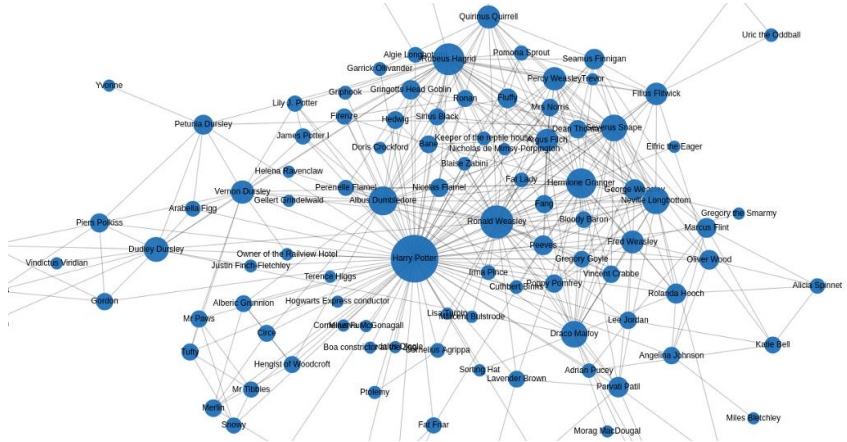
Data standards + Ontologies

Prescribed lists of fields, values, formats

International sources of standardized fields and terms

Vocabulary Lists
Good Guys
- Harry Potter
- Ron Weasley
- Hermione Granger

Bad Guys
- Voldemort
- Bellatrix Lestrange

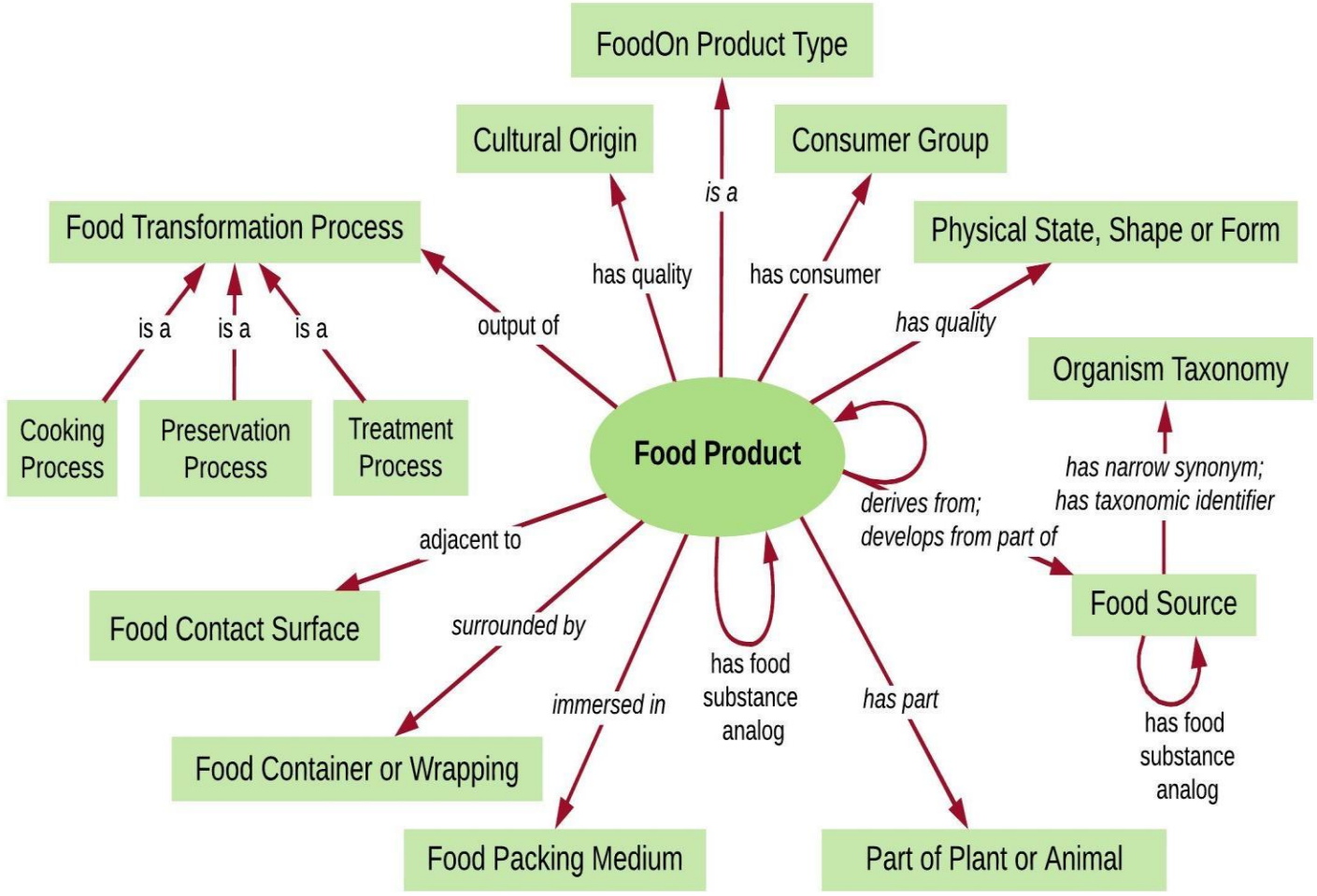


If you turned Harry Potter into a knowledge graph (linked data)….

The Food Ontology (FoodOn)

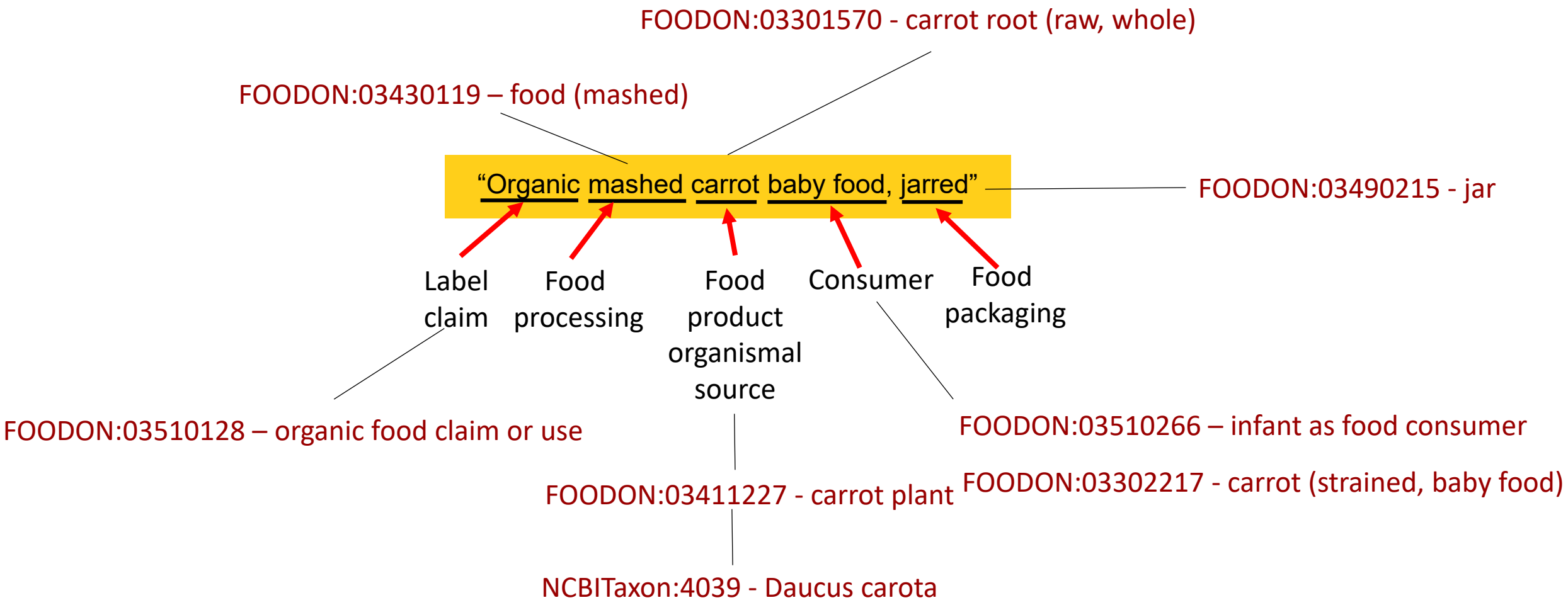
Dooley, Griffiths et al (2018), *Nature: Science of Food*

- **>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 → FDA Product codes)
- FoodOn consortia



https://github.com/FoodOntology/foodon www.foodon.org

Use of facets breaks down complex descriptions into mappable units



Suite of Ontology-based Tools

LexMapr: Automates mapping of free text to standardized terms

DataHarmonizer: Standards-based templates for data curation, validation, automated transformation

GEEM: Build ontology-based specifications

hAMRonization: harmonizes outputs from AMR detection tools (report)

Public Health Alliance for Genomic Epidemiology

https://github.com/pha4ge/hAMRonization

Tools: LexMapr Automating data transformations

Free text → **Ontologized** → **3rd Party Scheme**

Frz hamburger pattie

Hamburger Patty (frozen)

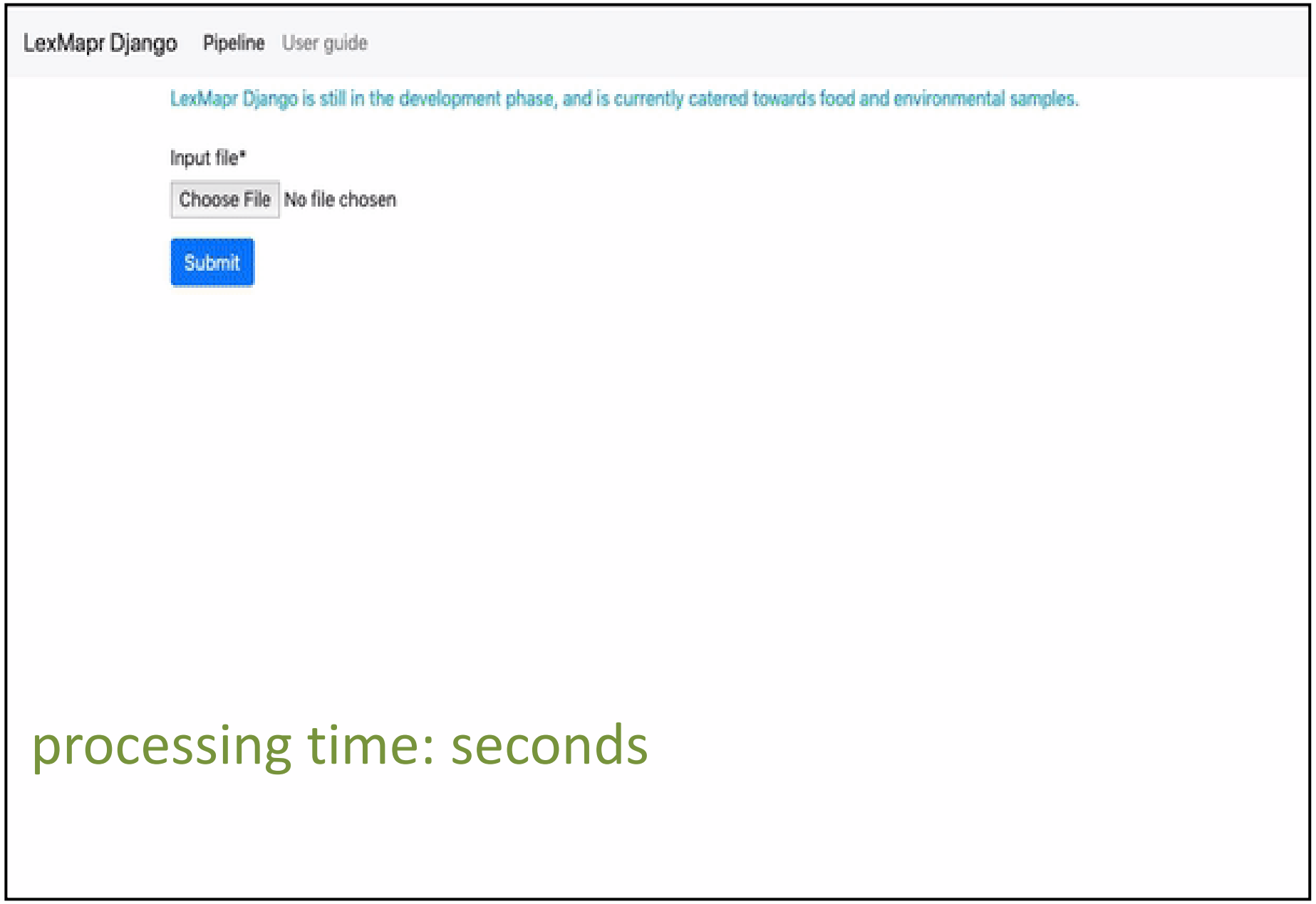
FOODON:03309571

Beef (IFSAC)

Data processing, mapping to ontologies

Map to 3rd party classification scheme

https://github.com/Public-Health-Bioinformatics/LexMapr



processing time: seconds

https://anaconda.org/bioconda/lexmapr

FDA's 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

BioProject	PRJNA215355 Listeria monocytogenes Retrieve all samples from this project
Submission	CFSAN; 2020-12-29

- Standardizing free text descriptions of sample sources (using ontologies – international terminology)
- More easily queried

Study illustrating improved analyses with standardized data: **Interpretative Labor and the Bane of Nonstandardized Metadata in Public Health Surveillance and Food Safety**

Clinical Infectious Diseases, Volume 73, Issue 8, 15 October 2021, Pages 1537–1539, https://doi.org/10.1093/cid/ciab615

"When standardized within an ontological framework, the metadata are interoperable across independent contributors from all over the world. This forms the basis of a successful One Health, open genomic epidemiology network "

EFSA's FoodEx2 Classification System

https://www.efsa.europa.eu/en/data/data-standardisation

- >20 000 terms, disambiguated by codes (different than ontology URIs)
- Basic food products (e.g. Rice flour, Gnocchi) organized by Food type hierarchies
- Facets (e.g. source, packaging material, production method)
- In some applications composite terms with own codes (products + attributes) are implemented

- Scoped for regulatory/investigation needs
- Tool for navigating/searching
- Related dictionary for non-food (hosts, anatomical parts, environmental materials)

- ***Mapping to FoodOn (universal translator between different terminologies)***

Institution-specific Dictionary 1 ⟷ FoodOn ⟷ Institution-specific Dictionary 2

Babel fish

Standards Summary

1. ISO 23418:2022 – WGS for food bacteria (international)

2. One Health AMR – collection & sharing (networks)

3. BioSample Packages (public sharing)
    - One Health/Pathogen
    - Food (NCBI, ENA)

4. FoodEx2 – (organization)

5. Ontologies (universal, linking data)
    - FoodOn
    - Many others (environments, anatomy, disease, geography, taxonomy etc)

How do these things all fit together???

There will never be one standard to rule them all.

Transformation

Data management tools

Transformation

Data management tools

Organizations use their own **dictionaries/systems**
Local

Assemble collections according to **genomics standards** e.g. ISO (recombine info)
Network

Submit to **public repos** using **BioSample** packages (subset)
International

Interoperability (ontologies)

What is needed for WGS data standards ecosystem?

- Data governance (**utility, permissions**)
- Harmonization approaches (**consensus**)
- Tools & platforms (**operationalize**)
- Community engagement (**use, design**)
- Sustainable funding (**maintenance**)

Thank you!

Centre for Infectious Disease Genomics and One Health - CIDGOH

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

ega12@sfu.ca
@griffiemma