Standardizing the Isolation Source Metadata for the Genomic Epidemiology of Foodborne Pathogens Using LexMapr

Balkey M.[1], Batz, M.[1], Gopinath, G.[2], Gosal, G.[3], Griffiths, E.[4], Tate, H.[2], Timme, R.[1]

[1] Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland, USA, [2] Center for Veterinary Medicine, U.S. Food and Drug Administration, Laurel, Maryland, USA, [3] Faculty of Health Sciences, SFU, Simon Fraser University, Canada,
[4] Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada

Abstract

Introduction: FDA's GenomeTrakr is a public/private genomic epidemiology network for foodborne pathogen surveillance, specifically targeting pathogens isolated from food or environmental sources. The raw genome plus a small set of associated metadata are made publicly available at the National Center for Biotechnology Information (NCBI). Metadata include organism name, geographical location, collection date, isolate contributor and isolation source. The isolation source field is currently a free text field, requiring no standard terminologies or structure. As the GenomeTrakr database grew to over 100K isolates and the diversity of isolation sources became more complex, this field became difficult to analyze and interpret using computational approaches.
Purpose: In order to maximize the use of GenomeTrakr data and make this resource FAIR (findable, accessible, interoperable and reusable), we have standardized the metadata for the isolation source of WGS data for publicly available GenomeTrakr records.
Methods: We evaluated and utilized LexMapr, a rule-based text-mining tool, to automate the curation of isolation source metadata and assign categories from the expanded source categorization schema Interagency Food Safety Analytics Collaboration (IFSAC+) based on IFSAC categories. LexMapr processes the text from the isolation source and extracts entities incorporate new standard descriptors for the isolation sour that are mapped to standard ontology terms from relevant ontologies such as: FoodON, ENVO, UBERON, among others.
Results: GenomeTrakr has a total of 9,452 unique isolation sources. LexMapr successfully processed 88% of these records, as determined by manual curation and verification. After the evaluation of LexMapr, 71,530 publicly available records were curated, assigned ontology terms, and categorized using the IFSAC+ categorization schema.
Significance: The use of standard terminologies in the context of metadata for WGS is essential to facilitate data exchange and generate machine-readable resources that can expand our understanding of the dynamics of pathogen transmission across the food chain.

Introduction

According to CDC, it is estimated that 48 million people get sick, 128,000 are hospitalized and 3,000 die from foodborne disease each year in the United States. Foodborne disease outbreak surveillance is essential for the identification of food commodities implicated in foodborne illness. Whole genome Sequencing (WGS) is the reference method for public health surveillance and outbreak investigation; it is used to verify implicated food vehicles associated to foodborne illness, identify resident and transient strains in food production environments, predict microbial phenotypes such as: virulence, pathogenicity and resistance to biocides, metals or antimicrobial drugs.

GenomeTrakr, a network of public health laboratories has successfully demonstrated the effective use of WGS in Food Safety across a diverse network of laboratories. It achieves this through standardized sequencing protocols in combination with rapid WGS data availability in public data repositories. The network depends on an open data model in which member laboratories submit WGS records obtained from a pure isolate to NCBI accompanied by detailed contextual metadata according to established criteria. Epidemiological analysis in foodborne pathogen surveillance is only as accurate as the available geographic information and source definition metadata. To facilitate better surveillance, GenomeTrakr requires submitters to include taxonomy name, country of origin, date of collection, isolate contributor and isolation source in each submission. As the network grows and the diversity of isolation sources becomes more complex, the implementation and enforcement of standardized terminology to describe the isolation source is imperative.

Free text input within the isolation source metadata requires extensive manual data preparation and curation to be suitable for automated attribution studies, which becomes prohibitively difficult at any meaningful scale of analysis. Thus, progress in WGS analysis techniques requires metadata described by standard terminologies that are both human and machine readable, allowing for automated processing and high-level machine learning applications.

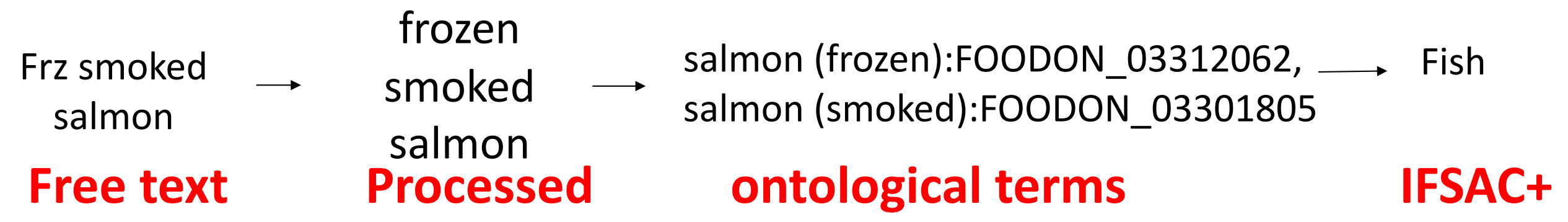
Controlled vocabularies/ontologies are fundamental for adding an interpretative level to WGS data, making information findable, accessible, interoperable, and reusable (FAIR data principles). To enhance GenomeTrakr metadata already published at NCBI, we aim to describe the use of the LexMapr application to curate GenomeTrakr records. By using this rule-based text-mining tool, we incorporate new standard descriptors at the GenomeTrakr database such as controlled terminologies from public ontologies and food categories for each source descriptor in the GenomeTrakr database.

Materials and Methods

LexMapr: Ontology driven tool, processes free text from isolation source descriptors and generates standard terminologies from controlled vocabulary/ontologies such as: FoodOn, GenEpiO, UBERON, ENVO, NCBI Taxon and specific food and environmental categories from IFSAC+.
https://github.com/Public-Health-Bioinformatics/LexMapr

- **Ontology:** Form of knowledge representation to describe terms and their relationships.
  - **FoodOn:** Food domain, its components, relationships and attributes.
  - **GenEpiO:** WGS-based food-borne pathogen investigations including genomics, lab, clinical and epidemiological data.
  - **UBERON**: Anatomical structures in animals.
  - **ENVO**: Environments encountered in ecological applications.
  - **NCBITaxon**: Organism classification.

- **IFSAC+:** Expanded categorization schema for food and environmental products. A reviewed version from the food categorization scheme developed for food implicated in outbreaks.



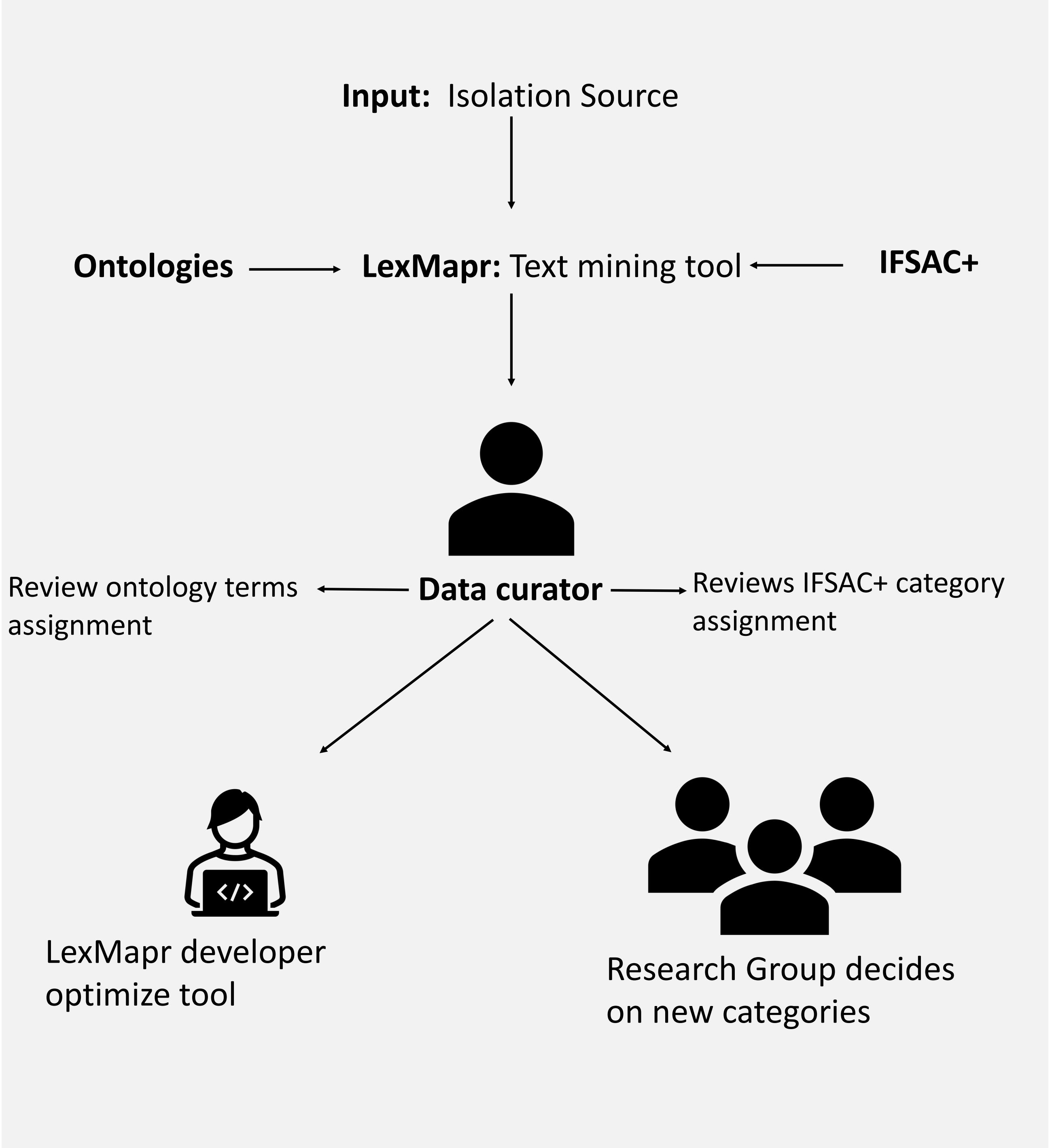

Input: Isolation Source
Ontologies
LexMapr: Text mining tool
IFSAC+
Review ontology terms assignment
Data curator
Reviews IFSAC+ category assignment
LexMapr developer optimize tool
Research Group decides on new categories

Figure 1. Evaluation of the use of LexMapr for standardization of isolation source metadata for food and environmental isolates.

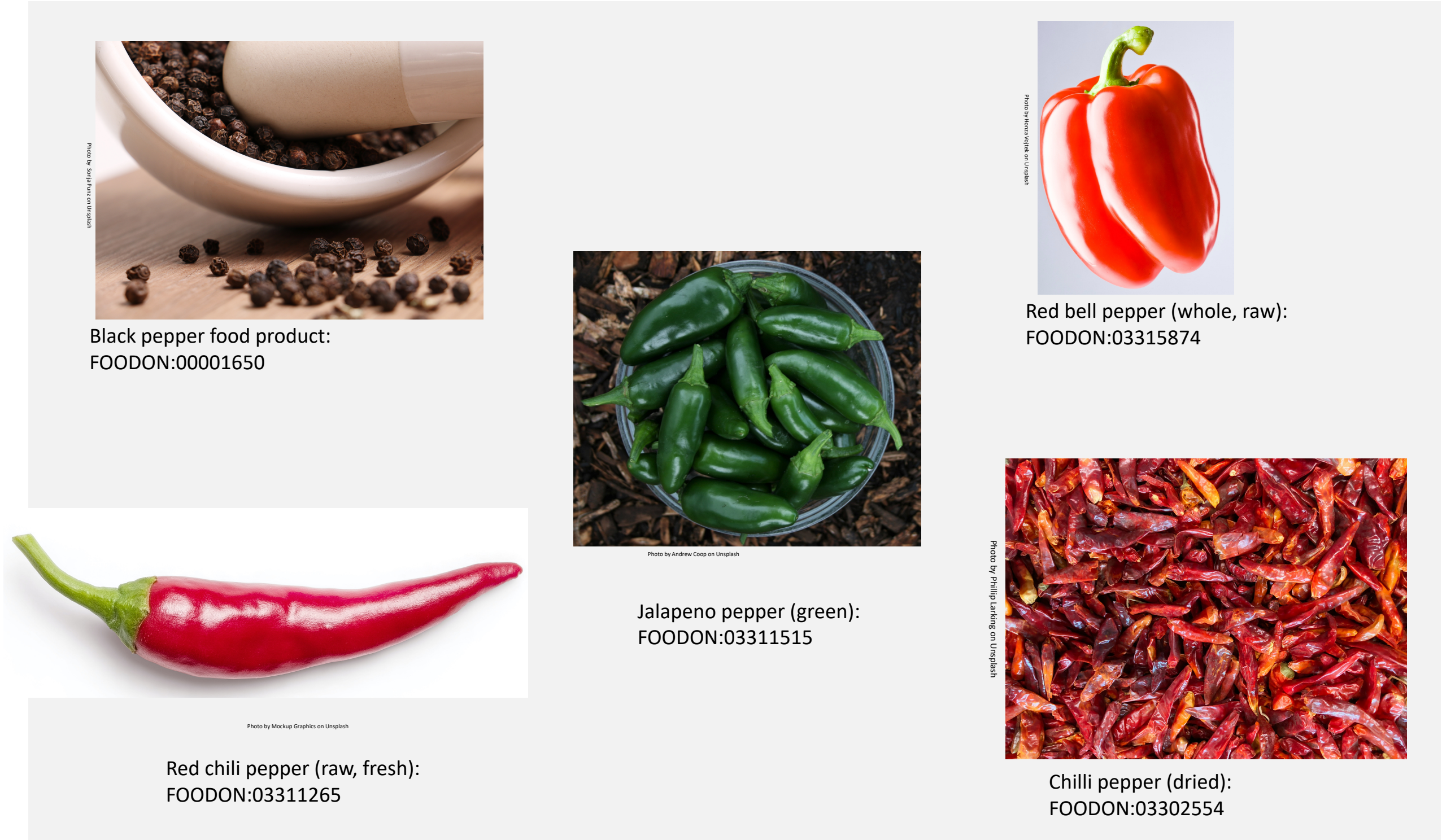
Results and Discussion

- Among 9,452 isolate source descriptors that were evaluated using LexMapr, 88% were categorized using IFSAC+ schema and translated into ontological terms; 12% of records lack an ontology definition or manual curation was required for IFSAC+ categorization assignment.

- Limitations for the automatic assignment of ontology terms and specific categories were found due two main reasons: incomplete or inconsistent information for the isolation source field and lack of standard descriptors for environmental sources specifically at the level of food facilities. As a result, GenomeTrakr Metadata guidelines were reviewed for clarification on metadata requirements, a GenomeTrakr Metadata Validation system was designed and a new effort for ontology development was established.

- IFSAC categorization schema was reviewed to include categories such as: Dietary supplement, food additive, food analog, preserves, sweetener, veterinarian clinical/research.

- After evaluation of LexMapr, 71,530 records were updated at NCBI and resources are publicly available.


Black pepper food product: FOODON:00001650

Jalapeno pepper (green): FOODON:03311515

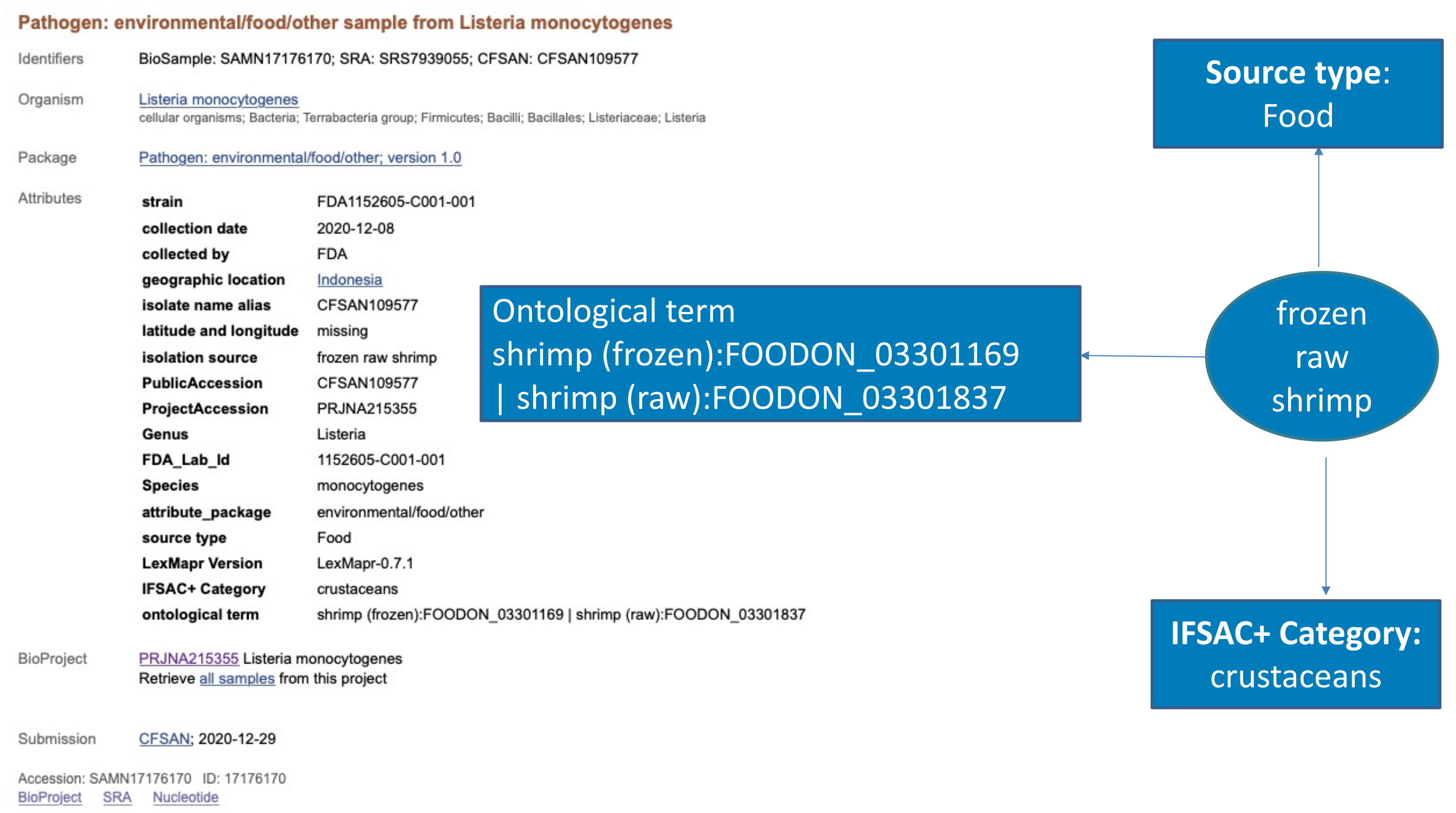
Red bell pepper (whole, raw): FOODON:03315874

Red chili pepper (raw, fresh): FOODON:03311265

Chilli pepper (dried): FOODON:03302554

Photo by Sonja Punz on Unsplash
Photo by Andrew Coop on Unsplash
Photo by Honza Vojtek on Unsplash
Photo by Mockup Graphics on Unsplash
Photo by Phillip Larking on Unsplash

Figure 2. The generic food descriptor 'pepper' can refer to multiple instances within the ontology and must be described in sufficient detail to allow for automatic assignment of ontology terms.


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	

Accession: SAMN17176170 ID: 17176170
BioProject SRA Nucleotide

Source type: Food

Ontological term shrimp (frozen):FOODON_03301169 | shrimp (raw):FOODON_03301837

frozen raw shrimp

IFSAC+ Category: crustaceans

Figure 3. Standardization of isolation source for food/environmental isolates at NCBI. The isolation source es described in terms of source type, ontological term, IFSAC+ category.

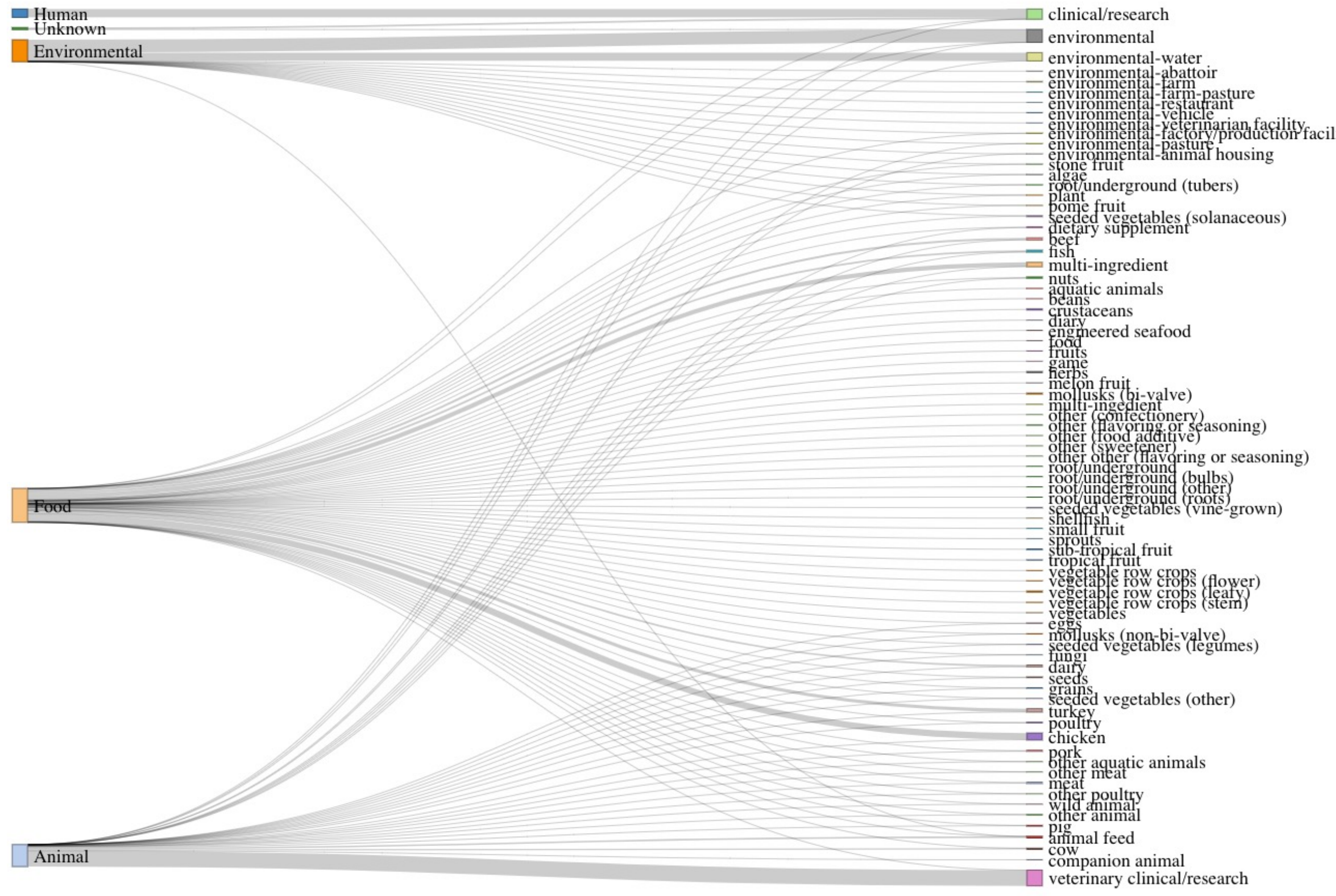


Figure 4. A Sankey diagram showing the primary IFSAC+ categorization schema for isolation sources by general source types.

Future Work

- Development of ontology for food production environment.

- GenomeTrakr Metadata Validation System <- Platform for the generation of metadata that conforms to metadata standards.

- Expansion of the requirements to describe food/environmental descriptors to include aspects of the food such as: packaging medium, food processing, food conservation, etc.

Conclusion

- LexMapr is a text-mining tool that standardizes isolation source attributes, by resolving name inconsistencies, translating source descriptors into semantic web ontology terms and providing defined IFSAC+ categories.

- Isolation source is a crucial metadata descriptor for WGS data from food and environmental isolates. The information provided is not limited to food general names, but it might include different aspects of the food product such as: preservation, capture method, transformation, foreign nomenclature, etc. LexMapr captures each of the source facets and generate standardize terms that are machine readable.

- Developing a domain ontology focused on food production environments will support the standardization efforts for WGS data from foodborne pathogens.

- GenomeTrakr is committed to strengthen WGS data resources for foodborne pathogen surveillance by implementing standards that better define the metadata associated to WGS records and making resources publicly available at NCBI.