

Adding IFSAC+ categories and FoodOn Ontology Terms to GenomeTrakr BioSamples

Maria D. Balkey
Staff Fellow

CFSAN – ORS – DM - MMSB
maria.balkey@fda.hhs.gov

Genomic Curiosity

Wunderkammer



Source: <https://www.artsy.net/article/artsy-editorial-eccentric-cabinets-curiosity-captivated-renaissance-europe>

Curio Cabinet

“repositories for all manner of wondrous and exotic - eclectic objects”

- *naturalia* (products of nature)
- *arteficialia* (the products of man)
- *scientifica* (testaments to man's ability to comprehend and dominate nature)

High quality Metadata guides the process of understanding genomic data

Facilitate an extensive curation of GenomeTrakr data by standardizing the source definition of bacterial genomes

1. Identify and resolve inconsistencies in the isolate source attribute for existing GenomeTrakr entries.
2. Translate the isolate source attribute into Semantic Web Ontology Terms to improve interoperability for the dissemination and integration of genomic data.
3. Categorize isolate source attributes into classes recognized by IFSAC and other standardized terminologies.
4. Create standard mechanisms to populate the isolate source field for this genomic data that will guide data contributors through the metadata collection steps.



A tool for transforming contextual data

Free text

Ontology Term

3rd Party Scheme

iqf green beans

↓
individually quick frozen
green bean

green bean (frozen):foodon_03310335



beans
(IFSAC+)

**Local Database =
93,659 entries**

Source types = 9569

**Data processing,
mapping to ontologies**

**Map to 3rd party
classification
scheme**



Will Hsiao
Boss stuff



Gurinder Gosal
LexMapr dev



Damion Dooley
Ontology curation



Ivan Gill
LexMapr dev



Dan Fornika
Dockerization



Emma Griffiths
*Curation,
User engagement*

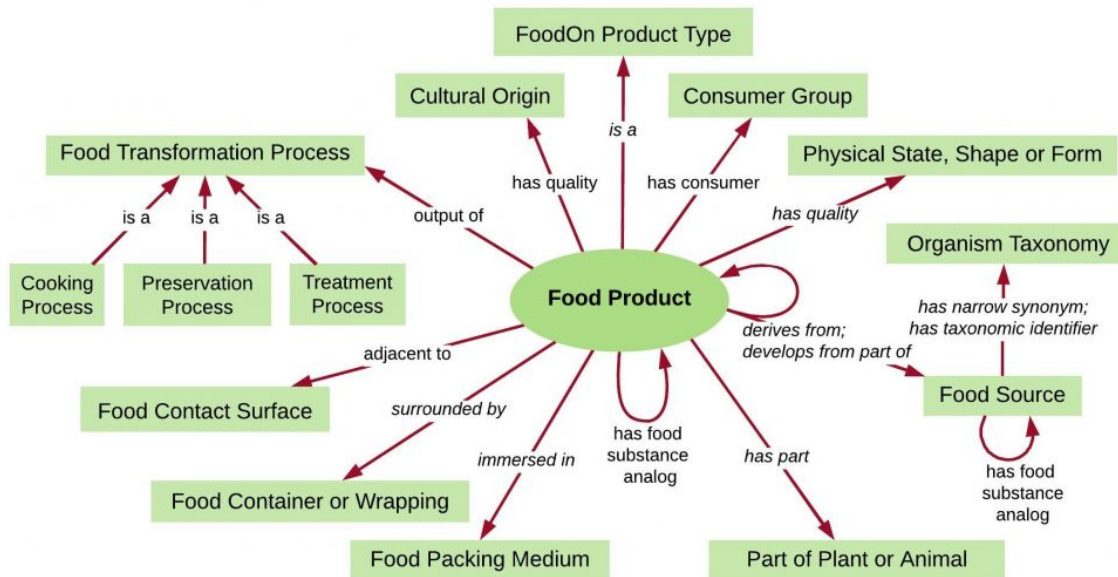
1. Identify and resolve inconsistencies in the isolate source attribute for existing GenomeTrakr entries.

Abbreviations	frz crab cakes -> frozen crab cake
Spelling mistakes	9oz. conatiner of guacamole -> container of guacamole
Character misuse	bnls /sknls chicken breast -> boneless skinless chicken breast avian liver -> avian liver
Plurality	Anaheim peppers -> Anaheim pepper
Multilingual phrases	queso fresco cheese -> Hispanic cheese Achiote molido (annatto) -> ground annatto
Use of redundant terms	Isolate, sample, enrichment
Use of ambiguous terms	Food, general, finished product, raw product
Use of irrelevant terms	Petri dish, patient, clinical, laboratory strain

2. Translate the isolate source attribute into Semantic Web Ontology Terms to improve interoperability for the dissemination and integration of genomic data.



Food product diagram

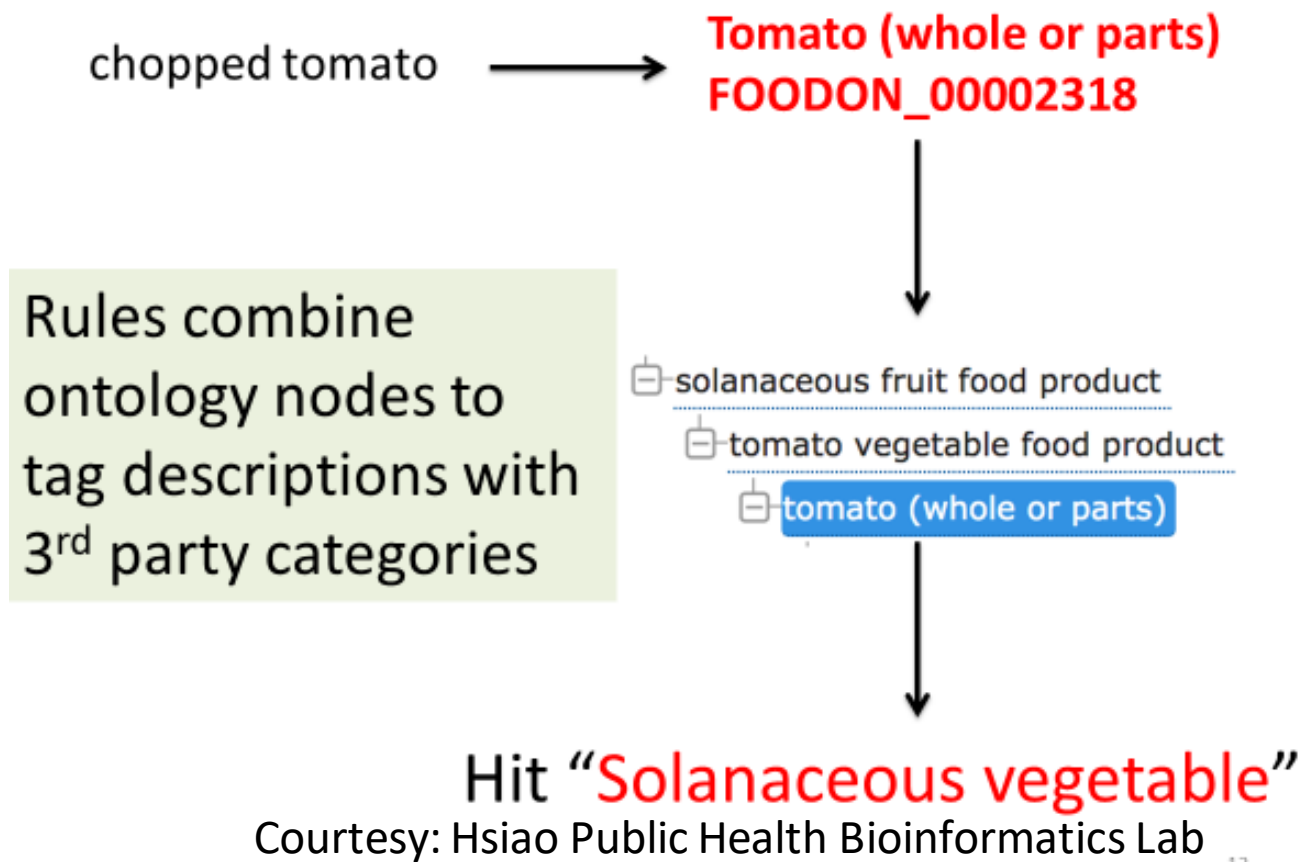


Dooley et al. npjscifood. 2018

- To index and retrieve genomic data from food and environmental sources.
- To build interoperability (communication) among software systems.

- **Ontology:** Formal representation of a domain modeling the logical relationships that exist across the members of such domain.

2. Translate the isolate source attribute into Semantic Web Ontology Terms to improve interoperability for the dissemination and integration of genomic data.



IFSAC Food Categories with Examples

(Not all food examples below have been linked to outbreaks, and many foods not shown below have been linked to outbreaks.)
This categorization scheme¹ updates a former scheme² to include more specific information.

Land Animals	DAIRY Solid/semisolid dairy products <i>hard and soft cheese</i> Fluid Milk <i>whole milk, skim milk</i>	
	EGGS Shell Eggs <i>whole shell eggs</i> Egg Products <i>egg whites in cartons</i>	
	MEAT-POULTRY Meat Beef <i>ground beef, steaks</i> Pork <i>bacon, hams</i> Other meat <i>lamb, goats</i> Poultry Chicken <i>whole chickens, chicken deli meats</i> Turkey <i>turkey hot dogs, whole turkeys</i> Other poultry <i>ducks, ostriches</i> Game <i>wild boars, venison</i>	
Aquatic Animals	FISH Siluriformes <i>catfish</i> Other Fish <i>grouper, tuna</i>	
	SHELLFISH Crustaceans <i>crabs, lobster</i> Mollusks Bivalve <i>clams, oysters</i> Non-bivalve <i>octopuses, squid</i> OTHER AQUATIC ANIMALS <i>frogs, jellyfish</i>	
Plants	OILS-SUGARS <i>olive oils, canola oils, sugars, honey</i> PRODUCE Vegetables Fungi <i>portabellas, button mushrooms</i> Sprouts <i>alfalfa sprouts, mung bean sprouts</i> Root-underground Roots <i>carrots, beets</i> Tubers <i>potatoes, yams</i> Bulbs <i>garlic, onions</i> Other <i>ginger, taro</i> Seeded vegetables Vine-grown <i>squashes, cucumbers</i> Solanaceous <i>tomatoes, peppers</i> Legumes <i>lima beans, snow peas</i> Other <i>okras, sweet corn</i> Herbs <i>basil, cilantro</i> Vegetable row crops Flowers <i>artichokes, broccoli</i> Stems <i>asparagus, celeries</i> Leafy <i>lettuce, spinach</i>	
	Fruits Melons <i>cantaloupes, watermelons</i> Pomes <i>apples, pears</i> Stones <i>apricots, cherries</i> Small <i>blueberries, strawberries</i> Tropical <i>bananas, mangoes</i> Sub-tropical <i>avocados, oranges</i>	
Other	GRAINS-BEANS Grains <i>rice, wheat</i> Beans <i>pinto beans, black beans</i> NUTS-SEEDS Nuts <i>peanuts, almonds</i> Seeds <i>sesame seeds, pumpkin seeds</i>	
	OTHER FOODS <i>ice, alcohol, dietary supplements</i>	

3. Categorize isolate source attributes into classes recognized by IFSAC and other standardized terminologies.

Source types = 9569

IFSAC+

Classes or Categories = 532

IFSAC -> Interagency Food Safety Analytics Collaboration (IFSAC) hierarchical scheme for categorizing foods implicated in outbreaks.

¹ Richardson LG, Bazaro MC, Parker CC, Dewey-Mattia D, Golden N, Jones K, Klontz K, Travis C, Zablotsky Kufel J, Cole D. An updated scheme for categorizing foods implicated in foodborne disease outbreaks: a tri-agency collaboration. *Foodborne Pathog Dis*. 2017 Sep 19. ahead of print.

² Painter JA, Ayers T, Woodruff R, Blanton E, Perez N, Hoekstra R, Griffin PM, Braden C. Recipes for foodborne outbreaks: a scheme for categorizing and grouping implicated foods. *Foodborne Pathog Dis*. 2009; 6:1259-64.

Updating isolate source field is in progress.

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

Identifiers BioSample: SAMN05462812; SRA: SRS1595264; CFSAN: CFSAN053054

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

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

Attributes

strain	FDA00010475
latitude and longitude	missing
collection date	2016-07-22
geographic location	Egypt
isolation source	individually quick frozen green bean
collected by	FDA
host disease	missing
isolate name alias	CFSAN053054
host	missing
FDA_lab_id	962768-1b
attribute_package	environmental/food/other
IFSAC+ Category	beans
FoodOn Ontology Term	green bean (frozen):foodon_03310335

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

Submission [CFSAN](#); 2016-07-29

Accession: SAMN05462812 ID: 5462812
[BioProject](#) [SRA](#) [Nucleotide](#)

41 categories reviewed
and updated

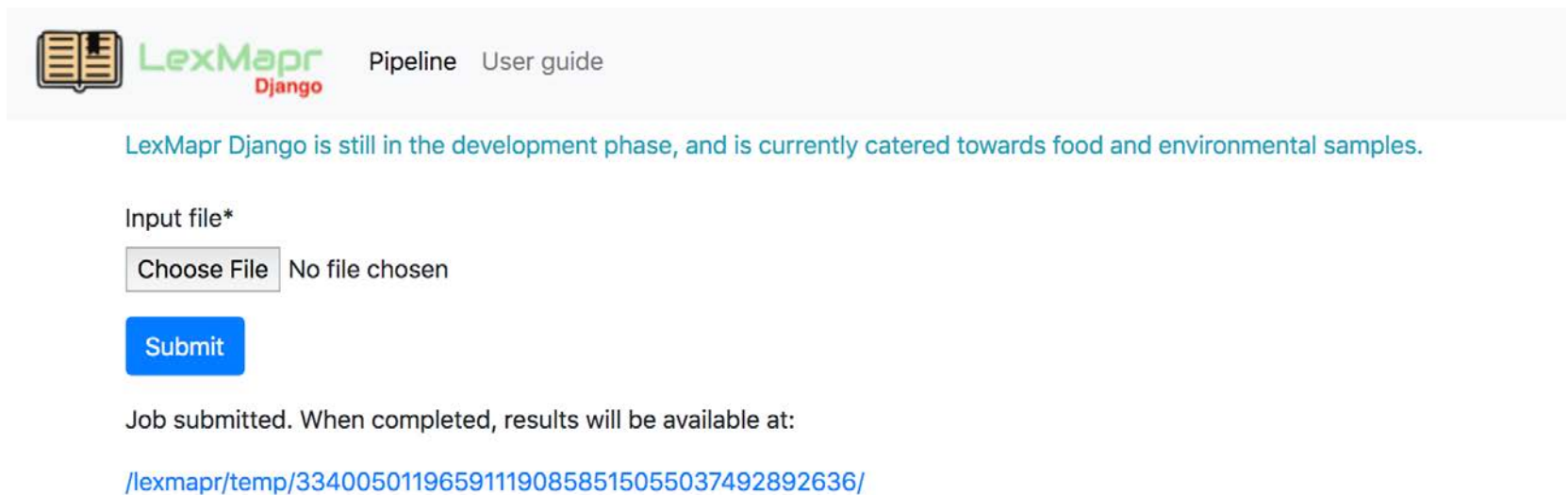
~~1qf green beans~~

4138 entries
updated at NCBI



4. Create standard mechanisms to populate the isolate source field for this genomic data that will guide data contributors through the metadata collection steps.

- LexMapr developers created a GUI interface.
- LexMapr Input: Upload csv file (any size)
 - Fields: **Sample_Id, Sample_Desc.**



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

Job submitted. When completed, results will be available at:

</lexmapr/temp/334005011965911190858515055037492892636/>

Under development! Feedback welcome.

<https://watson.bccdc.med.ubc.ca/lexmapr>

- LexMapr Output: TSV file
 - Fields: **Sample_Id, Sample_Desc,**
 - **“Matched components” = ontology terms,**
 - “Third Party Classification” = IFSAC+ categories**



This page will expire on Sept. 17, 2019, 2:11 p.m.

Sample_Id	Sample_Desc	Cleaned_Sample	Matched_Components	Third Party Classification
small_simple1	Chicken Breast	chicken breast	['chicken breast:foodon_00002703']	['chicken']
small_simple2	Baked Potato	baked potato	['potato (whole, baked):foodon_03302196']	['root/underground (tubers)']
small_simple3	Canned Corn	canned corn	['corn (canned):foodon_03302665']	['seeded vegetables (other)']
small_simple4	Frozen Yogurt	frozen yogurt	['frozen yogurt:foodon_03307445']	['dairy']
small_simple5	Apple Pie	apple pie	['apple pie:foodon_00002475']	['pome fruit']

Download (.tsv)

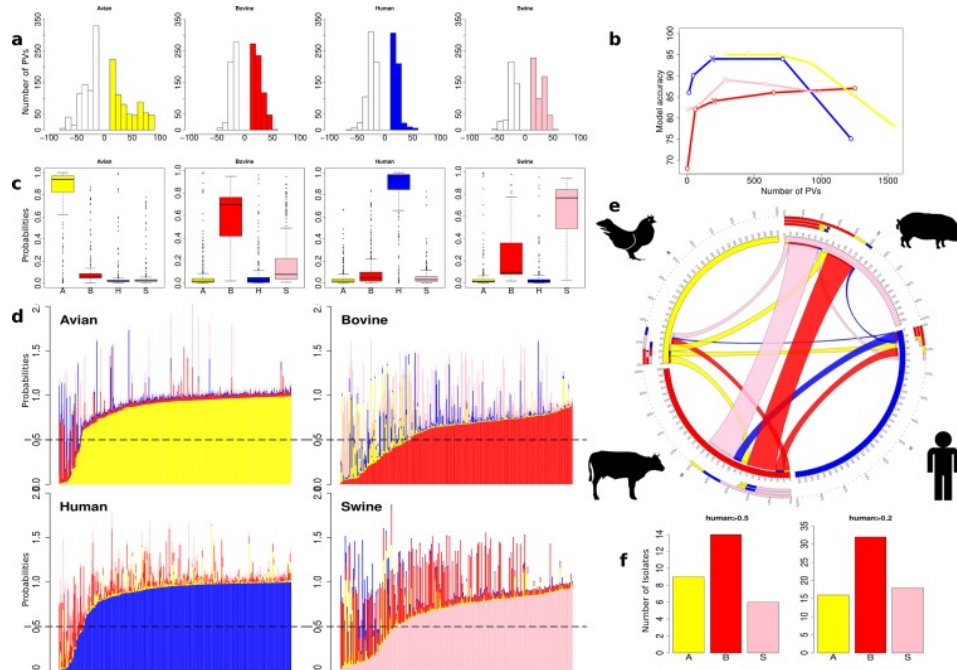
Why the isolate source is so important?

- Public Health Interventions, regulatory action and outbreak investigation depend on reliable sources.
- Predict bacterial sources using genomic features (source attribution).
- Investigate isolates that demonstrate host adaptability.
- Evaluation of transmission sources, persistence of genotypes across the food chain (risk assessment).
- Variation in stress-resistance capabilities across food and environmental sources.

Source attribution using Machine Learning

- Supervised Learning needs a labeled dataset to determine new labels from new data (e.g. Support Vector Machine (SVM) and Random Forest)

Host prediction by SVM

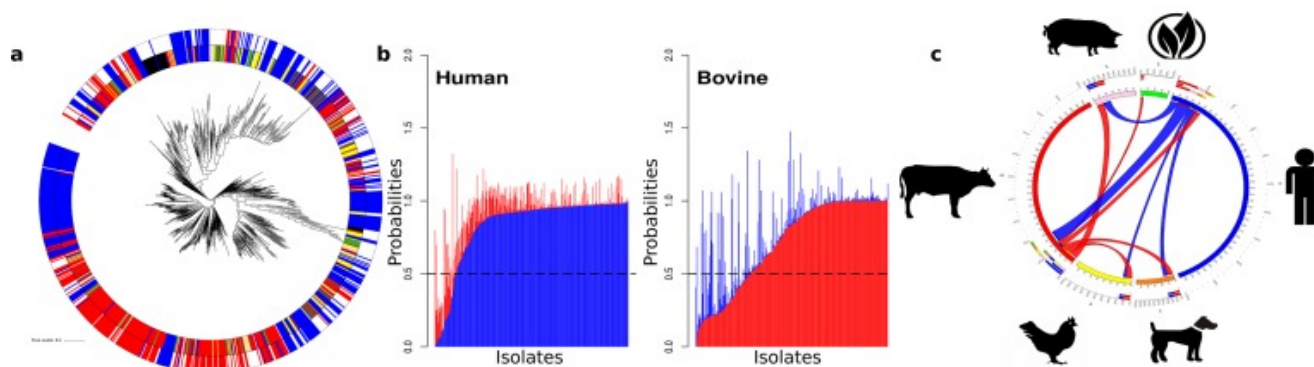


Lupolova et al. Microb Genom. 2017

- Data <- Salmonella sequences from clearly defined sources: **avian, bovine, human, swine.**
- Goal <- Prediction of isolate source for *Salmonella enterica* serovar Typhimurium
- ML approach <- SVM

Source attribution using Machine Learning

Accessory genome analysis and host prediction by SVM for *Escherichia coli*



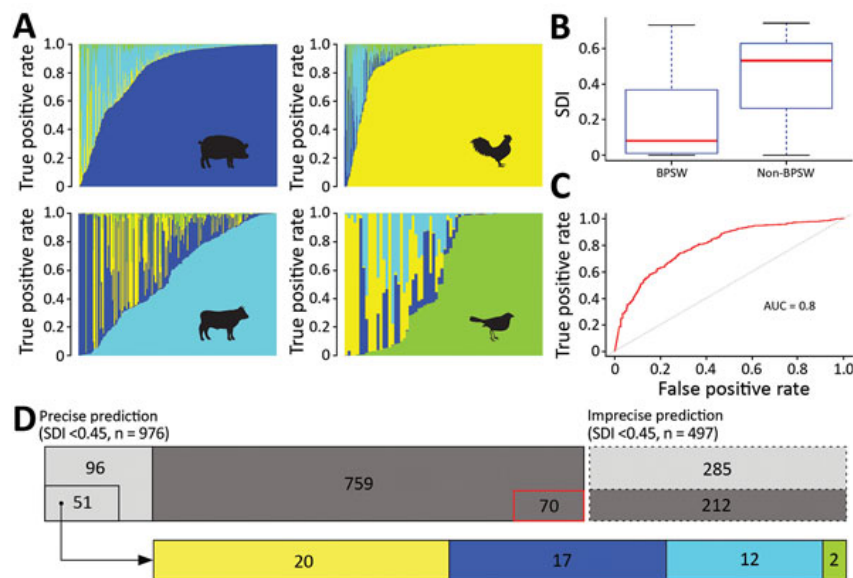
Lupolova et al. Microb Genom. 2017

- Host – habitat related sequences increases the accuracy of source predictors in case of *Escherichia coli* where the available data is less abundant than *Salmonella enterica*.
- ML supports inference in cases of multiple hosts and estimates likelihood of zoonotic events.

Source attribution using Machine Learning



Source prediction by Random Forest classifier



Shang et al. Emerg Infect Dis. 2018

- Data <- *Salmonella enterica* serovar Typhimurium sequences; sources: bovine, poultry, wild bird, swine, “miscellaneous food,” human.
- ML approach: Random forest.
- Source prediction of *Salmonella enterica* serovar Typhimurium giving genomic features.

Conclusions

- Isolate source metadata for the GenomeTrakr program is under extensive curation. Collaboration from participating labs might be requested to help accomplish this extensive task.
- Resolving inconsistencies for the isolate source field, adding IFSAC+ categories, FoodON ontology terms at our local DB and NCBI is a GenomeTrakr effort and a priority. Updates are in progress.
- LexMapr is the platform that we are currently using to accomplish this effort. Manual curation is essential to assess and improve the quality of GenomeTrakr metadata.
- LexMapr is currently available in a Web Graphical User Interface. Besides, it can be downloaded at your local machine through GitHub. Some command line interface experience is required.

Acknowledgements

- **FoodOn Consortium**
 - Simon Fraser University Department of Molecular Biology and Biochemistry
 - Fiona Brinkman
 - Emma Griffiths
 - University of British Columbia Department of Pathology and Laboratory Medicine
 - William Hsiao
 - Damion Dooley
- **LexMapr Team:** Hsiao Lab
- **GenomeTrakr Labs**
 - State Health and University Labs
 - FDA field Labs
- **U.S Food and Drug Administration**
 - CFSAN
 - Office of Regulatory Sciences
 - Office of Analytics and Outreach
- **National Institute of Health (NIH)**
 - National Center for Biotechnology Information (NCBI)

