Streamlining FoodOn Seafood Nomenclature using a Semi-automated ROBOT Template-driven Approach

Anoosha Sehar [1], Damion Dooley[1], Amanda Windsor[2] and William Hsiao [1*]

[1]*Centre for Infectious Disease Genomics and One Health (CIDGOH), Faculty of Health Sciences, Simon Fraser University, Burnaby, B.C, Canada, V5A 1S6*
[2] *United States Food and Drug Administration, Office of Regulatory Science, 5001 Campus Dr. College Park, MD 20740*

Abstract

The nomenclature of seafood species and their products is one of the very important areas which needs to be curated and regularly updated in FoodOn. In this paper, we present a semi-automated ROBOT template-driven approach we designed for aligning FoodOn with the FDA issued 'Seafood List', together with other established resources e.g., NCBI, ITIS, and Wikipedia. The basic data in the FDA Seafood List, which included Type, Common Name, FDA Law, FDA Acceptable Market Name(s) and Scientific Name was exported in an Excel format. FDA Seafood Labels (Scientific Name) were mapped against NCBITaxons and NCBI GenBank Names using ETE 3 toolkit and around 90% of labels were correctly matched. ITIS TSNs were available for over 85% of seafood labels which were fetched using a locally installed ITIS database. Wikipedia-URL was retrieved as a cross-referenced database using the FoodOn-Wikipedia tool. In some cases, Wikidata was also used as an interface to connect to NCBITaxon. The curated seafood data was then converted from a tab-delimited TSV template file to a Web Ontology Language OWL file format using ROBOT template. This method will not only help FoodOn to regularly update seafood organisms but will also help to maximize the seafood product coverage and data interoperability.

Keywords

Ontology, FoodOn, Seafood, Nomenclature, ROBOT

1. Motivation

FoodOn is a comprehensive ontology that provides a machine readable food product categorization system, but there are many branches of FoodOn that still need to be developed or refined [1]. The nomenclature of organisms and products related to seafood was recognized as one challenge which needs to be addressed in FoodOn because seafood products are often recognized by multiple names, which can differ by region, language and usage [2]. A standardized pattern to describe seafoods is needed in FoodOn in order to maximize data interoperability. Currently, FoodOn inherited seafood related organisms from facet B of LanguaL [3]. These entities have mostly been converted to NCBITaxon ontology terms or, in the absence of a matching NCBITaxon entry, FoodOn organism terms. In the case of filleted products derived from these organisms, these are now listed in the branch of FoodOn that contains entries from the LanguaL-indexed SIREN database. Due to the inconsistency between the use of common names and scientific names in seafood commodities around the world,

IFOW 2021: 2nd Integrated Food Ontology Workshop, held at JOWO 2021: Episode VII, The Bolzano Summer of Knowledge, September 11-18, 2021, Bolzano, Italy
EMAIL: anoosha_sehar@sfu.ca (A. Sehar); damion_dooley@sfu.ca (D. Dooley); Amanda.Windsor@fda.hhs.gov (A. Windsor); wwhsiao@sfu.ca (W. Hsiao)
ORCID: 0000-0001-5275-8866 (A. Sehar); 0000-0002-8844-9165 (D. Dooley); 0000-0002-5192-7047 (A. Windsor); 0000-0002-1342-4043 (W. Hsiao)

consistencies in taxonomic usage and organizational structure of seafood related entries needs to be examined.

2. Introduction

The United States Food and Drug Administration (FDA) has issued a guidance document called the "Seafood List" [4]. The Seafood List provides a searchable database for hundreds of species commonly sold as seafood commodities in the United States and serves as a guide for acceptable market names for seafood sold in interstate commerce. The list identifies the following information: the 'Type' of seafood as vertebrate, invertebrate, or invertebrate (crustacean); the 'Acceptable Market Name(s)' are names that FDA generally recognizes as a suitable "statement of identity" in the labeling of a species; the 'Common Name' is a name provided in scientific references for a species; the 'Scientific Name' is the Latin binomial for the genus and species of a fish, established by taxonomists; and 'Vernacular Names' are regional names commonly used in local markets [5]. An acceptable market name(s) may be a "common or usual name" established by either a history of common usage in the U.S. or more rarely, a name specifically coined as the market name for a species. In some cases, the market name and the common name are the same, e.g., "Giant pangasius" is both the market name and common name for *Pangasius sanitwongsei*. Almost all species in the FDA Seafood List have been assigned a unique common name and their use as market names (if there are no regulatory restrictions) has the advantage of limiting confusion about species identity in the marketplace [5].

It is imperative to connect and incorporate the information in an ontology in an automated fashion to improve consistency [6]. To address this challenge, FoodOn is endeavoring to integrate the FDA Seafood List by ROBOT [7] redesigning and mapping against multiple databases (e.g., National Center for Biotechnology Information (NCBI) [8], Wikipedia [9], and Integrated Taxonomic Information System (ITIS) [10].

3. Methodology

3.1. Data extraction from FDA

FDA has around 1980 fish terms in The Seafood List as of June 2021 [4]. The basic data in The FDA Seafood List was exported in an Excel .csv file format. The data was exported to the Google sheet 'FoodOn Robot Tables'; tab: 'Seafood' available at https://tinyurl.com/SeafoodList. The Google sheet 'FoodOn Robot Tables' currently provides Robot tables for anatomical part of the plant or animal (including seafood) from which the food product or its major ingredient is derived. Seafood list in 'FoodOn Robot Tables' will become location for all upcoming FoodOn seafood entries. The data in Seafood List exported from FDA included Type, Common Name, FDA Law, FDA Acceptable Market Name(s), and Scientific Name. Few annotations are renamed as shown in the first row as header, such that, scientific name is changed to FDA Seafood 'Label', Common name to 'FDA Alternative Term,'. Type and FDA Law are not included in FoodOn for now. Type is covered by distinctions made in the NCBITaxon hierarchy.

3.2. Mapping to NCBITaxon ID

The NCBITaxa function from the ETE 3 toolkit [11] was used to map FDA Seafood Label (Scientific Name) against the NCBI database in order to retrieve the NCBITaxon for every label using a custom python script.

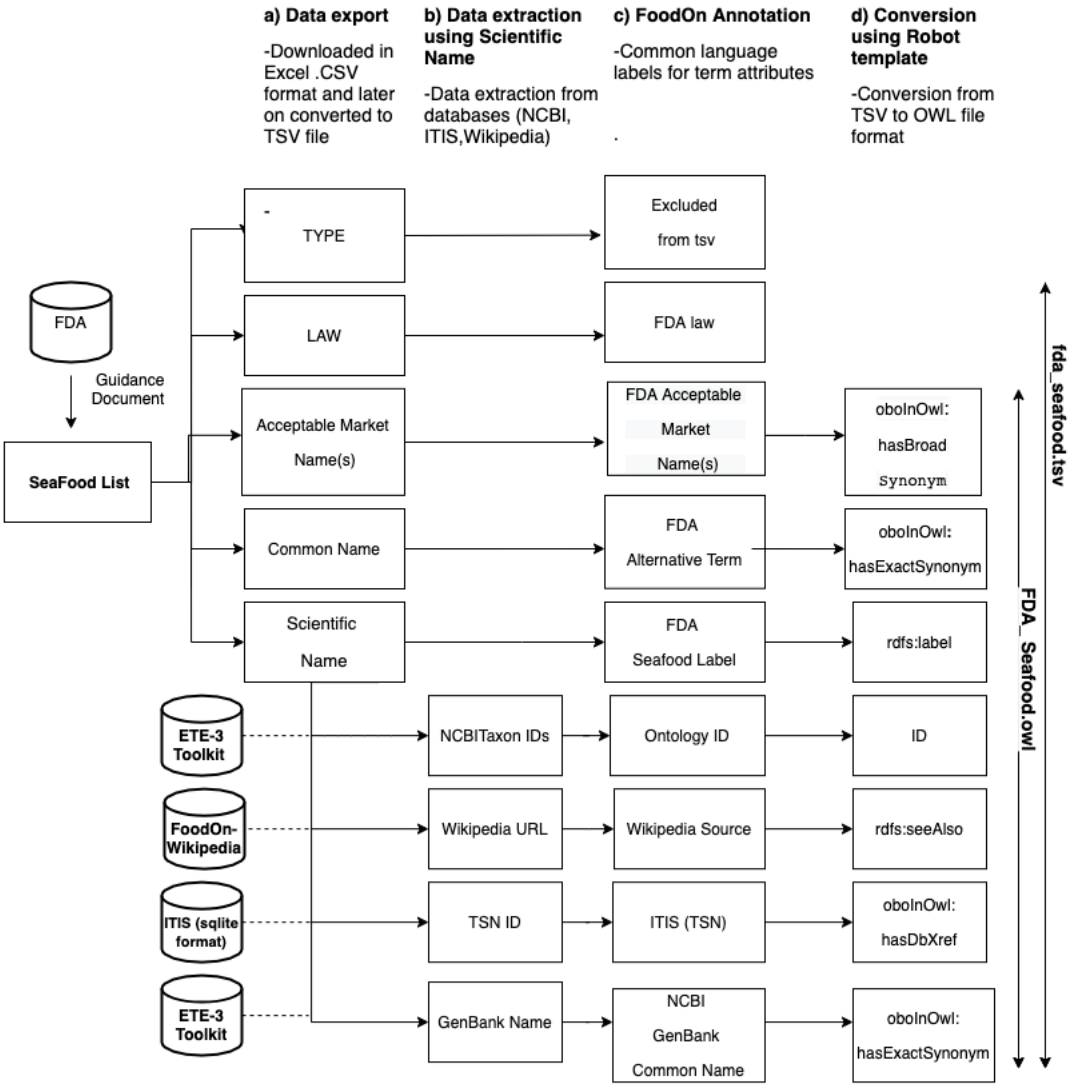

Figure 1: Workflow shows the FoodOn approach to synchronize with FDA Seafood list together with other resources (a) FDA Seafood List data was exported in an Excel (.CSV file format) which included the information about Type, Law, Acceptable Market Name(s), Common Name and Scientific Name. (b) Scientific Name was used to map and extract data from databases i-e-NCBI, Wikipedia and ITIS. NCBITaxon ID was fetched using ETE 3 Toolkit; Wikipedia URL was retrieved using FoodOn Wikipedia tool, TSN Number was retrieved from ITIS; GenBank Name was collected through ETE 3 Toolkit. (c) FoodOn provided common language labels for term attributes. (d) Finally, ROBOT was used for conversion from tab delimited TSV file format to OWL file format.

3.3. Extraction of TSN number from ITIS

The Integrated Taxonomic Information System (ITIS) has taxonomic information on plants, animals, fungi, and microbes of North America and the world [10]. Every entity has a Taxonomic Serial Number (TSN) which is a unique identifier. The tables in the database use the TSN number as a primary key and give information about taxonomic status, synonyms, taxonomic hierarchy, etc. The ITIS database was stored in a local directory in SQLite format and FDA Seafood Label (Scientific Name) was used as a target input to fetch the TSN number.

3.4. Wikipedia URL

Wikipedia is one of the most widely used cross-referenced databases. The Wikipedia URL was fetched for every FDA Seafood Label (Scientific Name) using the FoodOn-Wikipedia package, a tool for managing the collection of Wikipedia entry cross references, definitions and images for FoodOn entities [12]. In some cases where terms were not matched directly in NCBITaxon, Wikidata was used as an interface to connect to NCBITaxon. For example, NCBI does not have a record of *Acipenser multiscutatus* (FDA seafood label), but Wikidata has listed an alternative synonym *Acipenser Schrenckii* which was used to connect to NCBI [13].

3.5. NCBI GenBank Names

NCBI GenBank Names were extracted using the ETE 3 toolkit wrapped in a custom python script [11]. The difference between FDA common names and NCBI GenBank names were transformed into 'hasExactSynonym' annotation using ROBOT template [7].

3.6. ROBOT

ROBOT is a command-line tool for working with OWL ontology files [7]. The ROBOT template command was used to convert from a tab-delimited TSV template file of term labels, definitions and other attributes to an OWL file format. The first row of the template file has FDA Seafood List field names and other common language labels for term attributes, while the second row contains template ROBOT command string expressions used in the OWL conversion. The field 'FDA seafood label' was converted to 'label' which is a special keyword to specify an rdfs:label in ROBOT, and uniquely identifies the target term; 'FDA Market Name(s)' field was converted to 'oboInOwl:hasBroadSynonym' as FDA Acceptable Market Name (s) appears to be more generalized than species level; 'FDA alternative term' and 'NCBIGenBank Common Names' was transformed to 'oboInOwl:hasExactSynonym'; ITIS (TSN) number to 'oboInOwl:hasDbXref' and Wikipedia Source to 'rdfs:seeAlso'.

4. Results

FoodOn now has a semi-automated ROBOT template-driven process for synchronizing with the FDA Seafood List together with other established resources i.e., NCBI, ITIS, and Wikipedia. An initial bulk matching of FDA seafood label (scientific name) to NCBITaxon ID resulted in nearly 90% of FDA seafood labels being matched to NCBITaxon identifiers. Around 200 seafood labels were missing NCBITaxon labels initially, which revealed discrepancies between the databases. This is usually due to taxonomic revisions. When these discrepancies were addressed, an additional 100 NCBITaxon labels were able to be mapped. The remaining FDA seafood labels which do not have a NCBITaxon ID available will need to be looked up manually.

The curated Google sheet *(see Data Availability) also contains the terms which are potentially missing NCBITaxons and will be incorporated in FoodOn with the provision of FoodOn IDs. TSN numbers fetched from ITIS database were available for over 85% of seafood labels; nearly 65% of NCBI GenBank names were extracted from NCBI as synonyms, from which approximately 20% unique GenBank names have been kept in FoodOn; a Wikipedia link (as a cross referenced database) was retrieved by using the FoodOn-Wikipedia package [12], where it successfully mapped more than 75% of terms. The integrated Google sheet will be continuously updated as FDA adds new terms to their seafood list.

5. Data Availability

The TSV file is available at FoodOn GitHub: https://tinyurl.com/FDA-seafood-tsv

The robot converted OWL file can be accessed from: https://tinyurl.com/robot-fda-seafood-owl

* 'FoodOn Robot Tables'; tab: 'Seafood' is available at https://tinyurl.com/SeafoodList.

6. Discussion

A few discrepancies were observed while connecting data across other databases, which can be a hurdle for standardizing data. For example, the FDA Seafood List sometimes classifies a taxon as the genus and all species within that genus, for example *Arothron* spp., while NCBI provides a taxon ID at the genus level, i.e., *Arothron* (NCBITaxon: 50368). Secondly, a few terms with taxa rank at genus level in NCBI have an additional tag '<bony fishes>' attached to them, for example *Calamus* <bony fishes> (NCBITaxon:119695) which requires custom searching to match the terms which have the tag. This also creates an inconsistency in data representation and matching FDA Seafood Labels with NCBITaxon ID. A discrepancy between NCBI ids and FDA labels has also been observed. For example, the two separate labels in the FDA list *Scophthalmus maximus*, and *Psetta maxima* possess the same NCBITaxon ID because *Psetta maxima* is recognized as a synonym of *Scophthalmus maximus* (NCBITaxon: 52904) by NCBI.

7. Conclusion

A semi-automated ROBOT template-driven approach was designed for populating FoodOn with the FDA issued 'Seafood List' together with other mature resources e.g., NCBI, ITIS, and Wikipedia. The effort of FoodOn to update its seafood organism content by integrating labels from the FDA Seafood List along with synonyms derived from FDA acceptable market names and English common names is a first step towards a robust ontology of seafood terms that can underpin metadata interoperability across studies. This method will also help FoodOn to regularly update seafood organisms and maximize the seafood product coverage.

8. Future Work

The seafood labels which are missing NCBITaxon IDs will be looked up manually and will be provided with FoodOn IDs. Fishbase, a global biodiversity information system on finfishes and GBIF, the global biodiversity information facility can be added as an additional cross- referenced database [14]. FoodOn aims to have fillet terms for all FDA seafood labels where appropriate. Another goal would be the incremental addition of seafood items in FoodOn which are not in the FDA seafood list. One of the future FoodOn goals is to import the FAO-ASFIS list, which presently includes 12,871 fish species used as food worldwide [15] .

9. Acknowledgements

This work is primarily supported by the USDA Non-Assistance Cooperative Agreement 58-8040-8-014-F and Genome Canada Grant 286GET to W. Hsiao. Amanda Windsor thanks FDA colleagues for reviewing and improvements made to the manuscript.

References

[1] D. M. Dooley *et al.*, "FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration," *NPJ Sci Food*, vol. 2, p. 23, Dec. 2018.

[2] L. Towers, "US Seafood Naming Rules: Do They Provide Real Guidance for the Seafood Industry?," 22-Jun-2021. [Online]. Available: https://thefishsite.com/articles/us-seafood-naming-rules-do-they-provide-real-guidance-for-the-seafood-industry. [Accessed: 22-Jun-2021].

[3] J. D. Ireland and A. Møller, “LanguaL food description: a learning process,” *Eur. J. Clin. Nutr.*, vol. 64 Suppl 3, pp. S44-8, Nov. 2010.
[4] “The Seafood List.” [Online]. Available: https://www.cfsanappsexternal.fda.gov/scripts/fdcc/?set=SeafoodList. [Accessed: 22-Jun-2021].
[5] Center for Food Safety and A. Nutrition, “The Seafood List.” [Online]. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-industry-seafood-list. [Accessed: 25-Jun-2021].
[6] N. M. Ali, H. A. Khan, A. Y.-H. Then, C. Ving Ching, M. Gaur, and S. K. Dhillon, “Fish Ontology framework for taxonomy-based fish recognition,” *PeerJ*, vol. 5, p. e3811, Sep. 2017.
[7] R. C. Jackson, J. P. Balhoff, E. Douglass, N. L. Harris, C. J. Mungall, and J. A. Overton, “ROBOT: A Tool for Automating Ontology Workflows,” *BMC Bioinformatics*, vol. 20, no. 1, p. 407, Jul. 2019.
[8] “National Center for Biotechnology Information.” [Online]. Available: https://www.ncbi.nlm.nih.gov/. [Accessed: 25-Jun-2021].
[9] Wikipedia contributors, “Main Page,” *Wikipedia, The Free Encyclopedia*, 03-Feb-2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Main_Page&oldid=1004593520. [Accessed: 25-Jun-2021].
[10] C. A. Shaw, “ITIS (The Integrated Taxonomic Information System),” Jan. 2007.
[11] J. Huerta-Cepas, F. Serra, and P. Bork, “ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data,” *Mol. Biol. Evol.*, vol. 33, no. 6, pp. 1635–1638, Jun. 2016.
[12] *foodon-wikipedia*. Github.
[13] Wikipedia contributors, “Japanese sturgeon,” *Wikipedia, The Free Encyclopedia*, 02-Apr-2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Japanese_sturgeon&oldid=1015617395. [Accessed: 22-Jun-2021].
[14] “FishBase : A Global Information System on Fishes.” [Online]. Available: https://www.fishbase.se/home.htm. [Accessed: 22-Jun-2021].
[15] Food and Agriculture Organization of the United Nations (FAO-UN), “FAO Fisheries & Aquaculture ASFIS List of Species for Fishery Statistics Purposes, Overview,” 02-Nov-2006. [Online]. Available: http://www.fao.org/fishery/collection/asfis/en. [Accessed: 22-Jun-2021].