Development of a MIxS (Minimum Information about any (x) Sequence) Food Environmental Metadata Standard

Christopher J. Grim[1], Amanda M. Windsor[2], Brandon Kocurek[1], Susan R. Leonard[1], Taylor K. S. Richter[1], Gopal Gopinath[1], Maria Balkey[2], Padmini Ramachandran[2], Andrea Ottesen[3], Karen Jarvis[1], and Ruth Timme[2]

[1] U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition, Laurel, MD 20708

[2] U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition, College Park, MD 20740

[3] U.S. Food and Drug Administration, Center for Veterinary Medicine, Laurel, MD 20708

Extended Abstract

Keywords: ontology, food metadata, Minimum Information about any (x) Sequence, MIxS, food metagenome, food microbiome

Background

The increased accessibility and decreased cost of high throughput sequencing has enabled widespread use of this technology in the study of microbiomes. In the food industry, very few food commodities, if any, are produced as sterile products with certain products resulting from controlled microbial processes. In this regard, the biomass and composition of the microbiota associated with foodstuffs are extremely variable and influenced by factors such as cultivation and production environments, management practices, harvesting, packing, transport, animal and pest incursions, as well as hygiene of human workers. Additionally, the microbiological monitoring needs of the food industry is varied. Some sectors, for example, those involved in food safety and food quality, are only interested in a specific subset of microbes in the microbiome; surveillance has traditionally relied on time-intensive targeted cultivation of the microbes of interest. Metagenomics can provide microbiome characterization as well as culture-independent diagnostics, covering a wide range of needs for the food industry.

While metagenomics holds great promise to answer important biological questions pertaining to microbial composition of samples, these data have created a “Big Data” challenge for integrated microbiome research studies - especially with artificial intelligence and deep learning applications on the horizon. Comparative studies of metagenomic datasets are dependent upon robust but concordant metadata in which shared vocabularies are employed [1]. Standardized vocabularies to describe the features of samples and experiments following an interoperable metadata schema allows integration of datasets from varied sources which contain shared features. The Genomic Standards Consortium (GSC; https://gensc.org/) has defined a system for reporting “Minimum Information about any (x) Sequence" (MIxS). There is currently no MIxS environmental package to capture minimal metadata associated with food and its production and processing environments. Therefore, FDA researchers have begun the coordinated development

of a food metadata standard to describe attributes associated with microbiomes from food and food production environments.

Use Cases

The scope of what includes food and its production, processing and distribution, at first glance, can seem intractable. One of the first steps to developing a food metadata standard was to define the boundaries that encompass the sample types that would be covered, followed by the identification of metadata essential for each use case. An advisory group comprising government, regulatory, industry, and academic members with expertise in food safety and food production, was convened and use cases were defined from the various sectors. The scope of this metadata standard covers microbiomes from five general use cases: 1) finished and unfinished foods, 2) food production environment, 3) farm environment, 4) food production animals, and 5) functional foods. The finished and unfinished foods use case includes microbiomes associated with retail food commodities and their individual components and ingredients, whether in a ready-to-eat or raw state. The food production environment, which includes a subset of the built environment that is devoted to the production, processing, storage and distribution of food, includes samples specific to this environment such as production line food samples, raw ingredients, as well as environmental swabs and other sample types from environmental monitoring programs. Farm production environment samples include the food crop and any on-site processing stages, as well as components of environmental assessment studies designed to identify potential sources of contamination and the impact of farm management practices. The food-production animals use case includes livestock for human consumption as well as animal food and feed, including pet food. Lastly, functional foods, in which simple and complex microbial consortia are utilized under controlled conditions to produce foods such as the myriad of ethnic fermented foods, yogurt, probiotics, wines, and cheeses, were also included.

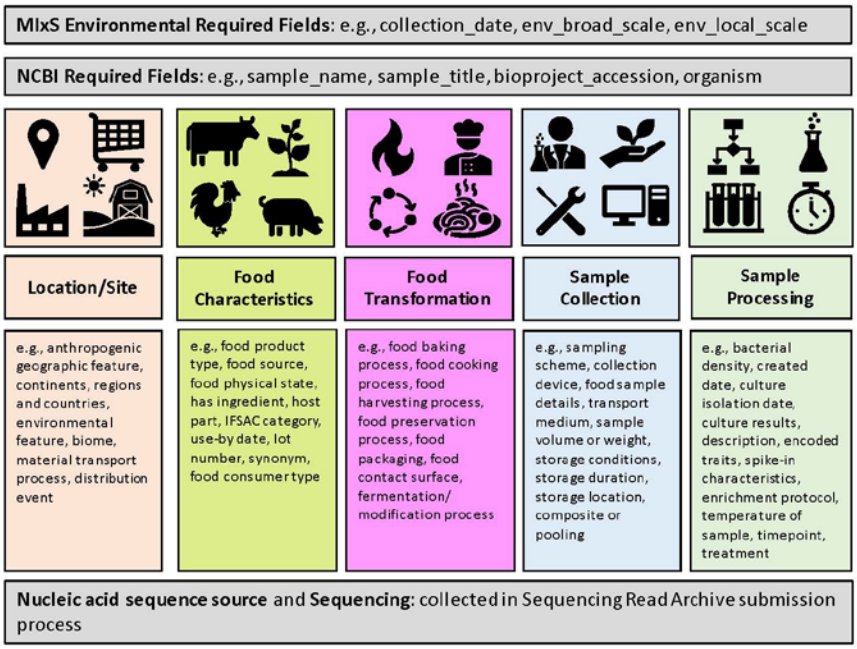

Figure 1. Categorization of contextual data contained in MIxS food metadata standard.

Categorizing the Metadata Standard

Using the recently described agricultural metadata standard [2] as a template, we organized the contextual information from our various use cases, or sub-packages, into several categories (Fig. 1). This was done to aid in both development of the various sub-packages, as well as to

benefit the submitter, by providing a logical organization of metadata. There are two categories of required metadata, that which is required by MIxS, for all environmental MIMS (Minimum Information about a Metagenome Sequence) submissions, and that which is required by NCBI and captured in the biosample and bioproject submission process. Further, the metadata standard includes nucleic acid sequence source and sequencing metadata variables, in compliance with MIxS standards, but this information is collected when the raw metagenomic sequencing data is submitted to the Sequence Read Archive (NCBI). The remaining metadata variables were binned into 5 categories, location, food characteristics, food transformation, sample collection, and sample processing. Some examples from each category are presented in Fig. 1. Required and core elements, those that are present in all use-cases, will be "mandatory" minimal metadata requirements. Those attributes that are use-case specific will be defined as "conditional mandatory" or "optional" depending on the use-case and the rationale to include the specific field.

Ontology

The accumulation of more and more data from environmental studies, especially those that encompass metagenomics and microbiome characterization, is driving research scientists to turn to data science analytical solutions. The use of controlled vocabularies and ontologies can improve the description, effective extraction and analysis of valuable information from within and across these complex environmental studies. Much in the same way that our binomial classification of living things provides a hierarchical view of all life, ontologies provide a shared language as well as define the relationship of all things in an environment. To advance this objective, we have begun the construction of a minimal metadata standard for food metagenomes that is comprehensively supported by existing ontological frameworks. For each use-case defined above, essential attributes were queried against the 251 ontologies contained in the Ontology Lookup Service [3, 4], provided by EMBI-EBI (https://www.ebi.ac.uk/ols/index) to identify candidate ontological terms. For example, 74% (72/97) of minimal metadata terms for the finished and unfinished food foundational use case are supported by ontologies. Not surprisingly, most of these terms originated from the FoodOn (https://foodon.org/) ontology [5]. Additionally, we pulled several categorical variables from the GenEpiO application ontology (https://genepio.org/), as our application is the food safety sector, and we also reused several terms from existing MIxS environmental packages. Attributes not defined in an existing ontology will be submitted for inclusion into FoodOn or other appropriate ontology.

Advantages

Ontologies and the use of controlled vocabularies continue to increase in usage in the biological sciences, building on the widespread usage of the Gene Ontology [6], and establishment of the Open Biomedical Ontologies (OBO) [7]. Their use is of extreme importance to generate machine-readable and searchable datasets, vital to the next decade of AI and machine empowered epidemiology and surveillance, to ensure ease of comparison across different studies and from different research groups. Here we describe the development of a minimal metadata requirement package, based on hierarchical ontologies, for food-related metagenomic research. This package aims to encompass a wide range of foods-based research such as outbreak or root-

cause traceback in a food production facility, or to satisfy a consumer complaint involving pet food, evaluating farm management practices that promote healthy soil microbiomes, or assessing perturbation of fermentation conditions on sensory analysis and nutritional content, just to name a few. Adoption of this metadata standard will greatly improve the utility and interoperability of food metagenomic research and significantly advance analytical capability across studies, with the goal to advance public health.

References

1. Hoehndorf, R., P.N. Schofield, and G.V. Gkoutos, *The role of ontologies in biological and biomedical research: a functional perspective*. Brief Bioinform, 2015. **16**(6): p. 1069-80.
2. Dundore-Arias, J.P., et al., *Community-Driven Metadata Standards for Agricultural Microbiome Research*. Phytobiomes Journal, 2020. **4**(2): p. 115-121.
3. Cote, R.G., et al., *The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries*. BMC Bioinformatics, 2006. **7**: p. 97.
4. Cote, R., et al., *The Ontology Lookup Service: bigger and better.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W155-60.
5. Dooley, D.M., et al., *FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration.* NPJ Sci Food, 2018. **2**: p. 23.
6. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
7. Smith, B., et al., *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nat Biotechnol, 2007. **25**(11): p. 1251-5.