Pacific Northwest
NATIONAL LABORATORY

PNNL-34342

Microbiome Metadata Management

May 2025

Beata Meluch

U.S. DEPARTMENT of ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-34342

Microbiome Metadata Management

May 2025

Beata Meluch

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Microbiome Metadata Management

Bea Meluch

1. Abstract

Tremendous amounts of microbiome omics data have been generated in recent years, which holds great potential for large-scale studies and metaanalyses. The utility of these datasets is unfortunately limited by their lack of searchability and interoperability. Thorough, standardized metadata is essential for unlocking the scientific discovery potential of disparate datasets. Several national and international initiatives exist to gather and standardize microbiome omics data to improve accessibility and reusability. Continuing to improve consistency in metadata will allow advances in automation and reduction in manual mapping between databases.

2. Introduction

As sequencing becomes cheaper, and proteomics and metabolomics methods improve, biological discovery is now limited by analysis rather than instrumentation. For example, the sequence of a given *E. coli* bacterium is now easy to obtain and not likely to be radically different from already sequenced bacteria. The source of the sample is therefore more meaningful for downstream analysis. Was it isolated from a healthy or sick person? Is there a current disease outbreak? (Field et al., 2011) This vital contextual information is known as *metadata.*

Metadata

In research science, metadata are data that describe the context of experimental data (L. Thompson et al., 2020). This can include sample information, preparation conditions, and analysis methods used, among many other options. Importantly, metadata also describe relationships between samples, which is what makes hypothesis testing possible (L. Thompson et al., 2020). A sample in a controlled experiment carries its experimental group and group conditions in its metadata.

Metadata enable searching in databases – for example, one could search for all samples from patients on a certain medication, or for all soil samples taken in Colorado. Detailed metadata also allows for reuse of experimental results in later studies, since the sample source and treatments are known. However, finding and comparing datasets is difficult when researchers record metadata idiosyncratically, as shown in Figure 1a. If one sample source is recorded as “urban” and the other as “public park”, are those samples comparable? Are they in the same group? This is where it becomes important to use controlled vocabularies (such as ontologies) in standardized metadata.

Standards

Standards, as they apply to metadata, are a set of instructions that specify what information needs to be included to describe a certain type of sample and experiment. The content of a standard will necessarily depend on the area of study in which it is used. Metadata standards are often created by a community recognizing a need to make their data more integrable.

One example of a community standards organization is the Genomic Standards Consortium (GSC), which is best known for creating and maintaining the Minimum Information about any x Sequence (MIxS) standard. The organization grew out of a conference in 2005 and membership remains open

worldwide (Yilmaz et al., 2011). Continuing development on GSC standards is predominantly done by volunteer scientists. Because the standards are developed through community engagement, they remain relevant to the scientists who will be using them. This encourages wider adoption and therefore improved interoperability (GSC, 2023).

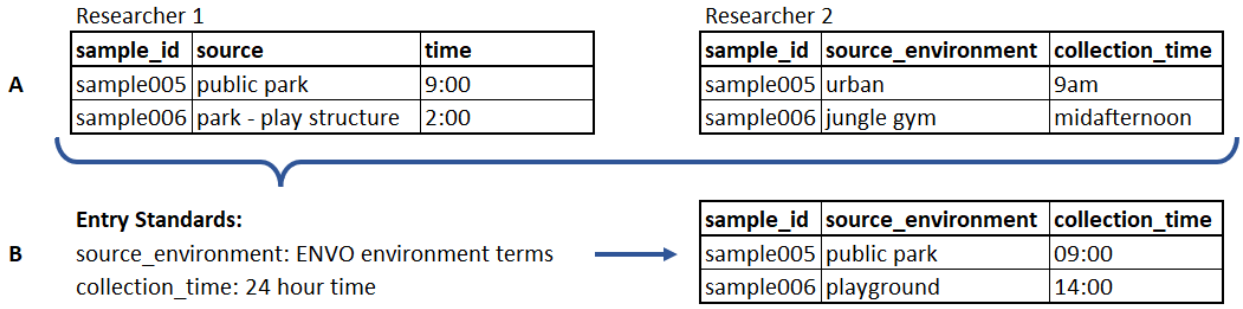

Figure 1. Metadata standardization improves readability and consistency in research. (A) Individual researchers in the same field may record sample information in ways that are intuitive to them but are not universal. This results in metadata that should be identical instead reading differently to both humans and machines. (B) Using standards to specify required fields and terms makes metadata consistent and searchable, which allows for easier reuse and better study design. Figure inspired by slides presented by Montana Smith and Lee Ann McCue at the Global Soil Biodiversity Conference 2023 (Smith, 2023).

Ontologies

Dictating the information needed for a given sample (often called “fields”, as in fields of a form) is only part of what makes a standard truly robust. Even if a metadata standard requires the field “source_environment”, in the absence of more specification, different users could reasonably enter “urban” or “public park” about the same sample as shown in Figure 1a.

One way to formalize possible entries into a field is to tie the field to an ontology. An ontology is a systematic description of the relationships between concepts in a particular area of knowledge (L. Thompson et al., 2020). An ontology uses a controlled vocabulary to describe its member objects, which is particularly useful in standardizing scientific descriptions. In the example above, if the standard specified that “source_environment” must use terms from the Environment Ontology (ENVO), both researchers would follow “geographic feature → anthropogenic geographic feature → park → public park” (Buttigieg et al., 2016). This allows all of the categorical information to be captured in a standardized way, as shown in Figure 1b.

Standards Development

When trying to record complete metadata for an experiment in a rapidly evolving scientific discipline, existing ontologies and standards may not cover all aspects of the new questions being explored. It can be tempting in this circumstance to proceed according to Figure 2 and establish a new customized standard that covers all of the old use-cases as well as the new one. This may solve the immediate problem but is challenging for the field as a whole. Competing metadata standards decrease interoperability between databases and datasets.

Fortunately, tools like OBO Foundry exist to help users find existing ontologies that fit their needs (Jackson et al., 2021). Experimental and environmental circumstances will still arise that are not

covered by existing ontologies or standards. In these cases, it is best to find a standard that is close to what is needed and use it as a foundation. Layering new information onto established standards helps the metadata retain some commonality with other datasets that use the shared source ontology.

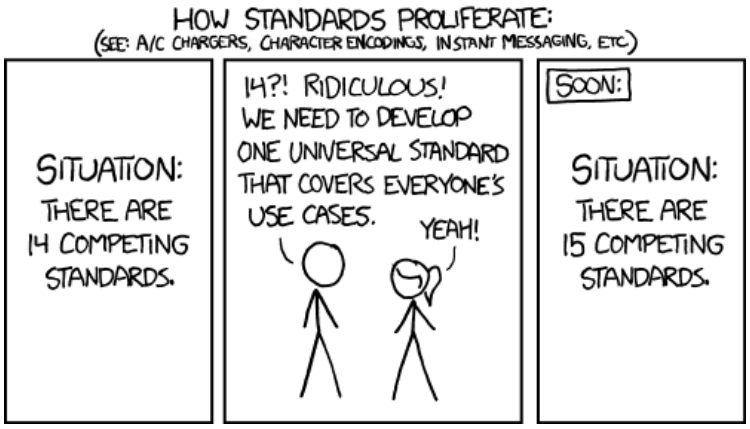

Figure 2. Standards are useful but never perfect. Standards need continuous maintenance to remain relevant and scientists instead often opt to develop an entirely new standard. To maintain interoperability, it is generally better to build upon established standards than to add a new competing standard into the mix. Figure reprinted from the webcomic xkcd (Munroe, 2011).

3. Motivation for Meaningful Metadata

The findability and interoperability of experimental data is therefore dependent on standardized, thorough metadata. At a deeper level, what is the inherent value of those characteristics? Why are researchers devoting so many resources to metadata management? The ultimate goal is to enable continuing discovery through the reuse of data.

Data reuse also lowers costs and increases the value of research funding. Funding agencies are becoming more vocal about data stewardship, often requiring data management plans as part of study design (Wilkinson et al., 2016). They hope to make and keep data more accessible through ongoing stewardship, allowing reuse into the future. Reuse of data means that the original award is now funding multiple studies without the cost of an entirely new experiment, and funders and researchers alike are of course interested in getting more for their money.

Furthermore, with sufficiently robust metadata, dataset reuse need not be limited to experiments within the same field. A 2022 paper reused data from a 2019 sleep study for a new investigation of food metabolomics (Gauglitz et al., 2022). The sleep study required participants to be on a restricted diet. Because the patient and diet-related metadata were thoroughly recorded, Gauglitz et. al. were able to use the dataset three years later to demonstrate a novel metabolomics annotation method. The time and labor costs of running a new study were saved and researchers were directly able to start a new analysis.

FAIR Data Principles

> "Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process." (Wilkinson et al., 2016)

The 2016 paper "The FAIR Guiding Principles for scientific data management and stewardship" by Wilkinson et al. eloquently lays out both the value of good data stewardship and the steps to achieve it. They break down the important considerations of data management into four principles: data should be Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016). These principles are heavily dependent on thorough, standardized metadata. The FAIR principles were developed with a focus on machine readability for future data mining and large-scale analyses but keep human usability in mind. The international community effort GO FAIR is now at the forefront of advocating for the adoption of FAIR principles across disciplines (Wilkinson et al., 2019).

Metadata in Microbiome Research

The exact focus of metadata standards varies by subject area, though the overall goal of reusability remains the same. Microbiome research has a particular focus on environmental metadata. Microbiomes are heavily influenced by hyperlocal environmental factors, so microbiome samples need to be accompanied by highly detailed sample collection metadata (Vangay et al., 2021).

Microbiome research is hindered by data compartmentalization and the inability to perform metaanalyses (Kyrpides et al., 2016). Huge amounts of sequencing data are technically publicly accessible, but are difficult to search or mine for further information. Microbiome scientists are pushing for FAIR data to stop this underutilization of resources (Kyrpides et al., 2016).

The pursuit of FAIR microbiome data is not limited to a niche of interested researchers. In 2016, the White House Office of Science and Technology Policy announced the National Microbiome Initiative to advance microbiome science (Handelsman, 2016). The second goal of the NMI, "Developing platform technologies", acknowledges the need for improved data sharing and accessibility. As mentioned above, government funding agencies are increasingly aware that the publicly funded research ecosystem could be getting more "bang for their buck" through better data integration.

4. Current Efforts in Microbiome Data Coordination

As microbiome research expands and accelerates, the scientific community is racing to keep up with FAIR data management practices. Domain-focused networks and organizations have arisen to facilitate collaboration in research and data management. Many organizations are focused on the United States due to funding sources, but similar initiatives exist internationally, such as the Australian Microbiome Initiative and the Brazilian Microbiome Project (*Australian Microbiome*, 2018; *Brazilian Microbiome Project*, 2018). These microbiome research organizations vary in mission, scope, and capabilities. The following are examples of large initiatives with different approaches to coordinating data from diverse sources.

Microbiome Centers Consortium

The Microbiome Centers Consortium (MCC) is a US-based network of research centers that focus on microbiomes, both host-associated and environmental (*Microbiome Centers Consortium*, 2023). The individual centers were generally founded to facilitate cross-disciplinary research as the field expands (Martiny et al., 2020). One goal of the MCC is to go a step further in avoiding siloing of knowledge by promoting resource sharing (*Microbiome Centers Consortium*, 2023). Additionally,

they aim to “share best practices…, help reduce redundancy…and become a communication hub” for advocacy (Martiny et al., 2020).

The MCC is not producing nor hosting research data. Rather, it works to improve communication between those already studying microbiomes. The MCC hosts annual meetings, which are a chance for researchers to work together on establishing common standards and best practices (*Microbiome Centers Consortium*, 2023). The MCC also promotes relevant educational events and job opportunities at other institutions. Recently, they have partnered with the American Society of Microbiology in advocating for the Office of Science and Technology to dedicate leadership and resources to microbiome research coordination (*Microbiome Centers Consortium*, 2023).

Earth Microbiome Project

In contrast, the Earth Microbiome Project (EMP) is a worldwide effort to understand microbial communities directly by analyzing thousands of samples according to standardized methods. The EMP network is composed of over 500 investigators who volunteered their samples, skills, and resources (*Earth Microbiome Project*, 2023). Their focus was on using metagenomics to describe and compare microbial diversity across environments, as published in their 2017 paper (L. R. Thompson et al., 2017).

The EMP is not currently accepting new samples, but they are still committing updates to their codebase and documentation. They are currently working on combining amplicon sequencing, metagenomics, and metabolomics into a multi-omic analysis (*Earth Microbiome Project*, 2012/2023). In the meantime, their original datasets are publicly available for download and data exploration through UCSD, although the user interface is challenging to navigate (*EMP Qiita*, n.d.).

Two existing standards were used for the EMP metadata: MIxS and ENVO. The EMP also layered on their own ontology (Earth Microbiome Project Ontology, EMPO) as they felt it was important to capture additional information about microbial environments (L. R. Thompson et al., 2017). The website and publications regarding the EMP do not explicitly mention FAIR data standards. However, by creating a publicly accessible database that adheres to consistent metadata standards, they have arguably still created a body of work that follows the FAIR principles.

National Microbiome Data Collaborative

The National Microbiome Data Collaborative (NMDC) is a Department of Energy-funded pilot initiative which seeks to “democratize microbiome data science” (Eloe-Fadrosh et al., 2022). The NMDC is developing three user-facing products: a data portal, where users can access curated microbiome study data; a submission portal, where users can submit their studies for review and addition to the data portal; and a set of microbiome omics workflows. Although the collaborative is US-based, studies including internationally gathered samples are included, so users can find microbiome data from around the globe (*NMDC*, 2019).

The data portal and submission portal organize study data according to the NMDC schema, a framework which defines the metadata for included samples and datasets. The schema describes the relationships between metadata elements. The elements themselves are primarily derived from the Environment Ontology (ENVO) and the GSC’s MIxS standard (Eloe-Fadrosh et al., 2022). By following standards common in the field, the NMDC keeps data easily searchable and interoperable.

The NMDC is developed with FAIR principles in mind, and is actively engaged with the GO FAIR community in advancing FAIR data in the microbiome field (Wood-Charlson et al., 2020).

The NMDC is one outcome of the OSTP National Microbiome Initiative mentioned above (Handelsman, 2016; *NMDC*, 2019). The NMDC team is composed primarily of staff from Lawrence Berkeley, Pacific Northwest, and Los Alamos National Laboratories (Wood-Charlson et al., 2020). The collaborative is initially focusing on data generated at the Joint Genome Institute (JGI) and the Environmental Molecular Sciences Laboratory (EMSL). These facilities are prioritized for three reasons: first, user data becomes publicly available after a certain interval; second, JGI and EMSL are associated with the national laboratories running the NMDC; and third, several types of omics data are covered between the two institutions (Wood-Charlson et al., 2020).

NEON

The National Ecological Observatory Network (NEON) is an example of an institution with a different mission resulting in functional overlap. NEON's mission is to study ecosystem change in the United States through ecological monitoring (Dalton, 2000). NEON is a distributed observation facility with 81 current sites that regularly measure a variety of environmental factors, which include water and soil microbiome sequencing (*NEON*, 2019). Data from all NEON sites is released into the public domain and is easily accessible through their data portal.

NEON explicitly provides data in the interest of following the FAIR data principles (*NEON*, 2019). When a NEON data package is downloaded, files containing sample metadata and metadata term definitions are bundled in with the download. NEON's metadata includes many custom fields as well as some terms taken from the Darwin Core, the Global Biodiversity Information Facility, the VegCore dictionary, and MIxS (*NEON*, 2019). The metadata files are provided in Ecological Metadata Language (EML) format, which is based on XML.

5. Metadata Harmonization in Practice

The initiatives listed above that do their own testing are able to define their data and metadata reporting to be internally consistent. Collective data efforts like the NMDC need to accommodate the metadata systems of a variety of data sources to present data in a findable and accessible manner. As an example, Figure 3 illustrates what needs to be done to harmonize metadata for adding NEON projects to the NMDC data portal.

The NMDC is currently working to incorporate JGI and EMSL studies that used samples from NEON sites. To upload the study information to the data portal, the NMDC schema must be able to accommodate the metadata provided with NEON samples. This means that the NMDC schema must be modified such that NMDC fields recognize their corresponding NEON terms, and if there is no match, the extra NEON terms must be added to the schema. Figure 3 shows a sampling of ways in which fields can connect.

Fields in one database may map to multiple fields in the other, as shown with the incubation information in Figure 3. Additionally, information can be grouped differently – where NEON breaks down chemical testing into separate tables, many aspects of a sample are grouped together in

the NMDC "Biosample" class. In the example, the carbon-nitrogen ratio fields are the only ones taken directly from the MIxS standard with no modification, reinforcing that sharing a common base standard improves interoperability.

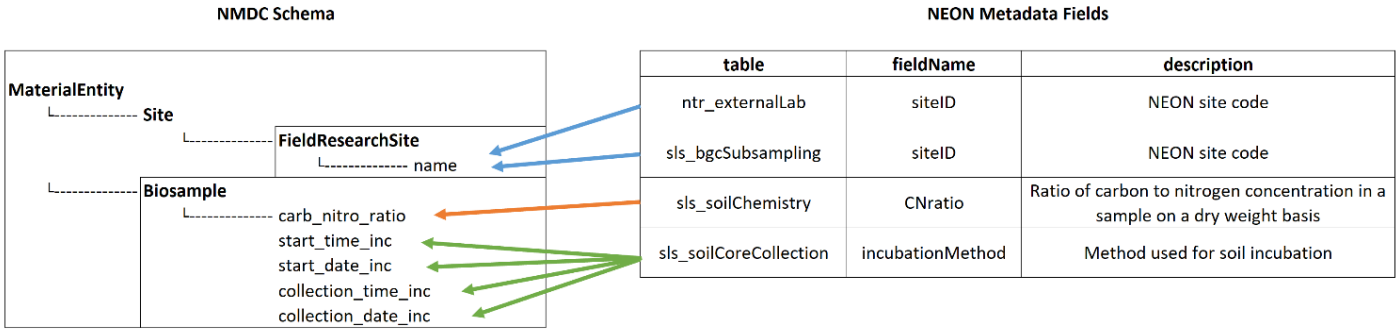

Figure 3. Mapping can be complex when similar concepts are stored differently across databases. In NEON's databases, the site code "siteID" is duplicated across data tables, whereas in the NMDC schema, the site code would be stored once as the attribute "name" of an object of class "FieldResearchSite". In contrast, the incubation method for a soil sample is saved in NEON as one field, whereas incubation details must be saved separately in the NMDC schema. The carbon-nitrogen ratio of a sample maps directly from the NEON field "CNratio" in the table "sls_soilChemistry" to the "carb_nitro_ratio" attribute of the NMDC class "Biosample" (*NMDC Schema*, 2023; NEON, 2023).

6. Conclusion

Metadata management may not be the flashiest topic, but it is increasingly important as biological data becomes easier and easier to generate. Data stewardship practices are scrambling to keep up with the production of microbiome omics data in particular. Computer automation promises additional capabilities for data mining and pattern recognition – if metadata is standardized to be machine-readable. Microbiome researchers have an opportunity to improve data stewardship by engaging with collaborations and networks and adhering to FAIR principles in their research.

Reusable data serves the best interests of the research community as well as funding organizations. Entities such as the DOE, NSF, and NIH have spent billions of dollars funding projects that may produce data that is used for only a single publication. With FAIR metadata, datasets can be discovered and reused by future researchers, which gives funding agencies more for their money. It also allows researchers to perform metaanalyses, assemble large comparative studies, and test additional hypotheses while saving them the time and effort of duplicate data generation. To fulfill the potential granted to scientists by continuing improvements in instrumentation, metadata management must become as much of a widespread scientific habit as recording experiments in a lab notebook.

7. Citations

Australian Microbiome. (2018). Bioplatforms. https://doi.org/10.25953/v12e-zq81

Brazilian Microbiome Project. (2018). Brazilian Microbiome. https://www.brmicrobiome.org

Buttigieg, P. L., Pafilis, E., Lewis, S. E., Schildhauer, M. P., Walls, R. L., & Mungall, C. J. (2016). The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics*, *7*(1), 57. https://doi.org/10.1186/s13326-016-0097-6

Dalton, R. (2000). NEON to shed light on environment research. *Nature*, *404*(6775), Article 6775. https://doi.org/10.1038/35005244

Earth Microbiome Project. (2023). [Jupyter Notebook]. biocore. https://github.com/biocore/emp (Original work published 2012)

Earth Microbiome Project. (2023). https://earthmicrobiome.org/

Eloe-Fadrosh, E. A., Ahmed, F., Anubhav, Babinski, M., Baumes, J., Borkum, M., Bramer, L., Canon, S., Christianson, D. S., Corilo, Y. E., Davenport, K. W., Davis, B., Drake, M., Duncan, W. D., Flynn, M. C., Hays, D., Hu, B., Huntemann, M., Kelliher, J., … Fagnan, K. (2022). The National Microbiome Data Collaborative Data Portal: An integrated multi-omics microbiome data resource. *Nucleic Acids Research*, *50*(D1), D828–D836. https://doi.org/10.1093/nar/gkab990

Field, D., Amaral-Zettler, L., Cochrane, G., Cole, J. R., Dawyndt, P., Garrity, G. M., Gilbert, J., Glöckner, F. O., Hirschman, L., Karsch-Mizrachi, I., Klenk, H.-P., Knight, R., Kottmann, R., Kyrpides, N., Meyer, F., Gil, I. S., Sansone, S.-A., Schriml, L. M., Sterk, P., … Wooley, J. (2011). The Genomic Standards Consortium. *PLOS Biology*, *9*(6), e1001088. https://doi.org/10.1371/journal.pbio.1001088

Gauglitz, J. M., West, K. A., Bittremieux, W., Williams, C. L., Weldon, K. C., Panitchpakdi, M., Di Ottavio, F., Aceves, C. M., Brown, E., Sikora, N. C., Jarmusch, A. K., Martino, C., Tripathi, A., Meehan, M. J., Dorrestein, K., Shaffer, J. P., Coras, R., Vargas, F., Goldasich, L. D., … Dorrestein, P. C. (2022). Enhancing untargeted metabolomics using metadata-based source annotation. *Nature Biotechnology*, *40*(12), Article 12. https://doi.org/10.1038/s41587-022-01368-1

GSC. (2023). *Genomic Standards Consortium*. Genomic Standards Consortium. https://genomicsstandardsconsortium.github.io/gensc.github.io//

Handelsman, J. (2016, May 13). Announcing the National Microbiome Initiative. *Whitehouse.Gov*. https://obamawhitehouse.archives.gov/blog/2016/05/13/announcing-national-microbiome-initiative

Home | NSF NEON | Open Data to Understand our Ecosystems. (2019). https://www.neonscience.org/

Home—NMDC Schema Documentation. (2023). https://microbiomedata.github.io/nmdc-schema/home/

Jackson, R., Matentzoglu, N., Overton, J. A., Vita, R., Balhoff, J. P., Buttigieg, P. L., Carbon, S., Courtot, M., Diehl, A. D., Dooley, D. M., Duncan, W. D., Harris, N. L., Haendel, M. A., Lewis, S. E., Natale, D. A., Osumi-Sutherland, D., Ruttenberg, A., Schriml, L. M., Smith, B., … Peters, B. (2021). OBO Foundry in 2021: Operationalizing open data principles to evaluate ontologies. *Database*, *2021*, baab069. https://doi.org/10.1093/database/baab069

Kyrpides, N. C., Eloe-Fadrosh, E. A., & Ivanova, N. N. (2016). Microbiome Data Science: Understanding Our Microbial Planet. *Trends in Microbiology*, *24*(6), 425–427. https://doi.org/10.1016/j.tim.2016.02.011

Martiny, J. B. H., Whiteson, K. L., Bohannan, B. J. M., David, L. A., Hynson, N. A., McFall-Ngai, M., Rawls, J. F., Schmidt, T. M., Abdo, Z., Blaser, M. J., Bordenstein, S., Bréchot, C., Bull, C. T., Dorrestein, P., Eisen, J. A., Garcia-Pichel, F., Gilbert, J., Hofmockel, K. S., Holtz, M. L., … Sachs, J. L. (2020). The emergence of microbiome centres. *Nature Microbiology*, *5*(1), Article 1. https://doi.org/10.1038/s41564-019-0644-x

Microbiome Centers Consortium. (2023, March). https://microbiomecenters.org/

Munroe, R. (2011, July 20). *Standards*. xkcd. https://xkcd.com/927/

NEON. (2023). *Soil physical and chemical properties, periodic (DP1.10086.001)* [Data set]. https://doi.org/10.48443/0phb-j505

NMDC. (2019). National Microbiome Data Collaborative. https://microbiomedata.org/

Qiita portal for the Earth Microbiome Project (EMP). (n.d.). Retrieved May 26, 2023, from https://qiita.ucsd.edu/emp/

Smith, M. (2023, March). METADATA STANDARDS AND DATA SUBMISSION TO THE NATIONAL MICROBIOME DATA COLLABORATIVE. *The 3rd Global Soil Biodiversity Conference*. Global Soil Biodiversity Conference, Dublin, IE. https://gsb2023.org/workshops/

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., Prill, R. J., Tripathi, A., Gibbons, S. M., Ackermann, G., Navas-Molina, J. A., Janssen, S., Kopylova, E., Vázquez-Baeza, Y., González, A., Morton, J. T., Mirarab, S., Zech Xu, Z., Jiang, L., … Knight, R. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, *551*(7681), Article 7681. https://doi.org/10.1038/nature24621

Thompson, L., Vangay, P., Blumberg, K., Christianson, D. S., Dundore-Arias, J. P., Hu, B., Timme, R., & Wood-Charlson, E. M. (2020). Introduction to Metadata and Ontologies. *National Microbiome Data Collaborative*. doi.org/10.25979/1607365

Vangay, P., Burgin, J., Johnston, A., Beck, K. L., Berrios, D. C., Blumberg, K., Canon, S., Chain, P., Chandonia, J.-M., Christianson, D., Costes, S. V., Damerow, J., Duncan, W. D., Dundore-Arias, J. P., Fagnan, K., Galazka, J. M., Gibbons, S. M., Hays, D., Hervey, J., …

Eloe-Fadrosh, E. A. (2021). Microbiome Metadata Standards: Report of the National Microbiome Data Collaborative's Workshop and Follow-On Activities. *MSystems*, *6*(1), e01194-20. https://doi.org/10.1128/mSystems.01194-20

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), Article 1. https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Dumontier, M., Jan Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2019). Addendum: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *6*(1), Article 1. https://doi.org/10.1038/s41597-019-0009-6

Wood-Charlson, E. M., Anubhav, Auberry, D., Blanco, H., Borkum, M. I., Corilo, Y. E., Davenport, K. W., Deshpande, S., Devarakonda, R., Drake, M., Duncan, W. D., Flynn, M. C., Hays, D., Hu, B., Huntemann, M., Li, P.-E., Lipton, M., Lo, C.-C., Millard, D., … Eloe-Fadrosh, E. A. (2020). The National Microbiome Data Collaborative: Enabling microbiome science. *Nature Reviews Microbiology*, *18*(6), Article 6. https://doi.org/10.1038/s41579-020-0377-0

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., … Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, *29*(5), Article 5. https://doi.org/10.1038/nbt.1823

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov