Title

Microbiome Metadata Standards: Report of the National Microbiome Data Collaborative's Workshop and Follow-On Activities

Journal

mSystems, 6(1)

ISSN

2379-5077

Authors

Vangay, Pajau

Burgin, Josephine

Johnston, Anjanette

et al.

Publication Date

2021-02-23

DOI

10.1128/msystems.01194-20

Peer reviewed

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

PERSPECTIVE
Applied and Environmental Science

Microbiome Metadata Standards: Report of the National Microbiome Data Collaborative's Workshop and Follow-On Activities

Pajau Vangay,[a] Josephine Burgin,[b] Anjanette Johnston,[c] Kristen L. Beck,[d] Daniel C. Berrios,[e] Kai Blumberg,[f] Shane Canon,[a] Patrick Chain,[g] John-Marc Chandonia,[a] Danielle Christianson,[a] Sylvain V. Costes,[e] Joan Damerow,[a] William D. Duncan,[a] Jose Pablo Dundore-Arias,[h] Kjiersten Fagnan,[a] Jonathan M. Galazka,[e] Sean M. Gibbons,[i,j] David Hays,[a] Judson Hervey,[k] Bin Hu,[g] Bonnie L. Hurwitz,[f] Pankaj Jaiswal,[l] Marcin P. Joachimiak,[a] Linda Kinkel,[m] Joshua Ladau,[a] Stanton L. Martin,[n] Lee Ann McCue,[o] Kayd Miller,[a] Nigel Mouncey,[a] Chris Mungall,[a] Evangelos Pafilis,[p] T. B. K. Reddy,[a] Lorna Richardson,[b] Simon Roux,[q] Lynn M. Schriml,[w] Justin P. Shaffer,[r] Jagadish Chandrabose Sundaramurthi,[a] Luke R. Thompson,[s,t] Ruth E. Timme,[u] Jie Zheng,[v] Elisha M. Wood-Charlson,[a] Emiley A. Eloe-Fadrosh[a]

[a]Lawrence Berkeley National Laboratory, Berkeley, California, USA
[b]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
[c]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
[d]IBM Almaden Research Center, San Jose, California, USA
[e]NASA Ames Research Center, Moffett Field, California, USA
[f]Biosystems Engineering Department, University of Arizona, Tucson, Arizona, USA
[g]Los Alamos National Laboratory, Los Alamos, New Mexico, USA
[h]California State University, Monterey Bay, California, USA
[i]Institute for Systems Biology, Seattle, Washington, USA
[j]Department of Bioengineering, University of Washington, Seattle, Washington, USA
[k]Center for Bio/Molecular Science & Engineering, Naval Research Laboratory, Washington, DC, USA
[l]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA
[m]Department of Plant Pathology, University of Minnesota, St. Paul, Minnesota, USA
[n]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[o]Pacific Northwest National Laboratory, Richland, Washington, USA
[p]Institute of Marine Biology Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Crete, Greece
[q]Department of Energy Joint Genome Institute, Berkeley, California, USA
[r]Department of Pediatrics, School of Medicine, University of California, San Diego, California, USA
[s]Northern Gulf Institute, Mississippi State University, Mississippi State, Mississippi, USA
[t]Ocean Chemistry and Ecosystems Division, Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, Miami, Florida, USA
[u]US Food and Drug Administration, Center for Food Safety and Applied Nutrition, College Park, Maryland, USA
[v]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[w]University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, Maryland, USA

ABSTRACT Microbiome samples are inherently defined by the environment in which they are found. Therefore, data that provide context and enable interpretation of measurements produced from biological samples, often referred to as metadata, are critical. Important contributions have been made in the development of community-driven metadata standards; however, these standards have not been uniformly embraced by the microbiome research community. To understand how these standards are being adopted, or the barriers to adoption, across research domains, institutions, and funding agencies, the National Microbiome Data Collaborative (NMDC) hosted a workshop in October 2019. This report provides a summary of discussions that took place throughout the workshop, as well as outcomes of the working groups initiated at the workshop.

Citation Vangay P, Burgin J, Johnston A, Beck KL, Berrios DC, Blumberg K, Canon S, Chain P, Chandonia J-M, Christianson D, Costes SV, Damerow J, Duncan WD, Dundore-Arias JP, Fagnan K, Galazka JM, Gibbons SM, Hays D, Hervey J, Hu B, Hurwitz BL, Jaiswal P, Joachimiak MP, Kinkel L, Ladau J, Martin SL, McCue LA, Miller K, Mouncey N, Mungall C, Pafilis E, Reddy TBK, Richardson L, Roux S, Schriml LM, Shaffer JP, Sundaramurthi JC, Thompson LR, Timme RE, Zheng J, Wood-Charlson EM, Eloe-Fadrosh EA. 2021. Microbiome metadata standards: report of the National Microbiome Data Collaborative's workshop and follow-on activities. mSystems 6:e01194-20. https://doi.org/10.1128/mSystems.01194-20.

Editor Vanni Bucci, University of Massachusetts Medical School

Address correspondence to Emiley A. Eloe-Fadrosh, eaeloefadrosh@lbl.gov.

Published 23 February 2021

[This article was published on 23 February 2021 with a byline that lacked Lynn M. Schriml. The byline was updated in the current version, posted on 30 March 2021.]

KEYWORDS data standards, metadata, microbiome, ontology

The National Microbiome Data Collaborative (NMDC) is a pilot initiative that was launched in July 2019 and is funded by the Department of Energy (DOE) Office of Science, Biological and Environmental Research Program, to support microbiome data exploration and discovery through a collaborative, integrative data science ecosystem (1). The NMDC team is building an open-source, integrated data science ecosystem that leverages existing data standards, data resources, and infrastructure in the microbiome research space. The NMDC initiative embraces the FAIR (findable, accessible, interoperable, and reusable) data principles (2) by incorporating community-driven data standards and quality measures to enable data integration and access in its science gateway. Understanding the current landscape of data standards for the microbiome research community is an important first step toward achieving the aims of the NMDC pilot initiative.

Information that contextualizes samples, including sample collection, sample preparation, data processing methods, and data products (3) (Fig. 1), also known as "metadata," is essential for the interpretation of measurements produced from a biological sample. Standardized metadata using common terms, such as from an ontology (a controlled vocabulary with logic linking between its terms), are essential for data sharing, synthesis, and reuse, and can enable the discovery of new insights (4). The Genomic Standards Consortium (GSC) (5) and the Open Biological and Biomedical Ontologies (OBO) Foundry (6) have made important contributions to the development of community-driven sample metadata standards. Yet, it is unclear how much of the microbiome research community are applying metadata standards, or whether there remain barriers to adoption.

To understand how data standards support microbiome science across research domains, institutions, and funding agencies, the NMDC team hosted 50 experts in microbiome research, data standards, genome annotation, bioinformatics, and community engagement for a 4-day workshop in October 2019 at the Lawrence Berkeley National Laboratory (https://microbiomedata.org/nmdc-ontology-workshop/). The workshop goals were to review how standards are currently used, explore approaches for improving community adoption of and compliance with standards, build consensus around the importance of metadata, and establish a network of key stakeholders to advocate for standards across their organizations and communities.

The main sessions of the workshop included (i) perspectives from repositories, infrastructure projects, metadata resources, and standards organizations (https://microbiomedata.org/nmdc-ontology-workshop/); (ii) group discussions on best practices, remaining challenges, and paths forward; and (iii) the initiation of working groups to evaluate current standards and their adoption, enhance existing standards, and identify training needs. Here, we summarize the workshop discussions on addressing barriers in microbiome data standards, and share outcomes from several working groups formed at the workshop.

ADDRESSING BARRIERS IN MICROBIOME DATA STANDARDS

Throughout the workshop discussions, two cross-cutting areas for improvement related to microbiome data and standards emerged: (i) encourage a culture that shares microbiome data, and (ii) understand and reduce barriers to (meta)data submission. We present a summary of the workshop discussions in the context of these two key themes.

Encourage a culture that shares microbiome data. Success in science is often measured by high-impact publications (7), creating pressure to be the first to make important discoveries and receive credit for the published contribution. Waiting until findings are published before making data available to others is not uncommon and remains a significant barrier to the provision of data to the broader community (8, 9). Even post publication, data sharing continues to be challenging due to a noted lack of

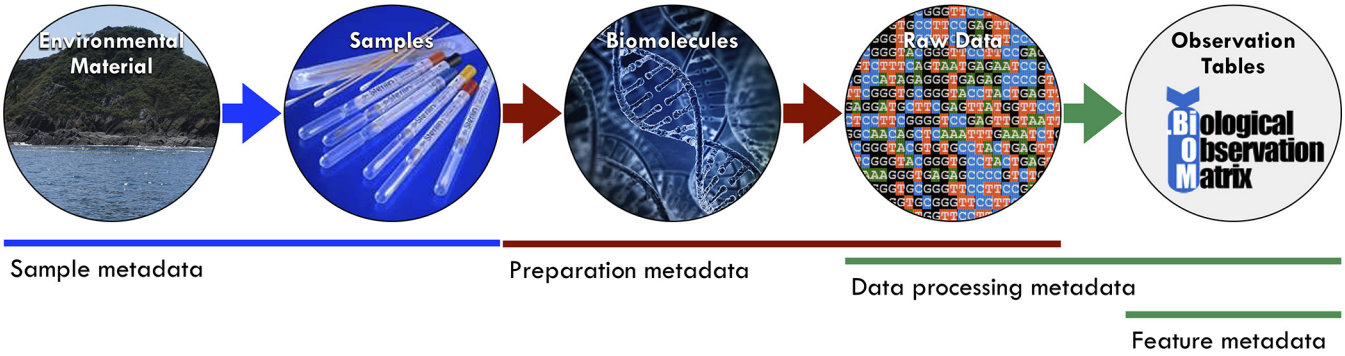

FIG 1 Examples of different types of metadata along the workflow from environmental samples to data and analysis tables. Submitting data to central repositories typically requires sample and preparation metadata. Sample metadata include information about when, where, and what sample was collected; preparation metadata describe how the sample was processed and turned into data products; data processing and feature metadata are generated by the repository or analysis software. Refer to Text S1 in the supplemental material for additional information.

time to prepare data for sharing and reuse, legal or privacy constraints, and concerns about misinterpretation or misuse of data (8, 10). As a result, researchers often cannot find data (11), or spend up to 50 to 80% of their time wrangling data into a more usable form (12). The current data revolution highlights the need to explore other measures of success (13–15), as researchers are producing massive quantities of data that could provide valuable context for questions far beyond their original intent. While funding agencies are discussing ways to mandate data sharing (16), the sharing of high-quality, well-curated data should also be driven by incentives. Other considerations include a mechanism to request permission to use data sets prior to publication by the data owner(s), as scientists would be more willing to share data with certain conditions on its use (8).

To encourage a culture that shares microbiome data, it is critical to develop incentives and promote ways to reward data stewardship. This workshop brainstormed several ways to encourage a culture that shares microbiome data, which the NMDC team is working to support.

(i) Establish digital object identifiers (DOIs) to enable data set citations. It has widely been reported that receiving credit through data set citations is important for data sharing (8, 17). Providing a method for citing data sets in published articles opens the door for data set reuse to be quantified and, therefore, easily incorporated as a new metric in the research incentives structure. Journals that publish data set papers, such as *Nature Scientific Data*, *Gigascience*, and *Microbiology Resource Announcements*, are an important start, and other publishers have started these discussions (18). Several organizations are able to issue and register DOIs for data sets, but determining the granularity of DOI assignment at the individual data set or project level, as well as tracking mechanisms, remain challenging. Further coordination with funders and additional publishers will be critical for defining, establishing, and promoting data citations and accurate citation metrics.

(ii) Host data analysis competitions to support training on FAIR data for early career researchers. Early career researchers, including graduate students, are seen as critically important for catalyzing the cultural shift toward sharing well-curated microbiome data. While they may not get to decide when their data are shared, early career researchers are often responsible for the experiments, data collection, data management, data formatting, and efforts needed to make experimental data reusable and publicly accessible. Because of the inherent data access and transparency challenges (19), meta-analyses can serve as important training for early career researchers to (i) understand the challenges in finding, accessing, and preparing data sets for analysis; (ii) recognize and appreciate data sets that are well curated and accessible; and (iii) thus, be motivated to prepare and share their own data. Hosting data competitions (e.g., DREAM challenges, http://dreamchallenges.org/) to encourage meta-analyses can

showcase data sharing and reproducible science, while also providing benefits for participants (training, professional development, funding) and making important contributions to science (20–23). Further, data competitions can showcase how aggregating multiple standardized, well-curated microbiome data sets can enable new discoveries (24) and, more importantly, forge new paths for optimizing data collection and applying data standards earlier in the research workflow.

(iii) Celebrate the value added by impactful meta-analyses. When exploring how to address the current grand challenges in microbiome science, novel approaches using large-scale data science applications are no longer a goal, but a necessity (25). For example, the increased application of machine learning to biological problems (26) has begun to expand how we think about data and data sharing (27). It used to be thought that researchers who published work using someone else's published data were considered "data parasites" (28, 29). Now, the Pacific Symposium on Biocomputing celebrates the impactful meta-analyses through their annual Research Parasite Awards (https://researchparasite.com/), which highlight important contributions of secondary analyses. Well-curated and FAIR microbiome data sets will be necessary for our field to explore applications of machine learning, automation, and secondary analyses (30, 31).

While making data accessible is an important first step, data sets with missing information, erroneous values, or inconsistent formats hinder reuse. The workshop participants also discussed ways to incentivize efforts for sharing reusable data.

(iv) Establish comprehensive and coordinated data management plan(s) in collaboration with funders, publishers, and research service centers. While funders and publishers have moved toward encouraging open access to data (32), the details of their data sharing policies vary (33, 34), and there are insufficient resources for enforcement (35). Data access remains a challenge for reproducible science (11, 34, 36, 37). A comprehensive data management plan that includes community standards should be supported by both funders and publishers, which would provide structure and guidelines for data management best practices throughout the scientific research process (38). In addition, a partnership with research service centers, such as sequencing and other omics centers, can provide an effective strategy for revisiting data management plans earlier in the data life cycle, before experimental data is generated.

(v) Provide training for a variety of learning styles. Data management best practices and data standards and ontologies are powerful tools in support of the FAIR data principles. However, even seasoned scientists are often overwhelmed by guidelines and intimidated by ontologies. It isn't enough to create a comprehensive data management plan. Making this material accessible to the diversity of individuals who participate in the research process will be critical for effective adoption. A "quick start" guide is often a more approachable entry point for a data management novice. Extensive, searchable documentation is key for veterans who just need a refresher. To allow understanding and exploration of these data types, access can be provided through interfaces that allow programmatic access and visual representation to support researchers with and without computational expertise. Further, the use of various formats, such as tutorial videos, interactive webinars, and in-person events, support a diversity of learning styles and enable bidirectional communication, which is critical for improving and updating training materials.

(vi) Establish a certification of "compliance." Despite the significant efforts already invested in defining minimum standards for microbiome data, such as the Minimum Information about any (x) Sequence (MIxS) packages (39), important work remains to ensure that the various standards and ontologies are interoperable and easily accessible to the research community. This entails working with researchers to identify metadata attributes that are valuable for data reuse within their respective communities, and defining community-specific benchmarks. Establishing a "certification of compliance" based on these benchmarks would enable designation of data sets ready for reuse, which encourages inclusion in follow-up studies and enhances their citation metrics (see section i above).

Understand and reduce barriers to data submission. In addition to encouraging a culture that shares microbiome data, the workshop participants also discussed infrastructure challenges that impede sharing. Current data submission processes to primary data repositories or analytic platforms can be difficult to navigate, creating barriers even for good data stewards. The workshop participants suggested the following as a starting point to understand and reduce barriers to data/metadata submission.

(i) Understand how communities are currently using MIxS packages. MIxS packages are available for a variety of sample types and environments, but comparing their usage across data repositories is challenging. Are certain domains using them more or less often than others? For example, identifying research areas (e.g., domain, geographic location) that rarely use MIxS packages, submit data with the minimal required fields, or use null values to represent more than one meaning (e.g., missing versus not collected) enables a more targeted approach to training and outreach.

(ii) Explore ways to harmonize data submission processes across platforms. Data submission portals, such as those involved in the International Nucleotide Sequence Database Collaboration (INSDC) (40), each have unique requirements and interfaces, some having more robust manuals or training documents than others. Enabling coordination through community standards and appropriate training materials will greatly enhance the availability of FAIR microbiome data.

(iii) Validate sample metadata with immediate, informative feedback. Using ontologies or MIxS packages requires the use of specific formats for sample metadata attributes. Most communities manage data in spreadsheets without use of controlled vocabularies or data standards, and reformatting entries is error-prone. Reducing barriers to reformatting spreadsheets using sample metadata validators provides immediate, informative, and targeted feedback (41). Efficient and effective data submission has a significant impact on researchers' likelihood to share well-curated data.

OUTCOMES FROM THE WORKING GROUPS

During the workshop, working groups were formed and tasked with identifying ways that the microbiome research community could achieve tangible progress to advance FAIR data principles (2). Three areas were targeted as initial steps that the NMDC team, in collaboration with the working groups, could promote to improve sharing and adoption of standards: (i) expanding and enhancing existing community-driven standards; (ii) understanding the current use of standards across research communities; and (iii) outlining a strategy for training and adoption of standards by the community.

Expanding and enhancing standards. In collaboration with the data standards community, the NMDC initiative is expanding and enhancing existing sample metadata standards for microbiome data. These efforts include closely collaborating with the GSC to convert the MIxS standard into machine readable formats (i.e., JSON-Schema, Web Ontology Language), reviewing and adding new terms for the next MIxS standard release (version 6), and engaging with new stakeholders to address domain-specific needs. While the NMDC pilot initiative does not currently support the migration of other packages or checklists to the MIxS standard, the team does encourage community-driven development of standards for emerging subfields through the GSC, such as an agricultural-focused metadata standard (42). The NMDC team is collaborating with the Environment Ontology (EnvO) (43) group to assist with the development of new terms, new relationships between terms, and training on EnvO, and is working with the Genomes Online Database (GOLD) (44), a manually curated metadata resource at the DOE Joint Genome Institute, team. As a result of these collaborative efforts, the NMDC initiative has established a schema (https://microbiomedata.github.io/nmdc-metadata/) for mapping core standards and ontologies to streamline the integration of diverse sample metadata spreadsheet formats. The NMDC metadata schema relies on Biosample information (https://microbiomedata.github.io/nmdc-metadata/) for linking complementary data originating from the same physical sample (e.g., 16S and

metagenomes), consistent with the National Center for Biotechnology Information (NCBI) and the European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI). While there are challenges in linking other data types beyond sequence data (e.g., geochemical analyses), the use of an International Geo Sample Number through the System for Earth Sample Registration (https://www.geosamples.org/overview) registry would support data linkages to unique biosamples and is being adopted by the NMDC.

Use of standards across research communities. In collaboration with representatives from NCBI and EMBL-EBI, this working group gathered MIxS environmental package usage data from the Sequence Read Archive (SRA) and European Nucleotide Archive (ENA), respectively. Examining the overall number of samples registered with MIxS environmental packages reveals similar rates of adoption across SRA and ENA (Fig. S1 in the supplemental material) (counts represent distinct samples submitted to each respective repository, and mirrored data are not double counted). Further evaluation of whether the MIxS packages are being applied as expected (Table S1) show noticeable differences between the two repositories (Fig. 2), which likely reflect distinct user communities. In ENA, usage of MIxS packages is higher across studies than across samples, suggesting that smaller studies are more regularly using MIxS. In SRA, human-associated packages are prominent, likely reflecting projects funded by the National Institutes of Health. While these statistics focus on baseline usage for MIxS packages, other checklist/packages, such as the "default ENA checklist" or the "NCBI metagenome package," are not necessarily incorrect, nor do they indicate poorly curated sample metadata. Some non-MIxS checklists/packages provide extensive metadata descriptors (e.g., the ENA sewage checklist), which may be unique to certain types of samples. The NMDC team will use these data as a baseline for assessing metadata standards adoption across communities, and to inform areas for targeted training or feedback collection. The NMDC team, in collaboration with the GSC, will report updates on MIxS standards usage in ENA and SRA, and incorporate this information into forthcoming training modules.

Training and adoption of standards by the community. In collaboration with international partners affiliated with the GO FAIR initiative (https://www.go-fair.org/), the NMDC team recently established the FAIR Microbiome Implementation Network, the first coordinated effort focused on FAIR data for the microbiome community (https://www.go-fair.org/implementation-networks/overview/fair-microbiome/). The Microbiome Implementation Network aims to promote discovery and reuse of microbiome data by formalizing core and domain-specific microbiome ontologies and establishing training on the NMDC data models. In addition, the NMDC team is building out a modular training strategy, in collaboration with the GSC and OBO Foundry, that will cover basic sample metadata, such as domain-specific characteristics (e.g., MIxS packages) and FAIR data best practices. As a high-level summary, this working group drafted *Introduction to Metadata and Ontologies: Everything You Always Wanted to Know About Metadata and Ontologies (But Were Afraid to Ask)* (Text S1).

Conclusions. The foundation for reusable data has been created by the standards community and data sharing is increasing throughout the microbiome community, but there are still barriers to making microbiome data truly FAIR. Workshop participants highlighted the need to encourage data sharing through changes in the incentive structure and research culture. They also stated the importance of providing researchers with sufficient tools, training, and infrastructure to lower the barriers to sharing well-curated, reusable data. The working groups provided valuable contributions to the NMDC initiative, which has fed into the development of the NMDC metadata schema linked to existing standards, evaluation metrics on the usage of the GSC MIxS environmental packages for targeted activities, and the design of training packages to complement available data standards. The NMDC pilot initiative will continue to work across the standards and microbiome research communities to reduce barriers to data sharing, recognize data contributions, and make microbiome data FAIR.

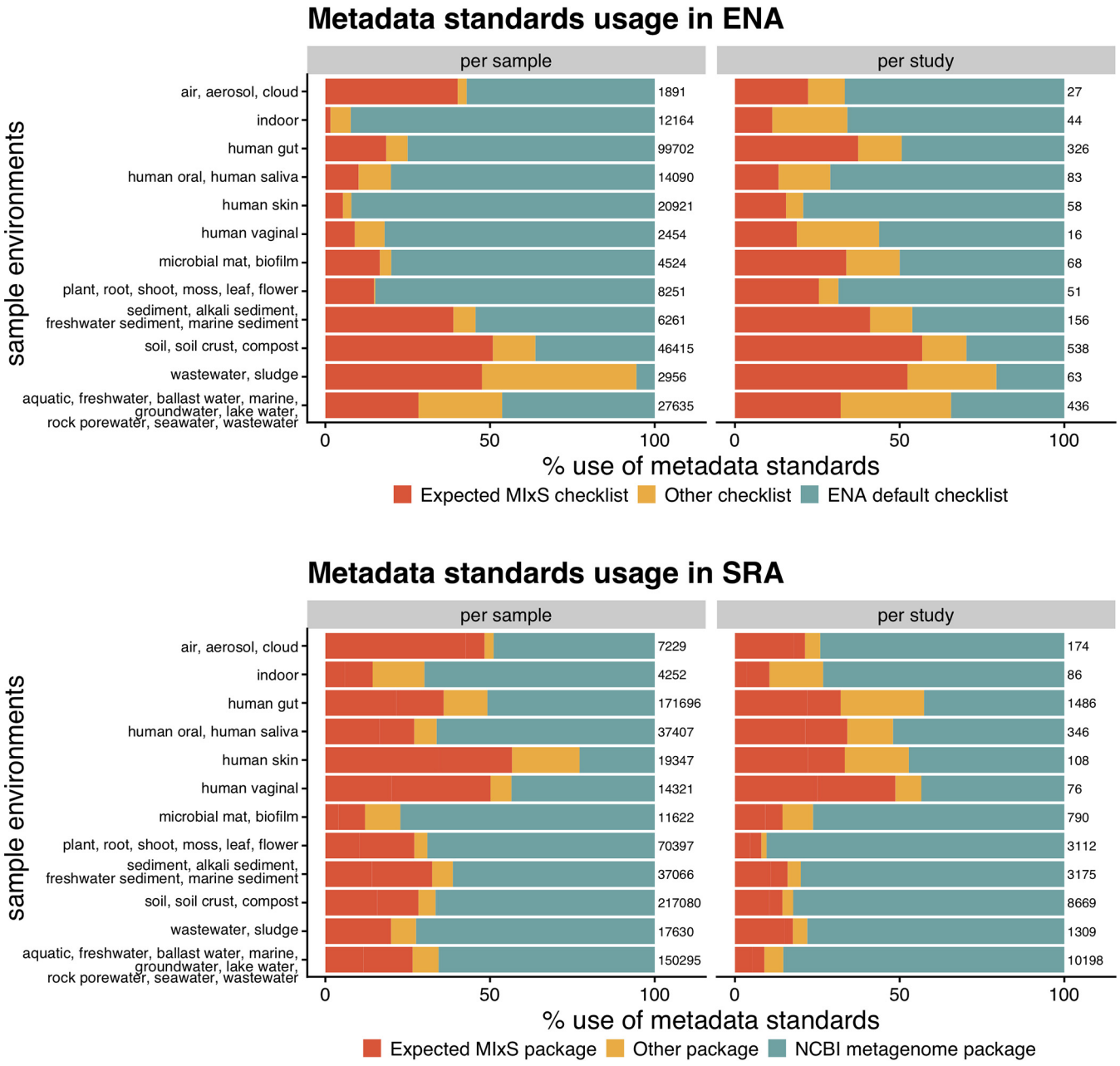

FIG 2 Usage of metadata standards across sample environments. For several MIxS packages, the working group identified representative metagenome organism name(s) for each package (see Table S1 for details) in order to inform how the MIxS packages were used across communities. The standards were evaluated as follows: (i) "Expected MIxS checklist/package," the chosen checklist/package used for sample registration was the most appropriate MIxS option based on the metagenome organism name provided (Table S1); (ii) "Other checklist/package," the chosen checklist/package used for sample registration may not have been the most appropriate MIxS checklist/package or followed an alternative set of standards; or (iii) "ENA default checklist or NCBI metagenome package," the chosen checklist/package used for sample registration was the ENA/NCBI defined minimum for samples/metagenome samples and did not use a specific sample metadata standard. Only public samples and their associated studies for raw read submissions of metagenomic and amplicon data (MIMS and MIMARKs survey) to ENA or SRA were included in the respective counts (counts reflect only submitted data to each repository and exclude mirrored data). Associated studies were counted once for each unique metagenome organism name represented in the study, and hence may have been counted more than once (i.e., a study associated with samples assigned with x unique metagenome organism names may be counted x times). Queries were run in fall 2020. ENA queries used the ENA Portal API with the respective taxon criteria and checklist ID (Table S1) (e.g., ENA sample counts with expected use of the Air MIxS checklist (https://www.ebi.ac.uk/ena/portal/api/search?result=read_run&query=(sample_accession=%22SAMEA*%22%20OR%20sample_accession=%22ERS*%22)%20AND%20(tax_eq(655179)%20OR%20tax_eq(1708701)%20OR%20tax_eq(1643811))%20AND%20checklist=%22ERC000012%22&fields=sample_accession). SRA queries used the NCBI Entrez Programming Utilities (e.g., SRA sample counts with expected use of the MIMS Air MIxS package, esearch -db biosample -query (biosample sra[Filter]) AND (((ncbi[Filter]) AND (air metagenome[Organism] OR aerosol metagenome[Organism] OR cloud metagenome[Organism]))) AND package mims metagenome/environmental, air version 5 0[Properties]).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TEXT S1, PDF file, 0.1 MB.

FIG S1, PDF file, 0.05 MB.

TABLE S1, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

We thank the Toolbox Dialogue Initiative representatives Stephanie E. Vasko, Marisa A. Rinkus, and Chet McLeskey for leading activities at the start of the workshop to prepare participants for open, respectful dialogue (see https://microbiomedata.org/nmdc-ontology-workshop/ for a report on the activities).

The NMDC is supported by the Genomic Science Program in the United States Department of Energy, Office of Science, Office of Biological and Environmental Research (BER) under contract numbers DE-AC02-05CH11231 (LBNL), 89233218CNA000001 (LANL), DE-AC05-00OR22725 (ORNL), and DE-AC05-76RL01830 (PNNL). S.V.C., D.C.B., and J.M.G. were funded by the Space Biology Program (Science Mission Directorate, Biological and Physical Sciences Division) of the National Aeronautics and Space Administration. J.H. was supported by internal Basic Research programs (Work Units 4888 and 1L73) at the U.S. Naval Research Laboratory. A.J. was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. The work of J.P.S. was supported by NIH-SD-IRACDA (5K12GM068524-17). S.M.G. was supported by a Washington Research Foundation Distinguished Investigator Award.

The opinions and assertions contained herein are those of the authors and are not to be construed as those of the Department of Defense, U.S. Navy, military service at large, or the U.S. Government.

REFERENCES

1. Wood-Charlson EM, Anubhav Auberry D, Blanco H, Borkum MI, Corilo YE, Davenport KW, Deshpande S, Devarakonda R, Drake M, Duncan WD, Flynn MC, Hays D, Hu B, Huntemann M, Li P-E, Lipton M, Lo C-C, Millard D, Miller K, Piehowski PD, Purvine S, Reddy TBK, Shakya M, Sundaramurthi JC, Vangay P, Wei Y, Wilson BE, Canon S, Chain PSG, Fagnan K, Martin S, McCue LA, Mungall CJ, Mouncey NJ, Maxon ME, Eloe-Fadrosh EA. 2020. The National Microbiome Data Collaborative: enabling microbiome science. Nat Rev Microbiol 18:313–314. https://doi.org/10.1038/s41579-020-0377-0.
2. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18.
3. National Microbiome Data Collaborative. 2021. Introduction to metadata and ontologies. https://microbiomedata.org/introduction-to-metadata-and-ontologies/.
4. Ponsero AJ, Bomhoff M, Blumberg K, Youens-Clark K, Herz NM, Wood-Charlson EM, Delong EF, Hurwitz BL. 2021. Planet Microbe: a platform for marine microbiology to discover and analyze interconnected 'omics and environmental data. Nucleic Acids Res 49:D792–D802. https://doi.org/10.1093/nar/gkaa637.
5. Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, Gilbert J, Glöckner FO, Hirschman L, Karsch-Mizrachi I, Klenk H-P, Knight R, Kottmann R, Kyrpides N, Meyer F, San Gil I, Sansone S-A, Schriml LM, Sterk P, Tatusova T, Ussery DW, White O, Wooley J. 2011. The Genomic Standards Consortium. PLoS Biol 9:e1001088. https://doi.org/10.1371/journal.pbio.1001088.
6. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann RH, Shah N, Whetzel PL, Lewis S, OBI Consortium. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25:1251–1255. https://doi.org/10.1038/nbt1346.
7. Casadevall A, Fang FC. 2015. Impacted science: impact is not importance. mBio 6:e01593–15–e01515. https://doi.org/10.1128/mBio.01593-15.
8. Tenopir C, Rice NM, Allard S, Baird L, Borycz J, Christian L, Grant B, Olendorf R, Sandusky RJ. 2020. Data sharing, management, use, and reuse: practices and perceptions of scientists worldwide. PLoS One 15:e0229003. https://doi.org/10.1371/journal.pone.0229003.
9. Schmidt B, Gemeinholzer B, Treloar A. 2016. Open data in global environmental research: The Belmont Forum's open data survey. PLoS One 11:e0146695. https://doi.org/10.1371/journal.pone.0146695.
10. Stuart D, Baynes G, Hrynaszkiewicz I, Allin K, Penny D, Lucraft M, Astell M. 2018. White paper: practical challenges for researchers in data sharing. Springer Nature.
11. Eckert EM, Di Cesare A, Fontaneto D, Berendonk TU, Bürgmann H, Cytryn E, Fatta-Kassinos D, Franzetti A, Larsson DGJ, Manaia CM, Pruden A, Singer AC, Udikovic-Kolic N, Corno G. 2020. Every fifth published metagenome is not available to science. PLoS Biol 18:e3000698. https://doi.org/10.1371/journal.pbio.3000698.
12. Lohr S. 2014. For big-data scientists, "janitor work" is key hurdle to insights. The New York Times.
13. Ravenscroft J, Liakata M, Clare A, Duma D. 2017. Measuring scientific impact beyond academia: an assessment of existing impact metrics and proposed improvements. PLoS One 12:e0173152. https://doi.org/10.1371/journal.pone.0173152.
14. Bollen J, Van de Sompel H, Smith JA, Luce R. 2005. Toward alternative metrics of journal impact: a comparison of download and citation data. Inf Process Manag 41:1419–1440. https://doi.org/10.1016/j.ipm.2005.03.024.
15. Seglen PO. 1997. Why the impact factor of journals should not be used for evaluating research. BMJ 314:498–502. https://doi.org/10.1136/bmj.314.7079.497.
16. National Institutes of Health. 2019. DRAFT NIH policy for data management and sharing. https://osp.od.nih.gov/wp-content/uploads/Draft_NIH_Policy_Data_Management_and_Sharing.pdf
17. Kratz JE, Strasser C. 2015. Comment: making data count. Sci Data 2:150039. https://doi.org/10.1038/sdata.2015.39.

18. Cousijn H, Kenall A, Ganley E, Harrison M, Kernohan D, Lemberger T, Murphy F, Polischuk P, Taylor S, Martone M, Clark T. 2018. A data citation roadmap for scientific publishers. Sci Data 5:180259. https://doi.org/10.1038/sdata.2018.259.
19. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. 2018. Meta-analysis and the science of research synthesis. Nature 555:175–182. https://doi.org/10.1038/nature25753.
20. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G, DREAM5 Consortium. 2012. Wisdom of crowds for robust gene network inference. Nat Methods 9:796–804. https://doi.org/10.1038/nmeth.2016.
21. Ghouila A, Siwo GH, Entfellner J-BD, Panji S, Button-Simons KA, Davis SZ, Fadlelmola FM, Ferdig MT, Mulder N, The DREAM of Malaria Hackathon Participants. 2018. Hackathons as a means of accelerating scientific discoveries and knowledge transfer. Genome Res 28:759–765. https://doi.org/10.1101/gr.228460.117.
22. Saez-Rodriguez J, Costello JC, Friend SH, Kellen MR, Mangravite L, Meyer P, Norman T, Stolovitzky G. 2016. Crowdsourcing biomedical research: leveraging communities as innovation engines. Nat Rev Genet 17:470–486. https://doi.org/10.1038/nrg.2016.69.
23. Bender E. 2016. Challenges: crowdsourced solutions. Nature 533:S62–4. https://doi.org/10.1038/533S62a.
24. Silge J, Robinson D. 2017. Text mining with R: a tidy approach. O'Reilly Media, Inc., Boston, MA.
25. Kyrpides NC, Eloe-Fadrosh EA, Ivanova NN. 2016. Microbiome data science: understanding our microbial planet. Trends Microbiol 24:425–427. https://doi.org/10.1016/j.tim.2016.02.011.
26. Jordan MI, Mitchell TM. 2015. Machine learning: trends, perspectives, and prospects. Science 349:255–260. https://doi.org/10.1126/science.aaa8415.
27. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow PM, Zietz M, Hoffman MM, Xie W, Rosen GL, Lengerich BJ, Israeli J, Lanchantin J, Woloszynek S, Carpenter AE, Shrikumar A, Xu J, Cofer EM, Lavender CA, Turaga SC, Alexandari AM, Lu Z, Harris DJ, DeCaprio D, Qi Y, Kundaje A, Peng Y, Wiley LK, Segler MHS, Boca SM, Swamidass SJ, Huang A, Gitter A, Greene CS. 2018. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 15:20170387. https://doi.org/10.1098/rsif.2017.0387.
28. Longo DL, Drazen JM. 2016. Data sharing. N Engl J Med 374:276–277. https://doi.org/10.1056/NEJMe1516564.
29. Duvallet C. 2020. Data detectives, self-love, and humility: a research parasite's perspective. Gigascience 9:giz148. https://doi.org/10.1093/gigascience/giz148.
30. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, Kosciolek T, Janssen S, Metcalf J, Song SJ, Kanbar J, Miller-Montgomery S, Heaton R, Mckay R, Patel SP, Swafford AD, Knight R. 2020. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 579:567–574. https://doi.org/10.1038/s41586-020-2095-1.
31. Sze MA, Schloss PD. 2016. Looking for a signal in the noise: revisiting obesity and the microbiome. mBio 7:e01018-16. https://doi.org/10.1128/mBio.01018-16.
32. Sheehan J. 2016. Increasing access to the results of federally funded science. The White House 22. https://obamawhitehouse.archives.gov/blog/2016/02/22/increasing-access-results-federally-funded-science.
33. Vasilevsky NA, Minnier J, Haendel MA, Champieux RE. 2017. Reproducible and reusable research: are journal data sharing policies meeting the mark? PeerJ 5:e3208. https://doi.org/10.7717/peerj.3208.
34. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA. 2011. Public availability of published research data in high-impact journals. PLoS One 6:e24357. https://doi.org/10.1371/journal.pone.0024357.
35. Couture JL, Blake RE, McDonald G, Ward CL. 2018. A funder-imposed data publication requirement seldom inspired data sharing. PLoS One 13:e0199789. https://doi.org/10.1371/journal.pone.0199789.
36. Miyakawa T. 2020. No raw data, no science: another possible source of the reproducibility crisis. Mol Brain 13:24. https://doi.org/10.1186/s13041-020-0552-2.
37. Stodden V, Seiler J, Ma Z. 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. Proc Natl Acad Sci U S A 115:2584–2589. https://doi.org/10.1073/pnas.1708290115.
38. 2017. Overcoming hurdles in sharing microbiome data. Nat Microbiol 2:1573. https://doi.org/10.1038/s41564-017-0077-3.
39. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, 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, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PSG, Charlson E, Costello EK, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman JA, Gallery RE, Gevers D, Gibbs RA, San Gil I, Gonzalez A, Gordon JI, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau AL, Kelley ST, Kennedy J, Knights D, Koren O, et al. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 29:415–420. https://doi.org/10.1038/nbt.1823.
40. Karsch-Mizrachi I, Takagi T, Cochrane G, International Nucleotide Sequence Database Collaboration. 2018. The international nucleotide sequence database collaboration. Nucleic Acids Res 46:D48–D51. https://doi.org/10.1093/nar/gkx1097.
41. Rideout JR, Chase JH, Bolyen E, Ackermann G, González A, Knight R, Caporaso JG. 2016. Keemei: cloud-based validation of tabular bioinformatics file formats in Google Sheets. Gigascience 5:27. https://doi.org/10.1186/s13742-016-0133-6.
42. Dundore-Arias JP, Eloe-Fadrosh EA, Schriml LM, Beattie GA, Brennan FP, Busby PE, Calderon RB, Castle SC, Emerson JB, Everhart SE, Eversole K, Frost KE, Herr JR, Huerta AI, Iyer-Pascuzzi AS, Kalil AK, Leach JE, Leonard J, Maul JE, Prithiviraj B, Potrykus M, Redekar NR, Rojas JA, Silverstein KAT, Tomso DJ, Tringe SG, Vinatzer BA, Kinkel LL. 2020. Community-driven metadata standards for agricultural microbiome research. Phytobiomes J 4:115–121. https://doi.org/10.1094/PBIOMES-09-19-0051-P.
43. Buttigieg PL, Pafilis E, Lewis SE, Schildhauer MP, Walls RL, Mungall CJ. 2016. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. J Biomed Semantics 7:57. https://doi.org/10.1186/s13326-016-0097-6.
44. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Sundaramurthi JC, Lee J, Kandimalla M, Chen I-MA, Kyrpides NC, Reddy TBK. 2021. Genomes OnLine Database (GOLD) v.8: overview and updates. Nucleic Acids Res 49:D723–D733. https://doi.org/10.1093/nar/gkaa983.