Title: FAIR enough? A perspective on the status of nucleotide sequence data and metadata on public archives

Authors: Christiane Hassenrück[1,2], Tobias Poprick[3], Véronique Helfer[3], Massimiliano Molari[4], Raissa Meyer[5], Ivaylo Kostadinov[6]

Affiliations:

[1] Leibniz Institute for Baltic Sea Research Warnemünde (IOW), Seestrasse 15, 18119 Rostock-Warnemünde, Germany

[2] MARUM - Center for Marine Environmental Sciences, University of Bremen, Leobener Straße 8, 28359 Bremen, Germany

[3] Leibniz Centre for Tropical Marine Research (ZMT), Fahrenheitstrasse 6, 28359 Bremen, Germany

[4] Max Planck Institute for Marine Microbiology, Celsiusstr. 1, 28359 Bremen, Germany

[5] Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany

[6] GFBio - Gesellschaft für Biologische Daten e.V., c/o Research II, Campus Ring 1, 28759 Bremen, Germany

Correspondence: christiane.hassenrueck@io-warnemuende.de

Abstract:

Knowledge derived from nucleotide sequence data is increasing in importance in the life sciences, as well as decision making (mainly in biodiversity policy). Metadata standards have been established to facilitate sustainable sequence data management according to the FAIR principles (Findability, Accessibility, Interoperability, Reusability). Here, we review the status of metadata available for raw read Illumina amplicon and whole genome shotgun sequencing data derived from ecological metagenomic material that are accessible at the European Nucleotide Archive (ENA), as well as the compliance of the primary sequence data (fastq files) with data submission requirements. While overall basic metadata, such as geographic coordinates, were retrievable in 98% of the cases for this type of sequence data, interoperability was not always ensured and other (mainly conditionally) mandatory parameters were often not provided at all. Metadata standards, such as the 'Minimum Information about any(x) Sequence (MIxS)', were only infrequently used despite a demonstrated positive impact on metadata quality. Furthermore, the sequence data itself did not meet the prescribed requirements in 31 out of 39 studies that were manually inspected. To tackle the most immediate needs to improve FAIR sequence data

management, we provide a list of minimal suggestions to researchers, research institutions, funding agencies, reviewers, publishers, and databases, that we believe might have a potentially large positive impact on sequence data and metadata FAIRness, which is crucial for further research and its derived applications.

Keywords: sequence data management, metadata standards, interoperability, reusability, ontology, digital sequence information (DSI), biodiversity, next generation sequencing

INTRODUCTION

Next generation sequencing has gained increasing popularity and is now firmly established as a routine tool in multiple fields of the life sciences, such as ecology (foremost microbial ecology), biodiversity research, and conservation biology. Furthermore, knowledge derived from nucleotide sequence data, also referred to as digital sequence information (DSI), is becoming increasingly relevant for decision-making in natural resource management (e.g. Sustainable Development Goals) and as part of international agreements (e.g. Convention on Biological Diversity). The amount of nucleotide sequence data has been and still is growing exponentially (Harrison et al., 2021). However, a string of nucleotides (ACTG) on its own does not contain much information - metadata and contextual parameters (Box 1) are required to describe sample origin and sequence generation for the data to be meaningful within and beyond the scope of the study, for which the sequence was obtained. Capturing and communicating not only the primary (sequence) data, but also its metadata and contextual data, is a crucial part of good data management. To promote sustainable data management and usage, the FAIR principles have been introduced. They offer guidance on how to make data Findable, Accessible, Interoperable, and Reusable (Box 1; Fillinger, de la Garza, Peltzer, Kohlbacher, & Nahnsen, 2019; Wilkinson et al., 2016), to prepare for a future of more automated analyses with the aim that the value of the data will not be restricted to a single study, but will extend to the reuse and integration across multiple studies over time. Recently, the trend of an increasing data volume, which is being more and more sustainably managed, has resulted in nucleotide sequence data being used more frequently in the emerging field of data science, answering new scientific questions with existing data, as such constituting a public good for the scientific community (Box 1). One prime example is the TARA Oceans data set, which has so far resulted in hundreds of publications making secondary use of the data – a number that is constantly increasing[1].

One key aspect of FAIR data is the implementation of standards for metadata (Box 1; Wilkinson et al., 2016). To this end, the Genomic Standards Consortium (GSC; Field et al., 2011) has established the 'Minimum Information about any (x) Sequence (MIxS)' family of standards, to describe sample collection and sequence generation in a consistent manner (Yilmaz et al., 2011; Fig 1). MIxS consists of several customized checklists tailored for a diverse set of sequencing applications and investigated environments, defining mandatory (always to be provided, core

[1] https://oceans.taraexpeditions.org/wp-content/uploads/2020/10/TARA_RA_EN_.pdf

parameters), conditionally mandatory (mandatory for a specific sequencing application), environment-specific, and optional parameters. In addition to standardizing metadata parameter names, MIxS suggests a consistent format for units and syntax for data values, thereby promoting the interoperability of the data provided in compliance with this standard. For instance, MIxS makes use of ontologies, such as the Environmental Ontology (ENVO; Buttigieg, Morrison, Smith, Mungall, & Lewis, 2013; Buttigieg et al., 2016) or the Experimental Factor Ontology (EFO; Malone et al., 2010), to describe the sampled environment or experimental conditions using a controlled vocabulary and to make the metadata more machine-actionable.

Box 1: Glossary

FAIR (paraphrased after Wilkinson et al. (2016):

- Findable: Data and metadata are linked and findable via a unique and persistent identifier (e.g. accession number). Metadata is further searchable.
- Accessible: Data and metadata are retrievable (by humans and machines) via their identifier. Metadata remains accessible even if associated data is not available anymore.
- Interoperable: Data and metadata use a common language for knowledge representation understandable by humans and machines.
- Reusable: Data are described by rich metadata to provide the context required for reuse.

Ontology: Ontologies impose a (machine-readable) hierarchical structure of relationships for the components of a given system, using a controlled and clearly defined vocabulary. In the case of the Environmental Ontology (ENVO), this increases the interoperability of environmental descriptions, helping (meta)data records achieve demonstrable FAIRness.

Metadata: Collection of parameters that describe the primary data, in this example nucleotide sequencing data. Most metadata parameters are intrinsic to the sampling or experimental design and the laboratory or analytical procedures. As such they are often known *a priori*, i.e. before the primary data collection. For instance, sampling location, experimental treatments, sequencing platform.

Contextual data: While often grouped together with metadata, contextual data is referring to parameters which are recorded alongside the primary data. For instance, temperature, salinity, inorganic nutrient concentrations. They are often primary data for other research fields.

Metadata standards: Metadata standards provide a structured framework for metadata documentation in compliance with the FAIR principles. They provide checklists of well-defined parameters to be reported and determine the vocabulary and units to be used to ensure consistent data across studies.

Data mining: In the context of this publication, data mining is referring to the retrieval and reuse of data sets that have been archived in open access data repositories.

The main hubs for long-term nucleotide sequence data storage are the three resources making up the International Nucleotide Sequence Database Collaboration (INSDC; Fig 1): the European Nucleotide Archive (ENA), the National Center for Biotechnology Information (NCBI), and the

DNA Data Bank of Japan (DDBJ). These databases are mirrored so that the same data is available in all three. The establishment of this infrastructure has been instrumental in propagating global standards for sequence data and metadata (e.g. fasta, and fastq data formats) and offers services far beyond the provision of access to such data (see e.g. Cook, Bergman, Cochrane, Apweiler, & Birney, 2018; Fukuda, Kodama, Mashima, Fujisawa, & Ogasawara, 2021; NCBI Resource Coordinators, 2018). Furthermore, as part of the GSC, the members of the INSDC have contributed to designing the MIxS standards, pioneering the implementation of such metadata standards in nucleotide sequence databases with their official adoption in 2011 (Yilmaz et al., 2011).

ENA offers a minimal metadata standard for sequence submissions, although the use of more extensive checklists, such as those based on MIxS, is strongly recommended[2]. Metadata on ENA is organized on several levels[3] (Fig. 1): Sequencing **runs** are associated with specific **experiments**, which are referring to individual nucleic acid extractions and/or library preparations. Experiments are collected into **studies**, which usually use a common methodological approach. Several studies can be summarized by an umbrella project that may correspond to the larger scientific project, for which the data was generated. **Samples**, referring to the biological material, can be associated with multiple studies through different experiments. This flexible metadata model allows representing complex experimental set-ups correctly, but can be hard for inexperienced submitters to navigate properly and provide the necessary information for checklist compliance.

When accessing sequence data as data consumer (Fig. 1), all provided metadata for each level (run, experiment, study, sample) can be retrieved as XML. To simplify metadata access, the ENA advanced search offers a collection of indexed parameters that use standardized names and are searchable (i.e. usable to restrict the search) and returnable (i.e. downloadable in a user-friendly TSV format; ENA Portal API; Fig. 1). On the run level, these indexed parameters are also inherited from sample and experiment metadata. At the moment, the implementation of indexed parameters is limited to metadata parameters that are mandatory for most checklists and/or most frequently provided. As such, many conditionally mandatory, environment-specific, and optional MIxS parameters, mainly due to a lack of consistent and widespread use, are not indexed and only accessible in XML format, where no standardized nomenclature, controlled vocabulary or

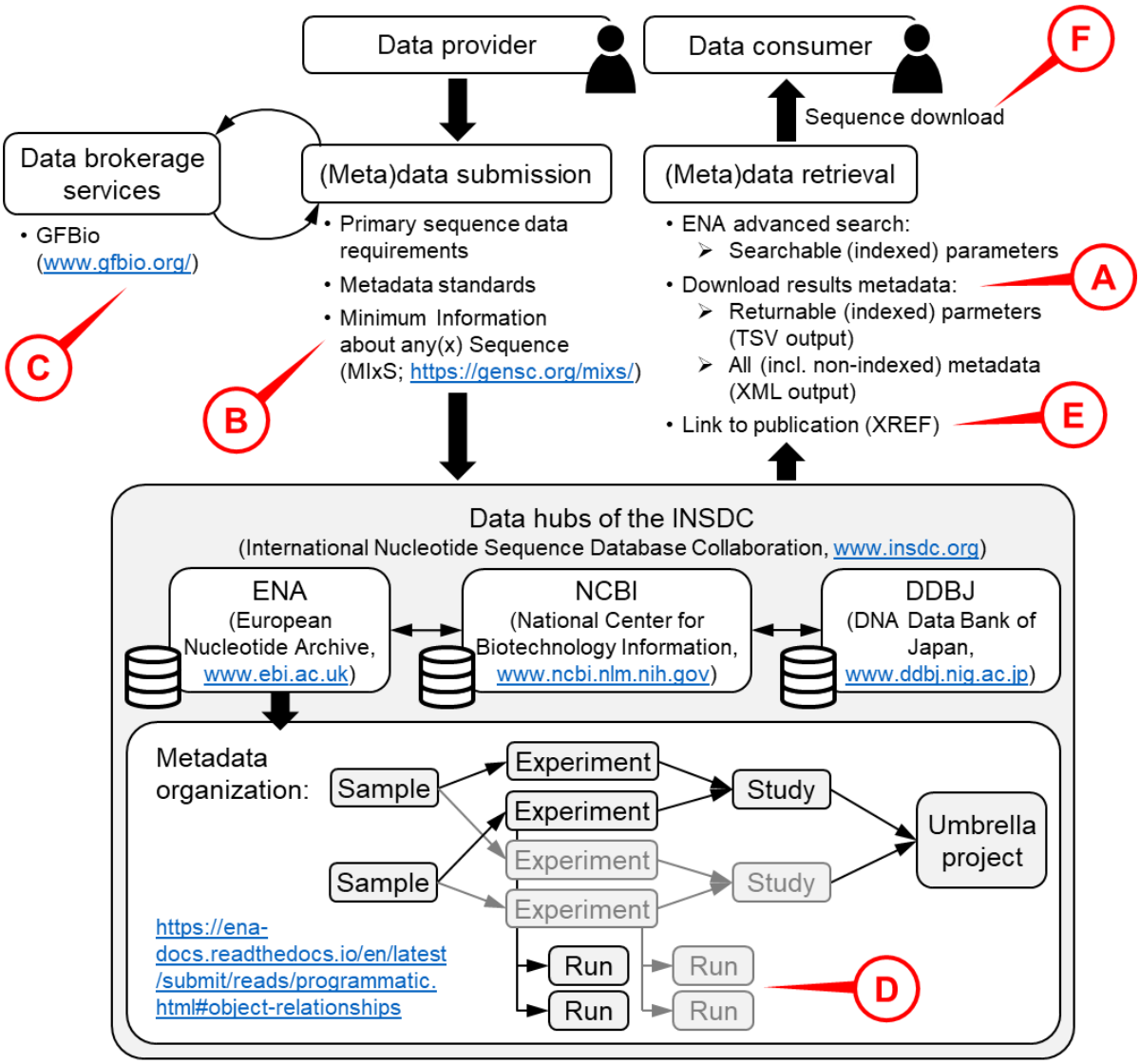
[2] https://ena-docs.readthedocs.io/en/latest/submit/samples.html#checklists
[3] https://ena-docs.readthedocs.io/en/latest/submit/reads/programmatic.html#object-relationships

specific data value format are enforced. Therefore, some of the value of MIxS is intrinsically lost, making non-indexed parameters not interoperable.

In addition to metadata requirements, ENA also standardizes the format of the submitted sequence data depending on the sequencing approach. For instance, paired-end Illumina raw reads have to be submitted as demultiplexed R1 and R2 files (fastq) without artificial sequences (e.g. adapters, linkers, barcodes/tags, primers) and prior to any quality trimming[4]. To provide sequencing data in such a format, initial sequence processing steps are necessary, starting from the multiplexed sequencer output. As bioinformatic sequence analysis pipelines vary, it is important to not deviate from the sequence format requirements by adjusting analysis workflows accordingly.

To support sustainable data management, data brokerage services, such as the German Federation for Biological Data (GFBio), have been established (Diepenbroek et al., 2014). Brokerage services offer a central entry point for data submissions, providing personal guidance (helpdesk) on FAIR data, supporting and often simplifying the data submission process, and ensuring data deposition on the most appropriate archive. As an additional checkpoint, brokerage services therefore constitute a valuable resource for each individual researcher to improve the FAIRness of their data, which is now becoming a strict requirement from many funding agencies.

[4] https://ena-docs.readthedocs.io/en/latest/submit/fileprep/reads.html?highlight=read%20format#fastq-format



A: Which metadata is available and how accessible is it?
Investigated parameters include geographic coordinates, environment description according to ENVO, target gene, nominal length, status as environmental sample.
B: How does use of MIxS checklists according to environmental packages affect (A)?
C: How does use of a brokerage service (GFBio) affect (A)?
D: Are sequences deposited in the correct format?
E: Can additional information be easily retrieved by linking the publication?
F: The analysis is sequence-centric, focusing on the metadata available per run.

Figure 1: Summary of the submission and retrieval process for nucleotide sequence data using the databases of the INSDC. The letters highlight the specific aspects in that process that were investigated here, focusing on how various factors (B, C) affect metadata quality (A) in a sequence-centric data mining approach (F), the compliance of the primary sequence data with data archiving requirements (D), and the retrieval of further information via associated publications (E).

To facilitate data reuse in data mining endeavors, access to and retrieval of the raw read sequencing data and metadata on run level, together with the inherited metadata parameters

153 describing the sample and sequencing experiment, are most crucial. However, despite the
154 available framework for FAIR data archiving, data interoperability and reusability are still limited
155 and often complicated by insufficient metadata (Eckert et al., 2020; Hoopen et al., 2016; Jurburg,
156 Konzack, Eisenhauer, & Heintz-Buschart, 2020; personal observation). Therefore, we decided to
157 conduct a review of the status of nucleotide sequence data and associated metadata accessible
158 through ENA to (i) identify deficits in metadata quality and (ii) provide suggestions for improving
159 FAIR data management (Fig. 1).
160
161 We restricted our analysis to a very popular example for biodiversity assessment in ecology:
162 paired-end amplicon (metabarcoding) raw read data generated from ecological metagenomes
163 (NCBI taxid: 410657) as source material on the Illumina platform. We focus on metadata
164 parameters, which are mandatory and/or crucial for the reuse of this kind of data, evaluating the
165 impact that the use of MIxS checklists (Fig. 1B) and the brokerage service GFBio (Fig. 1C) had
166 on metadata quality. We searched ENA on 13.12.2020 for raw read data using the following
167 search query: `tax_tree(410657) AND library_selection = "PCR" AND`
168 `library_strategy = "AMPLICON" AND library_layout = "PAIRED" AND`
169 `instrument_platform = "ILLUMINA" AND library_source = "METAGENOMIC"`. For
170 the resulting 413 849 search results on run level (Fig. 1F; hereafter referred to as cases), we
171 downloaded all available metadata parameters in TSV format as well as the sample and
172 experiment XML. Specifically, we checked for the following parameters if data was provided, ease
173 of access, correctness (if applicable), and compliance with standards (Fig. 1A): (i) geographic
174 coordinates, i.e. **latitude** and **longitude**, (ii) information about the **target gene** or subfragment or
175 primers, (iii) **nominal length**, i.e. insert size, and (iv) the parameters **environment_biome**,
176 **environment_material**, and **environment_feature**, which make use of ENVO to ensure a
177 standardized description of the sampled environment using a controlled vocabulary. As
178 comparison to the amplicon sequencing example, we repeated this assessment also for shotgun
179 metagenomic paired-end Illumina reads using the same query apart from the following changes:
180 `library_selection = "RANDOM" AND library_strategy = "WGS"` (Whole Genome
181 Sequencing; date accessed 08.01.2021; SI figures 1 + 2). The scripts to search ENA, download
182 the metadata, and calculate the summaries presented in this study are available on:
183 https://github.com/chassenr/ENA-metadata. The summaries of cases per year visualized in the
184 subsequent figures are further available in SI table 1.
185

Apart from metadata quality, the reusability of nucleotide sequence data - especially the automation thereof - strongly depends of the format of the submitted raw read data itself. Therefore, using a recent data mining effort focused on amplicon studies of the V3-V4 hypervariable region of the bacterial 16S gene (Molari et al. in prep.) as a **case study**, we evaluated if the raw read data (fastq files) had been archived according to the ENA guidelines (Fig. 1D). Furthermore, we checked the correct declaration as **environmental sample** and the availability of the associated **manuscript publication** comparing a manual search to the ENA XREF API (Fig. 1E), as these may provide further relevant information about the provenance of the data.

RESULTS

General trends: The number of cases (runs) retrieved by the above-mentioned query has been increasing continuously over the last decade, with more than 120 000 runs submitted in 2020 alone. In total, only 6.5% of runs (26 903 of 413 849) were linked to samples compliant with a MIxS checklist and environmental package, with 21% in 2015 and (with the exception of 2018) decreasing proportions since. On average 1.22 runs were submitted per sample, with a highly skewed distribution where only 8% of the samples were associated with more than one run. As such, most cases corresponded to a single sample, from which several of the investigated metadata parameters were inherited.

Geographic coordinates (Fig. 2 top): Latitude and longitude are mandatory MIxS parameters and essential for the reuse of sequencing data from environmental samples, especially in molecular ecology. However, these parameters are not necessarily enforced by ENA for submissions that do not use MIxS. Nevertheless, in our example the majority of all cases (80%) were archived with latitude and longitude available as decimal degrees in the TSV search output. For an additional 18% of the cases, latitude and longitude values were retrievable from the sample XML as part of non-indexed parameters. There, this data was stored under 23 different parameter names, and was therefore not easily accessible or interoperable. Considering only cases submitted according to a MIxS checklist and specifying a MIxS environmental package, latitude and longitude were always provided in some form and the proportion of cases with latitude and longitude available in the TSV output was slightly higher across all years (86%), although it has been declining from more than 99% to 61% since 2017.

Target gene (Fig. 2 middle): Missing information about which DNA region was targeted by the amplicon sequencing approach is one of the main obstacles for data interpretation and reuse. The ENA search output in TSV format includes the indexed parameter target_gene, which can be used to specify the amplified gene or locus name using free text. In MIxS, target_gene is further specified as a mandatory parameter for amplicon sequencing studies (MIMARKS survey), although its use is not enforced by ENA for data submissions. Additionally, the non-indexed parameters target_subfragment and pcr_primers are listed as conditionally mandatory parameters to supply additional metadata about the amplified gene region, and as such, should be supplied for all amplicon sequencing experiments. Among all cases investigated here, only 2% provided the target gene, with an additional 5% where some information about the amplified

region could be retrieved from non-indexed parameters in the sample XML (stored under 58 different parameter names) and from the library construction protocol included in the experiment XML. However, such entries were extremely inconsistent, ranging from gene and gene region names (or a combination of both) to primer names, primer sequences, and references for the applied PCR protocol. To reuse this data, each entry would have to be inspected manually, making available target gene information not only difficult to access, but also not interoperable. The proportion of cases with target gene information available among those submitted according to a MIxS checklist and environmental package was considerably higher with 35%, although still far from optimal bearing in mind that this is a mandatory parameter. The correct identification of the amplified region without any respective metadata is cumbersome and computationally expensive. Complete and correct metadata entries, using a standardized format or even controlled vocabulary preferably in accordance with existing ontologies, would drastically reduce computational and man-power requirements for post-deposition data curation and data reuse.

Nominal length (Fig. 2 bottom): Nominal length specifies the insert size, i.e. the length of the amplified fragment between the sequencing adapters (i.e. including primers) in the library. It is mandatory for all paired-end sequencing runs according to ENA, NCBI, DDBJ submission tutorials, and should have been enforced since 2014. However, in 86% of all cases this parameter was not provided. The use of a MIxS checklist and environmental package during the submission increased the percentage of cases providing nominal length to almost 100%, with only 55 cases in total in 2019 and 2020 lacking these values, suggesting that submitters who made the effort to be MIxS compliant were also more likely to provide metadata parameters outside of MIxS. Interestingly, peaks in the distribution of supplied nominal length values occurred at 250bp, 300bp, 500bp, and 600bp (data not shown), which correspond to the length of individual reads or the combined length of paired reads in the popular 2x250bp and 2x300bp sequencing approaches, and may therefore represent read length rather than insert size. This suggests that misconceptions exist about the definition of the parameter nominal length among sequence data submitters.

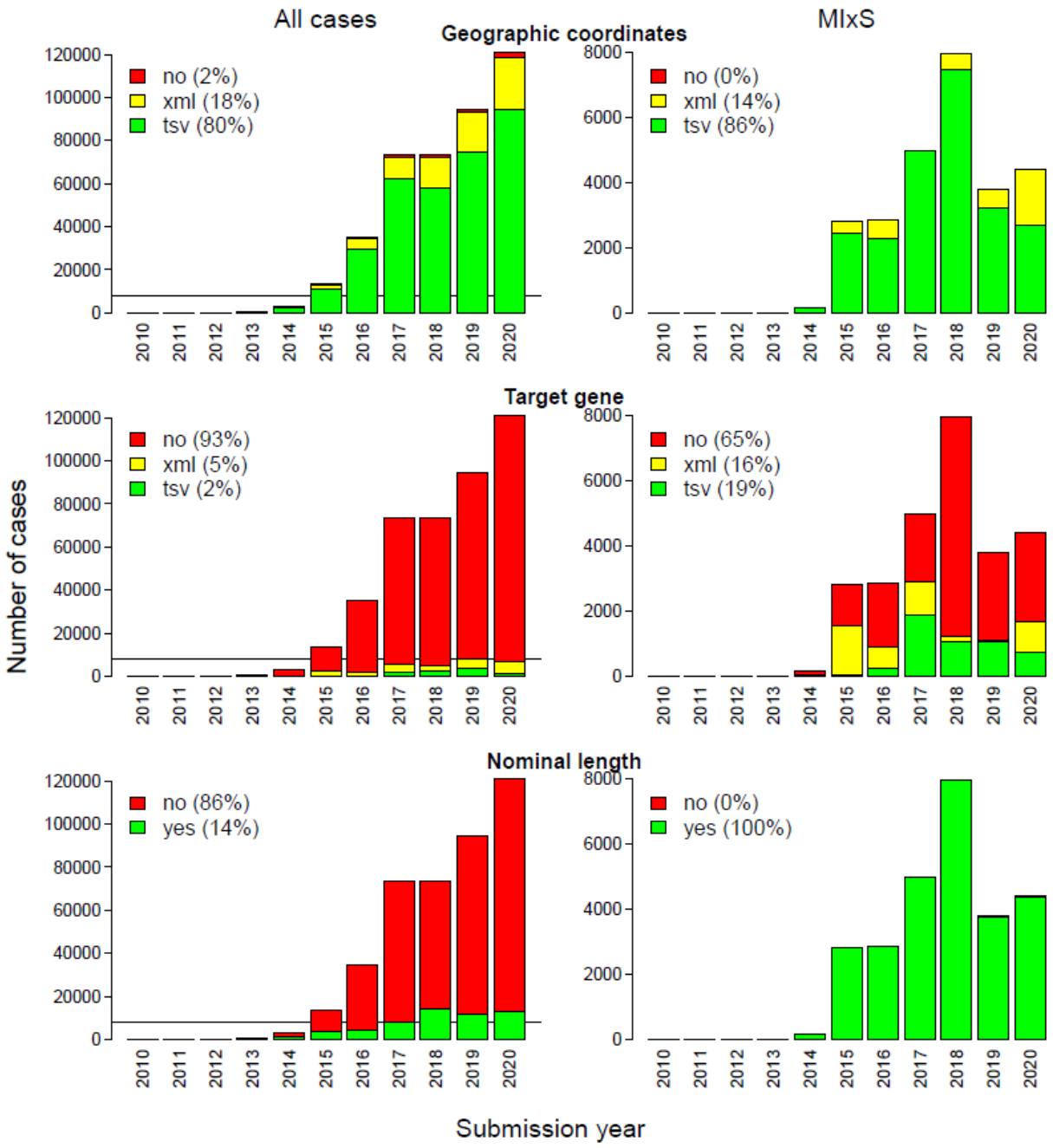


Figure 2: Number of cases (amplicon example) with and without metadata available for geographic coordinates (latitude, longitude), target gene (or related information, such as subfragment or pcr primers), and nominal length. For geographic coordinates and target gene, 'tsv' is referring to the information provided for indexed metadata parameters in the TSV search output, while 'xml' is referring to non-indexed metadata only accessible in the XML view of the ENA sample or experiment. The percentages summarize cases over all years and are rounded to integers. The plots on the right only show the cases with samples submitted according to a MIxS checklist and environmental package. The horizontal line in the plots on the left indicates the y-axes range of the plots on the right.

Environment description using ENVO (Fig. 3): The parameters environment_biome, environment_material, and environment_feature use ENVO terms to characterize various characteristics of the sample and the environment it originated from. They are mandatory parameters for any MIxS checklist. Across all cases 72-73% did not provide data for any of these three parameters. Conversely, values were supplied for all cases, which specified a MIxS checklist and environmental package, with the exception of 125 cases for the parameter environment_biome. However, ENVO terminology was inconsistently used despite the format required by MIxS. Especially for the parameter environment_biome, the majority of the provided values did not resemble any existing or even obsolete ENVO terms. Exact matches to ENVO term IDs, which are mandatory to be included according to the MIxS documentation, were only found for 8% of the cases, with an additional 2% with character string matches to ENVO term names. For environment_material and environment_feature, these proportions were considerably higher with 73% and 77% matches to ENVO term IDs and an additional 13% and 11% matches to ENVO term names, respectively. This demonstrates that the use of MIxS checklists drastically improved the availability of an environment description via the parameters environment_biome, environment_material, and environment_feature, but also that, if provided, the interoperability of this metadata is severely impaired by non-ENVO entries.

Interestingly, a considerable number of cases, none of which specified a MIxS checklist upon submission, provided metadata for the parameters env_broad_scale, env_local_scale, env_medium. Those parameters are included in a more recent version of MIxS (version 5.0) instead of environment_biome, environment_material, and environment_feature (version 4.0). As the new parameters are not (yet) indexed on ENA, we retrieved the values from the sample XML. Since 2018, 47 302 cases consistently provided values for all three parameters, corresponding to 16% of all cases submitted during that time period, and following an increasing trend with 27% of all cases in 2020. This trend counteracted the decreasing use of environment_biome, environment_material, and environment_feature over the investigated time period, resulting in stable proportions of approximately 35% of the cases per year archived, regardless of MixS checklist usage, with either of these two sets of environmental descriptors since 2016. Lack of compliance with ENVO terminology was also observed here, especially for the parameter env_broad_scale (data not shown).

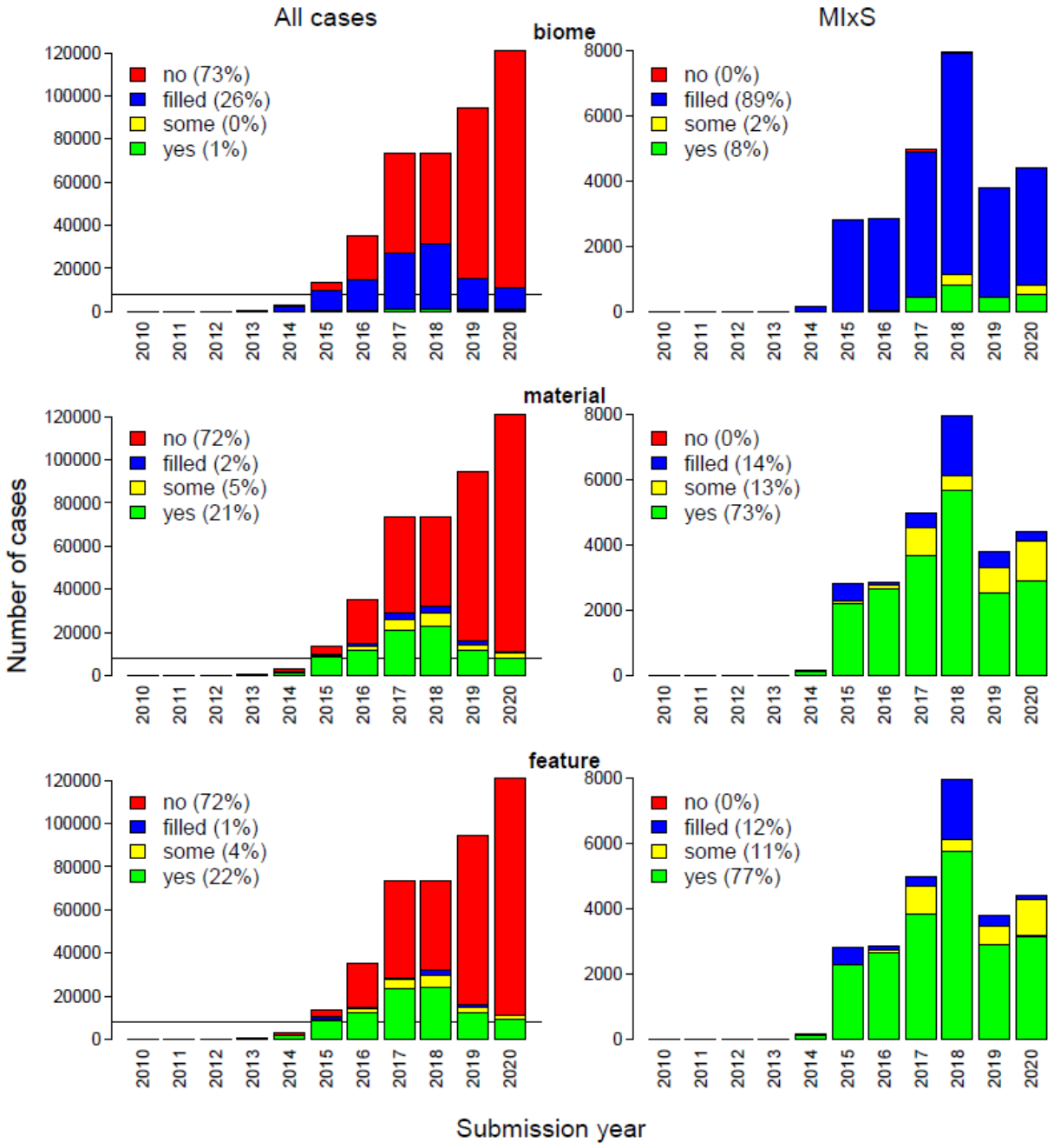


Figure 3: Number of cases (amplicon example) with values for environment_biome, environment_material, and environment_feature according to ENVO. Yes: exact match to ENVO term ID; some: character string match to ENVO term name (including matches after extensive character string manipulation); filled: a value is provided, but not ENVO-compatible; no: no entry. The percentages summarize cases over all years and are rounded to integers. The plots on the right only show the cases with samples submitted according to a MIxS checklist and environmental package. The horizontal line in the plots on the left indicates the y-axes range of the plots on the right.

Use of brokerage services: Since 2016, 1475 cases have been submitted via GFBio, making this brokerage service the most frequently used (as assessed by the number of cases) throughout the whole investigated time period, closely followed by CNSA (China Nucleotide Sequence Archive). Here, we briefly highlight the effect the use of this brokerage service had on the metadata supplied for the above parameters. We found that the quality of the metadata, specifically accessibility and interoperability, were considerably improved in submissions via GFBio (SI table 1): all cases (with the exception of 95 cases from one study) provided latitude and longitude data retrievable from the TSV search output of ENA and also used the correct format and terminology for the parameters environment_biome, environment_material, and environment_feature. Furthermore, all cases (without exception) contained nominal length data, although the suspicious peaks in the data distribution at 300bp and 500bp were still present. If information about the amplified gene was provided (26% of the cases), this was mainly accessible in the TSV search output of ENA (24%), and restricted to two values: 16S rRNA and 18S rRNA.

Metagenomic sequencing data: Until the end of 2020, 29 253 cases had been submitted with library_selection “RANDOM” and library_strategy “WGS” in an analogous example to the amplicon data sets that we explored above. The number of cases submitted each year has been consistently increasing, with a spike in 2019 caused by one study with an exceptionally high number of associated runs. Nonetheless, the proportion of WGS of the total number of amplicon and WGS cases in our particular example has been declining over the last decade, reaching a mostly stable value at approximately 5% since 2017 (SI table 1). Regarding the investigated metadata parameters, similar trends were observed in the WGS data compared to the amplicon example, although overall metadata quality was slightly higher (SI figures 1 and 2). Specifically, this improvement was related to the more frequent usage of MIxS checklists for in total 14% of all WGS cases (SI table 1). These cases associated with samples submitted in compliance with MIxS displayed an almost perfect track record for the parameters latitude and longitude, nominal length, environment_feature and environment_material in terms of metadata interoperability and consequently reusability.

Case study: In a recent study (Molari et al. in prep.), the raw reads archived on ENA from 39 studies using paired-end Illumina amplicon sequencing of the V3-V4 hypervariable region of the bacterial 16S rRNA gene were downloaded and bioinformatically processed to generate quality-trimmed merged fasta files to be used for oligotyping (Eren et al., 2014). This analysis required that the sequences were generated from the exact same gene region to be comparable.

Information about the primers used in the sequencing library preparation were obtained from associated publications or the submitters directly. After download, the raw read data was checked for compliance with ENA submission requirements, i.e. that paired-end Illumina reads were archived as demultiplexed, unmerged forward and reverse reads, without artificial sequences and prior to any quality trimming. Of the 39 inspected studies, only eight were submitted as required. The majority (28 studies) did not remove the primer sequences, eight studies contained already merged sequences, and one study even provided only the sequencer output prior to demultiplexing (sample barcode information had to be obtained from the author). This data mining experience showed that even if metadata that enables the findability and accessibility of the data is provided, the raw read data itself may often not be submitted as required, making manual checks mandatory and limiting the interoperability and reusability of the data.

As we investigated each of these studies in detail, comparing the information provided in the sample and study description as well as the associated publication (if available), we also checked the values for the logical (boolean) sample metadata parameter environmental_sample. This parameter "identifies sequences derived by direct molecular isolation from an environmental DNA sample" (ENA Portal API). This description applied to the samples from all 39 studies, however all samples were declared as non-environmental. Based on this observation, we revisited the metadata inspected in the current study: Of the 413 849 cases of amplicon data less than 2% were marked as originating from environmental samples. This seemed unlikely, although we were not able to check all submissions manually for the correctness of this parameter. Specifically, we found it paradoxical that none of the cases submitted according to a MIxS environmental package were actually declared as originating from environmental samples. If our suspicions about the incorrect use of this parameter were confirmed, it would make this metadata parameter unsuited for selecting data for reuse in data mining endeavors.

Lastly, we assessed the availability of a PubMed record retrievable via the ENA XREF API for the ENA study accessions in the case study, as such a publication may provide further information about a data set than available in the metadata on ENA. Associated publications were retrievable for 20 of the 39 investigated studies. Of the remaining 19, publications were found manually for 14. This shows the limitations of the automated approach using XREF. To improve metadata completeness, publications would have to be linked manually to the ENA study upon request by the author.

DISCUSSION

All investigated cases met the criteria for metadata findability and accessibility since those were a prerequisite of conducting this study. However, interoperability and therefore reusability remained a challenge. Overall, our results revealed a high variability in data and metadata interoperability and reusability on ENA for the particular examples relevant for molecular ecology. Laudably, with few exceptions, geographic coordinates (latitude and longitude) were always provided, and mostly available as indexed (searchable and returnable) parameters. Beyond the scientific relevance, geographic information about the sample and, by extension, sequence origin is essential for equitable Access and Benefit Sharing and the key parameter to linking sequence-based biodiversity observations to the Ocean Biodiversity Information System (OBIS), the Global Biodiversity Information Facility (GBIF), and other platforms for biodiversity assessment (Bax et al., 2019; Buttigieg et al., 2018; Canonico et al., 2019). However, other mandatory metadata parameters (nominal length) or those often crucial for the reuse of the data (target gene, classification as environmental sample, environment description) showed a very low interoperability, if values were provided at all.

While the MIxS metadata checklists have been established a decade ago (Yilmaz et al., 2011), they were only infrequently used despite their evident positive impact on metadata quality and consequently the reusability of the data. Furthermore, despite an increasing number of calls to action (Reiser, Harper, Freeling, Han, & Luan, 2018; Ryan et al., 2020; Stevens et al., 2020), the use of this community standard has been declining over the last years, especially for amplicon sequencing data in the selected example. The number of data sets being submitted each year is expected to continue to rise considering decreasing sequencing costs and the undiminished popularity of amplicon sequencing for multiple applications in molecular and microbial ecology and biodiversity research (e.g. environmental DNA studies). Therefore, it is even more worrisome that the use of standards in data submissions has not been following this same upward trend. We further noticed that even when data was submitted according to MIxS, this metadata standard was often not used as intended or to its full potential. In part, this situation may have arisen from inconsistencies in the documentation about metadata requirements provided by separate resources. For instance, the MIxS checklists are not implemented in their entirety by INSDC due to a lack of demand to archive such parameters. Especially conditionally mandatory parameters, such as target_gene, target_subfragment, and pcr_primers in the case of amplicon data (i.e. MIMARKS standard), are listed as optional for the MIxS checklists on ENA, the latter two also not

being indexed as searchable or returnable parameters. Furthermore, the description of the parameters environment_biome, environment_material, environment_feature in the ENA documentation specifies free text, whereas MIxS specifies the use of a controlled vocabulary and data syntax. In such cases, the more stringent standard should be communicated and adhered to in compliance with the original standard description.

Other issues, which we did not explore in more detail here, included duplicated data submissions, contradictory primer references, and conflicting metadata entries for the same run. The latter is especially difficult to track and often only detectable after a manual check of the data, metadata, and publication. In some such instances, we discovered contradictory entries for the sequencing method, instrument model, library selection, sample and library names, NCBI taxonomy ID (tax_id), and MIxS environmental package and checklist, often hidden only among the non-indexed parameters. Additional to deficient metadata, based on our case study (Molari et al. in prep.) we further suspect that a large proportion of raw read amplicon data (fastq files) has not been archived correctly according to the ENA submission guidelines. Our study adds another facet to the increasing body of work illustrating the deficits in nucleotide sequence databases (Eckert et al., 2020; Hoopen et al., 2016; Jurburg et al., 2020) with drastic consequences: Terabytes to Petabytes of data may not be readily interoperable and reusable, severely limiting their added value, long-term impact, and future relevance.

While ongoing development of standards and their integration across disciplines[5] is an essential endeavor to increase the added-value by standard-compliant (meta)data, we think that it is crucial to avoid further delays in improving (meta)data quality and FAIRness by making better use of existing standards. This task is up to each individual researcher, to voluntarily use more stringent checklists and provide optional parameters. Brokerage services, such as GFBio, fundamentally improved metadata quality and therefore data reusability, but are too personnel-intensive to solve all challenges described here. Luckily, the number of tools, platforms, and tutorials to inform and facilitate sustainable data management has increased rapidly over the last two years (Olsson & Hartley, 2019; Quiñones et al., 2020; Riginos et al., 2020; Sansone et al., 2019). However, to encourage such initiatives, primarily a shift in the recognition and scientific value system is required to provide incentives for proper data archival and publication, including standardized

[5] https://www.tdwg.org/community/gbwg/MIxS/

metadata, to enable long-term reuse of the data (Riginos et al., 2020; Sansone et al., 2019; Westoby, Falster, & Schrader, 2021).

In the following we provide a (non-exhaustive) list of suggestions to address some of the most acute deficits of data and metadata FAIRness for nucleotide sequence data from the perspective of molecular ecologists. We hope that they are easy to implement and can have a potentially large positive impact on the research fields relying on such data as well as derived applications in biodiversity and conservation policy and management strategies.

Suggestions @researchers:

- Make use of existing checklists and data brokerage services, and use checklists beyond mandatory parameters. For instance, the MIxS parameters target_gene, target_subfragment, and pcr_primers should be supplied for all sequencing read data that was generated with library_selection="PCR" AND library_strategy="AMPLICON".
- Enter data diligently and according to the specified format to facilitate interoperability. It is not only important **that** (meta)data is archived, but also **how**.
- Whenever possible, use ontologies and actively contribute to the improvement of ontologies by suggesting so far missing terms to the ontology developers. Many ontologies, like ENVO, have a low-threshold for suggesting a new term, e.g. opening an issue on GitHub, and provide guidance on using ontology terms in the context of MIxS[6].
- Update data submissions if additional information (manuscript DOI, accession numbers of related data sets) becomes available.

Suggestions @research institutions

- Invest in capacity development and training of early career researchers to avoid incorrect data (and associated metadata) submissions due to inexperience, and raise awareness for the persistently high value of FAIR data. Data archiving is not a trivial task, and it is at least as important as manuscript publications.
- Incentivization of good data management through recognition of data publications towards career progression metrics.

[6] https://github.com/EnvironmentOntology/envo/wiki/Using-ENVO-with-MIxS

Suggestions @funding agencies

- Funding as an incentive for good data management to promote a stronger inclusion of data and information stewardship in scientific projects (e.g. data management plan as prerequisite as already implemented by several funding agencies)
- Allocate additional funding for data technicians and data managers for the successful implementation of data management plans.

Suggestions @reviewers:

- Review the submitted data and metadata for a scientific manuscript as thoroughly as the manuscript text. FAIR data archiving should be as important a criterion for manuscript publication as scientific soundness.

Suggestions @publishers

- If feasible, make data availability statements (accession numbers) accessible outside access restrictions, so that publications can be more easily and automatically linked to the data sets.

Suggestions @databases:

- Implement automated checkpoints for data consistency, specifically related to the use of controlled vocabulary (ontologies) and empty entries for mandatory parameters (e.g. nominal length).
- Harmonize the documentation about the parameters that use a controlled vocabulary (ontologies) across the different resources (MIxS, ENA) by choosing the more stringent standard.
- Expand the list of indexed metadata (MIxS) parameters in a concerted effort with the scientific community to promote stronger adherence to more extensive checklists and standards.
- Facilitate easy access to embargoed data sets for reviewing purposes of data and metadata.

Acknowledgements

We would like to thank Pier Luigi Buttigieg for inspiring discussions about data management and data mining and the FAIRness of it all.

References

Bax, N. J., Miloslavich, P., Muller-Karger, F. E., Allain, V., Appeltans, W., Batten, S. D., … Tyack, P. L. (2019). A response to scientific and societal needs for marine biological observations. *Frontiers in Marine Science*, *6*(JUL), 1–22. doi: 10.3389/fmars.2019.00395

Buttigieg, P. L., Fadeev, E., Bienhold, C., Hehemann, L., Offre, P., & Boetius, A. (2018). Marine microbes in 4D — using time series observation to assess the dynamics of the ocean microbiome and its links to ocean health. *Current Opinion in Microbiology*, *43*, 169–185. doi: 10.1016/j.mib.2018.01.015

Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., & Lewis, S. E. (2013). The environment ontology: Contextualising biological and biomedical entities. *Journal of Biomedical Semantics*, *4*(1), 1–9. doi: 10.1186/2041-1480-4-43

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), 1–12. doi: 10.1186/s13326-016-0097-6

Canonico, G., Buttigieg, P. L., Montes, E., Muller-Karger, F. E., Stepien, C., Wright, D., … Murton, B. (2019). Global observational needs and resources for marine biodiversity. *Frontiers in Marine Science*, *6*(JUL), 1–20. doi: 10.3389/fmars.2019.00367

Cook, C. E., Bergman, M. T., Cochrane, G., Apweiler, R., & Birney, E. (2018). The European Bioinformatics Institute in 2017: Data coordination and integration. *Nucleic Acids Research*, *46*(D1), D21–D29. doi: 10.1093/nar/gkx1154

Diepenbroek, M., Glöckner, F. O., Grobe, P., Güntsch, A., Huber, R., König-Ries, B., … Triebel, D. (2014). Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform : The German Federation for the Curation of Biological Data (GFBio). *Informatik 2014 – Big Data Komplexität Meistern. GI-Edition: Lecture Notes in Informatics (LNI) - Proceedings*, 1711–1724.

Eckert, E. M., Di Cesare, A., Fontaneto, D., Berendonk, T. U., Bürgmann, H., Cytryn, E., … Corno, G. (2020). Every fifth published metagenome is not available to science. *PLoS Biology*, *18*(4), 1–7. doi: 10.1371/journal.pbio.3000698

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., & Sogin, M. L. (2014). Minimum entropy decomposition : Unsupervised oligotyping for sensitive partitioning of high- throughput marker gene sequences. *The ISME Journal*, *9*(4), 968–979. doi: 10.1038/ismej.2014.195

Field, D., Amaral-Zettler, L., Cochrane, G., Cole, J. R., Dawyndt, P., Garrity, G. M., … Wooley, J. (2011). The Genomic Standards Consortium. *PLoS Biology*, *9*(6), 8–10. doi: 10.1371/journal.pbio.1001088

Fillinger, S., de la Garza, L., Peltzer, A., Kohlbacher, O., & Nahnsen, S. (2019). Challenges of big data integration in the life sciences. *Analytical and Bioanalytical Chemistry*, *411*(26), 6791–6800. doi: 10.1007/s00216-019-02074-9

Fukuda, A., Kodama, Y., Mashima, J., Fujisawa, T., & Ogasawara, O. (2021). DDBJ update: Streamlining submission and access of human data. *Nucleic Acids Research*, *49*(D1), D71–D75. doi: 10.1093/nar/gkaa982

Harrison, P. W., Ahamed, A., Aslam, R., Alako, B. T. F., Burgin, J., Buso, N., … Cochrane, G. (2021). The European Nucleotide Archive in 2020. *Nucleic Acids Research*, *49*(D1), D82–D85. doi: 10.1093/nar/gkaa1028

Hoopen, P. Ten, Amid, C., Buttigieg, P. L., Pafilis, E., Bravakos, P., O-Tárraga, A. M. C., … Cochrane, G.

(2016). Value, but high costs in post-deposition data Curation. *Database*, *2016*, 1–10. doi: 10.1093/database/bav126

Jurburg, S. D., Konzack, M., Eisenhauer, N., & Heintz-Buschart, A. (2020). The archives are half-empty: an assessment of the availability of microbial community sequencing data. *Communications Biology*, *3*(1). doi: 10.1038/s42003-020-01204-9

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., … Parkinson, H. (2010). Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, *26*(8), 1112–1118. doi: 10.1093/bioinformatics/btq099

NCBI Resource Coordinators. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *46*(D1), D8–D13. doi: 10.1093/nar/gkx1095

Olsson, T. S. G., & Hartley, M. (2019). Lightweight data management with dtool. *PeerJ*, *2019*(3). doi: 10.7717/peerj.6562

Quiñones, M., Liou, D. T., Shyu, C., Kim, W., Vujkovic-Cvijin, I., Belkaid, Y., & Hurt, D. E. (2020). "mETAGENOTE: A simplified web platform for metadata annotation of genomic samples and streamlined submission to NCBI's sequence read archive." *BMC Bioinformatics*, *21*(1), 1–12. doi: 10.1186/s12859-020-03694-0

Reiser, L., Harper, L., Freeling, M., Han, B., & Luan, S. (2018). FAIR: A Call to Make Published Data More Findable, Accessible, Interoperable, and Reusable. *Molecular Plant*, *11*(9), 1105–1108. doi: 10.1016/j.molp.2018.07.005

Riginos, C., Crandall, E. D., Liggins, L., Gaither, M. R., Ewing, R. B., Meyer, C., … Deck, J. (2020). Building a global genomics observatory: Using GEOME (the Genomic Observatories Metadatabase) to expedite and improve deposition and retrieval of genetic data and metadata for biodiversity research. *Molecular Ecology Resources*, *20*(6), 1458–1469. doi: 10.1111/1755-0998.13269

Ryan, M., Schloter, M., Berg, G., Kinkel, L. L., Eversole, K., Macklin, J. A., … Sessitsch, A. (2020). Towards a unified data infrastructure to support European and global microbiome research- A call to action. *Environmental Microbiology*, *00*, 1–4. doi: 10.1111/1462-2920.15323

Sansone, S. A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A. L., & Thurston, M. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, *37*(4), 358–367. doi: 10.1038/s41587-019-0080-8

Stevens, I., Mukarram, A. K., Hörtenhuber, M., Meehan, T. F., Rung, J., & Daub, C. O. (2020). Ten simple rules for annotating sequencing experiments. *PLoS Computational Biology*, *16*(10), 1–7. doi: 10.1371/journal.pcbi.1008260

Westoby, M., Falster, D. S., & Schrader, J. (2021). Motivating data contributions via a distinct career currency. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1946). doi: 10.1098/rspb.2020.2830

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 1–9. doi: 10.1038/sdata.2016.18

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, 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), 415–420. doi: 10.1038/nbt.1823

595 **Supplement**

596

597 **SI table 1:** Number of amplicon and WGS cases (runs) submitted per year, separated by the
598 availability of various metadata parameters, use of MIxS environmental package and GFBio as
599 brokerage service. Metadata availability (Category) as explained in figures 1 and 2.
600
601 **SI figure 1:** Number of cases (WGS example) with and without metadata available for geographic
602 coordinates (latitude, longitude) and nominal length. For geographic coordinates, 'tsv' is referring
603 to the information provided for indexed metadata parameters in the TSV search output, while 'xml'
604 is referring to non-indexed metadata only accessible in the XML view of the ENA sample or
605 experiment. The percentages summarize cases over all years and are rounded to integers. The
606 plots on the right only show the cases with samples submitted according to a MIxS checklist and
607 environmental package. The horizontal line in the plots on the left indicates the y-axes range of
608 the plots on the right.
609
610 **SI figure 2:** Number of cases (WGS example) with values for environment_biome,
611 environment_material, and environment_feature according to ENVO. Yes: exact match to ENVO
612 term ID; some: character string match to ENVO term name (including matches after extensive
613 character string manipulation); filled: a value is provided, but not ENVO-compatible; no: no entry.
614 The percentages summarize cases over all years and are rounded to integers. The plots on the
615 right only show the cases with samples submitted according to a MIxS checklist and
616 environmental package. The horizontal line in the plots on the left indicates the y-axes range of
617 the plots on the right. The category 'filled' is overrepresented in the year of 2019 for 'biome' and
618 'material', mainly due to runs from a single study with more than 2000 runs.