OPEN LETTER

REVISED COPO - Managing sample metadata for biodiversity: considerations from the Darwin Tree of Life project [version 2; peer review: 1 approved, 1 approved with reservations, 1 not approved]

Felix Shaw[1], Alice Minotto[1], Seanna McTaggart[1], Aaliyah Providence[1], Peter Harrison[2], Joana Paupério[2], Jeena Rajan[2], Josephine Burgin[2], Guy Cochrane[2], Estelle Kilias[3], Mara K.N. Lawniczak[4], Robert Davey[1]

[1]Earlham Institute, Norwich, Norfolk, NR4 7UH, UK
[2]EMBL European Bioinformatics Institute, Hinxton, Cambridgeshire, CB10 1SD, UK
[3]Department of Zoology, University of Oxford, Oxford, Oxfordshire, OX1 2JD, UK
[4]Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1RQ, UK

v2 **First published:** 10 Nov 2022, **7**:279
https://doi.org/10.12688/wellcomeopenres.18499.1
Second version: 31 Jul 2023, **7**:279
https://doi.org/10.12688/wellcomeopenres.18499.2
Latest published: 10 Jun 2024, **7**:279
https://doi.org/10.12688/wellcomeopenres.18499.3

Abstract

Large-scale reference genome sequencing projects for all of biodiversity are underway and common standards have been in place for some years to enable the understanding and sharing of sequence data. However, the metadata that describes the collection, processing and management of samples, and link to the associated sequencing and genome data, are not yet adequately developed and standardised for these projects. At the time of writing, the Darwin Tree of Life (DToL) Project is over two years into its ten-year ambition to sequence all described eukaryotic species in Britain and Ireland. We have sought consensus from a wide range of scientists across taxonomic domains to determine the minimal set of metadata that we collectively deem as critically important to accompany each sequenced specimen. These metadata are made available throughout the subsequent laboratory processes, and once collected, need to be adequately managed to fulfil the requirements of good data management practice.

Due to the size and scale of management required, software tools are needed. These tools need to implement rigorous development pathways and change management procedures to ensure that effective research data management of key project and sample metadata is maintained. Tracking of sample properties through the sequencing process is handled by Lab Information Management

Open Peer Review

Approval Status ✓ ? ✗

	1	2	3
version 3 (revision) 10 Jun 2024			
version 2 (revision) 31 Jul 2023	✓ view		✗ view
version 1 10 Nov 2022	✓ view	? view	

1. **Katrina Exter**, Vlaams Instituut voor de Zee, Ostend, Belgium
2. **Birgitta König-Ries**, University of Jena, Jena, Germany
3. **Eric Darvish Crandall**, The Pennsylvania State University - University Park Campus, University Park, USA

Systems (LIMS), so publication of the sequenced data is achieved via technical integration of LIMS and data management tools.

Discussions with community members on how metadata standards need to be managed within large-scale programmes is a priority in the planning process. Here we report on the standards we developed with respect to a robust and reusable mechanism of metadata collection, in the hopes that other projects forthcoming or underway will adopt these practices for metadata.

Keywords
biodiversity, standards, sharing, metadata, samples, Darwin Tree of Life, taxonomic domains, data management, LIMS, metadata standards

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Tree of Life gateway.

Corresponding author: Felix Shaw (felixoshaw@gmail.com)

Author roles: Shaw F: Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Minotto A**: Investigation, Methodology, Project Administration, Software, Validation, Writing – Review & Editing; **McTaggart S**: Project Administration, Writing – Review & Editing; **Providence A**: Software, Writing – Review & Editing; **Harrison P**: Resources, Software; **Paupério J**: Data Curation, Software, Writing – Review & Editing; **Rajan J**: Data Curation, Software; **Burgin J**: Data Curation, Software, Writing – Review & Editing; **Cochrane G**: Data Curation, Writing – Review & Editing; **Kilias E**: Data Curation, Investigation, Methodology; **Lawniczak MKN**: Conceptualization, Funding Acquisition, Investigation, Methodology, Writing – Review & Editing; **Davey R**: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The Darwin Tree of Life Project is funded by Wellcome. This work was supported in whole, or in part, by Wellcome [206194, https://doi.org/10.35802/206194 and 218328, https://doi.org/10.35802/218328]. RPD, AM and FS are funded in part by the UK Biotechnology and Biological Sciences Research Council (BBSRC) Core Strategic Programme grant awarded to the Earlham Institute (BBS/E/T/000PR9817). COPO is hosted within CyVerse UK (https://cyverseuk.org), a research cloud computing platform funded through the National Capability for e-Infrastructure grant awarded to the Earlham Institute (https://www.earlham.ac.uk/national-capability-e-infrastructure). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Shaw F, Minotto A, McTaggart S *et al.* **COPO - Managing sample metadata for biodiversity: considerations from the Darwin Tree of Life project [version 2; peer review: 1 approved, 1 approved with reservations, 1 not approved]** Wellcome Open Research 2023, **7**:279 https://doi.org/10.12688/wellcomeopenres.18499.2

First published: 10 Nov 2022, **7**:279 https://doi.org/10.12688/wellcomeopenres.18499.1

REVISED **Amendments from Version 1**

This version has some minor changes to the text, a few clarifications, and an improved figure on the data model of COPO and Biosamples.

Any further responses from the reviewers can be found at the end of the article

Introduction

Datasets are fundamental assets in the life sciences, where knowledge is routinely extracted and interpreted on the genomic scale (International Society for Biocuration, 2018). The robustness of sequencing experiments relies on the accuracy of measurements regarding accompanying sample observations. Much effort goes into the production and validation of methods and models for data, with mechanisms for sharing this information being largely standardised. For example, consensus was quickly reached for including quality information into the file format used for sharing sequence data (FASTQ). Where human record-keeping is the deciding factor in the quality of information, there has been far less formal standardisation. The field of biocuration is an essential part of the research data lifecycle in order to cope with the quantities and complexities of modern data generation (Howe *et al.*, 2008). However, it can be an expensive and onerous process to produce, implement and adopt such a standard within a community, to apply *post hoc* curation, and to provide associated data storage and tools (Kennan & Markauskaite, 2015), (Whitmire *et al.*, 2015) (ten Hoopen *et al.*, 2016).

For software, data not provided in the correct format can cause runtime errors and erroneous results. Incorrect or absent metadata, though, may not necessarily break a system in the same way, but accurate metadata is a vital layer of context providing understanding for humans, and interoperability for machines. Therefore, metadata standardisation is equally as important as data standardisation. Still, there is often no requirement for tools, software, and analytical processes to check metadata quality, and little support given to scientists who have to carry out data curation.

The DToL project oversees the collection of diverse species from a wide range of habitats with the goal of producing draft genomes that are immediately available to the community. Recent articles have highlighted the importance of standardisation and formed a concise set of rules for researchers to follow in their sequencing experiments (Stevens *et al.*, 2020). Ensuring that future studies based on DToL reference genomes (such as resequencing), can be contextualised against a range of observable variables is essential. Therefore standardised and accurate recording of properties such as identifiers, taxonomic information, lifestage, body parts, labs and organisations involved, collectors and collection events, collection environments, identification uncertainties, hazard groups, preservation, barcoding, regulatory compliance and vouchering is key. In our sister paper, we describe the processes and procedures set up by DToL to ensure consistency and standardisation of metadata across the project (Lawniczak *et al.*, 2022).

Briefly, DToL samples are collected by and processed within the oversight of a Genome Acquisition Laboratory (GAL). Taxonomic experts who have in-depth knowledge of their research organisms, their respective habitats and/or historical collections, prepare the original specimen into sequencing-ready material. Whilst this may involve different laboratory processes, the same endpoint must be reached. A sample must be collected with the properties required for sequencing at the required depth and coverage to produce a reference quality genome. To do so, the DToL project set up a Samples Working Group to bring together researchers representing all GALs as well as six different eukaryotic taxonomic areas. These are plants, arthropods, lichens and fungi, Chordata, Protista, microalgae and other Metazoa (mainly comprising non-arthropod invertebrates). The members of the Samples Working Group meet twice monthly and are tasked with developing the target list of priority species, standardising metadata collection for the project, and developing, refining, and standardising collection procedures for these different taxonomic groups. This group also comprises researchers who will receive these physical samples for subsequent preparation into sequencing libraries, and is charged with developing protocols for sample storage, delivery, and compliance.

Sample manifest development

A project of the scale and ambition of DToL calls for a substantial effort in developing, monitoring and refining its metadata collection and management infrastructure. We informed our strategy through the alignment with existing specifications in the biodiversity domain, such as the Biodiversity Information Standards (formerly Taxonomic Databases Working Group, TDWG) Darwin Core. Bringing together past experiences in developing and implementing metadata standards from the Wellcome Trust (WT) Sanger, Collaborative OPen Omics (COPO) and European Nucleotide Archive (ENA) teams, in particular community-led collaborative work on such standards as MiXS (Yilmaz *et al.*, 2011) and CG Core (*GitHub - AgriculturalSemantics/cg-core: CG Core Metadata Reference Guide*, 2021)), the DToL team focused its efforts into its sample manifest infrastructure.

As outlined in our recent article (Lawniczak *et al.*, 2022), the DToL specimen collectors need to record information associated with the location and time a specimen was taken for the project. Each specimen is required to represent (if possible) a single genetic entity or individual, and digital identification of these specimens and subsequent samples must be accurately represented in data management tools. Multiple samples may be taken from the same specimen (*e.g.* different tissues from a large animal are put into different tubes) and this information must be tracked (see "Relationships between samples" below). Each tube containing a sample is assigned a SPECIMEN_ID, which is a unique identifier generated by the GAL, intending to reflect the genetic identity of the organism contained within it; symbionts, contamination and co-occurring cultures notwithstanding. Tubes also get a TUBE_OR_WELL_ID, which is the barcode stamped on the tube and represents a sample of that specimen. Some

organisms are too small to be collected in tubes and this is reflected in the name "TUBE_OR_WELL_ID". This also gives future scope for the more widespread collection and processing of samples in plates and wells. In either case, each individual sample can be identified by the concatenation of RACK_OR_PLATE_ID and TUBE_OR_WELL_ID. When two or more samples are taken from the same specimen (*e.g.* insect blood in one tube, a leg in another), any difference between these is captured in further fields. Cultured protists are presumed to be processed in the same way as other environmental samples. In this case, an ENV_SAMPLE_ID will be assigned to the sample, and this identifier will also be referred to in the derived single cell samples together with the SPECIMEN_ID. Unculturable protists and other organisms may be collected for both metagenomic analysis and single cell sequencing.

The DToL Samples Working Group identified the initial information that would describe the sample collection events to form a *sample manifest* specification. The sample manifest needed to be sufficient to cover common metadata across all taxonomic groups. This resulted in a single core schema document that comprises the set of metadata fields that need to be provided by a GAL before the sample material is accepted for sequencing (Lawniczak *et al.*, 2022). In the DToL project, the sample manifest is the initiating step in the tracking of the sample from collection to sequencing to release of the data in the public repositories (Cummins *et al.*, 2022) and further display on the DToL Data Portal (*Darwin Tree of Life*, 2022). In this way, the DToL sample manifest is suitable as a starting point for use in programmes associated with the Earth Biogenome Project, and potentially others, which will require significant organisation and collaboration to ensure comparison and reuse of the vast quantities of data that will be produced in the coming years.

To promote transparency, openness and to enable controlled versioning of the standards, the core manifest and standard operating procedure (SOP) are publicly available from the DToL GitHub repository.

When there are metadata divergences depending on the taxonomic groups, for example where samples are collected in wells or tubes depending on their physical size, these can often be mitigated by common fields. For example, we use PLATE_OR_RACK_ID and TUBE_OR_WELL_ID, respectively, to ensure that we have a consistent identification strategy.

In other cases where different fields are required for accurate metadata modelling, the DToL manifests can act as a basis to develop "extension" manifests to cover metadata fields that are specific to a single taxonomic group. For example, marine researchers may be concerned with the salinity, depth and pH of the seawater at the point a sample is collected, but avian researchers are unlikely to record these properties. Modelling these differences is important to meet community requirements in these varying scenarios, but it is equally important to have a firm idea of core fields that will be appropriate for all scenarios.

The workflow of agreed changes via the DToL Samples Working Group is documented in shared Google Documents, and the list of proposed changes is placed as a link to a separate document in the SOP. After the initial phase of the project it was agreed the manifest would be updated twice a year. This is both because the list of agreed metadata to collect is expected to become more stable with the project reaching maturity, and because a number of services need to implement and deploy changes in concert with each other. Depending on the type of changes required, updates can take from a few days for simpler issues to over a month for more complex changes. This makes frequent updates of the SOP problematic to manage, develop and deploy. The SOPs and Manifests are versioned and published on GitHub (*GitHub - darwintreeoflife/metadata*, 2023). When querying COPO for samples, the manifest version under which the samples were submitted is returned along with the sample metadata. This means users are able to determine which fields will be present. In some cases, samples may be retroactively updated in response to a manifest update, for example when there is a change of field name. However these types of changes are highly discouraged.

The Darwin Tree of Life project is affiliated with the Earth Biogenome Project (Lewin *et al.*, 2022). Other EBP projects such as ASG (Aquatic Symbiosis Genomics project) and ERGA (European Reference Genome Atlas) are currently building on DToL developments and are basing SOPs and processes on the DToL manifest. This will require the addition or removal of fields and controlled vocabulary terms as necessary. In both cases, it was agreed to follow the same timeline and to keep up to date with DToL updates given the overlap between the projects. This makes maintenance of software tools which handle these data types significantly less burdensome.

Once a version has been agreed upon, both the SOP and the sample manifest are openly published and versioned within the DToL GitHub repository so that changes can be implemented. Once a change is accepted and made to the SOP and/or manifest, developers of downstream information systems are informed and work can begin. Frequent changes include modification of column headers, addition of controlled vocabulary fields, and changing validation code so that verified submission to public repositories is maintained. The timeframe required by each type of change is approximately known so that there is a clear understanding of the process leading to a new release.

Where possible, we have mapped our manifest headers to TWDG Darwin Core terms in order to comply with the global standard for capturing occurrences and events around biodiversity monitoring. For example:

- DATE_OF_COLLECTION [http://rs.tdwg.org/dwc/terms/verbatimEventDate]

- DECIMAL_LATITUDE [http://rs.tdwg.org/dwc/terms/decimalLatitude]
- TUBE_OR_WELL_ID [http://rs.tdwg.org/dwc/terms/measurementID]

Such mappings will allow us, in the future, to submit our records into these international databases and start to link genomic sequence information alongside sample metadata and images.

EMBL-EBI ENA checklist

Data and associated metadata are archived in the ENA (Cummins *et al.*, 2022,). There is a minimum amount of information required for registering sample metadata in the ENA and that is defined through the sample checklists. Different sample checklists have been developed to validate the requirements on the minimum metadata needed to describe biological samples submitted by different research communities. ENA staff work actively with members of these communities – often through such initiatives as the Genome Standards Consortium (Field *et al.*, 2011) to capture appropriate requirements, understand working practices, and to integrate and/or map concepts across the different domains of life science. We produced an ENA sample checklist that reflects a subset of the DToL schema, called Tree of Life Checklist (ToL). The checklist was designed to validate all the metadata provided by the project which were deemed useful for public reuse and interpretation, and excludes metadata which are internal to project tracking. The checklist also aligns the metadata collected for the DToL manifest against other existing standards (for example the MIxS standards (Yilmaz *et al.*, 2011)) to allow comparison with samples outside of the project. Both metadata and data can therefore be represented in this public repository in a standardised way. This allows all ToL datasets to be coherent across sample collection environments, sequencing protocols and machines, and scientific institutions. This also provides a minimal standard for any other datasets based on the DToL manifests which are not part of DToL, and form the backbone of other important international projects, *e.g.* ERGA and EBP. This is a significant contribution to open and FAIR scientific data, and we hope that the checklist will help other groups and programmes develop submission pathways that follow similar guidelines, improving reproducibility and consistency across a wider range of biodiversity projects.

Metadata brokering

The timely and accurate submission of data and metadata to public repositories is a significant requirement for transparency in the life sciences (Gonzalez & Peres-Neto, 2015), and the discipline of Research Data Management is gaining traction with funders, publishers, and researchers themselves (Kennan & Markauskaite, 2015). However, researchers find it costly and difficult to transform information collected in the field or lab into consistent metadata that is suitable for meeting FAIR requirements (Wilkinson *et al.*, 2016). The generic nature of existing submission routes where much of the metadata is not mandatory or unvalidated (somewhat mitigated by repository checklists) is one hurdle. That the submission systems themselves are not tailored to specific communities and therefore can be difficult to navigate for new users is another. COPO is a mature, actively developed web-based brokering system for annotating and depositing datasets to a number of public repositories (Shaw *et al.*, 2020), and is able to be configured to fit with the needs of specific communities. COPO is used for brokering metadata between GALs and the ENA. Other services such as the LIMS at the Wellcome Sanger Institute pulls information from the COPO application programming interface (API) to insert into its own databases for downstream sample tracking in the lab. COPO is available for any researcher to use, but has also been developed to fulfil the needs of the DToL project, as well as other EBP-related projects such as ASG and ERGA.

DToL users are assigned one of two groups in COPO. The Sample Submitters group is for users to upload their manifests, and this group is commonly made up of sample collectors themselves. The Sample Supervisors group is for users who will provide supervisory oversight and will accept or reject samples based on human curation post-validation, and these users often work at the GAL where the samples will be received. They will be presented with a view of all DToL samples needing approval. Rejected samples are held back within the system. If samples are accepted, they will be queued for submission to ENA, where they will be further validated against the ENA ToL checklist, and assigned ENA and BioSample accessions (Courtot *et al.*, 2022).

DToL *Sample Supervisors* work with *Sample Submitters* and show them the sample manifest, assisting with any questions about its completion. The sample submitters/collectors fill in and submit the DToL Sample Manifest to COPO in the form of a tabular file, in Excel (CSV files are also accepted, but users prefer the more familiar and navigable format of spreadsheets). The file has many fields with drop down menus to help standardise spellings and terms as this is a common place for metadata tracking to diverge and can be difficult to resolve later. For example, collectors may refer to the sex of a specimen as F, f, Fem, FEMALE, etc. The use of data validation in the Excel helps avoid these issues for any term in which a relatively small set of defined entries is expected. We have used data validation for as many fields as possible, and the terms that are part of the subset of metadata submitted to ENA are mapped in the ENA ToL checklist.

When manifests are uploaded, they are checked for standards compliance to the SOP and multiple sets of validations are performed. Firstly, the manifest is validated against the NCBI Taxonomy for taxonomic integrity. Collectors must supply scientific names at the species level that match the main name in the taxonomy, and that are submittable to ENA (programmatic calls to the taxonomic query services at EBI return a "submittable" field). The taxonomic fields are then populated by COPO, as long as one of the fields SCIENTIFIC_NAME or TAXON_ID is provided, and

synonyms are converted into main names with a warning to the user the main name will be stored instead.

Provided the taxonomy validation passes, the manifest is then validated against the SOP specifications, whereby all mandatory values need to be present and formatted as described. As time has passed and a larger number of manifests have been submitted, we have included additional validation rules to mitigate against common human errors. A number of warnings are also shown in COPO where it is appropriate to invite the user to double check their entries (for instance for tube or well IDs which do not conform to the standard format). One such example of valuable validation are the checks for previous association of SPECIMEN_ID to a different TAXON_ID, and the trigger of an error if SPECIMEN_ID is found more than once when ORGANISM_PART is WHOLE_ORGANISM (as it would be impossible to have the entire individual collected in multiple tubes). Any validation errors are shown to the user (see Figure 1) and submission is aborted.

Upon successful validation, COPO sends an email to members of the Sample Supervisors group, to notify of a new manifest awaiting inspection. From this email, a supervisor will navigate to a view within the COPO system, which shows DToL profiles along with their samples. These samples may be filtered according to whether they are pending, accepted or rejected, and their metadata examined for correctness. Pending samples can then be selected either individually or in bundles and accepted or rejected. If accepted, the samples are placed in a queue and are then asynchronously submitted to the ENA for registering the BioSamples. In the background, the sample metadata is converted into a series of XML files required by the ENA to be referenced in subsequent data file uploads. In doing so, we hide significant complexity from the user, thus increasing accessibility and interoperability. Upon successful ingestion, the ENA creates BioSamples for each specimen and for each child sample, according to the data model shown in Figure 2. Within this data model, the individual sample-level BioSamples are linked to the specimen-level BioSample either by a "same as" 1:1 relationship for whole organisms only, a "sample derived from" many:1 relationship for organism parts, or a "sample symbiont of" many:1 relationship for symbionts. This is a simple relationship for single samples, but allows flexibility and granularity when considering more complex relationships, *e.g.* symbionts within a single specimen tube, which has been instrumental in modelling samples from the ASG project. All DToL samples and associated data are linked under ENA study PRJEB40665.

Once a specimen-level BioSample has been completely verified, it is allocated an associated public name (ToLID) which uniquely represents the source organism. This is generated by the Sanger Institute based on the species and the SPECIMEN_ID of the sample, and automatically retrieved by COPO during the submission process. The subset of fields that have been identified as relevant for ENA submissions become part of the Biosample metadata. Remaining metadata is kept within COPO and is available through its user and programmatic interfaces. If the submission is successful, each sample, including the "specimen level sample", is allocated a BioSample accession number that uniquely identifies it in ENA and the BioSample database. COPO then stores these accessions within its database for easy search and retrieval (Figure 3).

COPO implements an API for collaborators and developers to interact with manifest metadata. The API allows users

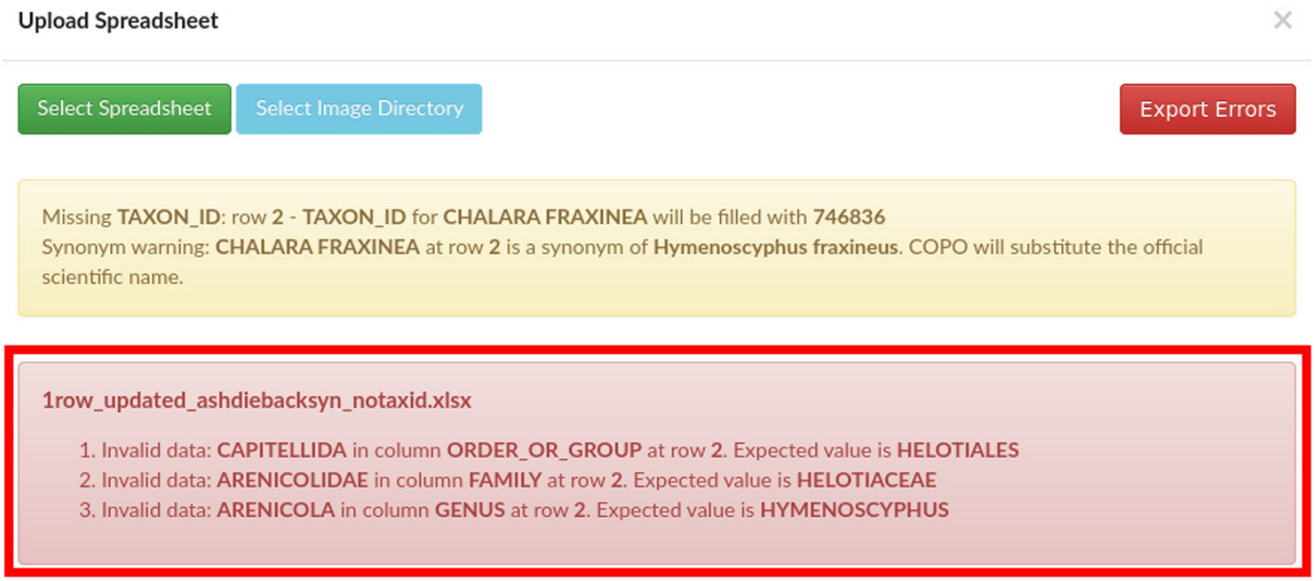


Figure 1. Collaborative OPen Omics (COPO) interface after an unsuccessful validation attempt, showing validation errors.

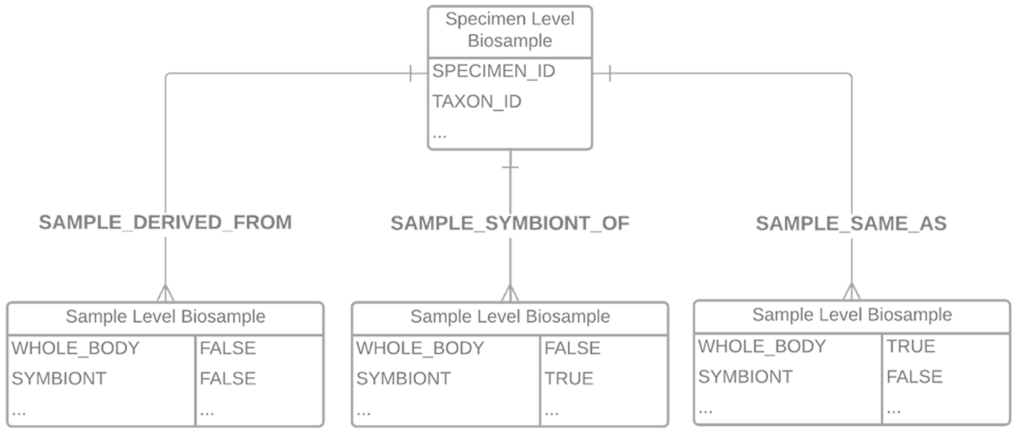
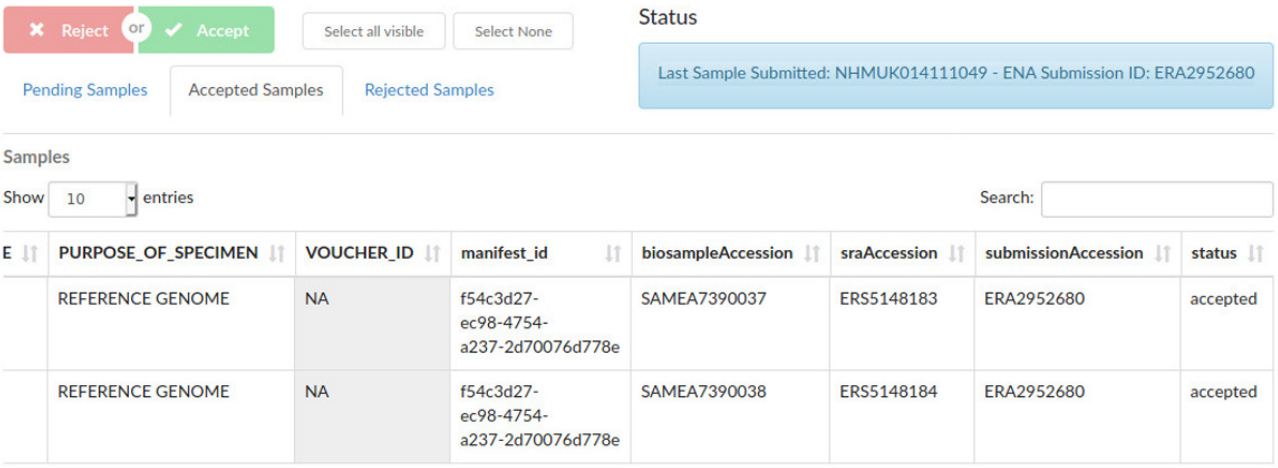


Figure 2. Data model of the relationship between specimens/samples within the BioSample database.



E	PURPOSE_OF_SPECIMEN	VOUCHER_ID	manifest_id	biosampleAccession	sraAccession	submissionAccession	status
	REFERENCE GENOME	NA	f54c3d27-ec98-4754-a237-2d70076d778e	SAMEA7390037	ERS5148183	ERA2952680	accepted
	REFERENCE GENOME	NA	f54c3d27-ec98-4754-a237-2d70076d778e	SAMEA7390038	ERS5148184	ERA2952680	accepted

Figure 3. Sample supervisor view, showing accepted sample metadata and European Nucleotide Archive (ENA) and BioSample accession numbers.

and other systems to get all manifests, all information about a particular manifest including submission status information, all samples, and all information about a particular sample. In the case of DToL, COPO supplies information to the Genomes on a Tree (GOAT) database (Sotero-Caio *et al.*, 2021) and STS (*Tree of Life Sample Management - Wellcome Sanger Institute*, no date), the Sanger Institute Lab Information Management System (LIMS) through these APIs.

More information about ToLIDs and the GOAT database can be found in our sister manuscript (Lawniczak *et al.*, 2022).

Imaging

Organism images are an important part of the sample documentation process, and need to be managed alongside metadata. As such, we are currently exploring the mechanisms within COPO with which we can submit images and the key unique identifier information into a suitable public repository. Currently a set of associated sample images can be uploaded. To associate each image with the correct sample metadata in COPO, collectors have to supply image files with names that contain the SPECIMEN_ID with the extension ".png" or ".jpg. Multiple images of the same organism can be submitted by appending an incrementing number to each e.g. "SP123_1.png", "SP123_2.png". The final destination of the images is the Bioimage Archive ('The BioImage Archive – Building a Home for Life-Sciences Microscopy Data', 2022).

Discussion

We believe we have arrived at a comprehensive set of policies and processes to capture and share rich metadata from a key biodiversity sequencing project which can be useful to others. Our manifests and SOP documents are openly available for other projects to use as a basis for their collection efforts. They represent a common and minimal set of standards, so we would recommend that any projects that wish to use them are free to do so, but any suggestions for changes

to the core set of fields and controlled vocabularies should be discussed with the DToL and EBP standards committees. The tools used to collect, validate, and submit this metadata are open source and/or freely available. We actively encourage other biodiversity projects to use these standards and tools and to provide feedback and extensions.

Below we discuss some issues that we experienced along the way so others can factor them into their biodiversity programmes:

Agreement on headings, valid terms, and ontologies to use for collection standardisation

Engaging with standards development as early as possible in a project is advised. It is inevitable that there will be differences in how a descriptive term represents a piece of information or a process when interacting with individuals or groups of researchers. As such, the field headers used in sample manifests and standardised documentation are vitally important to help researchers understand what information is required to complete the manifest. We engaged with the taxonomic groups over a period of months to ensure that we produced a common set of minimal fields that were easily understandable and clearly partitioned so that metadata collection was as easy as possible, yet still maintained required detail. Arriving at a minimal standard is a useful starting point but can take a large amount of discussion and development, so we would encourage and welcome other projects to use our standard and contact us for advice about potential changes and additions for other biodiversity projects.

Requirements to support various formats of collection (excel, CSV, phone/tablet app, **among others**)

Whilst many complaints are levelled at spreadsheet programs (Ziemann *et al.*, 2016), they remain a core tool in data collection for bench and field biologists. Many researchers are taught Excel, or at least learn the basics in school or university courses, and as such are fairly comfortable in its use. DToL needed to ensure the simplest route to collection was provided given that the number of required metadata fields in the manifest is now greater than 50. Whilst samples can be added into COPO through the user interface, this often is not the most useful way to input metadata in batches for larger numbers of samples, so it was agreed that fundamental data collection needed to be supported through spreadsheets.

Other methods of data collection were discussed, *e.g.* phone or tablet apps, and we have not pursued this avenue as yet, but they would be an area for future development. A mobile version of COPO would take a large effort, or at least a manifest completion tool which could then forward on the information via COPO's APIs, but could be useful to improve uptake.

A future goal is to set up continuous integration (CI) to automatically build new Excel and CSV files and place them in a separate folder in the GitHub repository to act as the definitive spreadsheet version that users can download and use for their sample metadata collection. Maintaining a strict spreadsheet layout and having a single point of reference for downloading and using it can be useful to promote compliance and versioning. It also allows validation at the version level, so submissions made against a previous version can be flagged to a user and the link to the new version can be provided.

Validation is often human-dependent, even with automated mechanisms

Despite all efforts, some human errors cannot be detected and/or fixed by automatic validation. For instance, it was not possible to make sure that the COLLECTION_LOCATION entered was referring to the latest collection (*e.g.* a museum or an aquarium), and not the original location. In this instance, we actually decided to implement specific fields for each situation to avoid validation problems or subsequent confusion about original versus actual physical collection location. It is often easier to adapt the standard to be unambiguous rather than relying on automated validation to highlight issues. However, changes to standards can result in extensive development time for the systems that implement it, so there is a balance to be made.

Validation can also be affected by unseen modifications in information sources and databases that are outside direct control of a project. For example, we have seen some cases of manifests passing taxonomy validation and subsequently failing soon after, as the taxonomy database is updated often. Therefore, it is vital for large biodiversity programmes to have good dialogue with taxonomic database managers as a result.

Tools and systems need to be developed closely with specific domains

Generalised tools can be useful in terms of coverage but less so in helping with community engagement and compliance. This is why COPO is developed as an open source and general data brokering platform, but we collaborate with research communities to develop specific brokering routes and user interfaces to match what a given community may expect or need. This can ease the perceived barriers to uptake. Looking to the future, as more sequencing will be carried out at the single-cell level, the importance of accepting sample metadata in plates, subsequent bulk validation, and user interfaces to make this easy to navigate will be required. As LIMS typically have this support, mapping this functionality to the sample collection and data brokering tools should be harmonised.

Core schemas plus extensions for taxonomic groups can help community uptake, showing that developers are listening to expert collectors

As before, we believe the DToL standards are a good minimal set of well-defined fields and terms for biodiversity projects. However, even within DToL there are corner cases where the core manifest needs extra information which is domain specific, *e.g.* protists. In this case, we worked with the DToL

protist groups to develop a handful of extra fields that are supplemental to the core set, *e.g.* salinity, water temperature, pH, among others. These fields can be validated in COPO to accommodate domain-specific requirements but would remain optional in the ToL checklist so when the sample reaches the ENA, it is validated against the core ToL standards and aligned across the project.

Even with tools and compliance, training and assistance with metadata collection is required

When projects instigate a metadata collection policy, even with the best intentions, a strong connection has to be made and maintained with sample metadata collectors to ensure that documentation is clear, the collection tools are user-friendly, and that people know where to go when errors and issues come up. Using a new software tool for the first time can be daunting, especially when complex information has to be provided. We regularly meet and speak with collectors (virtually through Slack or other online means), and feedback is discussed by the DToL Samples Working Group to understand where changes need to be made to a collection process or metadata standard, or where help can be given to improve uptake or to alleviate pain points.

Requirements to define update mechanisms for metadata and specifications

Given the size of the project, even with the best efforts and with the most attentive collectors, updates to metadata standards and specifications will always be necessary. A SOP is necessary to define the different scenarios for changes, with respect to important considerations such as regulatory compliance and accountability. Some of these will be actual updates to existing information within the public databases; others will be corrections or clarifications to individual samples or elements of the manifest or SOP itself. Tools and systems also need to be ready to integrate changes, test them, and ensure that updates are propagated to other dependent systems as appropriate. Change management is a vital part of software development to ensure compliance, so this should not be overlooked. As such, looking back to monitor how the manifests have aided uniformity of metadata richness across all submitted samples, as well as compliance to the SOPs and related guidance information, will be useful to inform future development.

When a new metadata upload occurs using updated manifests, it is crucial that the very same validation processes are triggered to avoid inserting metadata that does not respect the original SOP. All of these processes currently result in COPO keeping track of the changes, including the date the metadata was modified and by whom, so it provides an audit of metadata.

Summary

We have presented some key factors in our efforts to develop a standardised metadata framework, collection procedures, and technical software tools to facilitate the large-scale sample information management for an Earth Biogenome Project sequencing programme. We believe these are useful points of focus for subsequent efforts in this area, and we use our experiences within the UK Darwin Tree of Life project to demonstrate feasibility.

Data availability

Underlying data

No data are associated with this article.

Extended data

European Nucleotide Archive: Darwin Tree of Life Project: Genome Data and Assemblies; Accession number: PRJEB40665. https://identifiers.org/ena.embl:PRJEB40665

Zenodo: darwintreeoflife/metadata: Release for Wellcome Open Research, https://doi.org/10.5281/zenodo.7261393 (Shaw, 2022)

Analysis code

Analysis code available from: https://github.com/darwintreeoflife/metadata/tree/v2.4.1

Archived analysis code at time of publication: https://doi.org/10.5281/zenodo.7261393 (Shaw, 2022)

References

Blaxter M, Mieszkowska N, Di Palma F, *et al.*: **Sequence locally, think globally: The Darwin Tree of Life Project.** *Proc Natl Acad Sci U S A.* 2022; **119**(4): e2115642118.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Courtot M, Gupta D, Liyanage I, *et al.*: **BioSamples database: FAIRer samples metadata to accelerate research data management.** *Nucleic Acids Res.* 2022; **50**(D1): D1500–D1507.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Cummins C, Ahamed A, Aslam R, *et al.*: **The European Nucleotide Archive in 2021.** *Nucleic Acids Res.* 2022; **50**(D1): D106–D110.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Darwin Tree of Life. 2022; (Accessed: 9 June 2022).
Reference Source

Field D, Amaral-Zettler L, Cochrane G, *et al.*: **The Genomic Standards Consortium.** *PLoS Biol.* 2011; **9**(6): e1001088.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

GitHub: **darwintreeoflife/metadata.** GitHub, 2023; (Accessed: 19 July 2023).
Reference Source

GitHub: **AgriculturalSemantics/cg-core: CG Core Metadata Reference Guide.** GitHub, 2021; (Accessed: 30 September 2022).
Reference Source

Gonzalez A, Peres-Neto PR: **Act to staunch loss of research data.** *Nature*

Publishing Group UK, 2015; **520**: 436.
Publisher Full Text

Hartley M, Kleywegt GJ, Patwardhan A, *et al.*: **The BioImage Archive – Building a Home for Life-Sciences Microscopy Data.** *J Mol Biol.* 2022; **434**(11): 167505.
PubMed Abstract | **Publisher Full Text**

Howe D, Costanzo M, Fey P, *et al.*: **The future of biocuration.** *Nature.* 2008; **455**(7209): 47–50.
Publisher Full Text

International Society for Biocuration: **Biocuration: Distilling data into knowledge.** *PLoS Biol.* 2018; **16**(4): e2002846.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Kennan MA, Markauskaite L: **Research Data Management Practices: A Snapshot in Time.** *Int J Digit Curation.* 2015; **10**(2).
Publisher Full Text

Lawniczak MKN, Davey RP, Rajan J, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project [version 1; peer review: 2 approved with reservations].** *Wellcome Open Res.* 2022; **7**: 187.
Publisher Full Text

Lewin HA, Richards S, Aiden EL, *et al.*: **The Earth BioGenome Project 2020: Starting the clock.** *Proc Natl Acad Sci U S A.* 2022; **119**(4): e2115635118.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Shaw F, Etuk A, Minotto A, *et al.*: **COPO: a metadata platform for brokering FAIR data in the life sciences [version 1; peer review: 1 approved, 1 approved with reservations].** *F1000Res.* 2020; **9**: 495.
Publisher Full Text

Shaw F: **darwintreeoflife/metadata: Release for Wellcome Open Research (v2.4.1).** Zenodo. [code], 2022.
http://www.doi.org/10.5281/zenodo.7261393

Sotero-Caio C, Challis R, Kumar S, *et al.*: **Genomes on a Tree (GoaT): A centralized resource for eukaryotic genome sequencing initiatives.** In: *Pensoft Publishers.* Pensoft Publishers, 2021; e74138.
Publisher Full Text

Stevens I, Mukarram AK, Hörtenhuber M, *et al.*: **Ten simple rules for annotating sequencing experiments.** *PLoS Comput Biol.* 2020; **16**(10): e1008260.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

ten Hoopen P, Amid C, Buttigieg PL, *et al.*: **Value, but high costs in post-deposition data curation.** *Database (Oxford).* 2016; **2016**: bav126.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Tree of Life Sample Management. Wellcome Sanger Institute. (no date); (Accessed: 9 June 2022).
Reference Source

Whitmire AL, Boock M, Sutton SC: **Variability in academic research data management practices: Implications for data services development from a faculty survey.** *Program: electronic library and information systems.* 2015; **49**(4): 382–407.
Publisher Full Text

Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**: 160018.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Yilmaz P, Kottmann R, Field D, *et al.*: **Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications.** *Nat Biotechnol.* 2011; **29**(5): 415–420.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Ziemann M, Eren Y, El-Osta A: **Gene name errors are widespread in the scientific literature.** *Genome Biol.* 2016; **17**(1): 177.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✔ ? ✘

Version 2

Reviewer Report 05 January 2024

https://doi.org/10.21956/wellcomeopenres.21928.r72317

✘ **Eric Darvish Crandall**
The Pennsylvania State University - University Park Campus, University Park, Pennsylvania, USA

The authors describe metadata standards, a minimum information checklist and metadata brokering software (Collaborative Open Genomics; COPO) developed for the Darwin Tree of Life project. The standard, checklist and software are described as a contribution to making genomic datasets more findable, accessible, interoperable and reusable (FAIR).

Overall, I am glad to find like-minded researchers in Europe working toward a common goal of FAIR and open genomic data. I appreciate the care given to metadata curation, especially via Sample Supervisors. In full disclosure (see also my conflict of interest statement), I sit on the steering committee for the Genomic Observatories Metadatabase, (https://geome-db.org; see Deck *et al.*, 2017[1], Riginos *et al.*, 2020[2], Toczydlowski *et al.*, 2021[3] and Crandall *et al.*, 2023[4]).

I have some significant reservations about implementation as described in the current version of this letter.

1) New metadata standard seems redundant:

As with all academic endeavors, it is important to acknowledge and build upon previous work. It is a little unclear whether the authors are proposing a new standard or merely new software, but I will take the following sentence from the abstract as evidence of the former: "Here we report on the standards we developed with respect to a robust and reusable mechanism of metadata collection, in the hopes that other projects forthcoming or underway will adopt these practices for metadata."

In the first paragraph of the section entitled Sample Manifest Development, the authors mention existing standards that govern biodiversity data and genomic data: Darwin Core and MIxS respectively. In the next paragraph they already depart from these standards, using SPECIMEN_ID rather than materialSampleID (Darwin Core) or source_mat_id (MIxS) for the all important sample identifier field. Actually, as I peruse the linked DToL_SAMPLE_MANIFEST_v2.4.xlsx, most of the fields do not match with Darwin Core or MIxS. For example, there is DEPTH instead of

maximumDepthInMeters and minimumDepthInMeters (DarwinCore). Similarly with ELEVATION vs. minimumElevationInMeters and maximumElevationInMeters (Darwin Core). I suppose SCIENTIFIC_NAME is supposed to match up with specificEpithet (Darwin Core), since GENUS is a separate term, but it is unclear, as there are no definitions given in the manifest. Instead of yearCollected, monthCollected, dayCollected recommended by Darwin Core to avoid common ambiguities in dates, there is DATE_OF_COLLECTION, with a note to keep data in YYYY-MM-DD format (but we know that validation can't always be applied and some users will forget). There are others like DECIMAL_LATITUDE instead of decimalLatitude, that seem innocuous but could still impact data findability and interoperability. Similarly, TUBE_OR_WELL_ID would better be samp_well_name (MIxS), etc.

As the authors state in the discussion: "... the field headers used in sample manifests and standardised documentation are vitally important to help researchers understand what information is required to complete the manifest." Departing from existing standards defeats their attempt to make their data, and that of future users (ERGA and ASG are given as examples) findable, accessible, interoperable and reusable (FAIR). At the end of the section, there is a note about mapping terms to Darwin Core, but why not just use the Darwin Core terms in COPO directly to avoid any possibility of ambiguity? At GEOME we have found that neither standard covers all of the terms that we need, so we use a blend of the two, always providing clear definitions in the metadata template (our equivalent of the sample manifest). We only create new terms when it is clear that neither standard has a term that covers our needs.

I worry that creating a new metadata standard for genomic data where (admittedly imperfect) standards already exist will only add to the confusion around genomic metadata. For example, imagine trying to merge a COPO dataset with one found on GBIF (which uses Darwin Core) or the SRA (which is adopting MIxS). The GSC and TDWG have recently agreed to map their similar terms to one another's standard (http://www.gensc.org//news/2022/11/04/gsc_tdwg_mou.html), thereby reducing confusion around their overlap. I'd prefer if the authors not increase confusion again by creating new redundant terms for COPO and its users, and instead adhere to Darwin Core and MIxS whenever possible.

The authors provide examples of mapping their terms to Darwin Core, but nowhere is a complete mapping given. And why not just use accepted terms in the first place instead of creating a redundant standard?

It is this important reservation that causes me to not approve the letter.

2) Checklist alignment:

Reasonable people can disagree about what metadata should comprise a minimum set for a genomic dataset. In fact I have found that each time I engage in this exercise I arrive at a different answer, depending on the context of the data and who is part of the conversation. The authors provide another such list with the statement:

"The checklist also aligns the metadata collected for the DToL manifest against other existing standards (for example the MIxS standards (Yilmaz et al., 2011)) to allow comparison with samples outside of the project."

I see no evidence of this alignment in the linked checklist (at https://www.ebi.ac.uk/ena/browser/view/ERC000053). Moreover, the minimum information requirements (i.e. which terms are required, recommended, optional) differ significantly from the minimums defined by the Genomics Standards Consortium via MIxS (http://www.gensc.org/pages/standards/checklists.html). There should at least be an explanation for the differences, otherwise why should this checklist prevail over any other?

3) Citations to GEOME:

In a more commercial setting, GEOME and COPO might be considered competitors, but I hope that in an academic tradition we can be collaborators, both working towards FAIR genomic data. I think we ultimately want to both use our slightly different platforms to arrive at this same goal. I hope my comments can be taken in this spirit. I will seek to cite COPO going forward (and have already done so in a manuscript in review), and I hope this courtesy can be returned.

References

1. Deck J, Gaither MR, Ewing R, Bird CE, et al.: The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples.*PLoS Biol*. 2017; **15** (8): e2002925 PubMed Abstract | Publisher Full Text
2. Riginos C, Crandall ED, Liggins L, Gaither MR, et al.: 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.*Mol Ecol Resour*. 2020; **20** (6): 1458-1469 PubMed Abstract | Publisher Full Text
3. Toczydlowski RH, Liggins L, Gaither MR, Anderson TJ, et al.: Poor data stewardship will hinder global genetic diversity surveillance.*Proc Natl Acad Sci U S A*. 2021; **118** (34). PubMed Abstract | Publisher Full Text
4. Crandall ED, Toczydlowski RH, Liggins L, Holmes AE, et al.: Importance of timely metadata curation to the global surveillance of genetic diversity.*Conserv Biol*. 2023; **37** (4): e14061 PubMed Abstract | Publisher Full Text

Is the rationale for the Open Letter provided in sufficient detail?

Yes

Does the article adequately reference differing views and opinions?

No

Are all factual statements correct, and are statements and arguments made adequately supported by citations?

Yes

Is the Open Letter written in accessible language?

Yes

Where applicable, are recommendations and next steps explained clearly for others to follow?
Yes

Competing Interests: I sit on the steering committee for the Genomic Observatories Metadatabase, (https://geome-db.org), which does similar work to COPO. I receive no monetary compensation for my service.

Reviewer Expertise: Population genetics and genomics, Genomic Metadata, Open Science

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 13 May 2024

Felix Shaw

Reviewer Comment:

Point 1: “As with all academic endeavors, it is important to acknowledge and build upon previous work. It is a little unclear whether the authors are proposing a new standard or merely new software, but I will take the following sentence from the abstract as evidence of the former: "Here we report on the standards we developed with respect to a robust and reusable mechanism of metadata collection, in the hopes that other projects forthcoming or underway will adopt these practices for metadata." In the first paragraph of the section entitled Sample Manifest Development, the authors mention existing standards that govern biodiversity data and genomic data: Darwin Core and MIxS respectively. In the next paragraph they already depart from these standards, using SPECIMEN_ID rather than materialSampleID (Darwin Core) or source_mat_id (MIxS) for the all important sample identifier field. Actually, as I peruse the linked DToL_SAMPLE_MANIFEST_v2.4.xlsx, most of the fields do not match with Darwin Core or MIxS. For example, there is DEPTH instead of maximumDepthInMeters and minimumDepthInMeters (DarwinCore). Similarly with ELEVATION vs. minimumElevationInMeters and maximumElevationInMeters (Darwin Core). I suppose SCIENTIFIC_NAME is supposed to match up with specificEpithet (Darwin Core), since GENUS is a separate term, but it is unclear, as there are no definitions given in the manifest. Instead of yearCollected, monthCollected, dayCollected recommended by Darwin Core to avoid common ambiguities in dates, there is DATE_OF_COLLECTION, with a note to keep data in YYYY-MM-DD format (but we know that validation can't always be applied and some users will forget). There are others like DECIMAL_LATITUDE instead of decimalLatitude, that seem innocuous but could still impact data findability and interoperability. Similarly, TUBE_OR_WELL_ID would better be samp_well_name (MIxS), etc. As the authors state in the discussion: "... the field headers used in sample manifests and standardised documentation are vitally important to help researchers understand what information is required to complete the manifest." Departing from existing standards defeats their attempt to make their data, and that of future users (ERGA and ASG are given as examples) findable, accessible, interoperable and reusable (FAIR). At the end of the

section, there is a note about mapping terms to Darwin Core, but why not just use the Darwin Core terms in COPO directly to avoid any possibility of ambiguity? At GEOME we have found that neither standard covers all of the terms that we need, so we use a blend of the two, always providing clear definitions in the metadata template (our equivalent of the sample manifest). We only create new terms when it is clear that neither standard has a term that covers our needs. I worry that creating a new metadata standard for genomic data where (admittedly imperfect) standards already exist will only add to the confusion around genomic metadata. For example, imagine trying to merge a COPO dataset with one found on GBIF (which uses Darwin Core) or the SRA (which is adopting MIxS). The GSC and TDWG have recently agreed to map their similar terms to one another's standard (http://www.gensc.org//news/2022/11/04/gsc_tdwg_mou.html), thereby reducing confusion around their overlap. I'd prefer if the authors not increase confusion again by creating new redundant terms for COPO and its users, and instead adhere to Darwin Core and MIxS whenever possible. The authors provide examples of mapping their terms to Darwin Core, but nowhere is a complete mapping given. And why not just use accepted terms in the first place instead of creating a redundant standard? It is this important reservation that causes me to not approve the letter."

Author Response:

We appreciate the thoughtful feedback and acknowledge the concerns raised regarding the departure from existing standards in the initial stages of metadata development. In response to these concerns, we would like to offer the following rebuttal:

- Domain Specificity:
  - Darwin Tree of Life's metadata schema is designed from the bottom up, for its intended use case, mass bioscreening and biocuration. Therefore the field names were carefully customised by the Sample Working Group. It is not within COPO's remit to impose field names on the community. We instead take input from the community on how they would like to approach metadata collection most effectively for their project (in this case the generation of the sample manifest) then we make efforts, alongside the ENA to align the terminology and validation of the fields to existing standards to ensure interoperability. As the reviewer suggests, we do map our fields to existing standards, thus enabling the interoperability of DToL sample metadata with the community at large. COPO's API now allows any of the data collected to be exported to any of the commonly used standards, and in a variety of textual formats such as json, csv and ROCrate.
- Mapping to Darwin Core and MIxS:
  - Our metadata now aligns where possible, to MIxS and DWC, these mappings are available in the API. We acknowledge the importance of alignment with existing standards for the sake of interoperability and to ease potential integration with other datasets. The ongoing efforts to map our terms to Darwin Core reflect our commitment to harmonising our metadata schema with established standards.
- Immediate Benefits and Long-Term Alignment:
  - The decision to deviate from existing standards initially was driven by the needs of the community, and the immediate need to start collecting data

efficiently. We assure the reviewer that, as the project progresses, we are dedicated to aligning our metadata schema closely with Darwin Core and MIxS, ensuring long-term compatibility and adherence to widely accepted practices. Whilst this is an ongoing and imperfect process, we strongly believe that a working solution to 90% of the problem is better than a “perfect” solution which never reaches fruition.

- COPO's Adaptive Export Functionality:
  - We highlight that COPO's API is designed to facilitate data export in various standards, which now supports MIxS and Darwin Core. This means that while the initial data collection may use a specific schema, researchers can export data in formats that align with widely accepted standards, mitigating potential interoperability issues.
- Openness to Community Feedback:
  - We express our openness to feedback and collaborative input from the scientific community. We value the concerns raised and view them as valuable insights that can contribute to the ongoing refinement of our metadata schema. The iterative nature of the project allows for continuous improvement based on feedback and emerging best practices.

In addition to the points mentioned above, we would like to emphasise the following aspects regarding COPO's role as a data broker and our philosophy on standards development:

- COPO as a Data Broker:
  - Our primary goal with COPO is to act as a data broker, facilitating seamless exchange and management of research data. Recognising the diverse needs and workflows of researchers, we aim to serve as an intermediary that eases the burden on researchers during data collection, curation, and sharing processes.
- Bottom-Up Standards Development:
  - We strongly advocate for a bottom-up approach to standards development. We believe that standards should evolve organically from the research community, taking into account the specific needs and practices of individual researchers and projects. COPO's design reflects this philosophy, empowering researchers to contribute to the development of standards based on their unique requirements and preferences. It is then our job to ensure data alignment.
- Easing the Burden on Researchers:
  - COPO is designed to alleviate the challenges faced by researchers in adopting and adhering to standards. Our platform is user-friendly and adaptable, recognising the dynamic nature of research projects. By allowing researchers to work with familiar terms and map them to established standards, we aim to strike a balance between flexibility and adherence to best practices.
- Continuous Improvement and Community Involvement:
  - We are committed to continuous improvement based on community feedback. COPO's development is an ongoing process, and we invite researchers to actively engage with COPO, share their experiences, and contribute to the platform's evolution. This collaborative approach ensures that COPO remains responsive to the evolving needs of the research community.

Reviewer Comment:

Point 2: “Reasonable people can disagree about what metadata should comprise a minimum set for a genomic dataset. In fact I have found that each time I engage in this exercise I arrive at a different answer, depending on the context of the data and who is part of the conversation. The authors provide another such list with the statement: "The checklist also aligns the metadata collected for the DToL manifest against other existing standards (for example the MIxS standards (Yilmaz et al., 2011)) to allow comparison with samples outside of the project." I see no evidence of this alignment in the linked checklist (at https://www.ebi.ac.uk/ena/browser/view/ERC000053). Moreover, the minimum information requirements (i.e. which terms are required, recommended, optional) differ significantly from the minimums defined by the Genomics Standards Consortium via MIxS (http://www.gensc.org/pages/standards/checklists.html). There should at least be an explanation for the differences, otherwise why should this checklist prevail over any other?”

Author Response:

Thank you for your comments and for raising these concerns regarding the ENA checklist alignment to standards. We understand that we have not been clear in describing this alignment. As mentioned in the manuscript, the ENA sample checklists capture the requirements on the minimum metadata needed to describe biological samples. The ENA holds a number of different checklists, including the full set of MIxS environmental extensions to core packages, that are currently being updated to version 6.2. In the majority of cases, the MIxS terms are added as specified by the GSC, however sometimes there needs to be some slight adjustments to (a) fit the ENA data model (e.g. some terms such as the sequencing method are directly associated to the data and not to the sample, therefore this may not be mandatory for sample submission) or (b) if another standard already uses a pre-defined term for the same use, efforts are made to align these to enhance interoperability across the use in different ENA checklists. As mentioned above, as the ToL metadata schema was designed bottom-up to respond to the specific needs of the community, it is not totally compliant with MixS. The ENA Tree of Life Checklist was designed to validate the ToL sample metadata with terminology aligned to existing ENA submitting standards, including the ENA implementation of the MIxS standards to allow interoperability. We have corrected the sentence in the text so it reflects more accurately what has been done for the ToL checklist. It now reads “The terminology used in the checklist also aligns syntactically the metadata collected for the DToL manifest against other existing standards (including the MIxS standards (Yilmaz et al., 2011)) to allow comparison with samples outside of the project.”

Reviewer Comment:

Point 3: “In a more commercial setting, GEOME and COPO might be considered competitors, but I hope that in an academic tradition we can be collaborators, both working towards FAIR genomic data. I think we ultimately want to both use our slightly different platforms to arrive at this same goal. I hope my comments can be taken in this spirit. I will seek to cite COPO going forward (and have already done so in a manuscript in review), and I hope this courtesy can be returned”.

Author Response:

Thank you for your thoughtful comments, and we genuinely appreciate the insights you provided. We acknowledge the oversight in omitting a reference to GEOME in our initial manuscript and have addressed this in the resubmitted version. We are pleased to learn about the significant work carried out by GEOME and fully concur with your perspective on collaboration in academic endeavours. The shared objective of advancing FAIR genomic data is paramount, and we believe that the collaboration of GEOME and COPO can contribute meaningfully to this mission. We are enthusiastic about the potential for collaboration and sincerely thank you for your openness to such an engagement. Moving forward, we commit to citing GEOME in our future work, and we hope for a similar consideration in return. Your positive approach resonates with our mission, and we anticipate fruitful joint efforts in advancing genomic data standards. Your constructive feedback is invaluable, and we deeply appreciate your contribution to our shared goals.

Competing Interests: No competing interests were disclosed.

Reviewer Report 01 August 2023

https://doi.org/10.21956/wellcomeopenres.21928.r64298

Katrina Exter
Vlaams Instituut voor de Zee, Ostend, Flanders, Belgium

No further comments.

Is the rationale for the Open Letter provided in sufficient detail?
Yes

Does the article adequately reference differing views and opinions?
Yes

Are all factual statements correct, and are statements and arguments made adequately supported by citations?
Yes

Is the Open Letter written in accessible language?
Yes

Where applicable, are recommendations and next steps explained clearly for others to follow?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Open science; data managment; marine biodiversity

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 12 July 2023

https://doi.org/10.21956/wellcomeopenres.20513.r57911

?

Birgitta König-Ries
[1] University of Jena, Jena, Germany
[2] University of Jena, Jena, Germany

This article describes the developments towards standardisation, collection and preservation of metadata describing the samples that form the basis for sequencing in the Darwin Tree of Life project. I believe this to be a valuable resource for people involved in this project in some way, interested in working with the data or aiming to set up similar projects (on the same or even on smaller scales).

In my opinion, the paper is very definitely worth to be indexed.

I believe, though, that it would profit from some reorganisation. I found the flow of information not always easy to follow and had the impression that different threads that could well be separated were intertwined in the story line.

To me, the description of the metadata standard, the SOPs, and so on (so basically the artefacts that were developed in the project to support metadata management) should be described separately from the workflow used when entering a sample into the system and adding metadata to it. For this workflow description, an overview figure would be very helpful.

Some more detailed remarks:

Introduction

- paragraph 2: "providing human understanding": A major aim of FAIR data and metadata is to provide not only human but also machine understanding (or machine actionability or whatever you want to call it). I assume that is also what you aim for here.

- paragraph 3: "Therefore, standardised and .... " That list seems more exemplary than complete. If that is true you may want to rephrase.

Sample manifest development

- paragraph 2, 3rd line from the bottom:
  - RACK_OR_PLATE_ID and TUPE_OR_WELL_ID: From the text it is not quite clear whether that and means one of the two or both

page 4, left column, bottom paragraph:

- Are the different versions of the manifest developed in the google docs persisted somewhere and can be referred to via a persistent identifier. If not: Why is that not needed?

- If manifests or SOPs change: What happens to metadata/samples processed and described with older versions? Is there some reprocessing?

page 7, figure 2:

- A legend needs to be provided and/or some standard form of diagram (ER, UML Class, ...) should be used.

page 7. left column, first paragraph:

- "names that contain...": Does that mean there is only one image file per specimen possible or can the filenames contain additional information. If so: how is that standardized to ensure interpretabilitiy?

Is the rationale for the Open Letter provided in sufficient detail?

Yes

Does the article adequately reference differing views and opinions?

Yes

Are all factual statements correct, and are statements and arguments made adequately supported by citations?

Yes

Is the Open Letter written in accessible language?

Yes

Where applicable, are recommendations and next steps explained clearly for others to follow?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: research data management, biodiversity informatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 25 Jul 2023

Felix Shaw

I thank the reviewer for their time and thoughtful criticism. I have taken on board these and have altered the manuscript, with some rewording where things were not clear, clarifications and additions where asked for and a refactor of the figure of the COPO/Biosamples data model (figure 2).

Competing Interests: No competing interests were disclosed.

Reviewer Report 14 June 2023

https://doi.org/10.21956/wellcomeopenres.20513.r59555

Katrina Exter
[1] Vlaams Instituut voor de Zee, Ostend, Flanders, Belgium
[2] Vlaams Instituut voor de Zee, Ostend, Flanders, Belgium

Making the data and knowledge arising from genomics-based biodiversity studies fully FAIR is a vital endeavour. The importance of these studies to our understanding of biodiversity loss, and how to mitigate that loss, is enormous: smoothing the pathway to the rapid uptake of the results of these studies, and making it easier for the outputs from the numerous studies taking place around the world to be meaningfully combined and compared, is crucial. FAIR data management, starting from the field scientists collecting the data and continuing through to the publication and re-use of the data, is a necessary for any of this to happen. Emphasising the interoperability, understandability, and provenance of these data, is key.

The DToL project will produce extremely useful data and science for genomics, biodiversity, and taxonomy specialists. Explaining how the data and metadata are managed by DToL is important: (1) it explains how the data are created, standardised, and shared, creating trust in the project's outputs, (2) it contributes to the world-wide efforts that are being made to do FAIR data management. This paper, together with its related paper quoted therein, gives a good overview of the approach taken with the management of the metadata collected from the field scientists and curated by the DToL, and the submission of the (meta)data to ENA. It explains the steps undertaken, and very usefully shares the lessons learned. Links are given to their GitHub repository where the reader can investigate the metadata SOP and template in more detail; the software used are open source: taken together, the same methodology could be ad(o)(a)pted by others.

The formatting and standardisation of the sample metadata (the logsheets) are well described: fixed columns and formatted entries are required. One comment I would add here is that these

metadata would be even more interoperable (especially to machines/developers) if more semantics were added. There *is* a short paragraph in the text about "mapping" the column names to terms from controlled vocabularies, and when this is complete I would recommend that this mapping file is also made available in GitHub. This would also help with backward compatibility (it was mentioned that column titles can change or be added). Where terms cannot be found, they could be created within a DToL vocabulary. Exposing these now semantically-annotated metadata as RDF would go even one step further in making the metadata and the described data more usable by others.

Another comment I have is that it would also have been useful (especially for the data-specialist readers) to have included a diagram of the DToL (meta)data model.

Looking into more detail of the (meta)data, I have a few suggestions to the project

- It would improve the machine-interoperability and tracking of the SOP(s) if it was described in a machine-accessible way. This is not common practice, but nonetheless useful to do: see https://github.com/BeBOP-OBON for one such approach.

- With respect to the terminology used for indicating which parts of the body samples were taken from: have DToL considered adopting (or creating) an ontology here?

- Information about the SOPs and/or standardisation of the sample preparation and sequencing is missing: no link to a current or future article where this is explained is given. As sample preparation can have an impact on the sequences subsequently obtained, and the methodology of sequencing is even now still an evolving field, the steps taken to ensure the provenance of the genomics results are important to share. This does not detract from this article, however I would recommend that DToL considers such a publication.

Is the rationale for the Open Letter provided in sufficient detail?

Yes

Does the article adequately reference differing views and opinions?

Yes

Are all factual statements correct, and are statements and arguments made adequately supported by citations?

Yes

Is the Open Letter written in accessible language?

Yes

Where applicable, are recommendations and next steps explained clearly for others to follow?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Open science; data managment; marine biodiversity

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 25 Jul 2023

Felix Shaw

I thank the reviewer for their valuable insight and time in reviewing this article. With regards to the specific points raised, the details regarding sequencing are published in genome notes very regularly on wellcome open research, and I believe these contain the particulars e.g. https://wellcomeopenresearch.org/articles/8-319/v1. With regards to this paper, we are talking about metadata, so the specifics of genomic sequencing and library prep are out of scope. An ontology for the body parts was not considered, but I agree this would be a step in the right direction. I will look at the BeBOP-OBON link with interest. We are currently looking at ROCrate as a way of making outputs from COPO's API more machine readable (https://www.researchobject.org/ro-crate/).

Competing Interests: No competing interests were disclosed.