Recording Sample Metadata for the Darwin Tree of Life Project

Sample Manifest Standard Operating Procedure

Version: 2.3

Published Date: May 2021

Authors: Mara Lawniczak and Rob Davey

Correct and comprehensive recording of sample metadata is critical to the long-term utility of the work we do in the Darwin Tree of Life project: these metadata will link our genome sequences to their origins, and weave our work into the rich fabric of understanding of British and Irish, and global, biodiversity. Please read this Standard Operating Procedure (SOP) in full before completing the Sample Manifest as it contains detailed guidance on how to record metadata. Also contained is generic guidance on how to process specimens. Taxon-specific SOPs are available from each taxonomic working group to provide guidance on sample processing and regulatory compliance. Specific guidance on sample submission is available in the Sanger DToL Sample Submission SOP V2.3 and the Earlham DToL Sample Submission SOP V1.1.

Purpose: DToL aims to generate high quality genome sequences from samples and to embed these sequences into the landscape of biodiversity science. To do this we must adhere to correct physical handling of the specimens, and correct collation of rich metadata describing the specimens. This SOP contains specific instructions for filling in the metadata manifest. The project will not accession and process samples that do not have complete associated metadata.

Additional related SOPs are available describing (1) how to prepare samples for different taxonomic groups, which helps to assure delivery of high-quality samples that are more likely to be transformed into high quality genomes, (2) how to submit and ship samples to Sanger or Earlham, and (3) how to submit samples for molecular barcoding. The latest versions of these SOPs can be found in the DToL Shared Drive.

Future plans for this SOP: This SOP will be reviewed on a biannual basis by the Samples Working Group to incorporate feedback from the community. Metadata are currently collected manually using a defined spreadsheet, referred to as the DToL SAMPLE MANIFEST V2.3. This is enhanced by the COPO system (http://copo-project.org), a data management and brokering platform that allows metadata to be collected either in an online interface or through the downloading of partially filled and re-uploading of fully-filled spreadsheets. COPO links to a database that tracks all samples and their associated metadata as they progress from collection to genome assembly. COPO will produce a pipeline to update metadata for uploaded samples (see visual COPO documentation for more information) and are currently updating samples by email (see the “Changes to Uploaded Sample Metadata” section below). Finally, the data are archived in the ENA (https://www.ebi.ac.uk/ena/browser) for all sequenced samples.

Raising issues: We are still developing best practice, and elements of this SOP are subject to change. We expect that there will be questions to answer and lessons learned to share. If you are comfortable sharing in real time, please use the DarwinTreeOfLife Slack Workspace. If you do not have access to this, email Sophie Potter sp27@sanger.ac.uk. Otherwise, please raise specific issues by emailing the Samples Working Group at DTOL_SWG@sanger.ac.uk.

Document History

Major Version	***Date***	***Changes***	***Contributors***
1.0	2019-12-01	Draft version	SamplesWG and Sanger only
2.0	2020-06-20	Further clarifications on metadata	Mara Lawniczak, Nick Salmon, Nancy Holroyd, Seanna McTaggart, Jeena Rajan, Rob Davey
2.1	2020-07-01	Some turnover in terms, mapping to ENA checklist, incorporating barcode fields	Jeena Rajan, Rob Davey, Mara Lawniczak, Lyndall Pereira-da-Conceicoa
2.2	2020-09-04	Replaced most spaces with underscores; changed the field that was capturing difficulty to collect to also encompass high priority specimens; some clarifications on terms.	Mark Blaxter, Mara Lawniczak
2.3	2021-03-19	Addition of new fields/columns: ORIGINAL_GEOGRAPHIC_LOCATION, ORIGINAL_COLLECTION_DATE and BARCODE_HUB. ORGANISM_PART and TISSUE_FOR_BARCODING terms added: MOLLUSC_FOOT, UNICELLULAR_ORGANISMS_IN_CULTURE or MULTICELLULAR_ORGANISMS_IN_CULTURE. PURPOSE_OF_SPECIMEN term added: R&D. SYMBIONT changed to ASG format (multi-row entries for targets and symbionts). DIFFICULT_OR_HIGH_PRIORITY_SAMPLE term added: FULL_CURATION. Clarification on terms (“NOT_PROVIDED”); post-submission changes through COPO.	Alice Minotto, Lyndall Pereira-da-Conceicoa, Nancy Holroyd, Rob Davey, Jeena Rajan, Radka Platte, Kenneth Haug, Felix Shaw, Mara Lawniczak, Nicola Chapman, Josephine Burgin

Completing the Sample Manifest: Overview

Scope of this document

Specific guidance on preparing samples is not covered by this SOP. Please refer to the guidance for the specific taxonomic group you are working on.

Submission of samples is also not covered by this SOP. Please refer to the Sanger DToL Sample Submission SOP, the Earlham DToL Sample Submission SOP, the NHM DToL Sample Submission Barcoding SOP, and/or the RBGE DToL Sample Barcoding SOP (document to be added) as appropriate. These can all be found in the “1. SOPs and Sample Manifest – CURRENT VERSIONS” folder on our DToL Shared Google Drive.

The importance of “SPECIMEN_ID”

The SPECIMEN_ID must reflect the genetic identity of the individual, serving to link the various samples, images, vouchers, DNA barcodes, etc. that derive from one individual organism together. The SPECIMEN_ID also allows the laboratory team to resample the same individual specimen (and thus the same haplotypes) if needed, e.g. in the case of requiring more DNA to create a library. For example, ten different individual specimens each in their own tube would have ten distinct SPECIMEN_IDs, even if they are all from the same species. However, a single specimen split across ten tubes would result in each of those ten tubes having the same SPECIMEN_ID. This unique SPECIMEN_ID has two critical functions: identifying the GAL that holds responsibility for the specimen, and also declaring the genetic uniqueness of the specimen.

Each DToL specimen must be linked to a standardized, auto-generated sequence of numbered SPECIMEN_IDs that begin with a prefix unique to the GAL submitting the specimen. SPECIMEN_IDs must be unique to an individual (e.g., Ox0001 cannot be used again after it has been assigned to a specimen). SPECIMEN_IDs must follow the format specific to each GAL as listed below.

Table: GAL Specimen Codes

GAL	***Code Model***	***Number of digits***	***Contact person* Email address***
NHM	NHMUK000000000	9	Heather Allen H.Allen@nhm.ac.uk
RBGE & UofE	EDTOL00000	5	EDTOLnumbers@rbge.org.uk
Kew	KDTOL00000	5	Ester Gaya e.gaya@kew.org
EARLHAM	EI_00000	5	Seanna McTaggart seanna.mctaggart@earlham.ac.uk
MBA	MBA-000000-000A	5	Rob Mrowicki robmro@mba.ac.uk
OXFORD/WYTHAM OXFORD (Protist)	Ox000000 Ox500000	6 6	Liam Crowley liam.crowley@zoo.ox.ac.uk Estelle Kilias estelle.kilias@zoo.ox.ac.uk
SANGER	SAN0000000	7	Nancy Holroyd neh@sanger.ac.uk

* As of October 2020

Other "_ID"s

A sample can represent a set of specimens as well as multiple parts of the same specimen, and so the GAL_SAMPLE_IDs and COLLECTOR_SAMPLE_IDs can refer to an individual organism or something else (e.g., a soil sample could be represented by the COLLECTOR_SAMPLE_ID and a specimen taken from within that collection of soil be assigned a SPECIMEN_ID). The COLLECTOR_SAMPLE_ID is the identifier assigned by the collector to the specimen or the sample, hence the use of the term SAMPLE rather than SPECIMEN in this metadata field. The same is true of the GAL_SAMPLE_ID. For example, if a collector collects a sample that could have mixed genotypes or species, this will have a single COLLECTOR_SAMPLE_ID, and will need to be split further into specimens, each of which is assigned a unique SPECIMEN_ID.

It is permitted to have identical names for any or all of three categories (COLLECTOR_SAMPLE_ID, GAL_SAMPLE_ID and SPECIMEN_ID). The SPECIMEN_ID is the only one that is required for sequencing to commence.

Management of COLLECTOR_SAMPLE_ID, GAL_SAMPLE_ID and their relationship to SPECIMEN_ID is the responsibility of the collector and the GAL providing the samples.

Manifest Validation Process

Choose whether you prefer to use the Sample Manifest from the google spreadsheet or another option (e.g., Epicollect5, ARCGIS). We recommend that you retain a copy in Excel (XLS/XLSX) or Google spreadsheet form so as not to lose the data validation given the likelihood that further edits will be required.

Google spreadsheet: The Google spreadsheet can be used by ***making a copy*** and using it as an online spreadsheet, or by downloading it and entering data locally. If you choose to do the latter, please download as an XLS/XLSX (Microsoft Excel format) file to ensure that the data validation fields are retained.

Please carefully read the guidance in this SOP for each field, and attempt to get your submitted manifests as close to the guidance as possible. If your sample requires metadata fields or terms that are not present in the manifest, please contact DTOL_SWG@sanger.ac.uk to discuss and define new fields or terms.

Once you have completed entering all metadata, the initial check **upon submission to COPO** will confirm that each TAXON_ID maps to the correct species name. If mismatches are found, this will require the submitter to examine the mismatches and determine the nature of the problem. Please read the guidance on TAXON_ID below carefully as you should be able to ensure that each TAXON_ID precisely and accurately matches a species name in advance of submitting your manifest. There are too many possibilities to enumerate them all here, but three of the most common issues are a misspelling in the SCIENTIFIC_NAME or the TAXON_ID fields, a species for which no TaxonID is available in the NCBI TaxonomyDB, or a change in the taxonomy not reflected in NCBI TaxonomyDB. These will need to be addressed before the manifest can be validated. More information on how to fix these issues is below in the discussion of the TAXON_ID field.

Once you have ensured that your manifest is ready for validation, follow the guidance in the Sanger DToL Sample Submission SOP or Earlham DToL Sample Submission SOP. If any other issues with the information provided within the sample manifest are identified (e.g., missing mandatory entries, duplicate rows, incorrect date formats) the sample manifest will be returned to you to resolve these issues.

Once this process is complete and every sample has a TAXON_ID together with complete metadata, the manifest is considered to be "validated". However, prior to samples being accepted at the sequencing institute, DNA barcoding data is required. Manifests can be submitted to COPO while awaiting barcoding results and relevant fields (e.g., SCIENTIFIC_NAME, PUBLIC_NAME, TAXON_ID) to be updated. The process for "updating" a validated manifest will be developed over the coming months.

For all samples , if there is any possibility of species misidentification (SPECIMEN_ID_RISK = Y), only after DNA barcoding data is returned and samples are confirmed as the species they were declared will the samples be accepted (see **Sanger DToL Sample Submission SOP V2.3**). At this stage, each sample will be allocated a "PUBLIC_NAME" that reflects both the species and the SPECIMEN_ID (i.e., the genetic identity of the sample).

When data are submitted to ENA for release (as part of BioSamples, raw data and assembly submissions), the submissions will include all of the fields below indicated by **ENA_submission**. If the field name is in **turquoise**, then an entry for each specimen is mandatory for that field, even if only to declare why the information is missing. For all other fields, we strongly encourage data entry but it is not mandatory if it has not been collected.

Changes to Uploaded Sample Metadata

Any updates or changes to any fields for uploaded specimens should be sent as an email request to EI.COPO@earlham.ac.uk specifying the BioSamples accession, the field to update and the new value. For taxonomic changes, only the BioSamples accession and the new SCIENTIFIC_NAME is needed to update the taxonomy of a sample/specimen. COPO will produce a pipeline to update metadata for uploaded samples (see visual COPO documentation for more information on manifest submission and process updates).

Vouchers of Specimen or Sample

Every submitted specimen should be accompanied by voucher material. This material should be accessioned by a registered collection for permanent storage. Physical voucher material may be separated on collection, and be submitted directly to the designated collection organisation, or material remaining after processing may be returned to the designated collection from the sequencing centre. In cases where the entire specimen is consumed by processing, we request that digital images are recorded and submitted in lieu of physical samples. We regard it as good practice to record digital images of all specimens and samples destined for DToL processing, whether or not physical vouchers are retained, as this provides a close-to-life record of the organism sampled (see below).

Photographs of Specimen or Sample

Every submitted specimen should be accompanied by a photograph with explicit labelling as described below. Currently, we do not have a production repository available for storing DToL sample photographs. In the meantime, please store your photos in the DToL shared Google drive in the appropriate GAL or Taxon Working Group directory.

In preparation for linking images to metadata, please name images using the following format: SPECIMEN_ID-X.Y where X is a numerical identifier for the number of photographs you have taken of the same individual, and Y is the file format, e.g. NHMUK014110995-1.png and NHMUK014110995-2.png for two photos of the same specimen provided. When uploading photographs, please use PNG or JPG format.

File names must exactly match the SPECIMEN_ID in order to match photographs to samples automatically.

Detailed instructions for filling in the Sample Manifest

I. The manifest has three tabs. Please only fill in the **Metadata Entry** tab. If you discover a missing attribute in the drop-down menus, new attributes can be suggested by raising a request to the Samples Working Group at DTOL_SWG@sanger.ac.uk. Please only do this if absolutely required (i.e. no available term is a good proxy, and the absence of the attribute likely to affect many samples).

II. **Information must be entered for all fields below with** **turquoise bold names** [in the Google spreadsheet version of the manifest, these fields are represented by cells with a green fill. The fill will go white when an entry has been made to help you identify where mandatory fields still require data.] For all mandatory fields with **turquoise bold names**, even if information is unavailable, they must be populated with the appropriate term describing why this information is missing. The acceptable missing value terms are:

A. **NOT_APPLICABLE** = information is inappropriate to report. This can also indicate that the standard itself fails to model or represent the information appropriately.

B. **NOT_COLLECTED** = information was not given because it has not been collected.

C. **NOT_PROVIDED** = information of an expected format was not given but a value may be given at the later stage (this may be a particularly useful missing information term for VOUCHER_ID).

Fields that are named in **BOLD** without color do not require an entry describing why the information is missing because we expect that many samples will not have information for these fields (e.g., most samples will not have DEPTH information). However, if you have collected the information related to these terms, please do enter it.

Many terms will have the data released publicly as part of the ENA record. For every field for which this is true, you will find “**ENA_submission**” next to the name of the term.

III. **All dates** in the manifest must be formatted consistently as **YYYY-MM-DD** (ISO8601).

IV. In fields that are “free text” we ask that you use only the core alphanumeric characters, plus full stop “.”, hyphen “-”, underscore “_” and spaces (summarised in coding parlance as ” **-_.a-zA-Z0-9**”). Please avoid “|” (the vertical pipe symbol) except where we indicate it should be used to separate elements in a list. Please **do not** use “special characters” (such as other punctuation and “logical” marks: “#″`;:?!@*()[]{}/\,=+”, etc.).

Column by column instructions for the Metadata Entry tab

A. SERIES: This field holds the name of the series of samples this particular one belongs to. Genome Acquisition Labs (GALs) are expected to ship samples for processing in batches, labelled uniquely. We remind GALs that at least 48 samples should be accumulated prior to validation submission or sample shipment. We also encourage a minimum of 10 species where possible.

B. RACK_OR_PLATE_ID: The barcode identifier of the rack or 96 well plate that holds the samples when submitted. Partners should use barcoded racks (or plates where relevant) for samples. These should be scanned in and not manually entered.

C. TUBE_OR_WELL_ID: This field should record the FluidX barcode for each tube in a rack (or each well in a plate, where relevant). Barcodes must be entered using a barcode scanner in advance of preparing samples to reduce errors – do not enter barcodes manually.

D. SPECIMEN_ID**:** (**ENA_submission**) This is a unique identifier that refers to the genetic identity of the supplied material. It is assumed that the SPECIMEN_ID refers to a singular genetic individual. If the same individual specimen is split into several samples submitted in separate tubes, the SPECIMEN_ID for these samples would be the same. If multiple individuals of a species are sampled (e.g. from the same population), they must be placed in multiple, individual tubes, each with a unique SPECIMEN_ID. If sampling from organisms where distinguishing genetic individuals is difficult (e.g. mat-forming species like mosses or bryozoans), tease out individual units as far as is possible (e.g. single strands from a moss mat), and place each in a separate specimen tube with a unique SPECIMEN_ID.

- Each GAL maintains their own register of SPECIMEN_IDs for the project. Please ensure that you do not use IDs that have already been used, and that you stick to the format required by the GAL you are submitting on behalf of.

E. ORDER_OR_GROUP: The taxonomic Order into which the Family is placed or (if this is not defined) the monophyletic group to which the Family or Genus belongs. This should correspond to the taxonomy as represented in the NCBI Taxonomy Database. If you or your taxonomist have a disagreement with the taxonomy represented on NCBI Taxonomy Database, please raise this with the NCBI TaxonomyDB curators as described below.

F. FAMILY: The taxonomic Family into which the Genus is placed. This should correspond to the taxonomy as represented in the NCBI Taxonomy Database. If you or your taxonomist have a disagreement with the taxonomy represented on

NCBI Taxonomy Database, please raise this with the NCBI TaxonomyDB curators as described below

G. **GENUS**: The taxonomic Genus to which the Species belongs. This should correspond to the taxonomy as represented in the NCBI Taxonomy Database, and with the generic component of the scientific name given below. If you or your taxonomist have a disagreement with the taxonomy represented on NCBI Taxonomy Database, please raise this with the NCBI TaxonomyDB curators as described below.

H. **TAXON_ID**: (**ENA_submission**) A valid NCBI TAXON_ID to the species level is mandatory in order to submit data to public repositories. The species name in the manifest must be identical to that listed in the "current name" box in the Taxonomy Browser for that species. If this is not the case, you must write to ena-dtol@ebi.ac.uk to request the change.

If there is another taxon database for your group, e.g. EukRef, please fill in the NCBI TAXON_ID, and then use the TAXON_REMARKS field to specify the taxon database and the ID/accession/URL.

- TAXON_IDs can be looked up based on the species at the following links: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi or https://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi.

- If no TAXON_ID exists, or a credible TAXON_ID exists that likely is a synonym of the species name the collector or submitter would use (through differential usage, error or lack of currency of the NCBI taxonomy), please ask for assistance by writing to ena-dtol@ebi.ac.uk, providing the full taxonomy, scientific name and authority for the chosen name where possible. If required, a new TAXON_ID should be available within 14 days. In the case of conflict, the sample submitter will be contacted and may be required to provide further information. Please note that the final species name on submission of the data to INSDC will be the one associated with the TAXON_ID in NCBI Taxonomy.

- When a sample is provided that requires DNA barcoding before a species ID is possible, please provide the SCIENTIFIC_NAME and TAXON_ID of the most likely species identity and be sure to select SPECIMEN_ID_RISK = Y, this can be updated in COPO after DNA barcoding has confirmed identification.

I. **SCIENTIFIC_NAME**: (**ENA_submission**) The latin binomial/combined genus and species name with a space in between.

- See TAXON_ID above if you or the taxonomic expert have substantive issues with the species name present for the taxon in the NCBI TaxonomyDB.

- Any changes to SCIENTIFIC_NAME post manifest submission to COPO (due to species re-identification or other taxonomic change), should be requested by email to COPO.EI@ealrham.ac.uk and should include the new SCIENTIFIC_NAME and BioSamples accession (other related taxonomic fields will be auto-filled by COPO). If applicable, please include information for the fields COMMON_NAME, TAXON_REMARKS and INTRASPECIFIC_EPITHET, otherwise these will be overwritten and left blank.

J. **TAXON_REMARKS**: Free text to summarise any known issues with the mapping of TAXON_ID to SCIENTIFIC_NAME or add other taxon database identifiers here, e.g., EukRef. Here you can also comment on STRAIN availability, if the specimen is a representative of a living and accessible strain/colony/culture. If there are no issues, leave this field **blank**.

K. **INFRASPECIFIC_EPITHET**: Where the sample is from a formally named infraspecific taxon, give the infraspecific name here, with prefixes in the following format: ssp. (for subspecies), var. (for variety), cv. (for cultivar), br. (for breed). Entries in this field should reflect organisms that can be found living outside of laboratories (see next attribute for lab strains). If there is no epithet here, leave this field **blank**.

L. **CULTURE_OR_STRAIN_ID:** (**ENA_submission**) Please give the reference ID from the source culture collection, such that the culture accession can be found in the collection's database. This is only relevant if the sequenced material is derived from a living, culturable, named laboratory strain (e.g. *Anopheles coluzzii* N'Gousso strain). This field should not be used to record a variant or type that has been collected anew from the wild: such information should be placed in **OTHER_INFORMATION**. Leave this field **blank** if it is not relevant.

M. **COMMON_NAME**: Vernacular name, if the species has one. If multiple names are required, separate names with a | (vertical pipe) character. If you are unsure of or the species has no vernacular name leave this field **blank**.

N. **LIFESTAGE**: (**ENA_submission**) The life stage of the specimen from which the sample was derived. This field has a controlled vocabulary: use the drop-down menu or look at the available terms on the second tab to complete. Please note that there are currently curated attributes for animals, for plants/fungi/macroalgae, and for some protists.

- If these do not fit your taxa, please contact DTOL_SWG@sanger.ac.uk. Please enter **NOT_PROVIDED** if your proposal for a lifestage term has not yet been accepted.

O. **SEX**: (**ENA_submission**) The sex of the specimen from which the sample was derived. This field has a controlled vocabulary: use the drop-down menu. If the sex of the organism is not known, use **NOT_COLLECTED**. The sex may be determined at a later date using the genome sequence data, but this will be captured in a different field, so this field should refer solely to the sex as determined by morphological examination of the specimen or strong inference (e.g., the species is from a clade that is always hermaphroditic/monoecious).

P. **ORGANISM_PART**: (**ENA_submission**) A description of the exact tissue(s) in the tube or well. Accurate information here is important for downstream analyses on the symbiome, chromosomal diminution, RNAseq, etc. There is a tab in the DToL Sample Manifest that defines the terms that can be used for ORGANISM_PART. This tab lists definitions for the full tissue, but pieces of that tissue are acceptable (e.g., LUNG is defined as 'the lung of a vertebrate', but the whole lung is not expected and a small piece of lung is expected).

- Please combine tissues by entering multiple terms from the ontology using the | (vertical pipe) symbol (e.g. for head + abdomen of an insect enter "HEAD | ABDOMEN"). When using multiple body parts, there will be a data validation error that arises, but these can be ignored as long as the spelling and capitalization of the terms is identical to the provided list.

- If the tissue or organism part you are providing is not present in the drop-down menu, please choose the best generic category (these start with **) and add the name of the tissue that you have put into the tube in the "OTHER_INFORMATION" free text field. Please also email the Samples Working Group at DTOL_SWG@sanger.ac.uk to request the necessary additions. We will update attributes quarterly.

- If the sample is shipped as a DNA or RNA extract, select the tissue from which this was extracted and add further information in the OTHER_INFORMATION field regarding quality, quantity, etc. Note that any shipment of DNA should be discussed in advance as tissue is expected.

Q. **SYMBIONT**: This is to indicate whether the sample contains a known symbiont (i.e. you have metadata for it and a species-level and ENA-submittable TAXON ID). Select "TARGET" if only the "host" metadata is known OR if it is a symbiont-only culture. Select "SYMBIONT" if you have a known symbiont in the sample and you have metadata (including, critically, a species-level identification supported by a valid taxon ID) for the symbiont. If you need to select "SYMBIONT" you will then need to copy and paste your "TARGET" row and amend the following fields to reflect the symbiont data:

- ORDER_OR_GROUP, FAMILY, GENUS, TAXON_ID, SCIENTIFIC_NAME, TAXON_REMARKS, INFRASPECIFIC_EPITHET, CULTURE_OR_STRAIN_ID, COMMON_NAME, LIFESTAGE, SEX, ORGANISM_PART

If there is no explicit information on potential symbionts, this field should be left **blank**.

R. **RELATIONSHIP:** (**ENA_submission**) This is a free text field to permit declaration of any known parental, child, or sibling relationship between the specimen and any other specimens that are submitted for the DToL project, OR to declare if the specimen is a “barcode exemplar” for another specimen.

- If there are known genetic relationships between submitted specimens, please concisely state the relationship: “Full sibling to SPECIMEN_ID1”, “Mother to SPECIMEN_ID2”, “Maternal half sibling to SPECIMEN_ID1, SPECIMEN_ID2, and SPECIMEN_ID3”, or “Trio child of SPECIMEN_ID1 and SPECIMEN_ID2”. If knowledge of the relationships is not confident but suspected, do not add anything here and instead add this information to the “OTHER_INFORMATION” field (e.g., “suspected full or half sibling to SPECIMEN_ID2”).
- If the specimen is acting as a barcoding exemplar for another specimen because the entire organism must be used for reference genome sequencing and it is not possible to take a sample for DNA barcoding (e.g., midges from the same swarm where one is submitted for sequencing and 5 are submitted individually for DNA barcoding), then add “barcode exemplar for SPECIMEN_IDx” and insert the SPECIMEN_ID for the specimen that is going for reference genome sequencing, potentially without its own DNA barcoding.
- If there is no relationship to note, this field can be left **blank**.

S. **GAL:** (**ENA_submission**) Use the drop-down menu to select the Genome Accession Lab (GAL) responsible for this sample. If the GAL is also the collector, then this will be the same affiliation as the COLLECTOR_AFFILIATION.

T. **GAL_SAMPLE_ID**: (**ENA_submission**) This is the unique name assigned to the sample by the GAL. This will include an abbreviation for the GAL and a simple shorthand identifier. This is a free text field, but please do not use spaces or special characters, e.g. #, !, ^, *, etc. It is fine for the GAL_SAMPLE_ID to be the same as the COLLECTOR_SAMPLE_ID and the SPECIMEN_ID if warranted.

U. COLLECTOR_SAMPLE_ID: This is the unique name assigned to the sample by the COLLECTOR or COLLECTOR_AFFILIATION. This is a free text field, but please **do not use spaces or special characters**, other than hyphens and underscores ("-" and "_") i.e do not use #, !, ^, *, etc.

- In some cases, you will be splitting a single specimen across multiple tubes (see SPECIMEN_ID), and you will want to consider what kind of information you want in your unique sample names for this. For example, if the specimen is a butterfly with SPECIMEN_ID = Ox000005, and you put the head in one tube and the thorax in another, your COLLECTOR_SAMPLE_IDs might reflect this with one tube called Ox000005-h and the other called Ox000005-t. Likewise, the COLLECTOR_SAMPLE_ID may be the name given to a collection consisting of a 'clump' from a mat-forming species, which may then be subdivided into different specimen tubes, each given a unique SPECIMEN_ID.

V. COLLECTED_BY: (**ENA_submission**) Enter the name of the person or people who collected the sample using all CAPITALS, and separate names with "|" (vertical pipe symbol), e.g., "CAROLUS LINNAEUS | JEAN_BAPTISTE LAMARCK".

- We note that storage of names with affiliations in a database brings the DToL system under the aegis of the GDPR regulations, and we must ask GALs and collaborators to agree to their data being stored in COPO and to those data being propagated to secondary databases (including ENA and the final collections of record).

W. COLLECTOR_AFFILIATION: (**ENA_submission**) Free text field to supply the university, institution, or society that is responsible for the collected specimen. This is typically the society or institution of the person(s) specified in the COLLECTED_BY field. If multiple people are specified in COLLECTED_BY, ensure that their institutional affiliations are also separated by a vertical pipe symbol. Position in the list of affiliations should match the person in the same position in the list of names (e.g., PERSON A | PERSON X | PERSON C will have their affiliations as: (INSTITUTE A | INSTITUTE X | INSTITUTE C). If multiple people are listed but all from the same affiliation, no need to repeat the affiliation. For people unaffiliated with any institution or society, please list as "PRIVATE".

X. DATE_OF_COLLECTION: (**ENA_submission**) The date the sample was collected, with year, month and day specified (**YYYY-MM-DD**).

- If the specimen is from a zoo, botanic garden, culture collection and has a known date of collection from the wild or acquisition from another collection, please note this information in ORIGINAL_FIELD_COLLECTION_DATE and only include **here** the date when the sample was taken from its location (e.g., "London Zoo", "Millennium Seed Bank", etc.).

Y. **COLLECTION_LOCATION**: (**ENA_submission**) General description of the location where the sample was taken. This must start with the country (United Kingdom, or look up other accepted country names here https://www.ebi.ac.uk/ena/browser/view/ERC000053), but also include more specific locations (e.g. "Barton's Pond") ranging from least to most specific and separated by | character, e.g. "United Kingdom | East Anglia | Norfolk | Norwich | University of East Anglia | UEA Broad". It is important to give the name of the site here if possible.

- If the specimen is from a zoo, botanic garden, culture collection and has a known origin elsewhere, please note this information in ORIGINAL_GEOGRAPHIC_LOCATION and **only** include here information about the location of the specimen at the time from which a sample was taken (e.g., "London Zoo", "Millennium Seed Bank", etc).

Z. **ORIGINAL_COLLECTION_DATE**: (**ENA_submission**) If the specimen is from a zoo, botanic garden, culture collection and has a known date of collection **from a known origin elsewhere** (e.g. the wild), please record the date here in as much detail as possible, with year, month and day specified (**YYYY-MM-DD**). YYYY-MM and YYYY is acceptable where further detail is not known. This information is important for regulatory compliance checks. Leave this field **blank** if it is not applicable.

AA. **ORIGINAL_GEOGRAPHIC_LOCATION**: (**ENA_submission**) If the specimen is from a zoo, botanic garden, culture collection and has a **known origin elsewhere**, please record the general description of the original location here. This should start with the country (United Kingdom, or look up other accepted country names here https://www.ebi.ac.uk/ena/browser/view/ERC000053), but also include more specific locations (e.g. "Barton's Pond") ranging from least to most specific and separated by | character, e.g. "United Kingdom | East Anglia | Norfolk | Norwich | University of East Anglia | UEA Broad". It is important to give the name of the site here if possible. This information is important for regulatory compliance checks. Leave this field **blank** if it is not applicable.

BB. **DECIMAL_LATITUDE**: (**ENA_submission**) The geographic location where the specimen or sample was taken in decimal degrees, between -90 and 90. We advise that locations are specified to a minimum of 3 decimal places (https://en.wikipedia.org/wiki/Decimal_degrees).

CC. **DECIMAL_LONGITUDE**: (**ENA_submission**) The geographic location where the specimen or sample was taken in decimal degrees, between -180 and 180. We advise that locations are specified to a minimum of 3 decimal places (https://en.wikipedia.org/wiki/Decimal_degrees).

DD. **GRID_REFERENCE**: Information to geolocate the sample area, preferably with a map or standardised geolocation reference, e.g. OS GRID REF: SP45998 08751. https://osmaps.ordnancesurvey.co.uk/ is useful to map lat-long to grid references. This field is optional and can be left **blank**.

EE. **HABITAT**: (**ENA_submission**) Any comments about the location, habitat or substrate, *e.g. damp mossy ground in moderate shade.* If substrate is living and there is a chance that it is included in the sample, add this to the SYMBIONT category. We recommend using terms from the ENVO ontology. If the specimen is from a zoo or botanic garden, you can add its original habitat to "OTHER_INFORMATION" but here, please only capture its habitat at the time of collection (e.g. "reptile cage at London Zoo").

FF. **DEPTH**: (**ENA_submission**) Depth below water body surface, supplied in metres. This is not the absolute depth of the water body. Do not supply the unit, e.g. use 200 for 200 m below sea level, 100-200 for 100-200 m range below sea level, etc. Leave this field **blank** if the depth was not recorded or it is not an applicable field.

GG. **ELEVATION:** (**ENA_submission**) Altitude above sea level, supplied in metres. Do not supply the unit, e.g. use 200 for 200 m above sea level, 100- 200 for 100-200 m range above sea level, etc. Please supply elevation of water surface for inland water bodies. Leave this field **blank** if the elevation was not recorded or it is not an applicable field.

HH. **TIME_OF_COLLECTION**: Time of day of sample collection in 24-hour clock format, with hours and minutes separated by colon e.g. 13:35, 04:53, etc. This should be in GMT/UTC. This field may be particularly relevant for RNAseq but it is not mandatory. Leave this field **blank** if the time was not recorded.

II. **DESCRIPTION_OF_COLLECTION_METHOD**: A detailed as possible description of the sample collection methods, e.g. "*caught with fibre net within densely wooded area, and immediately placed into the collection container*".

JJ. **DIFFICULT_OR_HIGH_PRIORITY_SAMPLE**: Drop down menu to flag species/samples that are difficult to collect (rare), elected for FULL CURATION as a family representative or high priority to push through sequencing for any reason.

KK. **IDENTIFIED_BY**: (**ENA_submission**) Enter the name of the person or people who identified the sample to species level. Use ALL CAPs, and separate names with |

(vertical pipe symbol), e.g., “CAROLUS LINNAEUS | JEAN-BAPTISTE LAMARCK”.

- We note that storage of names with affiliations in a database brings the DToL system under the aegis of the GDPR regulations, and we must ask GALs and collaborators to agree to their data being stored in COPO and to those data being propagated to secondary databases (including ENA and the final collections of record).

LL. **IDENTIFIER_AFFILIATION**: (**ENA_submission**) Free text field to supply the university, institution, or society that is responsible for the collected specimen. This is typically the society or institution of the person(s) specified in the IDENTIFIED_BY field. If multiple people are specified in IDENTIFIED_BY, ensure that their institutional affiliations are also separated by a vertical pipe symbol. Position in the list of affiliations should match the person in the same position in the list of names (e.g. “Person A | Person X | Person C” will have their affiliations as: “Institute A | Institute X | Institute C”. If multiple people are listed but all from the same affiliation, no need to repeat the affiliation.

MM. **IDENTIFIED_HOW**: Indicate what method(s) were used to identify the specimen to the nominal species (e.g., morphology, ITS barcoding). This is free text and should include reference to an authoritative key if possible. If the identification is by a taxon expert, note that here and ensure the name of that person is in the IDENTIFIED_BY column.

NN. **SPECIMEN_ID_RISK**: Y/N field to indicate if there is any risk that the SPECIMEN_ID provided does not reflect a single genetic entity OR the species names it has been submitted under. Examples of this include 1) a clump of tissue or cells that could comprise multiple individuals; 2) a species that is part of a species complex or group where it can be difficult to be certain of species identity. Please make every effort to ensure this field is N if possible (e.g., by taking single strands of clumpy organisms that are most likely to reflect a single genetic entity or ensuring molecular barcode data support the species name provided).

OO. **PRESERVED_BY**: Name of person that carried out the preservation (including specimen dissection and tissue removal), supplied in CAPITALS. Multiple preserver names should be separated by a | character.

- We note that storage of names with affiliations in a database brings the DToL system under the aegis of the GDPR regulations, and we must ask GALs and collaborators to agree to their data being stored in COPO and to those data being propagated to secondary databases (including ENA and the final collections of record).

PP. **PRESERVER_AFFILIATION**: Free text field to supply the university, institution, or society that is responsible for the collected specimen. This is typically the society or institution of the person(s) specified in the PRESERVED_BY field. If multiple people are specified in PRESERVED_BY, ensure that their institutional affiliations are also separated by a vertical pipe symbol. Position in the list of affiliations should match the person in the same position in the list of names (e.g., Person A | Person X | Person C will have their affiliations as: (Institute A | Institute X | Institute C). If multiple people are listed but all from the same affiliation, there is no need to repeat the affiliation.

QQ. **PRESERVATION_APPROACH**: e.g. snap frozen, dry ice, ethanol/dry ice slurry, in RNALater, lyophilised, air dried, etc.

RR. **PRESERVATIVE_SOLUTION**: Suspension liquid used to preserve the sample, e.g., RNALater, RLT Buffer, DESS. If no preservative was used, this field should be left **blank**.

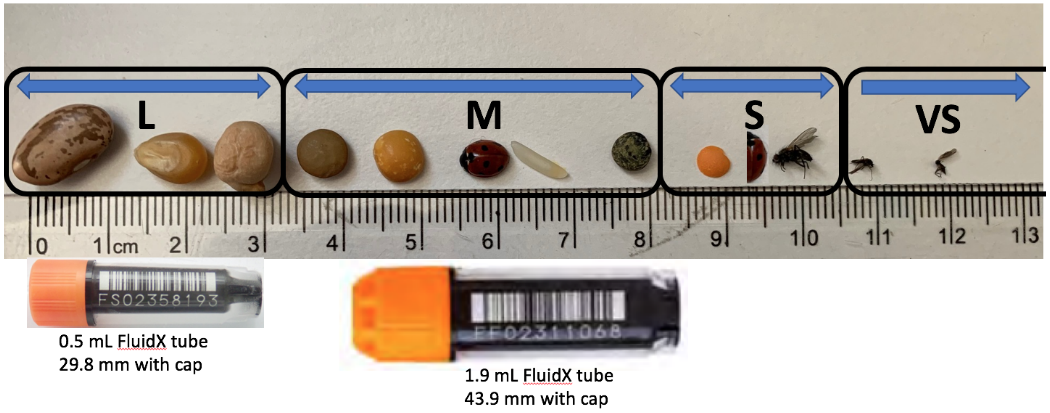
SS. **TIME_ELAPSED_FROM_COLLECTION_TO_PRESERVATION**: some organisms may be held living in collection for a period of time for starvation or other factors. This entry should be specified in hours, but no unit, e.g. 0.5 for half an hour, 3 for 3 hours, etc.

TT. **DATE_OF_PRESERVATION**: Date on which the species was preserved. Please use **YYYY-MM-DD** format.

UU. **SIZE_OF_TISSUE_IN_TUBE**: How large is the sample in the tube. We aim for one lentil-sized piece per tube but sometimes adding more or less tissue than this will be necessary. Please note the approximate size of the piece or pellet: use the following shorthand:

- “VS” for very small
- “S” for small (~red lentil sized)
- “M” for medium (~yellow lentil/ladybird sized/5mm)
- “L” for large (>5mm, chickpea/bean sized)
- If the specimen is a single cell, use “SINGLE_CELL”
- Aim for single lentil sized (S or M) pieces in tubes whenever possible. If the sample is L, then wherever possible process this into multiple tubes of S or M sized pieces (up to 10 tubes per specimen is welcomed). See visual guidance below.
- If the sample has been shipped as extracted DNA please enter “NOT_APPLICABLE”. Note that we expect that all samples will be

extracted at Sanger.



Guidance for "Size_of_tissue_in_tube"
L = popcorn kernel or dried chickpea sized and larger
M = green, yellow lentil sized, whole ladybird size
S = red lentil, half a ladybird size
VS = smaller than half a red lentil
SC = single cell

VV. BARCODE_HUB: (ENA_submission) Drop down menu to flag the GAL responsible for DNA barcoding of the submitted taxa.

WW. TISSUE_REMOVED_FOR_BARCODING: State "**Y**" or "**N**". See the appropriate Molecular Barcoding SOPs for detailed instructions, noting that barcoding requires materials in specific tube or plate types so the SOP must be referred to. If you are collecting across different taxonomic groups, ensure you know which GALs will receive material so that you allocate your samples into different plates depending on their destination (as of October 2020, marine fungi and seaweeds go to MBA, plants go to RBGE, and everything else goes to NHM).

XX. PLATE_ID_FOR_BARCODING: This is the barcode number on the side of the tissue plate. Barcoding sites will provide pre-labelled plates and tubes. If you are submitting plant tissue, these will not be submitted in plates, so this is not necessary and you can put NOT_APPLICABLE.

YY. TUBE_OR_WELL_ID_FOR_BARCODING: This is either the well number on a plate (there are 96 wells per tissue plate) OR the barcode/unique identifier on the tube containing the tissue sample.

ZZ. TISSUE_FOR_BARCODING: Please state what part of the organism was dissected for DNA barcoding (e.g. leg, soft-body tissue etc.). Muscle tissue is ideal for barcoding. This list is a repeat of the attributes available for "ORGANISM_PART" with one addition of "DNA_EXTRACT"

AAA. BARCODE_PLATE_PRESERVATIVE: Guidance is found in the barcoding SOPs. Typically, animal samples will be submerged in 70% ethanol, plant tissue will be preserved in silica gel, and fungal tissue will be frozen or lyophilized. Record the volume, concentration, and type of preservative/method of preservation used here.

BBB. **PURPOSE_OF_SPECIMEN**:

- The majority of specimens will be for “REFERENCE_GENOME”. All samples listed for REFERENCE_GENOME sequencing are assumed to also need DNA BARCODING and RNA-SEQUENCING, and the term “REFERENCE GENOME” encompasses all three things (reference genome, barcoding, RNA-seq) wherever samples allow. Please use REFERENCE GENOME for all specimens / samples of a particular species unless they should be destined for an alternative use only.

- If a particular tissue is needed solely for RNAseq use “RNA-SEQUENCING”

- If the specimen is intended for population genetics or resequencing please use “SHORT_READ_SEQUENCING.

- If a particular tissue or specimen is intended for research and development, for example as part of an R&D diversity panel, or as part of a preservation trial,please use “R&D”. These samples may not progress to reference genome sequencing and may be used for protocol testing.

- The drop-down option for DNA_BARCODING_ONLY is reserved for those specimens submitted solely for DNA barcoding (e.g., when the sample is too small to provide material for both reference genome and barcoding and genome paratype / other specimens must be used as proxies, or when the specimen was identified to species level but died before being preserved, or is otherwise unsuitable for HMW DNA, but the material is valuable for barcoding).

CCC. **HAZARD_GROUP**: If the specimen needs to be processed in a containment level 1, 2, or 3 lab. Please note that any specimens above Hazard Group 1 must be discussed prior to shipping samples. To determine if the species is above HG1, please check both the HSE “Approved List of Biological Agents” and the SAPO list of animal pathogens. If the species is not listed on either of these lists, then it is HG1.

DDD. **REGULATORY_COMPLIANCE**: Please enter Y (yes), NOT_APPLICABLE or N (not known). Note that Sanger ToL will not be able to process further any samples where N is entered.

- Enter Y if you have affirmed that the necessary regulatory compliance documents have been obtained and are available to or at your GAL.

These may include landowner permission, restricted area (SSSI, Nature Reserve, etc.) permission, BAP, CITES or other endangered species permission, Home Office Licencing for sampling for specified animals (vertebrates, cephalopods), phytosanitary permissions, veterinary pathogen sampling permissions.

- If you have determined that no regulatory permissions or documents are required (for example where the sample is from a long-established culture) please enter NOT_APPLICABLE.
- This is an important “per species” check that ensures that permissions were granted to collect and transfer the specimen for this research purpose. The sample provider should ensure this documentation is obtained, and that copies of the relevant paperwork are shared with the sequencing institution where necessary and as stipulated, for example, by regulations/approvals or licencing authorities. Please see the Sanger DToL Sample Submission SOP for details on occasions where it is necessary to provide electronic copies of such documentation to the sequencing institution at the point of Sample Manifest submission.

EEE. **VOUCHER_ID**: (**ENA_submission**) Accession number of voucher material from the sequenced specimen. This ID should be prefixed by the name of the collection (e.g. ATCC:12345) and refers to the physical voucher of the specimen that is accessioned and curated into a collection. This field can be updated in COPO at a later date if accession numbers are not available at the time of sample preparation (e.g., moth bodies sent to Sanger, moth wings to be accessioned and curated at Oxford Museum of Natural History and given a museum accession or acquisition number = Voucher_ID). In such cases please use **NOT_PROVIDED** as a placeholder, allowing for update at a later time.

- In some cases, voucher material will need to be made from a specimen that is different than the one being submitted for sequencing (e.g., a midge is too small to provide both a voucher and a specimen for sequencing, so another midge from the same swarm may provide a para-genomotype voucher). When this is the case, it should be noted.

FFF. **OTHER_INFORMATION**: Free text field for further relevant information not captured by the other fields. This is a place also for partners to flag species that should be prioritized in the sequencing queue. If this species represents one of the two family representatives submitted for the project, please note this here. If there is nothing else to add here, this field should be left **blank**.

[end of document]