Recording Sample Metadata for Darwin Tree of Life - Standard Operating Procedure

Version: 2.1

Published Date: July 2020

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 V2.0 and the Earlham DToL Sample Submission SOP V1.0.

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 quarterly 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 V2.1. In the near future, this will be enhanced by the COPO system (http://copo-project.org), a data management and brokering platform that will allow metadata to be collected either in an online interface or through the downloading of partially filled and re-uploading of fully-filled spreadsheets. COPO will then link to a database that tracks all samples and their associated metadata as they progress from collection to genome assembly. Finally, the data will be 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 required.	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

Completing the Sample Manifest

Overview

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.

Submitting 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 Barcoding SOP, and/or the RBGE DToL Sample Barcoding SOP as appropriate.

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:

NHM: NHMUK000000000 (9 numeric digits) [Heather Allen H.Allen@nhm.ac.uk]

RBGE & UofE: EDTOL00000 (5 numeric digits) [EDTOLnumbers@rbge.org.uk]

KEW: KDTOL00000 (5 numeric digits) [e.gaya@kew.org]

EARLHAM: EI_00000 (5 numeric digits) [seanna.mctaggart@earlham.ac.uk]

MBA: MBA00000 [Rob Mrowicki robmro@mba.ac.uk]

OXFORD/WYTHAM: Ox000000 (6 numeric digits) [liam.crowley@zoo.ox.ac.uk]

SANGER: SAN0000000 (7 numeric digits) [neh@sanger.ac.uk]

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., epicollect, 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 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 may be required. Manifests can be validated and held until barcoding results are back and relevant fields (e.g.,

SCIENTIFIC_NAME, PUBLIC_NAME, TAXON_ID) can be updated. The process for "updating" a validated manifest will be developed over the coming months.

For all samples sent to Sanger, if there is any possibility of species misidentification (SPECIMEN_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 V2.0**). 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 BioSample, 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 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.

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 DToI 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.

Filling in the Sample Manifest

1. The manifest has two 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).

2. **Information must be entered for all fields below with turquoise bold names** [in the Google spreadsheet version of the manifest, these fields also have purple cells that will go white when an entry has been made to help you identify where mandatory fields still require data.] For all fields with **turquoise bold names**, even if information is unavailable, these fields still need to be populated with the appropriate term for why this information is missing. The acceptable missing value terms are as follows:

   NOT APPLICABLE = information is inappropriate to report, can indicate that the standard itself fails to model or represent the information appropriately.

   NOT COLLECTED = information was not given because it has not been collected.

   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 for why the information is missing because we expect that many samples will not have information for these fields (e.g., most samples won't have Depth information). However, if you have collected the information related to these terms, please do enter it. Finally, 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.

3. **All dates** in the manifest **must be** formatted consistently as **YYYY-MM-DD** (ISO 8601).

4. Column by column instructions for the **Metadata Entry** tab are as follows:

   a. **SERIES**: This is to 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**: Use barcoded racks (or plates where relevant) for your samples and enter the barcode of the rack (or plate) here. This should be scanned in and not manually entered.

   c. **TUBE_OR_WELL_ID**: This is 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 between 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 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 (e.g. single strands from a mat), and place each in a separate specimen tube with a unique SPECIMEN_ID. Each GAL has their own bank 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**: Self explanatory

f. **FAMILY**: Self explanatory

g. **GENUS**: Self explanatory

h. **TAXON_ID**: (**ENA_submission**) A valid NCBI TAXON_ID to the species level is mandatory in order to submit data to public repositories. 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.

   You can also check if the name is submittable to ENA here: www.ebi.ac.uk/ena/taxonomy/rest/scientific-name/

   If no TAXON_ID exists or the TAXON_ID exists but the name is incorrect or outdated, please request a new one by writing to ena-dtol@ebi.ac.uk, providing the full name, authority, and publication where possible. 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 lowest taxonomic rank identification as possible (ORDER_OR_GROUP/ FAMILY/ GENUS) and leave SCIENTIFIC_NAME blank.

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

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. Can also comment on STRAIN availability, if specimen is a representative of a living and accessible strain/colony/culture. If there are no issues, leave this field empty.

k. **INFRASPECIFIC_EPITHET**: Where the sample is from a formally named infraspecific taxon, give the infra-specific 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 empty.

l. **CULTURE_OR_STRAIN_ID:** (**ENA_submission**) This is only relevant if the sequenced material is derived from a living, culturable, named laboratory strain. This should not reflect a specific strain that can currently be collected

in the wild but rather that there is a formal collection in which the strain is referenced by a specific ID (e.g. *Anopheles coluzzii* N'Gousso strain). Leave this field empty if it is not relevant.

m. COMMON_NAME: Vernacular name, if the species has one. If multiple names are required, separate names with a | character. If you are unsure of or the species has no vernacular name leave this field empty.

n. **LIFESTAGE**: (**ENA_submission**) Use the drop down menu or look at the available terms on the second tab. Please note that there are attributes for animals, for plants/fungi/macroalgae, and for protists.

o. **SEX**: (**ENA_submission**) Use the drop down menu. If 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. If this field is not appropriate for your organism, please use NA. 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 every 3-6 months.

q. **SYMBIONT**: Scientific names of expected other species present in this specific sample, taking consideration of the ORGANISM_PART. If multiple species expected, separate using |. If the sample is a parasite collected from a host, then this term should be used to record the host species, if known. If there is no information on potential symbionts, this can be left blank.

r. RELATIONSHIP: (**ENA_submission**) Free text field to declare if the specimen has a known parental, child, or sibling relationship to 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, please concisely state "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" etc.. If the relationships are 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 xSPECIMEN_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 can be left blank.

s. **GAL:** (**ENA_submission**) Use the drop down menu to select the GAL responsible for this sample. If the GAL is 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. Ideally 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**, e.g. #, !, ^, *, etc. In some cases, you will be splitting the same 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 = Ox0005, and you put the head in one tube and the thorax in another, your COLLECTOR_SAMPLE_IDs might reflect this with one tube called Ox0005-h and the other called Ox0005-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 CAPs, and separate names with | (vertical pipe symbol), e.g., DOUGLAS BOYES | LIAM CROWLEY. We note that storage of names with affiliations in a database brings the DToL system under the aegis of the GPDR 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.

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

y. **COLLECTION_LOCATION**: (**ENA_submission**) General description of the location. This should start with the country, 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 OTHER_INFORMATION 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. **DECIMAL_LATITUDE**: (**ENA_submission**) In decimal degrees, between -90 and 90. We advise that locations are specified to minimum of 3 decimal places (https://en.wikipedia.org/wiki/Decimal_degrees)

aa. **DECIMAL_LONGITUDE**: (**ENA_submission**) 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)

bb. **GRID_REFERENCE**: Information to geolocate the sample area, preferably with a map or standardised geolocation reference, e.g. OS GRID REF: SP 45998 08751. https://osmaps.ordnancesurvey.co.uk/ is useful to map lat-long to grid references. This field is optional and can be left blank.

cc. **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).

dd. **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 empty if the depth was not recorded or it is not an applicable field.

ee. **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 empty if the elevation was not recorded or it is not an applicable field.

ff. **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 blank if the time was not recorded.

gg. **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 collection container*.

hh. **EASE_OF_SPECIMEN_COLLECTION**: Drop down menu describing how easy it is likely to be to collect the species in future, e.g. *easy, easy but seasonal, moderate, moderate but seasonal, difficult, difficult and seasonal.*

ii. **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 |, e.g., DOUGLAS BOYES | LIAM CROWLEY. We note that storage of names with affiliations in a database brings the DToL system under the aegis of the GPDR 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).

jj. **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.

kk. **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.

ll. **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).

mm. **PRESERVED_BY**: Name of person that carried out the preservation, 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 GPDR 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.

nn. **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.

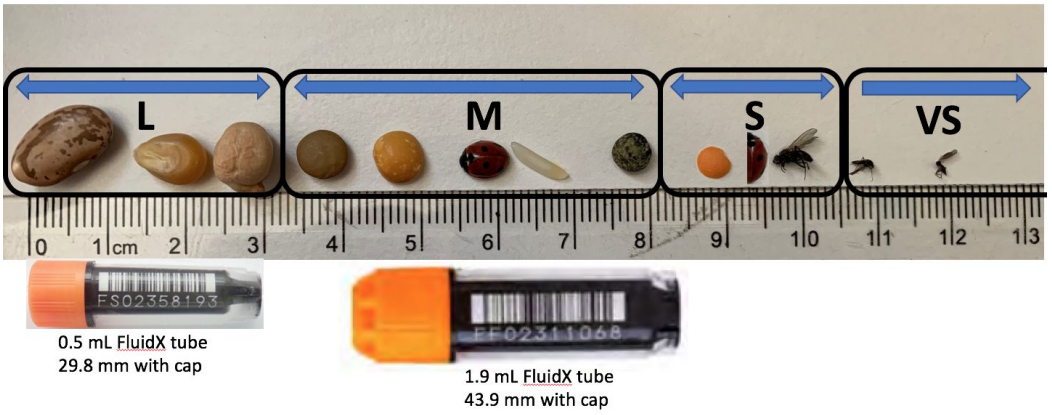
oo. **PRESERVATION_APPROACH**: e.g. snap frozen, dry ice, ethanol/dry ice slurry, in RNALater, lyophilised, air dried, etc.

pp. **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 empty.

qq. **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.

rr. **DATE_OF_PRESERVATION**: Date at which the species was euthanized. Please use YYYY-MM-DD format.

ss. **SIZE_OF_TISSUE_IN_TUBE**: Aim for one lentil-sized piece per tube but sometimes adding more or less tissue than this will be necessary. Note the approximate size of the piece or pellet: use “VS” for very small, “S” for small (~red lentil sized), “M” for medium (~yellow lentil/ladybird sized/5mm), or “L” for large (>5mm, chickpea/bean sized). 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). If the specimen is a single-cell, use “SINGLE CELL(S)”. See visual guidance below.



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

tt. **TISSUE_REMOVED_FOR_BARCODING**: State “Y” or “N”. See the appropriate Molecular Barcoding SOPs for detailed instructions.

uu. **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.

vv. **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.

ww. **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 "Fungal DNA extraction"

xx. **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.

yy. **PURPOSE_OF_SPECIMEN**: The majority of specimens will be for "REFERENCE GENOME" but if a particular tissue is needed solely for RNAseq or if a specimen is intended for "RESEQUENCING(popgen)", please note this here. All samples listed for REFERENCE GENOME sequencing are assumed to also need DNA BARCODING and RNA-sequencing, so the term "REFERENCE GENOME" encompasses three things (ref genome, barcoding, rnaseq wherever samples allow). 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 other specimens must be used as proxies, or when the specimen has been identified to species level but died before being preserved (or is otherwise unsuitable for HMW DNA), but the material is valuable for barcoding).

zz. **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.

aaa. **REGULATORY_COMPLIANCE**: (Where applicable) use Y or N to denote that the appropriate regulatory compliance documents for this collection event are held by the sample provider. 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.

bbb. **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 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 voucher). When this is the case, it should be noted.

ccc. **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. Please add “High Priority” to the field here if this is true. 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 empty.