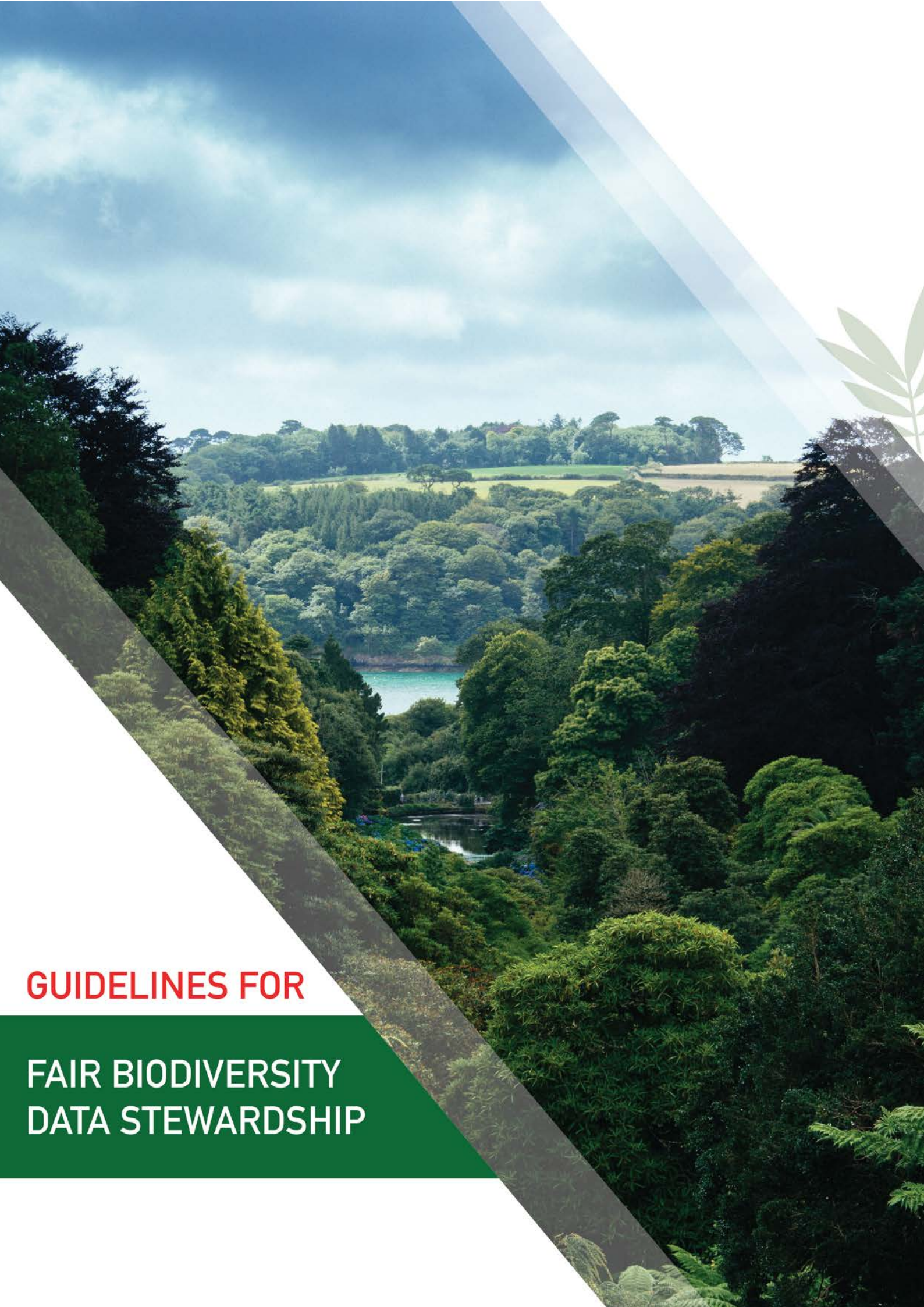


# GUIDELINES FOR

# FAIR BIODIVERSITY DATA STEWARDSHIP

# GUIDELINES FOR FAIR BIODIVERSITY DATA STEWARDSHIP


2024

**THE GUIDELINES FOR FAIR BIODIVERSITY DATA STEWARDSHIP**

These guidelines are developed by the subject matter experts under the FAIR Data Stewardship Guidelines for Reproducibility in Biodiversity Research (Phase I) project. This document is prepared as a guide for biodiversity communities in Malaysia to adopt best practices in biodiversity data management, digitisation and data sharing of biodiversity specimen collections.



Academy of Sciences Malaysia
Level 20, West Wing, MATRADE Tower
Jalan Sultan Haji Ahmad Shah off Jalan Tuanku Abdul Halim
50480 Kuala Lumpur, Malaysia

Perpustakaan Negara Malaysia Cataloguing-In-Publication Data
GUIDELINES FOR FAIR BIODIVERSITY DATA STEWARDSHIP
e ISBN 978-629-7712-02-4

# TABLE OF CONTENTS

# PREFACE

The FAIR Data Stewardship Guidelines for Reproducibility in Biodiversity Research (Phase I), a project supported by the Academy of Sciences Malaysia and the International Science Council Asia Pacific Region (ISC ROAP), has been tasked to draft a guideline on biodiversity data sharing and management practices that meets the Findable, Accessible, Interoperable and Reusable (FAIR) principles. The task is in line with *Projek Pendigitalan Spesimen dalam Program Inventori Koleksi Saintifik Biodiversiti Kebangsaan* coordinated by the Ministry of Natural Resources, Energy and Climate Change (NRECC, formerly known as the Ministry of Energy and Natural Resources). The Guideline was drafted by subject matter experts in collaboration and consultation with various stakeholders through a series of participations in webinars and conferences, meetings, and engagements.

Undoubtedly, in order to make sure that biodiversity research is reproducible, one important aspect of biodiversity data sharing and management practices is to fulfil the FAIR principles. Challenges to support FAIR biodiversity data arise from the various stages of the lifecycle, starting from data creation, data processing, data analysis, data preservation, and data sharing to data reuse. To address such challenges, these Guidelines will describe practical guidance and recommendations to ensure that biodiversity specimen collections are managed and digitised using standardised processes and workflows. The Guideline will also explain tools and checklists for biodiversity collection centres, museums, and depository institutions to achieve high data quality and fitness-for-use of biodiversity data. More importantly, these Guidelines will provide a reference for data collectors, data curators, data custodians, data aggregators and data users on their roles in the data-sharing ecosystem for biodiversity.

As one of the most megadiverse countries in the world, the implementation of best practices and recommendations for biodiversity data stewardship and management at each collection centre, museum and depository institution will help Malaysia to maximise protection and conservation efforts, to inform policymakers on policy measures for good governance of biodiversity in the country and many other benefits. It is hoped that the adoption of these Guidelines will lay the foundation for an exciting Open Science journey for biodiversity research in Malaysia.

## CONDUCT OF STUDY

These Guidelines aim to address challenges related to data sharing, digitisation and management in each stage of the data lifecycle for data collectors, data curators, data aggregators, data custodians and data users. By adopting the Guideline, it is hoped that biodiversity data management, digitisation, and data-sharing practices among biodiversity communities in Malaysia will be more unified, systematic, and streamlined to facilitate the Open Science journey for biodiversity in the country.

The methodology for these Guidelines encompasses comprehensive strategies to strengthen data stewardship in biodiversity research in Malaysia. Hence, this study is guided by three focal objectives, namely, (1) crafting a landscape assessment to determine and meet researchers' data needs, enhancing data stewardship awareness and readiness, (2) establishing a clear hierarchy and understanding the data journey to streamline data flow and stewardship roles within the lifecycle; and (3) formulating comprehensive data stewardship guidelines for various stakeholders, including researchers and IT professionals. These Guidelines are drafted based on inputs from discussions with subject matter experts, stakeholder engagements, participation in conferences, and workshops with data curators from collection centres and museums in Malaysia.

# GUIDELINE AUTHORS

Associate Professor Dr Liew Thor Seng
Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah

ChM Dr Ong Song Quan
Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah

Dr Phon Chooi Khim
Forest Research Institute Malaysia

Dr Nurzatil Sharleeza Mat Jalaluddin
Faculty of Science / Centre for Research in Biotechnology for Agriculture, Universiti Malaya

Suhaila Azhar
Centre for Research in Biotechnology for Agriculture, Universiti Malaya

Sheikh Muhammad Arif Shaifuddin Sh Zulrushdi
Centre for Research in Biotechnology for Agriculture, Universiti Malaya

# CONTRIBUTORS

Dr Hwang Wei Song
Lee Kong Chian Natural History Museum of National University of Singapore

Ong Su Ping
Forest Research Institute Malaysia

Yasser Mohamed Arifin
Ministry of Natural Resources, Energy and Climate Change

Associate Professor Dr Liew Chee Sun
Faculty of Computer Science & Information Technology, Universiti Malaya

Dr Nada Badruddin
Forest Research Institute Malaysia

Dr Izfa Riza Hazmi
Centre for Insect Systematics, Universiti Kebangsaan Malaysia

Dr Siti Nurlydia Sazali @Piksin
UNIMAS Insect Reference Collection, Universiti Malaysia Sarawak

Dr Tan Boon Chin
Centre for Research in Biotechnology for Agriculture, Universiti Malaya

Lim Kooi Fong
Biovis Informatics SDN BHD

Dr Yong Kien Thai
Faculty of Science, Universiti Malaya

Dr Ng Ting Hui
Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah

Professor Dr Abrizah Abdullah
Faculty of Arts & Social Sciences, Universiti Malaya

Associate Professor Dr Sarinder Kaur A/P Kashmir Singh
Faculty of Science, Universiti Malaya

# ABBREVIATIONS AND ACRONYMS

ACE - Access Database Engine

APS-C - Advanced Photo System type-C

ASM - Academy of Sciences Malaysia

AWS - Amazon Web Services

BOLD - Barcode of Life Data Systems

BORNEENSIS - Reference Collections of ITBC, UMS

CF - Compact Flash

CMOS - Complementary Metal-Oxide-Semiconductor

DBMS - Database Management System

DDBJ - DNA Databank of Japan

DMP - Data Management Plan

DSLR - Digital Single-Lens Reflex

DwC - Darwin Core

ENA - European Nucleotide Archive

FAIR - Findable, Accessible, Interoperable, and Reusable

FRIM - Forest Research Institute Malaysia

GBIF - Global Biodiversity Information Facility

HDMI - High-Definition Multimedia Interface

INSDC - International Nucleotide Sequence Database Collaboration

ISC ROAP - International Science Council Regional Office for Asia and the Pacific

ITBC - Institute for Tropical Biology and Conservation

JPEG - Joint Photographic Experts Group

MFT - Micro Four Thirds

MOSP - Malaysia Open Science Platform

MS - Microsoft

MyBIS - Malaysia Biodiversity Information System

NoSQL - Not Only Structured Query Language

OCR - Optical Character Recognition

SD - Secure Digital

SDBMS - Specimen Database Management System

SQL - Structured Query Language

TDWG - Taxonomic Databases Working Group

ToT - Training of Trainers

TTL - Through-The-Lens

UID - Unique Identifier

UIRC - UNIMAS Insect Reference Collection

UMS - Universiti Malaysia Sabah

UNIMAS - Universiti Malaysia Sarawak

USB - Universal Serial Bus

WIFI - Wireless Fidelity

# GLOSSARY

| | |
|---|---|
| **Analogue data** | Data that is represented in physical media |
| **Biodiversity data management workflow** | The five stages of the biodiversity data management workflow: (1) cataloguing, (2) labelling, (3) curating and storing, (4) retrieving and analysing, (5) disseminating |
| **Biodiversity information facilities** | Web services are offered for the purpose of making scientific data on biodiversity available online |
| **Biodiversity researchers** | Biodiversity researchers are involved in all or some of the following: collecting, processing, curating, using, and managing biological samples; cataloguing, curating, validating, interpreting, analysing, and aggregating and publishing specimen information (including information on sampling, collection facilities and taxonomy) |
| **Camera lens** | A tool used to bring light to a fixed focal point |
| **Camera sensor** | A piece of hardware inside the camera that captures light and converts it into signals, which result in an image |
| **Collections** | Specimens that were accessioned by depository institutions |
| **Collection facilities** | Facilities in depository institutions that are designated for the long-term storage of the specimens collected from the field |
| **Curated data** | Data of the collections that may change or update at any of the stages of the data lifecycle, for example, taxonomy information |
| **Darwin Core (DwC)** | A standard for exchanging information about biological diversity |
| **Data management** | Data management is the process of collecting, organising, securing, and storing an organisation's information in order to be analysed and reused |
| **Data stewardship** | The collection of practices that ensure an organisation's data is accessible, usable, safe, and trusted |

| | |
|---|---|
| **Data cleaning** | The process may include format checks, completeness checks, reasonableness checks, limit checks, review of the data to identify outliers (geographic, statistical, temporal or environmental) or other errors, and assessment of data by subject area experts (E.g., taxonomic specialists) |
| **Data integration** | A process of merging data from several sources into a single, unified view |
| **Data lifecycle** | The series of steps that a certain data unit undergoes after being created |
| **Data migration** | A process of moving data from a digital data spreadsheet or physical logbook to a database management system |
| **Data quality assessment** | Data quality assessment evaluates the data's fitness for use in a specific context |
| **Depository institutions** | The institutions with collection facilities to store and manage the collection long-term |
| **Digital single-lens reflex camera** | An advanced type of digital camera that provides high-level image quality |
| **Digital object identifiers (DOI)** | A unique alphanumeric string of characters assigned by a registration agency (the International DOI Foundation) to identify content and create a permanent link to its location on the Internet |
| **Digitisation** | Conversion and capture of data associated with physical specimens and artefacts into electronic formats and databases. This process may involve imaging |
| **Existing specimens** | Specimens that were collected and deposited in the collections of depository institution |
| **FAIRification/FAIRify** | A process aims at addressing the translation of raw datasets into FAIR datasets |

| | |
|---|---|
| **Gazetteer** | An index of geographical features and their locations, often with geographic coordinates |
| **Georeference** | The process (verb) or product (noun) of interpreting a locality description into a spatially mappable representation using a georeferencing method |
| **Herbarium** | A collection of plant specimens preserved, labelled, and stored in an organised manner that facilitates access |
| **Holotype** | A single specimen upon which the description and name of a new species is based |
| **Kit lens** | A lens that comes with the camera upon purchase |
| **Macro lens** | A lens that allows focusing extremely close to the subject |
| **Megadiverse country** | Country that has the largest indices of biodiversity |
| **Metadata** | Data about data. Metadata is the descriptors used for describing, tracing, using, and managing the deposited item. Metadata describes characteristics such as content, quality, format, location and contact information |
| **Newly collected specimens** | Specimens that were collected recently after the implementation of the biodiversity data management |
| **Normalisation** | Procedures or processes to reduce data redundancy and to ensure data dependencies make sense |
| **Primary biodiversity data** | Primary biodiversity data is obtained from a specimen that has been held in a collection facility and contains all or some of the following information about the specimen: Collection facility information (collection name, reference number and status of the specimen), sampling information (collector, method, location and date of collection where the specimen was collected), taxonomy information and other information derived from the specimen |

| | |
|---|---|
| **Reproducibility of research** | Consistent results of research are obtained using the same data and code as the original study |
| **Scholarly publication** | The published findings of researchers who have acquired new information in their respective fields by applying scientific ideas and procedures |
| **Specimens** | Organisms collected from the field research |
| **Specimen Database Management System (SDBMS)** | A Microsoft Access template developed for managing specimen data |
| **Static data** | Data of the collections that will not change throughout the data lifecycle, for example, sampling information for the specimens |
| **Taxonomy** | A scheme of classification for organisms |
| **Voucher specimen** | A preserved and archived whole or part of the specimens of a plant or animal |

# EXECUTIVE SUMMARY

"The FAIR Data Stewardship Guidelines for Reproducibility in Biodiversity Research (Phase I)" project is an important work aiming toward strengthening data stewardship support for the Open Science ecosystem in Malaysia. The project will lay the foundation of the biodiversity data journey by ensuring that all challenges and issues along the biodiversity data life cycle are addressed. The project started on 1st October 2021 and has since then engaged with a number of local and international biodiversity communities and participated in several conferences and webinars. The project has also successfully submitted two deliverables: (1) Data Gap-Need Analysis and (2) Draft Guidelines on FAIR Biodiversity Data Stewardship; both have been instrumental in achieving the project objectives. For this exercise, the project will submit the third deliverable of the project, which is the Final Activity Report and the FAIR Biodiversity Data Stewardship Guidelines.

The FAIR Biodiversity Data Stewardship Guidelines were developed by invited subject matter experts who have extensive experience, expertise, and knowledge in three aspects: (1) Biodiversity Data Management, (2) Digitisation, and (3) Data Quality. The inputs were consolidated based on best practices and extensive references from scholarly publications, seminars, conferences, and relevant guidelines. Contents of the Guidelines were also enriched by feedback from biodiversity communities that were obtained during a stakeholder engagement workshop. In essence, the Guidelines describe recommendations for best practices on workflows, tools and software, as well as roles and responsibilities of data collectors, data curators, data aggregators and data users across the stages of the data life cycle. The Guidelines will also be supplemented with training materials in the form of modules, manuals and video tutorials that will be used in capacity-building activities. It is anticipated that the Guidelines and training materials will contribute to ensuring proper and streamlined workflows for Biodiversity Data Management, Digitisation and Quality Control, which are important for achieving FAIR Biodiversity Data Stewardship in Malaysia.

# LIST OF FIGURES

# LIST OF TABLE

# CHAPTER 1

# INTRODUCTION

## 1.1 OPEN SCIENCE, FAIR PRINCIPLES, AND DATA STEWARDSHIP

Data stewardship and management are very much synonymous with Open Science, which is an initiative to make research output (such as data, publications, etc.) more transparent and accessible. Specifically, Open Science is about extending the principles of openness to the whole research cycle (Figure 1.1) based on cooperative work and new ways of diffusing knowledge through digital technologies and new collaborative tools[1]. The ultimate goal of Open Science is to achieve the Findable, Accessible, Interoperable and Reusable (FAIR) principles through the use of machine-actionability tools and harmonisation of data structures and semantics. The outcome, when implemented, will result in better data stewardship and management[2]. In Malaysia, Open Science is introduced as a national initiative through the implementation of the pilot project Malaysia Open Science Platform (MOSP)[3]


**Figure 1.1** The various stages of the data lifecycle.

Adapted from UK Data Archive and British Ecological Society (Data Management)

[1] European Commission 2016, Open Innovation, Open Science, Open to the World, Publications Office of the European Union, viewed 02 June 2022, < https://op.europa.eu/en/publication-detail/-/publication/3213b335-1cbc-11e6-ba9a-01aa75ed71a1>
[2] Wilkinson MD et al 2016, The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*. No. 3,160018
[3] Akademi Sains Malaysia, n.d., *Malaysia Open Science Platform*. Akademi Sains Malaysia, viewed 02 June 2022, <https://www.akademisains.gov.my/mosp/>

Good data stewardship and management are preconditions for supporting knowledge discovery, data sharing and reuse. Data stewardship is defined as "...the long-term preservation of data so as to ensure their continued value, sometimes for unanticipated uses. Stewardship goes beyond simply making data accessible. It implies preserving data and metadata so that they can be used by researchers in the same field and in fields other than that of the data's creators. It implies the active curation and preservation of data over extended periods, which generally requires moving data from one storage platform to another. The term "stewardship" embodies a conception of research in which data are both an end product of research and a vital component of the research infrastructure..."[4]

Data stewards play a major role in advising, supporting, and training researchers on data lifecycle and good data management practices, from initial planning to post-publication. This includes storing, managing and sharing research outputs, such as data and images. From a bigger perspective of data sharing, data stewards will advise and educate researchers on practices that support Open Science and reproducibility of research, ethical, policy and legal considerations during data collection, processing and dissemination. In other words, data stewards are accountable for the quality of the organisation's data assets and the curation process (i.e., data cleaning and preservation), they should work with researchers to develop naming standards, data definitions, and metadata to be used, and provide advice on or assistance for writing research data management plan with researchers[5].

### 1.1.1 Biodiversity Data Management

This Biodiversity Data Management Guidelines focuses on the management of specimens and specimen data from the field to depository institutions, leading to data sharing in accordance with the lifecycle of the data and the principles of FAIR. Traditionally, specimen management and specimen data management have been treated as separate processes within depository institutions, while biodiversity researchers, usually the specimen and data collectors, manage specimens and data outside the organisation of depository institutions. Data collectors – especially researchers outside a depository institution, should have a data management plan as well (see page 25, Chapter 3.1) - and have decided where to deposit specimens collected from a research project. Once the institution is identified and approved by the research permission granting authorities, data collectors should follow the depository institution's policy and guidelines for biodiversity data management. The depository

[4] National Academy of Sciences (NAS) 2009, Ensuring the integrity, accessibility, and stewardship of research data in the digital age, The National Academies Press, Washinton, DC.
[5] Malaysia Open Science Alliance Working Group on Capacity Building and Awareness 2020, *New Career Pathway Acknowledging Open Science Practices*, Academy of Sciences Malaysia, viewed 02 June 2022, < https://www.akademisains.gov.my/mosp/new-career-pathway-via-open-science/>

institution and data collectors outside the depository institution should work closely together in accordance with the biodiversity data management workflow and data lifecycle (Figure 1.1) to ensure seamless integration of data and specimens to avoid loss of data. Therefore, a software application is needed, i.e., a database management system (DBMS), to manage the databases, allowing different users to use them throughout the lifecycle of the data. In most cases, each institution has its own, sometimes unique, database environment – a collective system of components that includes data, software, hardware, people and procedures.

A discussion of the pros and cons of the various databases and software applications is beyond the scope of these Guidelines. However, the requirements of data standards and workflows in the data cycle outlined in these Guidelines can be adopted directly by each data custodian into their preferred DBMS and software application.

### 1.1.2 Digitisation

Digitisation is a process of converting analogue data to digital. This includes transcribing text data from specimen labels and other specimen-related documents into digital records, producing reasonable quality digital images from physical and dried specimens, converting analogue audio and video recordings to digital recordings, converting textual location descriptions into digital georeferences, and converting other specimen-related data into digital format.

For collection centres, museums and depository institutions, digitisation of specimen collections will increase the availability and searchability of specimens when published on online platforms. The richness of digitised specimen collections allows researchers to be more efficient in searching for specimen collections, as well as identifying and annotating specimens. Additionally, occurrence data from collections' digital information may help scientists assess species abundances and ranges, as well as monitor climate change effects, for conservation research and forecasting studies[6,7]. Specimen digitisation takes time, and it is an ongoing process. Therefore, it is necessary to set priorities for specimen digitisation, especially for collections that are vulnerable to threats of degradation, such as rare or sensitive fossils, and rare or threatened species and habitats. Information about these rare or threatened taxa is critical to have a conservation management plan as soon as possible to ensure the survival of these taxa.

[6] Lister AM 2011, Natural history collections as sources of long-term data sets. *Trends in Ecology and Evolution*, vol. 26, no. 4, pp. 153–154.

[7] MacDougall AS, Loob JA, Claydenc SR, Goltzd JG & Hindse HR 1998, Defining conservation priorities for plant taxa in southeastern New Brunswick, Canada using herbarium records. *Biological Conservation*, vol. 86, no. 3, pp. 325–338.

## 1.2 CURRENT STATE OF BIODIVERSITY DATA STEWARDSHIP AND MANAGEMENT IN MALAYSIA

Biodiversity data and datasets are generated in different hierarchical biodiversity levels, from the ecosystem and species to genetic diversity. While the maximum use of these biodiversity data enables a better understanding and management of biodiversity, the data are poorly managed, archived, integrated, shared, and preserved in the country. Our data gap-need analysis highlighted several weaknesses in data stewardship and management for biodiversity research (Figure 1.2):

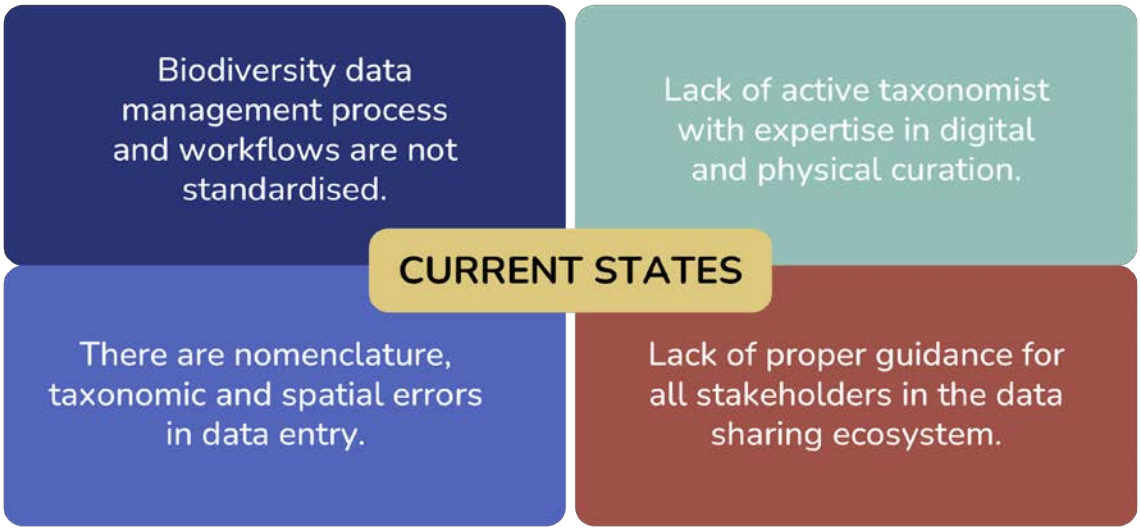



**Figure 1.2**. Findings from the Data Gap-Need Analysis

Another survey[8] has been carried out to gauge biodiversity data management practices and digitisation activities among local biodiversity researchers and data curators. Several key findings are:

i. Respondents use various tools to manage their specimen collections, such as Microsoft Excel, Google Sheets and Microsoft Access, with 60% of them using the databases for less than 5 years.
ii. Digitisation is a relatively new practice among local biodiversity communities, and almost 80% have digitised less than 20% of their respective collections.
iii. 38% of the respondents have deposited digitised specimen data in national or international depositories, 28% have deposited them in internal departments / institutional networks or depositories of their institution, while the other 72% have stored the digitised sample data on physical external drives and hard drives.

[8] The survey responses are also the participants of the virtual stakeholder engagement workshop that was organized under this project on the 4th of July 2022.

## 1.3 THE RATIONALE AND OBJECTIVES FOR THE GUIDELINES

Malaysia is a biodiversity-rich region ranked as the 12th most megadiverse country[9]. The 5th National Report to the Convention on Biological Diversity (2015) reported Malaysia hosts about 15,000 species of vascular plants in Malaysia, with an estimated 8,300 species in Peninsular Malaysia and 12,000 in Sabah and Sarawak. Fauna diversity includes 307 known species of mammals, 785 species of birds, 242 species of amphibians and 567 species of reptiles, as well as 2,068 species of freshwater and marine fishes[10]. Although Malaysian biodiversity accounts for a high flora and fauna diversity, the availability of findable and accessible biodiversity data from the country is significantly lower than in other less-rich biodiversity regions such as Europe and North America[11]. In Malaysia, these primary biodiversity data are mostly undigitised and written in field notebooks, logbooks and specimen labels. Only a limited volume of data is transcribed, integrated, and aggregated into a system, such as the Malaysia Biodiversity Information System (MyBIS).

Challenges in biodiversity data sharing arise from the various stages of the lifecycle, starting from data creation, data processing, data analysis, data preservation, and data sharing to data reuse. To address these challenges, it is important for collection centres, museums, and depository institutions to adopt more unified, systematic, streamlined workflows and standardised metadata formats and schema vocabularies using appropriate tools and software to manage biodiversity data. As there are few experts in taxonomy in our region, it is important to catalogue all the data and make it discoverable and accessible with an appropriate standard so that experts can link data from different sources in the future and improve the quality of the data.

The FAIR Biodiversity Data Stewardship Guidelines (hereafter, 'Guidelines') have been developed to provide best practices for workflows, tools, and software, as well as guidance on the roles and responsibilities of data collectors, data curators, data custodians, data aggregators, and data users that are appropriate to the practical realities of the country. More importantly, it is hoped that the adoption of these Guidelines will lay the foundation for an Open Science journey for biodiversity research in Malaysia.

[9] Malaysia Biodiversity Information System (MyBIS) n.d., *Malaysia – Background*, Malaysia Biodiversity Information System (MyBIS). viewed 02 June 2022, <https://www.mybis.gov.my/art/33>

[10] Malaysia Biodiversity Information System (MyBIS). n.d., *Dataset*, Malaysia Biodiversity Information System (MyBIS), viewed 02 June 2022, <https://www.mybis.gov.my/one/analysis.php>

[11] Stephenson PJ & Stengel C 2020, An inventory of biodiversity data sources for conservation monitoring. *PloS ONE*, vol. 15, no. 12, e0242923.

The main objectives of the Guidelines are to provide:

i. Best practices that are practical to facilitate standardised processes and workflows for managing and digitising biodiversity data.
ii. Tools and checklists to achieve high data quality of biodiversity data.
iii. A reference for data collectors, data curators, data custodians, data aggregators, and data users on their roles in the data-sharing ecosystem.

### 1.3.1 Scopes of Guideline

The Guidelines focus on describing best practices of the workflows to manage biodiversity data throughout the process of archiving specimens from field to shelf. It recommends the specimen data to be digitised so that the texts and images can be processed by a computer. The Guidelines further recommend the use of a DBMS to manage the large volume of digitised texts and images using a database and therefore, allowing different users to have access and manage the data throughout the different stages of the data lifecycle. In these Guidelines and the accompanying manual, we provide a template for a Specimen Database Management System (SDBMS) using Microsoft Access, which requires no special hardware other than a computer and Microsoft Office license which is available in most institutions. We also provide metadata guides, requirements, and recommended tools for data cleaning and data quality assessment that depository institutions can adopt.

In addition, these Guidelines provide recommendations to carry out specimen digitisation works, primarily focusing on producing reasonable quality digital images from physical and dried specimens. Key contents under this chapter will highlight important aspects of planning for digitisation, appropriate workflows, and recommended tools to be used when digitising specimens. However, these Guidelines focus on the acquisition of photos used for general reference purposes and do not include more sophisticated digitisation methods like multiple focus and stacking techniques or 3D photo geometry. The overview of the FAIR Biodiversity Data Stewardship Guidelines is summarised in Figure 1.3.

| | | | |
|---|---|---|---|
| Deliverable | FAIR DATA STEWARDSHIP GUIDELINES | | |
| Domains | **Biodiversity Data Management** | **Digitisation** | **Quality Control** |
| Scopes | 1. Cataloguing<br>2. Labelling<br>3. Curating & Storing<br>4. Retrieving & Analysing<br>5. Disseminating | 1. Speciment Preparation<br>2. Speciment Image Capture<br>3. Speciment Image Processing | 1. Data Quality Assessment<br>2. Data Cleaning |
| Key Elements | **Roles & Responsibilities** | **Workflows** | **Tools** **Software** |
| Outcomes | 1. Stardardized biodiversity data management & digitalization processes and workflows<br>2. High quality and fitness-for-use of biodiversity data<br>3. Clear roles and responsibilities for data custodians, data curators, data aggregators and data users in the data sharing ecosystem for biodiversity | | |

**Figure 1.3.** Overview of the FAIR Biodiversity Data Stewardship Guidelines

# CHAPTER 2

# DATA STANDARDS

All data accompanying the specimens shall be digitised to allow the information to be processed by a computer. This information is then to be put in a database management system (DBMS) that follows a global standard (i.e., Darwin Core) to ensure the data is stored locally and, at the same time, can be retrieved or disseminated effectively from local storage to different data aggregator platforms. The DBMS manages data throughout the lifecycle of the data and allows different users to use it.

Using a standard metadata format can facilitate future digitisation and data integration work. We adopted the Darwin Core Standard (DwC), a metadata format developed by the Biodiversity Information Standards (TDWG) community[12] in these Guidelines. The DwC standard is used for most of the fields in the five data tables in SDBMS and Digitisation process: (1) Collection information, (2) Personnel profile information, (3) Sampling information, (4) Taxonomic information, and (5) Image data information. These five tables consist of 77 fields (i.e., attributes) of biodiversity-related data. Description of the image data information datasheet is adopted from the Image Submission Protocol of the Barcode of Life Data Systems (BOLD)[13]. Each of the tables and fields is described in the Metadata Quick Reference Guide (Appendix 1). User is encouraged to use and reference the Darwin Core Data Standard whenever possible[14,15].

## 2.1 CATEGORY REQUIREMENTS

The 67 fields are grouped into four categories of requirements, namely (1) Required, (2) Required when available, (3) Strongly recommended, and (4) Recommended when available. Rules for category requirements for new specimen data collected from the field and the incomplete data from the existing collection are different, in which a stricter requirement can possibly be imposed for assessing the 'completeness' of newly collected specimens, but it is not realistic to apply similar requirements to

[12]Darwin Core Standard (DwC) is a biodiversity metadata standard adopted by several main biodiversity databases such as Global Biodiversity Information Facility (GBIF), The Atlas of Living Australia (ALA), Ocean Biogeographic Information System (OBIS), FishNet2, Vernet, Encyclopedia of Life (EOL)

[13]Barcode of Life Data Systems (BOLD) System 2013, *Barcode of Life Data Systems Handbook*. Barcode of Life Data Systems (BOLD) System, viewed 02 June 2022, < https://www.boldsystems.org/>

[14]Baker ME, Rycroft S & Smith VS 2014, Linking multiple biodiversity informatics platforms with Darwin Core Archives. *Biodiversity Data Journal*, no. 2, e1039

[15]Darwin Core Task Group 2009, Darwin Core Terms: A quick reference guide, Biodiversity Information Standards (TDWG), viewed 02 June 2022, <https://www.tdwg.org/standards/dwc/>

backlog (existing) specimens. Although the data from existing old collections may be incomplete, the data are still valuable for many purposes, and therefore, some flexibility should be allowed to prevent the data from being lost. In this case, the incompleteness can be compensated by clearly stating the variable completeness and precision levels. The category requirements recommended by these Guidelines are listed in Appendix 2.

## 2.2 DATA QUALITY ASSESSMENT (DQA)

A data quality assessment can be carried out by measuring the completeness, correctness, conformity, consistency, and coherence of the data to determine whether the data meets defined standards. However, the grading scheme for DQA differs for both newly collected and existing specimens. The DQA template recommended in these Guidelines is attached in Appendix 3. The suggested DQA template is a minimum requirement but is not static and will depend on the depository institution's policy on data quality and standards. We recognise that an institution may change the status of the data requirement to a higher level, from '*Recommended when available*' to '*Strongly recommended* ' to '*Required when available* 'or to '*Required*', but we recommend that the level should not be below the recommendations described in these Guidelines.

## 2.3 DATA CLEANING

Data cleaning is a process of refining and improving data quality by correcting errors and omissions that have been identified as inaccurate, incomplete, or inappropriate data during the data quality assessment. A general step-by-step guide for manual data cleaning[16] is:

i. Define and determine the types of errors.
ii. Identify and search for occurrences of the error.
iii. Make the necessary corrections.
iv. Document the error types and instances.
v. Change data entry procedures to reduce the likelihood of future errors.

[16] Chapman AD (2005). Principles and Methods of Data Cleaning, Primary Species and Species-Occurrence Data. Global Biodiversity Information Facility

The following is a list of available data-cleaning tools for biodiversity data[17]:

i. Georeferencing tools, such as Georeferencing Calculator[18,19], Canadensys[20] coordinate conversion, Google Maps and InfoXY to validate geographic data.

ii. Tools or databases for verifying scientific names or taxon rank such as Global Names Resolver for resolving lists of scientific names, Catalogue of Life for references, Taxonomic Name Resolution Service (TNRS) to standardise taxonomic names, and other taxonomically oriented database for checking the validity of the species list or taxon list [21].

iii. Dealing with cluttered data using OpenRefine[19,,22]

iv. Tool to parse dates into its components with Canadensys date parsing.

vi. Tools to parse name strings into their components with GBIF Name Parser[23].

[17] The list of data cleaning tools can be found from the Biodiversity Data Mobilization Course organised by the GBIF Secretariat

[18] Wieczorek C & Wieczorek J 2019, Georeferencing Calculator, Rauthiflor LLC, viewed 02 June 2022, <http://georeferencing.org/georefcalculator/gc.html>

[19] Bloom DA, Wieczorek JR & Zermoglio PF 2020, *Georeferencing Calculator Manual*, Global Biodiversity Information Facility. Copenhagen

[20] Université de Montréal Biodiversity Centre, n.d., Search for records in Canadensys explorer. viewed 27 July 2022, <http://data.canadensys.net/explorer/#tab_simpleSearch>

[21] Penev L, Mietchen D, Chavan V, Hagedorn G, Smith V, Shotton D, Ó Tuama É, Senderov V, Georgiev T, Stoev P, Groom Q, Remsen D & Edmunds S 2017, Strategies and guidelines for scholarly publishing of biodiversity data. Research Ideas and Outcomes 3, no. e12431.

[22] Verborgh R & De Wilde M 2013, Using OpenRefine, Packt Publishing Ltd, Birmingham, UK

[23] Conti M, Nimis PL & Martellos S 2021, Match algorithms for scientific names in floritaly, the portal to the flora of Italy. *Plants*, vol. 10, no. 5, pp. 974.

# CHAPTER 3

# ROLES AND RESPONSIBILITIES

The quality of the data depends on all the personnel involved in any stage of the data lifecycle, as shown in Figure 3.1. All biodiversity researchers generate primary biodiversity data based on specimens collected in the field are responsible and accountable for processing and storing the physical voucher specimens and digitally cataloguing the data[24]. Once a collection has been created; the same biodiversity researchers, personnel in depository institutions, any organisations and other researchers, even members of the public, may play roles as data curators, data custodians, data aggregators, and data users.

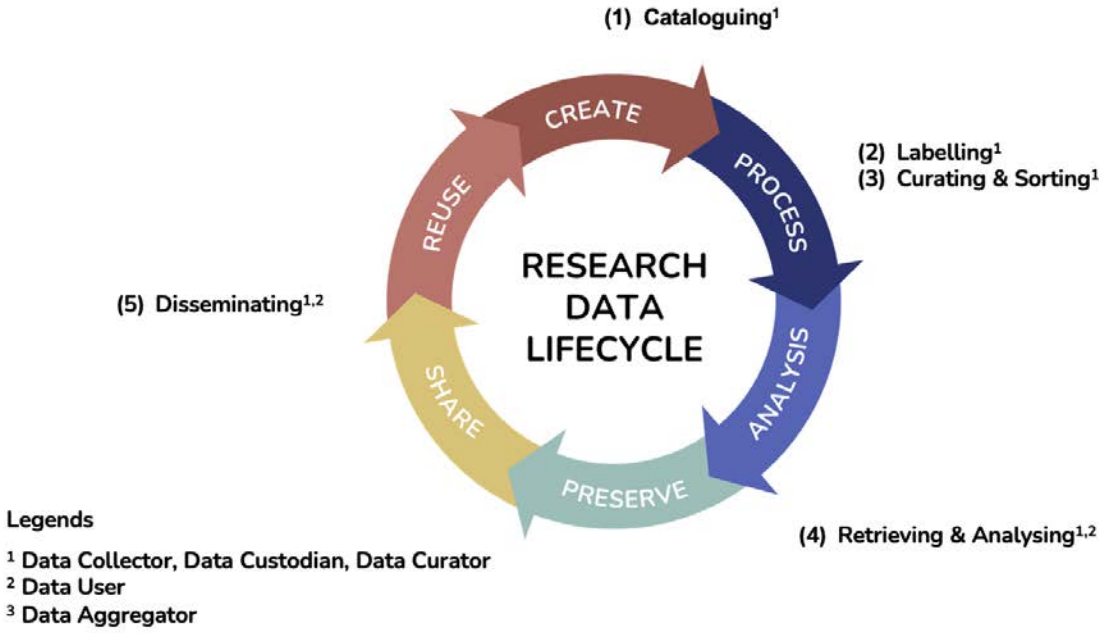



**Figure 3.1.** The combination of data lifecycle, roles, and biodiversity data management workflow. Adapted from UK Data Archive[25], Malaysia Open Science Alliance Working Group on Capacity Building and Awareness[26] and British Ecological Society (Data Management)[27]

[24] Nelson G & Ellis S 2019, The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B*, no. 374.
[25] Van den Eynden V 2013, Data Life Cycle and Data Management Planning, UK Data Archive, viewed 02 June 2022, <https://dam.ukdataservice.ac.uk/media/187718/dmplanningdm24apr2013.pdf>
[26] Malaysia Open Science Alliance Working Group on Capacity Building and Awareness 2020, *MOSP Data Lifecycle*, Academy of Sciences Malaysia, viewed 02 June 2022, <https://www.akademisains.gov.my/mosp/mosp-data-lifecycle/>
[27] British Ecological Society 2018, Data Management, British Ecological Society, viewed 02 June 2022, <https://www.britishecologicalsociety.org/wp-content/uploads/2019/06/BES-Guide-Data-Management-2019.pdf>

## 3.1 DATA COLLECTOR

Data collectors are any individuals who collect the specimens and associated data. All biodiversity studies with different research objectives generate at least standard data, i.e., presence data with taxonomic and sampling information[28]. Therefore, it is necessary to collect as detailed information as possible in the field and enter it into the database as quickly as possible, using the most common and complete standard vocabulary for biodiversity data and other data, for example, multimedia resources associated with specimens[29,30].

Once the information has been entered (typed) into a computer for the first time, it becomes data that will be curated and updated by the biodiversity researcher, curators, or users, regardless of their role. The curated and updated data of the specimens must be complete and accurate. The same data should never have to be re-entered (re-typed) throughout its lifecycle[31] , and it should be used in the entire biodiversity data management workflow—from labelling voucher specimens, generating reports from collection facilities, retrieving and formatting data for analysis, to compiling data for dissemination via online biodiversity information aggregators or scientific publications.

If researchers working on biodiversity wish to incorporate the principles of FAIR into their workflows for managing biodiversity data, they shall plan accordingly and state this clearly in their research proposal, which can be submitted to research permission-granting authorities. At the same time, researchers shall decide which collection facilities to deposit their specimens after reviewing the depository institutions' policies for accepting specimens. Ideally, a depository institution practising the FAIR principles should be chosen. Researchers should include, in any publications produced, information on the depository institution of their specimens. In any circumstances, data collectors should do their best to preserve the physical specimens and labels that accompany the specimens in optimal (long-lasting) conditions before depositing the specimens in collection facilities that are managed by depository institutions.

[28] Mandeville CP, Koch W, Nilsen EB & Finstad AG 2021, Open data practices among users of primary biodiversity data. *BioScience*, vol. 71, no. 11, pp. 1128-1147.
[29] Darwin Core Task Group 2009, Darwin Core Terms: A quick reference guide, Biodiversity Information Standards (TDWG), viewed 02 June 2022, <https://www.tdwg.org/standards/dwc/>
[30] Biodiversity Information Standards (TDWG) 2018, *Audiovisual Core*, Biodiversity Information Standards (TDWG), viewed 02 June 2022, <https://www.tdwg.org/community/ac/>
[31] Sikes DS, Copas K, Hirsch T, Longino JT & Schigel D 2016, On natural history collections, digitized and not- a response to Ferro and Flick. *ZooKeys*, no. 618, pp. 145 – 58.

Finally, it is important to discuss biodiversity data management with each collaborator and each member of the research team so that FAIRification of the data can be agreed[32]. This is an important step before collecting data and specimens, for example, in Sabah, the researcher shall not share the biological resources (voucher specimens and data) with anyone other than the institution named in the application form without the permission of each local staff member and authorities[33].

## 3.2 DATA CURATOR

Data curators would typically be the curator or collections manager of depository institutions, who manage the specimens and the data. While the individual researcher collects specimens and creates data to begin the lifecycle of biodiversity data, the data curator (i.e., collections manager or curator at a depository institution) also plays an indispensable role in subsequent stages of the data lifecycle. Data curators manage collection facilities to preserve specimens and data, and these specimens shall be made accessible to the public upon request by following the policies of depository institutions.

## 3.3 DATA CUSTODIAN

A data custodian would be a single agency or institution (e.g., depository institution) that is most knowledgeable about the content of a dataset and the corresponding management requirements. A data custodian is involved in the governance of every stage of the data lifecycle, including creating, processing, analysing, preserving, sharing, and reusing, particularly for data collectors, data curators and data users.

The number of biodiversity researchers applying the principles of Open Science and FAIR is increasing. Therefore, depository institutions as data custodians should ride the wave of the paradigm shift from collection ownership to collection stewardship by revising their collection management policies to incorporate the principles of Open Science and FAIR[34]. Such revised policies should cover all stages of the data lifecycle, from donation and deposit of specimens to researcher access to specimens, curation of data, and annual reporting of data and metadata.

[32] National Science Council 2020, The Malaysian Code of Responsible Conduct in Research 2nd Edition. Academy of Sciences Malaysia, Kuala Lumpur.
[33] *Sabah Biodiversity Enactment 2000*
[34] Colella JP, Stephens RB, Campbell ML, Kohli BA, Parsons DJ & Mclean BS 2021, The open-specimen movement. *BioScience*, vol. 71, no. 4, pp. 405-414

Data custodians have the role of ensuring that important datasets are created, preserved, and made available according to their predetermined guidelines. Setting up a person or organisation to be in charge of regulating various facets of data management helps prevent datasets from being compromised. The stated data policy that applies to the data, as well as any other applicable data stewardship requirements, should guide how these components are managed[35]. According to Burley and Peine[37], a data custodian's typical duties may include adhering to appropriate and pertinent data policies and data ownership guidelines, making sure that the dataset is accessible to the right users, maintaining appropriate levels of dataset security, performing basic dataset maintenance, such as data storage and archiving, dataset documentation, including updates to documentation, and ensuring the quality and validity of any additions to the dataset.

Depository institutions providing collection facilities must publicise their respective institution's requirements. They should also communicate and coordinate with research permission-granting authorities and key research funders to avoid conflicting provisions in policies so that requirements for data collectors are standardised. While policies and guidelines are set by the FAIR principles to enable FAIRification of new specimens and data that will be deposited in the future, depository institutions need to have a clear strategy for dealing with the existing collection, especially for the existing specimens that have not yet been digitally or even physically catalogued. The respective institutions should have a roadmap for digitising the existing collections[36].

## 3.4 DATA AGGREGATOR

Data aggregators play a role as centralised depository that connects all collection data from different institutions via data aggregator platforms. If data collectors, data curators and data custodians implement the FAIR principles in the first part of the biodiversity data lifecycle, namely collection, processing, analysis and curation at individual and institutional levels, data aggregators play a key role in the later part of the lifecycle, namely sharing and reusing. Indeed, data collectors and data curators share and reuse data, but most of the time, within or in close proximity to their circle, it is usually limited to users who have access to researchers, research groups or depository institutions. User may need to contact different researchers and depository institutions separately to obtain the data and integrate it themselves. The data from collectors, curators and custodians can reach wider

[35] Burley, TE & Peine, JD 2009, *NBII-SAIN Data Management Toolkit. Online*: US Geological Survey Open-File Report 2009, viewed 31 May 2009, <http://pubs.usgs.gov/of/2009/1170/>.
[36] Nelson G & Ellis S 2019, The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B*, no. 374.

audiences through a standard platform created by data aggregators. These data, therefore, can be re-used in future.

It is inevitable that different aggregators will have different objectives. In addition to the basic data standards and information provided by a general platform, a data aggregator platform may be taxa-specific, geographically specific to a country and specific to a project to serve different user groups, each with specific requirements for the platform. Data aggregators should always participate in consortia to ensure that the standard protocol for data exchange is followed.

Data quality and its reliability are important. However, there is no simple diagnosis for data quality deficiencies that data aggregators can perform[37]. Therefore, data standards and the quality of it should be applied by data collectors, curators, and custodians as much as possible. Data users can further improve the quality of data when they examine specific data for their research.

Therefore, the data aggregator should focus on providing a function in the platform that allows users, especially those with relevant expertise, to correct and validate the data aggregated from different sources. This function should allow the expert to communicate with the data collector, curator or custodian about the corrections and validations made to the data. It is more effective and efficient to practise good data management and, when necessary, correct the data at the source than to correct errors downstream[38].

## 3.5 DATA USER

Data users, besides re-using biodiversity data from different sources integrated by data aggregators, curators, and custodians, can improve the quality of data during their research. Data users need to be aware and use the FAIR principles according to the guidelines and requirements of the depository institutions and licensing authorities. This will ensure that the user has permission to modify and reuse the data for scientific, educational, or commercial purposes. Data users must also equip themselves with modern digital tools and skills in using databases to catalogue biodiversity data, use data aggregator platforms, and integrate and analyse biodiversity data[39].

[37] Franz NM & Sterner BW 2018, To increase trust, change the social design behind aggregated biodiversity data. *Database*, vol. 2018, no. bax100.

[38] Nelson G & Ellis S 2019, The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B*, no. 374.

[39] Orr MC, Ferrari RR, Hughes AC, Chen J, Ascher JS, Yan YH, Williams PH, Zhou X, Bai M, Rudoy A, Zhang F, Ma KP & Zhu CD 2021, Taxonomy must engage with new technologies and evolve to face future challenges. *Nature Ecology & Evolution*, vol. 5, no. 1, pp: 3-4.

Whenever possible, data users should publish the research results that reuse these biodiversity data with journals whose publishers practise the FAIR principles and provide a modern publishing scheme. Finally, data users need to acknowledge and cite the data in a responsible way by giving proper credit to the data collectors, curators, custodians, and aggregators.

# CHAPTER 4

# BIODIVERSITY DATA MANAGEMENT WORKFLOW

## 4.1 BIODIVERSITY DATA MANAGEMENT WORKFLOW AND DATA LIFECYCLE

A paradigm shift is proposed for biodiversity data management workflows by involving all stakeholders in data management, regardless of their constant or changing role, as data collector, data curator, data custodian, data aggregator or data user, throughout the data lifecycle by following the same data standards (refer Figure 3.1). These five stages of the biodiversity data management workflow have been described in the context of the data lifecycle: (1) cataloguing, (2) labelling, (3) curating and storing, (4) retrieving and analysing, (5) disseminating.

The workflow proposed in these Guidelines, starting with processing specimens, then creating a comprehensive database of records and subsequently further steps needed to process the next specimens, is not too far from the usual practice in depository institutions [40,41,42]. However, in Malaysia, this workflow is not usually done systematically in many depository institutions. Thus, it is hoped that the detailed workflows and steps as described in these Guidelines, as well as the accompanying example of SDBMS, which is a tool that can be accessible to all stakeholders, will clearly demonstrate the entire process. The aim is not to have perfect and complete data at all stages, but to ensure that a minimum standard of work is achieved at each stage of biodiversity data management according to the principles of FAIR.

[40] Nelson G, Paul D, Riccardi G & Mast AR 2012. Five task clusters that enable efficient and effective digitization of biological collections. *ZooKeys*, no. 209, pp. 19-45.
[41] Integrated Digitized Biocollections (iDigBio) 2016, Digitization Workflows and Protocols, Integrated Digitized Biocollections (iDigBio), viewed 02 June 2022, <
https://www.idigbio.org/wiki/index.php?title=Digitization_Workflows_and_Protocols&mobileaction=toggle_view_desktop>
[42] Hackett RA, Belitz MW, Gilbert EE & Monfils AK 2019, A data management workflow of biodiversity data from the field to data users. *Applications in Plant Sciences*, vol. 7, no. 12, e11310.

## 4.2 DATA MANAGEMENT WORKFLOW GUIDELINES

These Guidelines focus on the workflow of managing biodiversity data throughout the process of archiving specimens from field to shelf[43]. The quality and accessibility of data depend on the quality of the archived specimens. Depository institutions shall maintain a transparent and high standard in terms of policies or procedures for the acquisition of the specimens.

The specimens are permanent references for all derived data and must be made available for inspection so that existing data can be verified and updated (e.g., taxonomic information) and new data can be generated (e.g., genetic data and morphological measurement data).

### 4.2.1 Acquisition and Accession of the Specimens

In the management of biodiversity data and specimens, especially during field collection, the task of ensuring that these data and specimens are acquired following proper procedures, such as obtaining permission from the authorities, is becoming more important these days. Without this procedure, ownership of the data and specimens is questionable and sharing of the data might be restricted. Some enactments and regulations can be referred to, e.g., the Access to Biological Resources and Benefit Sharing Act 2017 [Act 795], the Sabah Biodiversity Enactment 2000 and the Sarawak Biodiversity Centre Ordinance, 1997 with amendments 2014.

After the collector completes research and if the collector wishes for the specimens to benefit others, the collector is encouraged to donate their specimens to a depository institution. The ownership of the specimens can then be transferred from the collector to a depository institution by signing ownership transfer documents. The procedure varies between each depository institution, but the basis of such documents is to establish institutional control over the specimens legally.

All incoming dry specimens should be isolated from existing collections in the depository to avoid introducing pests. The incoming dry specimens shall then be treated to ensure all pests are killed before being incorporated into the existing collection. The NPS Museum Handbook by the National Park Service and many other references provide detailed procedures for treating incoming specimens. Basically, freezing, heating, or creating a low-oxygen environment can be applied to eradicate pests that may be present in specimens. If these methods failed to eradicate pests, the last option would be

[43] The proposed workflow in this Guidelines is adapted from the current practices in the BORNEENSIS, Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah

chemical treatment. For other specimens preserved in fluid, it is important to confirm the type of fluid used and replenish the fluid before incorporating it into the existing collection. During the transportation of the specimens, all measures should be taken to minimise damage to the specimens.

Depository institutions shall obtain as much information as possible about the specimen from the data collector, in particular, the required sampling and taxonomic information. The list of required data can be found in Appendix 2. Specimens must not be catalogued until all acquisition and accession procedures have been completed. Some collectors use codes to indicate different localities. It is the responsibility of depository institutions to obtain information from the collector's field notes to decipher the codes.

### 4.2.2 Database Management System (DBMS)

Microsoft Access is the recommended software for database management systems (DBMS) as it is one of the best entry-level tools for database management and it is also a powerful augmentation compared to Microsoft Excel. Microsoft Access combines the relational Access Database Engine (ACE) with a graphical user interface and software development tools that allow users to quickly acquire basic skills and gradually adapt the DBMS to their specific needs.

The Microsoft Access template that we recommend is hereafter referred as the Specimen Database Management System (SDBMS). The SDBMS consists of four basic tables, namely Personnel Profile Information, Sampling Information, Taxonomic Information, and Collection Information (Figure 4.1). Each table stores a particular type of object/information and is linked with other tables by relationships. Depending on the routines in each depository institution, other tables can be added to the SDBMS, such as the Specimen Loan Information table to record all specimens on loan, the Publication Information table to record all specimens used in a published article, the Project Information table to identify all the specimens collected or recorded in a specific research project and etc.

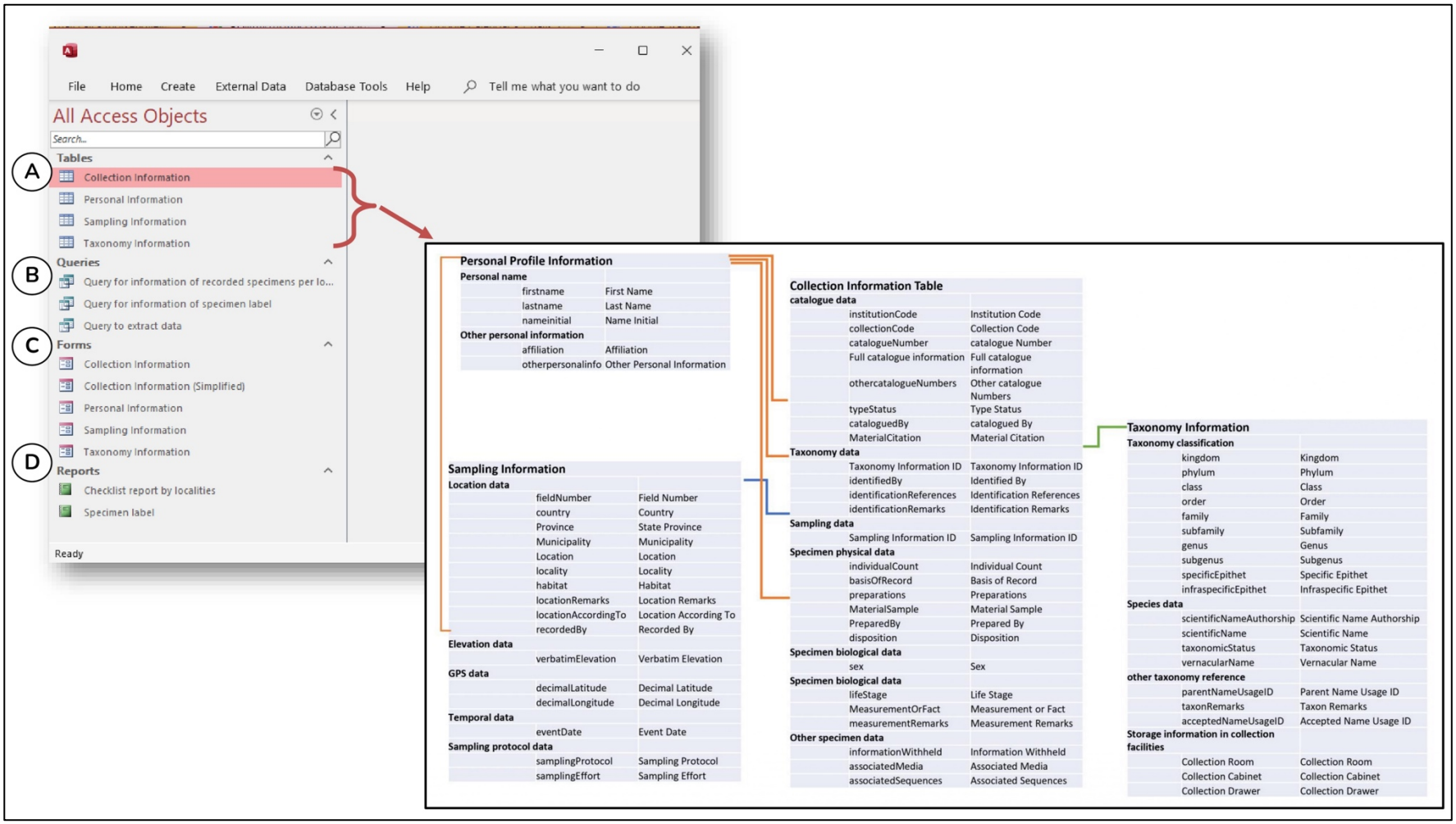



**Figure 4.1** An overview of the objects in the Specimen Database Management System (SDBMS), including (A) four tables, (B) three queries, (C) four forms with one additional form that is a simplified version of the collection information, and (D) two reports. The inset shows the fields for each table and the relationship of the fields between the table

.

In addition to adding tables to enhance the functionality of the SDBMS in handling routines other than specimen data management, user can write a query using SQL to join the information between the tables in the SDBMS and tables in other databases or stand-alone tables (i.e., datasheets). For example, the image data consists of 10 fields, adopted from the Image Submission Protocol of the Barcode of Life Data Systems (BOLD), including the field "Sample ID", which contains the same information as "Full Catalogue Information" of the collection information dataset in the SDBMS. This unique identifier can be used to link the image data information dataset and the collection information dataset to obtain up-to-date taxonomic and sampling information.

A total of 67 fields are included in the four tables (Figure 4.1). These fields are grouped into four categories of requirements, namely (1) Required, (2) Required when Available, (3) Strongly Recommended, and (4) Recommended when Available (Table 4.1). Depending on the depository institution's policy on data quality and standards, the institution may change the status of the data request to a higher level, from Recommended when Available > Strongly Recommended > Required when Available > Required. The original field shall not be removed from the tables in the SDBMS, as many of these fields are the minimum requirement in many other databases. The name of each field shall be maintained as it is to facilitate data export and import into data aggregators.

**Table 4.1** List of fields in the four tables in the Specimen Database Management System (SDBMS) and the status of the fields for each table. If there are differences in the status of the fields for existing and newly collected specimens, these are indicated in brackets. For details of the fields in each table, see Appendix 2[44]

| List | Tables | | | | Total |
|---|---|---|---|---|---|
| | Collection Information | Personnel Profile Information | Sampling Information | Taxonomy Information | |
| Required | 9 | 4 (0) | 14 (0) | 6 | 33 (15) |
| Required when Available | 2 | (4) | 1 (15) | 9 | 12 (30) |
| Strongly Recommended | 10 | 1 (1) | 1 (1) | 2 | 14 (14) |
| Recommended when Available | 5 | - | - | 3 | 8 (8) |
| **Total** | **26** | **5** | **16** | **20** | **67** |

For each table, there is a ready-made form that provides the user with a user-friendly interface for entering or updating data (Figures 4.1C and 4.2B). The form and table carry the same fields (Figure 4.2). Since it is tedious and sometimes confusing to enter data directly into a table with many fields and records, the form is an important tool for the user to arrange the fields in a logical format and make only the relevant fields visible in the form (Figure 4.2). In addition to the data type and input mask settings in the table, which set rules to minimise data entry errors, such as "Short Text", "Date", "Number" and others, user can create a look-up field in the form with a predefined list, e.g., the place name, from a list of values or a table. With this lookup function, user need not re-type the same data again but choose from the predefined list.

[44] See Appendix 2: The number represent the number of fields with designated status for data entry requirement in the collection of depository institution in SDBMS for newly collected specimens and for existing specimens (the number in parenthesis)

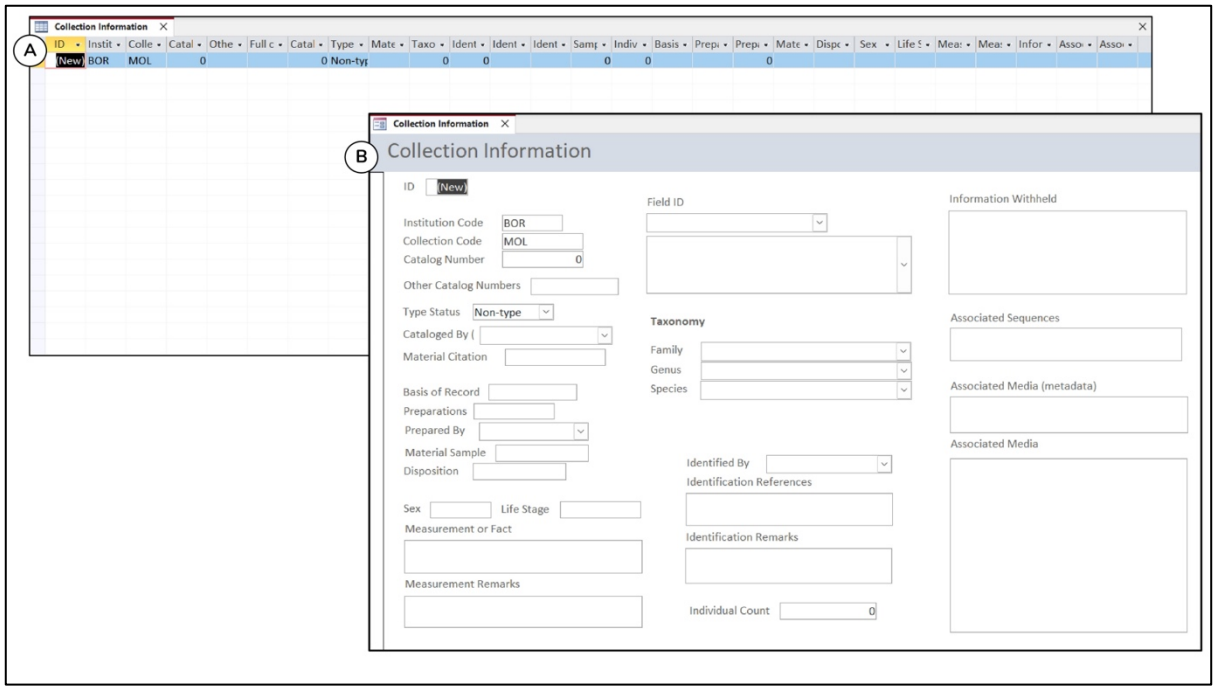



**Figure 4.2**. An overview of (A) the collection information table and (B) the collection form. The form's interface is more user-friendly for data entry than a table with a large number of fields and records

There are two pre-built reports for the user to create the specimen label and the specimen checklist based on the data and records they select using the two queries that correspond to each of the reports (Figures 4.1B, 4.1D and 4.3). User can customise the reports according to their needs. They have the option of structuring the data in a document with conditional formatting, creating statistics on the data and grouping or sorting the data. Finally, there is a template for a query to extract data needed for further analysis, e.g., a data table with a specific row and column format for data analysis or a data sheet with collection information for mapping. Examples can be found in Section 4.2.5.

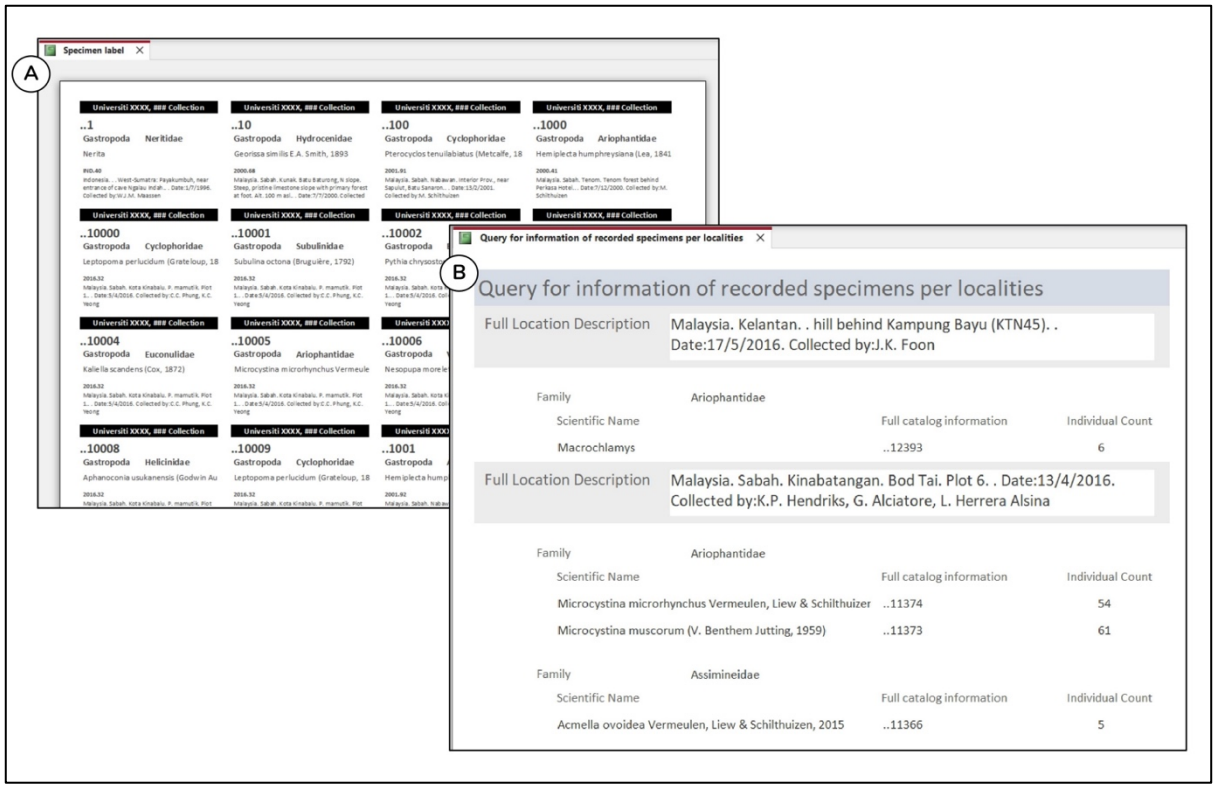



**Figure 4.3.** An overview of (A) the specimen label report and (B) the specimen checklist report. The reports can be customised by user according to their needs

The depository institution must designate a specific person as authorised personnel to manage the SDBMS, who can change the structure of the SDBMS in order to maintain the integrity of the SDBMS. The depository institution shall adapt, integrate, or develop procedures for handling various tasks, including the management of specimens and specimen data, based on these guidelines, so that the database can be integrated and used to facilitate routine and task completion. This SDBMS is not an entirely new procedure, once the depository institution adapts and establishes SDBMS in their routines, it shall improve and speed up the process of digitisation. Depository institutions shall refer to 'the Manual Specimen Database Management System (SDBMS) for FAIR Biodiversity Data Stewardship' or seek advice from subject matter experts to prepare and modify the SDBMS correctly and efficiently. Those institutions that have specimen data in other DBMS but do not fully follow Darwin Core (DwC) terms and standards, or have digital data in Microsoft Excel and wish to incorporate the existing data into the SDBMS (i.e., data migration), should consult subject matter experts as this will require additional steps, including normalising the existing data, i.e., splitting the data into smallest units of information (i.e., table fields) as per DwC[45], moving different types of data

[45] Baker ME, Rycroft S & Smith VS 2014, Linking multiple biodiversity informatics platforms with Darwin Core Archives. *Biodiversity Data Journal*, no. 2, e1039

according to the elements (i.e., tables), and cleaning up data that is inconsistent in terms of notation, text description and data format. Institutions can refer to one of the case studies in the manual for processing and importing existing data into the SDBMS.

### 4.2.3 Cataloguing data from specimens

At the cataloguing stage, the data lifecycle of specimens begins when the data is created and catalogued in the Specimen Database Management System (SDBMS), even if the specimens have not been identified. At least two types of information are available at this stage, namely collection information (e.g., institution code, collection code, catalogue number) and sampling information (e.g., location data, collectors, date), all of which are static data and will not change throughout the lifecycle of the data. Once a specimen has been catalogued, a collection has been created, and the depository institutions play the role of data custodian.

The data of the newly collected specimens from the field and the data from the existing collection are to be treated differently. While it is possible to impose a stricter requirement for completeness of data for new data, it is not realistic to apply the same requirement to backlog specimens as many old collections sometimes do not have complete data. However, the data attached to old specimens are still valuable for many purposes. It is important to remember that suboptimal data is better than none at all, and that the perfect should not be the enemy of the good. As long as the level of variable completeness and precision is clearly stated, the data can still be useful.

Data that were recorded in a physical logbook or are only available on the label of the specimen can be manually entered into the database[46]. If the data are available in a digital spreadsheet, normalisation of the data in the spreadsheet is required (see Section 4.2.2). If the normalisation processes cannot be done, the data can still be entered manually into the SDBMS. This comes at an additional cost to the depository institutions, data collectors, curators, and custodians, but this is key to getting the data up to standard and in a structured way as best as possible[54,47].

[46] Escribano N, Galicia D & Ariño AH 2018, The tragedy of the biodiversity data commons: a data impediment creeping nigher?. *Database (Oxford)*,vol. 2018, no. bay033

[47] Costello, MJ, Michener, WK, Gahegan, M, Zhang, ZQ, Bourne, P & Chavan, V 2012, *Quality assurance and intellectual property rights in advancing biodiversity data publications ver. 1.0,* Global Biodiversity Information Facility, Copenhagen, Pp. 33.

#### 4.2.3.1 Personnel profile information

Throughout the workflow of managing biodiversity data (Figure 4.1), many personnel play different roles: as collectors of specimens from the field ("Recorded By" in the Sampling Information Table) and as curators who prepare and preserve the specimens (“Prepared By"), identify the specimens to species level ("Identified By") and then catalogue the specimen data in a database ("Catalogued By"), all of which should be recorded in the Collection Information Table. One person can play all these different roles, therefore, it is important to have a master list that contains the information of all people involved in the workflow. The Personnel Profile Information Table in the SDBMS is used to store and manage all the profiles of the personnel in terms of their personal information, namely, "First Name", "Last Name", and "Name Initial". It is important that there is only one entry in the table for each person. Other profile information such as "Affiliation" and "Other Personal Information" must also be filled in the Profile Information Table, which can provide additional unique identification of users.

#### 4.2.3.2 Sampling information

The lifecycle of biodiversity data begins with the acquisition of data, primarily through sampling, which generates the sampling information. Sampling information is static data that is finalised after sampling is completed and does not change during the lifecycle of the data. The sampling data must be entered into the database as soon as possible to avoid loss of information. In the SDBMS, there are 16 fields in the sampling information table (Figure 4.4, Appendix 2) relating to GPS data, location data, elevation data, temporal data, and sampling protocol data, of which 14 fields are required for newly collected specimens, one field is required when available, and one field is strongly recommended.

A. Newly collected specimens from the field

First, enter the "field number", which is a unique identifier for the sampling event. This information can be the reference of the sampling event recorded in the logbook of the data collector, e.g., ‘2022.Ali.01’, which means the first sampling event made by Ali in 2022. The field number must be as short but as informative as possible so that entering the sampling information for the specimen can be done more effectively by simply selecting the field number rather than reading and searching for the location details.

Next, enter the verbatim description of the site: "Country", "State Province", "Municipality", "Location", "Locality", and "Habitat". The first four fields can be standardised by providing the standard value list to avoid confusion due to different spellings or different names for the same place (Figure 4.4). For the

"Locality", the description must be as detailed as possible; if possible, it should include not only the name of the place, but also additional information that helps to identify the place.

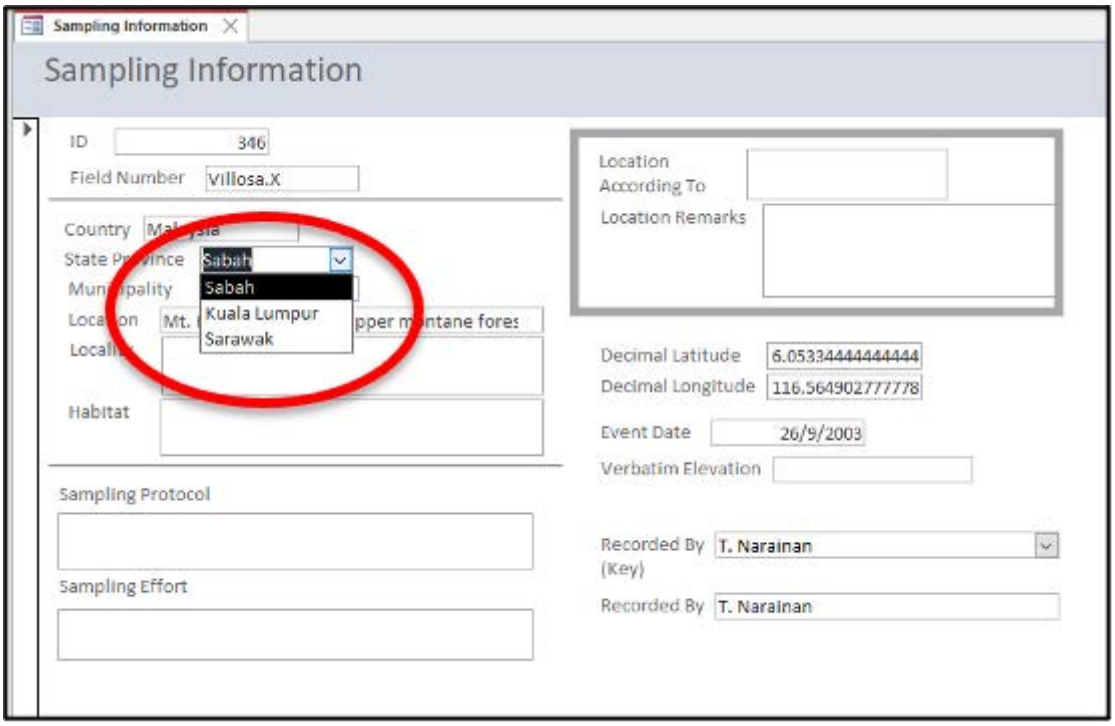



**Figure 4.4.** An example of a standard value list to avoid confusion due to different spellings or different names for the same place

In addition to the textual description of the location, it is important to include the coordinates "Decimal Latitude" and "Decimal Longitude" of the location or area, as the name of the location may change over time. If samples have been taken from more than one location in an area, the coordinates of the centroid of the area may be used, and a note of this needs to be made in the "Location Remarks". It is also recommended that the "Verbatim Elevation" of the area be entered in the sampling information.

It is recommended that the data collector enters information such as "Location Remarks" and "Location According To", especially when the coordinates of GPS are not available or the location description is based on the collector's own interpretation of the information on the specimen's label. Finally, the information on "Event Date", "Recorded By", "Sampling Protocol", and "Sampling Effort" are required. It is important that the sampling design, methods, and effort are specified to allow better inferences and to improve the re-use of data and the reproducibility of the analysis[48,49].

[48] Dobson AD, Milner-Gulland EJ, Aebischer NJ, Beale CM, Brozovic R, Coals P, Critchlow R, Dancer A, Greve M, Hinsley A, Ibbett H, Johnston A, Kuiper T, Comber SL, Mahood SP, Moore JF, Nilsen EB, Pocock MJO, Quinn A, Travers H, Wilfred P, Wright J & Keane A 2020, Making messy data work for conservation. *One Earth*, vol. 2, no. 5, pp. 455-465.
[49] Foster SD, Vanhatalo J, Trenkel VM, Schulz T, Lawrence E, Przeslawski R & Hosack GR 2021, Effects of ignoring survey design information for data reuse. *Ecological Applications*, vol. 31, no. 6, e02360.

B. Existing specimens in collections with incomplete sampling data

Usually, sampling information can be completed for new data or existing collections with complete data. However, for existing collections with incomplete data or data lacking accuracy in the data sheet or on the specimen label, it is important to digitise all the available sampling information. Do not dispose of the specimen as it can still be very useful for taxonomy and larger scale inventory research for coarser resolution data, such as general location description[50]. Georeferenced the textual locality can still be estimated based on the gazetteer.

#### 4.2.3.3 Taxonomy information

When user use the SDBMS for the first time, they can either enter the taxonomy information manually into the SDBMS or import tabular taxonomy information that follows the field names in the same format as the taxonomy information table directly into the SDBMS. Thereafter, the new taxa name can be added to the table from time to time by manual entry in the SDBMS. In addition to the default taxonomic classification level fields in the Taxonomy Information Table, user can add fields of other relevant classification levels according to the classification scheme for organisms. Whenever possible, the taxonomic information of each taxon (i.e., each entry) shall be provided at the lowest possible taxonomic level and at least at the "Kingdom", "Phylum" and "Class" levels.

In addition to taxa for which full taxonomic information is not available, it is worthwhile to establish a provisionally circumscribed genus or species name for a morphospecies in the Taxonomy Information Table for specimens that could not be identified to species or genus level during the cataloguing phase. In any case, the taxonomic information is considered curated data, as the taxonomic information may change, whether due to changes in the taxonomic classification or misidentification of the collection[51,52,53]. These are normal situations for many taxa, especially the invertebrates, whose classification is less stable compared to vertebrates. Therefore, the taxonomic information is not static and should be updated from time to time, it is important to jot down the reasons causing the changes.

[50] Smith AB, Murphy SJ, Henderson D & Erickson KD 2021, Imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth. *Global Ecology and Biogeography*, vol. 32, no. 3, pp. 342 – 55.
[51] Chapman, AD 2005, *Principles of Data Quality*, Global Biodiversity Information Facility, Copenhagen
[52] Goodwin ZA, Harris DJ, Filer D, Wood JR & Scotland RW 2015, Widespread mistaken identity in tropical plant collections. *Current Biology*, vol. 25, no. 22, pp. R1066–R1067.
[53] Mesibov R 2018, An audit of some processing effects in aggregated occurrence records. *ZooKeys*, no. 751, pp. 129-146.

#### 4.2.3.4 Collection information

A. Catalogue data

After the specimens collected from the field have been processed and stored, the collection information is then created. In the SDBMS, there are 26 fields in the collection information table, of which nine fields are required for newly collected specimens, two are required when available, 10 are strongly recommended, and five are recommended when available, that related to catalogue data, specimen collection data, specimen biological data, other specimen data, sampling data and taxonomic data. All of these are entered as new information by the data collector and data curator, except for the sampling and taxonomic data, which come from the Taxonomic Information Table and the Sampling Information Table.

For the catalogue data element, "Institution Code", "Collection Code" and "Catalogue Number" are required. The concatenation of these three data into "Full Catalogue Information" must be a unique value. This is static data generated by the institution and serves as a unique identifier for the specimens and the other information associated with the specimens. The specimens cannot be catalogued until they have been given an accession number by the depository institutions. Reserving catalogue numbers should be avoided as it will mess up the running catalogue number sequence and create unnecessary confusion in future.

For the specimen with existing collection numbers (e.g., donated specimens from other depository institutions), the original voucher specimen number shall be entered in the field "Other Catalogue Numbers". If the specimen is a type specimen, record this data in the field "Type Status". It is strongly recommended to enter the details of the person who catalogued the specimen in "Catalogued By". If the specimen is used in a publication, the citation of the scientific publication shall be indicated under "Material Citation".

B. Linked to sampling and taxonomy information

Once the tables are linked, information stored in other tables can be searched from a drop-down list at each field in the Collection Information Table. We do not need to re-type country, state, locality, Kingdom, Phylum, Class, etc. again. It is strongly recommended to enter the data on "Identified By", "Identification References", and "Identification Remarks". If the collection cannot be identified to species level, it is still important to label it at a higher taxonomic level for storage in depository

institutions, as most institutions arrange and organise collections according to taxonomy (see page 45, Chapter 4.2.5).

C. Specimen collection data

Next, the descriptions of the specimen collection lots are required in terms of "Individual Count", "Preparations", and "Material Sample". At the same time, other information such as "Basis of Record", "Prepared By", and "Disposition" are strongly recommended to be entered into the database.

D. Specimen biological data

Some biological data of the specimens might have been collected in the field or during the processing of the specimens, for example, measurements, colours and other fields of the morphological features of the specimens. If these data are available, they shall be entered into the SDBMS. These data have the status "recommended when available", including "Sex", "Life Stage", "Measurement or Fact", and "Measurement Remarks". Much of this biological data obtained from specimens immediately after their collection or preservation is valuable as some of the characteristics may be lost after the death of the organisms and the preservation process.

#### 4.2.3.5 Other specimen data

It is important to enter the information into the database when the “Associated Media” and “Associated Sequences” are available. For “Associated Media”, data users and curators can enter the data on the Digital Object Identifiers (DOI) of the images or videos in publications or online repositories[54]. Images can be deposited in general repositories such as Zenodo[55], figshare[56] or Flickr[57] or in specialised repositories for biodiversity images such as Morphbank[58], iNaturalist[59], and for 3D data such as Morphosource[60].

Similarly, once research is concluded and the manuscript is under preparation, DNA sequences shall be uploaded to GenBank®[61] to obtain the GenBank accession number. This number shall be entered

[54] Penev L, Mietchen D, Chavan V, Hagedorn G, Smith V, Shotton D, Ó Tuama É, Senderov V, Georgiev T, Stoev P, Groom Q, Remsen D & Edmunds S 2017, Strategies and guidelines for scholarly publishing of biodiversity data. Research Ideas and Outcomes 3, no. e12431.
[55] More information about Zenodo can be found on the website, https://zenodo.org/
[56] More information about Figshare can be found on the website, https://figshare.com/
[57] More information about Flickr can be found on the website, https://www.flickr.com/
[58] More information about Morphbank can be found on the website, https://www.morphbank.net/
[59] More information about iNaturalist can be found on the website, (2022). https://www.inaturalist.org/
[60] More information about Morphosource can be found on the website, https://www.morphosource.org/
[61] More information about GenBank can be found on the website, https://www.ncbi.nlm.nih.gov/genbank/

into "Associated Sequences". In addition to GenBank, genetic data can also be deposited in the two other repositories of the International Nucleotide Sequence Database Collaboration (INSDC), namely the European Nucleotide Archive (ENA)[62] and the DNA Databank of Japan (DDBJ)[63]. User can also deposit genetic data via a connected depository such as Barcode of Life Data Systems (BOLD)[64].

For some specimens, some information may be available but not entered in the designated fields of the tables. This information and the explanation for this decision shall be entered in the field "Information Withheld". Special attention must be paid to sensitive data to prevent potential threats to biodiversity[65,66].

### 4.2.4 Labelling

In the labelling stage, specimens must be labelled based on the information created in the cataloguing stage, while initial preservation and at least preliminary sorting of specimens are completed, and specimens may be temporarily held in the processing laboratory for further examination and identification.

Label should be attached to specimen at any time. There is a function in the SDBMS that enables the user to print labels once the specimens are catalogued (Figure 4.3A). The data collector and curator can select the information in SDBMS, customise the format, including text font format and label size, and then create a label that is saved in PDF format for printing. In any circumstances, three important pieces of information must be included on the label, namely (1) "Full Catalogue Information" consisting of "Institution Code", "Collection Code" and "Catalogue Number", (2) "Field Number" and as much detail of the sampling information as can be accommodated on the label, and (3) taxonomic information. A label indicating that the specimens are donated shall be attached to the specimens. Data curators, data users and data custodians shall bear in mind that any original labels with the specimens shall not be removed.

[62] More information about European Nucleotide Archive (ENA) can be found on the website, https://www.ebi.ac.uk/ena/browser/home/

[63] More information about DNA Data Bank of Japan can be found on the website, https://www.ddbj.nig.ac.jp/index-e.html

[64] More information about Barcode of Life Data Systems (BOLD) can be found on the website, https://www.boldsystems.org/

[65] Chapman, AD & Wieczorek, J 2006, *Guide to best practices for georeferencing*, Global Biodiversity Information Facility, Copenhagen

[66] Orr MC, Ferrari RR, Hughes AC, Chen J, Ascher JS, Yan YH, Williams PH, Zhou X, Bai M, Rudoy A, Zhang F, Ma KP & Zhu CD 2021, Taxonomy must engage with new technologies and evolve to face future challenges. *Nature Ecology & Evolution*, vol. 5, no. 1, pp: 3-4.

One of the most important aspects of labelling is the materials used to label the specimens, as some materials can damage the collections, especially for specimens preserved in fluid. Printing using pigment ink on archival, acid-free paper without whitening/bleaching agents is encouraged. Discussion of labelling materials for different type of organisms can be found on Duckworth et al.[67], Elkin and Norris[68], Huxley et al.[69] and National Park Service [70].

### 4.2.5 Curating and Storing

During the curating and storing stages, specimens are identified to the lowest taxonomic level possible, and digitisation of the specimens (See page 52, Chapter 5) may be undertaken. Once the specimens have been catalogued, preserved, labelled, identified, and photographed, they can be permanently stored in the collection facilities. During this and the previous stages (i.e., labelling, curating and storing), the data collector, data curator and data custodian play an important role in establishing the permanent record of the specimens and data following a standard format with the required minimum information that suitable for analysis later. At these stages, too, the data is refined and updated.

It is very important to keep the collections in a secure environment. Even if the collections are not stored in permanent collection facilities and are still in the laboratory workrooms for further processing and analysis, depository institutions must have a policy in place to ensure that specimens are handled carefully and that collections are traceable at all times. The collections shall be deposited in the collection facilities as soon as the processing and analysis of the specimen collection is completed.

If the collections are not used immediately, they must be stored in the designated collection facilities as soon as possible. Normally, the collections are stored in drawers and cabinets according to the taxonomic classification. Therefore, the information on storage in the collection facilities, which consists of the information "Collection Room", "Collection Cabinet" and "Collection Drawer", is included in the Taxonomic Information Table of the SDBMS. All collections of the same taxonomic unit, e.g., the same species, shall be stored in the same storage room adjacent to the other species of

[67] Duckworth WD, Genoways HH & Rose CL 1993, Preserving natural science collections: chronicle of our environmental heritage. Mammalogy Papers: University of Nebraska State Museum. pp. 271.
[68] Elkin, L & Norris, CA (eds) 2019, Preventive Conservation: Collection Storage. Society for the Preservation of Natural History Collections, New York.
[69] Huxley, R, Quaisser, C, Butler, CR & Dekker, WRJ 2020, Managing Natural Science Collections: A Guide to Strategy, Planning and Resourcing. Routledge. Oxon, OX.
[70] Bacharach J (Ed) 2005, Museum Handbook Part I: Museum Collections. National Park Service Museum Management Program, Washington, DC, viewed 02 June 2022, <https://www.nps.gov/museum/publications/mhi/mhi.pdf>

the same genus. The collection facilities shall be in optimal condition to preserve the preserved collection according to the standard of conservation and preservation. We recommend that data curators and data custodians refer to Duckworth et al.[71], Elkin and Norris[72], Huxley et al.[73] and National Park Service[74] for further guidance on best practices in collecting specimens and managing collection facilities.

At this stage, specimen collection data has met the prerequisite to start the FAIRification process – data has been created and is findable, i.e., collections with unique identifiers "Full Catalogue Information", and specimens with rich data and metadata that conform to a commonly accepted data standard (i.e., DwC), (see page 21, Chapter 2). Indeed, continuous curation (see page 38, Chapter 4.2.3) of collection data is necessary to refine the collection of taxonomic information or location data of doubtful locations that have not been georeferenced[75].

At the same time, the collections need to be digitised by taking images of the specimens and labels with measuring scale (see page 59, Chapter 5). Since the collections already have associated data, this collection data can be shared through a data aggregator. This is an important step for the data custodian and the data curator to ensure that the specimen data is accessible to the data user, who can also act as a data curator, with permission of data custodians, to correct and add information to the specimen collection data, in order to connect data and expertise[76].

### 4.2.6 Retrieving and Analysing

In the retrieving and analysing stages, the data is ready for use. The data collector, data curator, data user and data custodian can retrieve not only the data of the specimens that have just been entered into the database, but also the data of other specimens that were already in the database in order to perform data analysis. In this analysis phase of the data lifecycle, the data can be used to produce reports on the state of the collections for depository institutions' administrative purposes and analysis

[71] Duckworth WD, Genoways HH & Rose CL 1993, Preserving natural science collections: chronicle of our environmental heritage. Mammalogy Papers: University of Nebraska State Museum. pp. 271.
[72] Elkin, L & Norris, CA (eds) 2019, Preventive Conservation: Collection Storage. Society for the Preservation of Natural History Collections, New York.
[73] Huxley, R, Quaisser, C, Butler, CR & Dekker, WRJ 2020, Managing Natural Science Collections: A Guide to Strategy, Planning and Resourcing. Routledge. Oxon, OX.
[74] Bacharach J (Ed) 2005, Museum Handbook Part I: Museum Collections. National Park Service Museum Management Program, Washington, DC, viewed 02 June 2022, <https://www.nps.gov/museum/publications/mhi/mhi.pdf>
[75] Chapman, AD 2005, *Principles of Data Quality*, Global Biodiversity Information Facility, Copenhagen
[76] Hobern D, Baptiste B, Copas K, Guralnick R, Hahn A, van Huis E, Kim E, McGeoch M, Naicker I, Navarro L, Noesgaard D, Price M, Rodrigues A, Schigel D, Sheffield C & Wieczorek J 2019, Connecting data and expertise: a new alliance for biodiversity knowledge. *Biodiversity Data Journal* , vol. 7, e33679.

in scientific research projects as part of the manuscript submitted to scholarly publication. In the preserving stage of the data lifecycle, the data collector, data curator, data user and data custodian shall ensure that the database is securely stored. Good data management should include a strategy for effectively backing up and storing the data.

#### 4.2.6.1 Reporting

While managing biodiversity data according to the lifecycle of the data is an ongoing process, it is important for depository institutions to produce regular reports to document the status of their collections. This is the first step in enabling a paradigm shift from specimen and data ownership to specimen and data stewardship through explicit integration of specimens into existing data management plan guidelines and annual reporting[77]. This regular reporting needs to be made public and include the basic statistics on the collection in terms of new specimen collections added and the frequency of data sharing on the data aggregator platform[78].

There is a query function in the SDBMS to facilitate the documentation process in preparing the reports based on localities, i.e., "Query for information of recorded specimens per localities" and "Checklist report by localities". This function provides templates to retrieve the required data and generate a report summarising the records (collections) by localities. Both the query and report templates in the SDBMS are editable, and the status of the collections can be displayed by grouping the data by taxonomy, location, and year.

#### 4.2.6.2 Analysing

In addition to institutional reporting purposes, researchers who collected the data can use the SDBMS to retrieve the data in the format they need for data analysis in their research. The most common format used in various biodiversity analyses is a matrix of biodiversity community data with samples as rows and species as columns[79] .

Data users and data collectors can use a query template in the SDBMS (Query to extract data) to retrieve data such as "Scientific name" from the Taxonomy Information Table, "Field number" from the Sampling Information Table, "Full Catalogue Information" and "Individual Count" from the Collection

[77] Colella JP, Stephens RB, Campbell ML, Kohli BA, Parsons DJ & Mclean BS 2021, The open-specimen movement. *BioScience*, vol. 71, no. 4, pp. 405-414

[78] Chapman, AD 2005, *Principles of Data Quality*, Global Biodiversity Information Facility, Copenhagen

[79] Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'hara RB, Simpson GL, Solymos P, Stevens MHH & Wagner H 2013, vegan: Community ecology package, Software. <http://CRAN.R-project.org/package=vegan>

Information Table of selected records from SDBMS. The query result can be exported to Microsoft Excel spreadsheet format as a list of records that can be converted into a biodiversity community data matrix using the Pivot Table function in Microsoft Excel – a data summary tool that automatically sorts, counts, and sums up data and displays the summarised data.

Another common request from researchers is to create a species distribution map. Similar to creating the biodiversity community data matrix, researchers can add additional fields from the Sampling Information Table, such as "Decimal Latitude" and "Decimal Longitude", to create a map.

Researchers do not need to keep separate data sheets for different needs in the research process for the information they have collected and curated from the sample collections. As suggested in these Guidelines, they only need to enter the data into the SDBMS once and then retrieve it for analysis. This workflow makes it easier for researchers or data collectors to manage the same dataset in different ways for different purposes and minimises the errors in data consistency (e.g., entering the same information in different ways) that could arise from many different versions of datasets.

### 4.2.7 Disseminating of Biodiversity Data

In the disseminating stage, the data collector, data curator, data user and data custodian can share the standardised data through a data aggregator (e.g., biodiversity information facilities) either through a scholarly publication or a biodiversity data aggregation platform. At this stage, the data is in the sharing and reuse phase of the data lifecycle.

One of the challenges in biodiversity research is that the specimen data is not openly available and awareness of making such data available is very low[80]. While Sections 4.2.3 to 4.2.6 (pages 38 to 46) have described the important actions to produce good quality data in line with biodiversity data standards and documentation, the next step is to adopt best practices for the dissemination of biodiversity data outside the depository institutions, in particular through data aggregator platforms and scholarly publications, to improve the reusability and replicability of biodiversity research.

[80] Mandeville CP, Koch W, Nilsen EB & Finstad AG 2021, Open data practices among users of primary biodiversity data. *BioScience*, vol. 71, no. 11, pp. 1128-1147.

#### 4.2.7.1 Disseminating data via data aggregator platforms

While different depository institutions could disseminate data through institutional reports or online platforms, the interconnectivity of the data depository between institutions can be challenging[81]. A centralised depository that connects all collection data from different institutions is important and data aggregator platforms, commonly referred to as biodiversity information facilities, have an important role to play. Data curator can use a query template in the SDBMS (Query to extract data) and add the required fields to the different tables according to the requirements of the different dataset classes in the selected biodiversity information facilities.

Biodiversity information facilities, such as the Global Biodiversity Information Facility (GBIF)[82] and the Malaysia Biodiversity Information System (MyBIS)[83] should be considered data aggregators, and the quality of the data depends on the depository institutions as they create and manage the data under their care. In addition, biases, errors, and limitations may occur in the preparation, publication and long-term maintenance of the data (for the data aggregator) and in the field sampling, cataloguing, labelling, curating, reporting and disseminating (for the management of the biodiversity data).

Data users should not assume that the quality of the data is guaranteed, as described in the previous chapters, because most of the data in the biodiversity collections is curated. It is impossible for the data aggregator to provide quality assurance so that any user can find, access, and use the data without prior knowledge. A certain level of expertise is required to examine the quality of all data before use. Nevertheless, there are various tools and workflows that user can use to check and improve the quality of the data[84,85,86,87].

[81] Andreone F, Boero F, Bologna MA, Carpaneto GM, Castiglia R, Gippoliti S & Minelli A 2022, Reconnecting research and natural history museums in Italy and the need of a national collection biodepository. *ZooKeys*, no. 1104, pp.55-68.
[82] More information about GBIF can be found on the website, https://www.gbif.org
[83] More information about MyBIS can be found on the website, https://mybis.gov.my/one/
[84] Jin J & Yang J 2020, BDcleaner: a workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. *Global Ecology and Conservation*, vol. 21, no. e00852.
[85] Zizka A, Carvalho FA, Calvente A, Baez-Lizarazo MR, Cabral A, Coelho JFR, Colli-Silva M, Fantinati MR, Fernandes MF, Ferreira-Araújo T, Moreira FGL, Santos NMC, Santos TAB, dos Santos-Costa RC, Serrano FC, Alves da Silva AP, de Souza Soares A, Cavalcante de Souza PG, Tomaz EC, Vale VF, Vieira TL & Antonelli A 2020, No one-size-fits-all solution to clean GBIF. *PeerJ,* vol. 8, e9916
[86] Owens HL, Merow C, Maitner BS, Kass JM, Barve V & Guralnick RP 2021, occCite: Tools for querying and managing large biodiversity occurrence datasets. *Ecography*, vol. 44, no. 8, pp. 1228 – 35.
[87] Ribeiro BR, Velazco SJE, Guidoni-Martins K, Tessarolo G, Jardim L, Bachman SP & Loyola R 2022, *bdc*: A toolkit for standardizing, integrating and cleaning biodiversity data*. Methods in Ecology and Evolution*, vol. 13, no. 7, pp. 1421 – 28.

### 4.2.7.2 Publishing data in scholarly publications

Up to this point, the shared data and published report, as described above, have not gone through the peer review process, which is the accepted standard for scientific publications. This is one way to ensure that the quality of the dataset is of high standards as it would be reviewed and scrutinised by other experts. There are some ways to publish biodiversity data through scholarly publishing[88].

Many scientific journals publish datasets, including prestigious journals, most of which are open access with Creative Common Licence (see https://www.gbif.org/data-papers). Publishing data soon after a study is completed would have a positive impact on the scientific careers of researchers/data users/collectors[89].

In addition to publishing datasets, researchers should also publish the underlying specimen data they have used to answer scientific questions in biodiversity and ecology research. Not all journals require authors to make all data publicly available without restriction at the time of publication. Nevertheless, all researchers are encouraged to make the data available as a necessary step in the research process because if researchers follow the workflow for managing biodiversity data described in this guideline, the process of preparing and publishing datasets of specimen records is not a time-consuming process[90].

Finally, researchers are encouraged to publish biodiversity-related papers in journals where the publisher has advocated and implemented the principles of Open Access and Open Data while providing tools to enhance the dissemination of biodiversity data, e.g., Pensoft Publishers[91] and ARPHA Writing Tool[92,92,93]. In addition, the information published in the traditional way of journals, such as the sample data on the website and in PDF format, has its limitations as this data may not be machine-readable[94]. Using semantic technologies to improve publications is the future of scientific

[88] Penev L, Mietchen D, Chavan V, Hagedorn G, Smith V, Shotton D, Ó Tuama É, Senderov V, Georgiev T, Stoev P, Groom Q, Remsen D & Edmunds S 2017, Strategies and guidelines for scholarly publishing of biodiversity data. Research Ideas and Outcomes 3, no. e12431.
[89] Lortie CJ 2021, The early bird gets the return: The benefits of publishing your data sooner. *Ecology and Evolution*, vol. 11, no. 16, pp. 10736 – 40.
[90] Escribano N, Galicia D & Ariño AH 2018, The tragedy of the biodiversity data commons: a data impediment creeping nigher?. *Database (Oxford)*,vol. 2018, no. bay033
[91] More information about Pensoft can be found on the website, https://pensoft.net/
[92] More information about Arphahub can be found on the website, https://arphahub.com/
[93] Penev L, Georgiev T, Senderov V, Dimitrova M, & Stoev P 2019b, The Pensoft data publishing workflow: The FAIRway from articles to Linked Open Data. Biodiversity Information Science and Standards 3, no. e35902.
[94] Sikes DS, Copas K, Hirsch T, Longino JT & Schigel D 2016, On natural history collections, digitized and not- a response to Ferro and Flick. *ZooKeys*, no. 618, pp. 145 – 58.

publishing[95,96]. However, publishers and journals offering such services may charge article processing fees. Many authors may not receive sufficient financial support from their institutions to do so.

[95] Penev L, Dimitrova M, Senderov V, Zhelezov G, Georgiev T, Stoev P & Simov K 2019a, OpenBiodiv: a knowledge graph for literature-extracted linked open data in biodiversity science. Publications, vol. 7, no. 2.
[96] Bénichou L, Guidoti M, Gérard I, Agosti D, Robillard T & Cianferoni F 2021, European Journal of Taxonomy: A deeper look into a decade of data. *European Journal of Taxonomy*, vol. 782, no. 1, pp.173-196.

# CHAPTER 5

# DIGITISATION EQUIPMENT AND WORKFLOW

## 5.1 DIGITISATION EQUIPMENT AND SPECIFICATION

Appendix 4 provides a list of itemised equipment budgets and anticipated costs for a digitisation project. The examples in this guide can be used as a reference, but as technology improves, equivalent or higher specifications can be achieved. All depository institutions may have one or more of the devices listed in the examples.

### 5.1.1 Digital Single-lens Reflex Camera (DSLR) and Camera Lens

For advanced camera sensors for DSLR, we recommend the Micro Four Thirds system (MFT or M4/3) and complementary metal-oxide-semiconductor (CMOS). The equipment of a DSLR for digitising specimens should have the following minimum specifications:

i. 12-megapixel (MP) camera resolution,
ii. Advanced photo system type-C (APS-C) CMOS Camera Sensor (Figure 5.1)
iii. 32GB Secure Digital (SD) or Compact Flash (CF) card storage capacity
iv. Fit to connect wired or wireless to a larger monitor or computer

For sharp photos of small specimens, a macro lens or a kit lens is required. The lens must have a maximum zoom range of 200mm, a minimum focal length of 30cm and a magnification factor of 1:1. With the normal zoom lens, you can shoot a variety of specimens, from large to macro specimens. The camera and lens should be stored in a dry cupboard or airtight container with dry reagents (calcium chloride) to prevent fungal growth.


**Figure 5.1.** Example of APS-C COMS camera. Left-hand side: Canon 4000D; Right-hand side: Nikon D3100

### 5.1.2 Photo Studio Light Box

Good lighting is essential to produce a high-quality image. For small collections of specimens, the use of a photo studio light box is recommended (Figure 5.2). The size of the photo studio light box depends on the size of the specimens; commercially available light boxes have dimensions ranging from 20cm × 20cm × 20cm to 80cm × 80cm × 80cm. At least 25lm (lumens) of light output is required. Additional light output can be added after calibrating the light intensity. It is recommended to use a photo studio light box that allows for interchangeable backdrops and backgrounds. A natural background colour such as white, grey or black is advisable, depending on the colour of the sample. The use of a professional studio set-up is only recommended for large specimens (larger than 100cm ×100cm ×100cm), e.g., vertebrates (e.g., amphibians, reptiles and birds) and herbarium specimens (Figure 5.3).

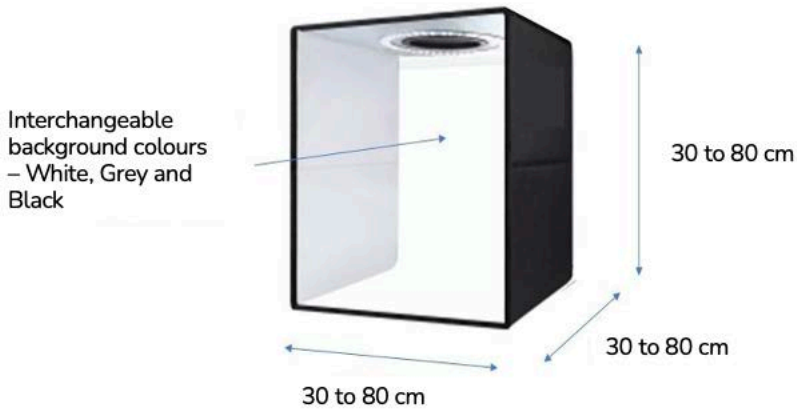



**Figure 5.2.** Example of a photo studio light box ranging from 30cm to 80cm and equipped with a light source, internal reflector, and interchangeable background

**Figure 5.3.** Example of two professional studio set-ups

### 5.1.3 Camera Stand

To maintain the stability and position of the camera, a good camera tripod is necessary for the entire digitisation process. A tripod with a 3-way swivel head or a stable ball head that can hold the camera statically once it is locked in place (Figure 5.4). Be sure to check the load capacity of the camera tripod. For an APS-C CMOS camera with a kit lens, for example, the tripod head should be able to handle loads of at least 2kg.

Depository institutions often use the copy stand (Figure 5.5) to digitise their specimens, as the stand provides an inverted position for the camera. However, the stand is relatively inflexible as it can only hold the camera in one position. If a digitising station needs to be transported from one place to another, a copy stand is also not suitable for mobile use, not to mention the relatively higher cost compared to a tripod.


**Figure 5.4.** Example of a tripod with a 3-way pan head camera stand

**Figure 5.5.** Example of a copy stand

### 5.1.4 Larger Monitor Screen

You will need a computer with processing units on which software or an application can be installed to control the camera remotely. The Canon EOS Utility or Nikon Camera Control Pro 2 are examples of remote camera control software that needs to be installed. The computer must be able to connect directly to the camera via cable (HDMI) or wireless (WIFI). Be sure to check that the power supply to the display is independent of the camera so that you can work effectively. A computer or laptop with a screen size of at least 21 inches is recommended.

### 5.1.5 Storage

It is recommended to use at least a 32GB SD/CF card with Speed Class 10 (write speed of 10MB/s) as camera memory and at least 1TB with internal backup software as external memory. Examples of camera memory cards are SanDisk 32GB SD and Kingston 32GB SD. Examples of external storage

are WD MyPassport (1TB – 4TB), WD My Book (3TB – 10TB), Seagate Backup Plus Portable (2TB – 5TB) and Seagate Backup Plus Hub (4TB).

## 5.2 DIGITISATION WORKFLOW

This Guideline describe three stages of digitisation of dry specimens. The first stage is pre-production, which includes planning and the preparation checklist; the second stage is production, which consists of the details of data/image capture and processing; and the third stage is post-production, which includes tools and guidelines for transcribing the physical data, storage and backup, and database creation.

### 5.2.1 Digitisation pre-production stage

To make the digitisation process more effective, the depository institution needs to make a plan for the specimens it wants to digitise. For example, for holotypes, rare specimens or according to the priority of the specimens. The overall process for the pre-production phase is shown in Figure 5.6.

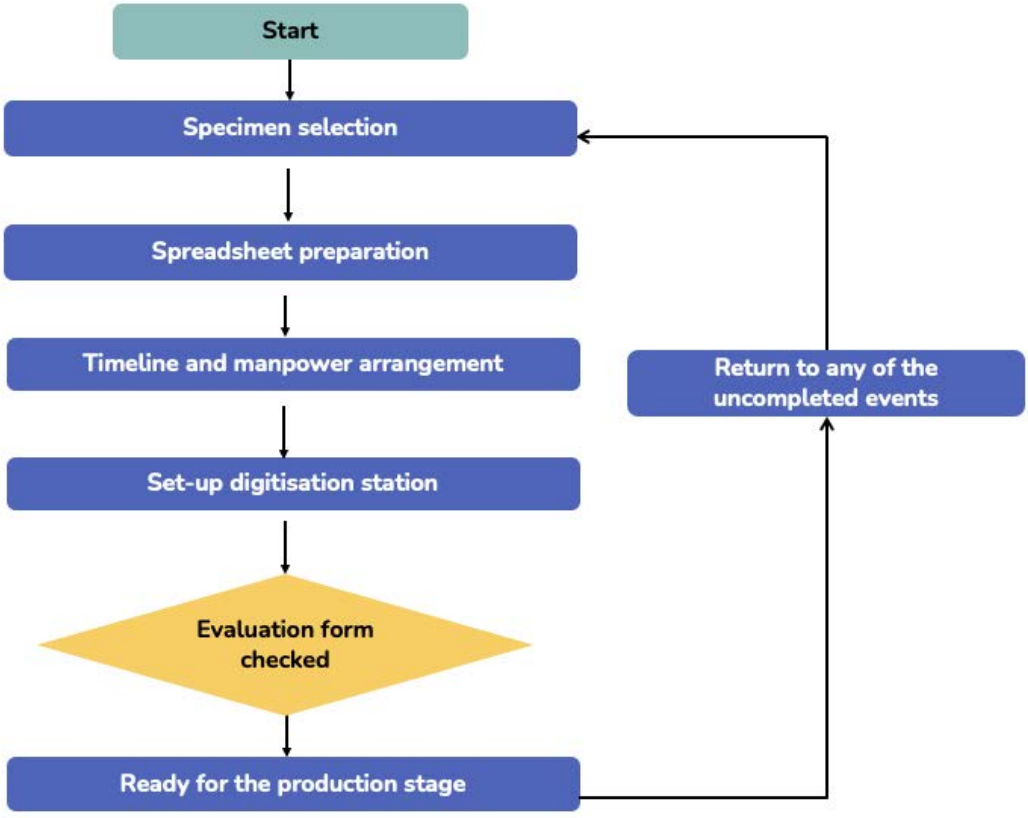



**Figure 5.4.** Workflow for pre-production stage

There are four processes to be applied in the pre-production phase:

A. Selection of specimens to be digitised: This process begins with the identification of groups of specimens for which a plan must be made with a timetable and manpower for digitisation. Next, the cabinets and trays containing the species to be digitised are identified. The data curator will need to prioritise digitisation (i.e., identify whether some images will be on display, on loan, etc.)
B. Spreadsheet preparation: Key in the information for the image by using the datasheet in Appendix 5. The explanation and requirement for each table field can be found in Appendix 1E and Appendix 2E.
C. Timing and staffing: For the digitisation process to be completed within a specific timeframe, the depository institution will need to prepare a timeline for the expected completion of each group of specimens. For example, the depository institution could get an overview of a staff member's workload and assign them the digitisation task based on duty or committee or time shift mechanisms.
D. Digitising station setup: This section illustrates the layout and position of the camera for imaging samples of different sizes (Figures 5.7, 5.8 and 5.9). It is recommended that an evaluation be made at the end of this phase (Appendix 6).

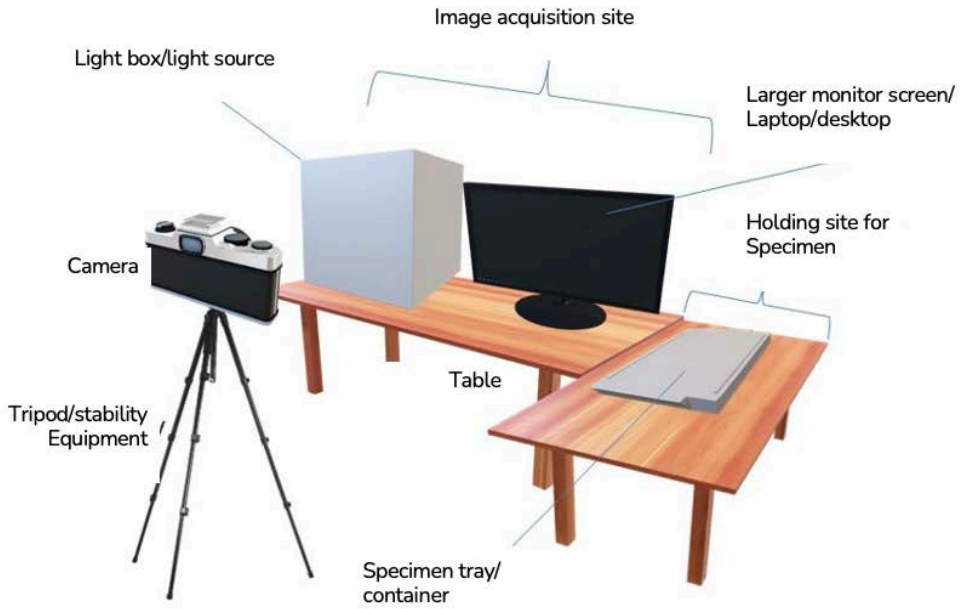



**Figure 5.7.** General set-up for a digitisation station

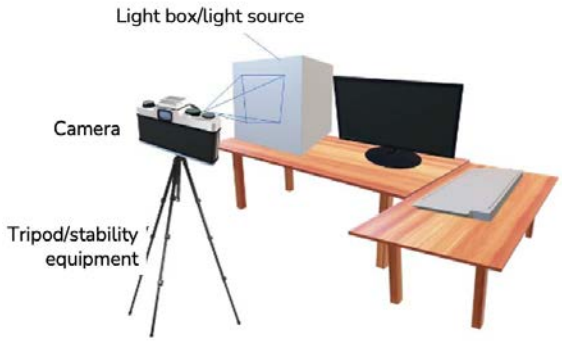



**Figure 5.8.** Left: Camera position for lateral/anterior axis/direction of the specimen

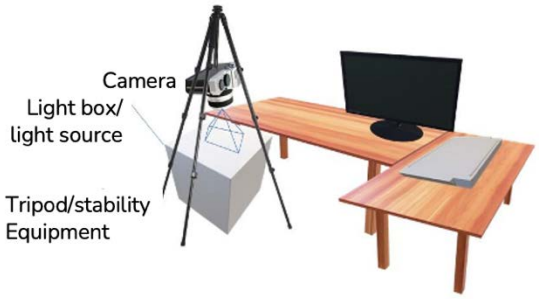



**Figure 5.9.** Right: Camera position for the dorsal-ventral direction of the specimen

### 5.2.2 Digitisation Production Stage

Figure 5.10 shows the general workflow for the production phase, focusing on the details of sample placement, technical camera setup, light or colour calibration and an image checklist before the post-production phase.

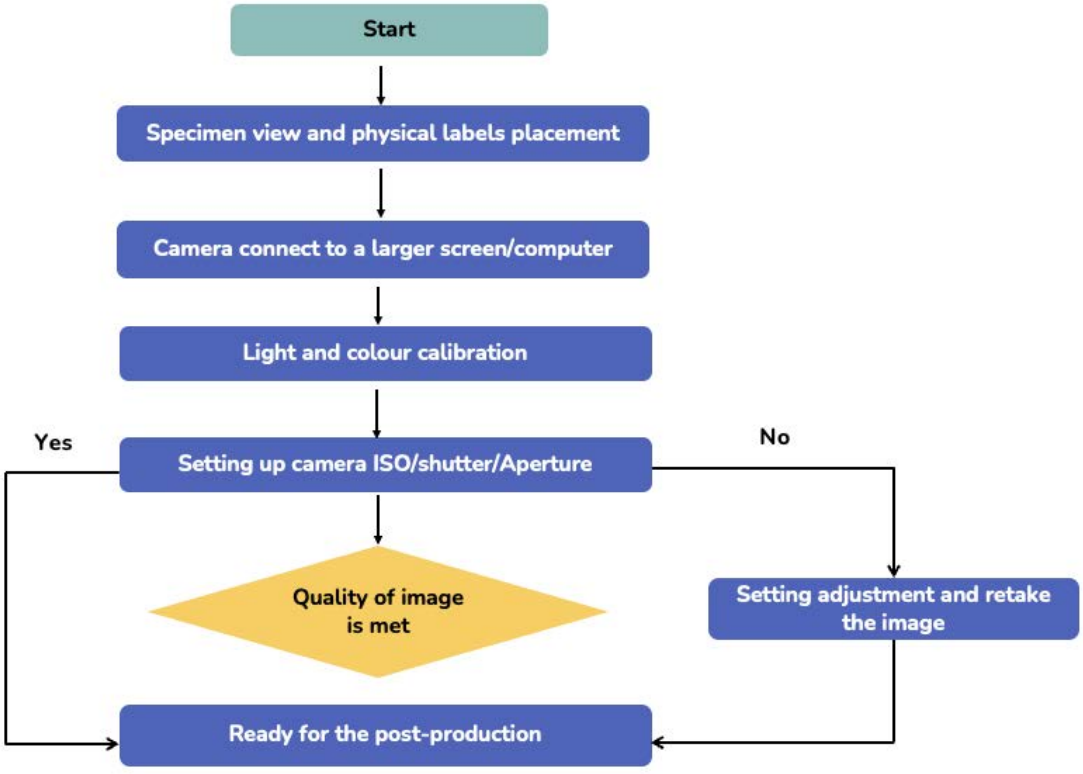



**Figure 5.10.** Workflow for the production stage

A. **Specimen view and physical label placement:** Once you have selected a specimen to be digitised, it is important to check the condition and quality of the specimen, which in most

cases must be free of pests. If the specimen is infested with fungi or covered in dusk, the specimen must go through a cleaning process. If the specimen needs to be taken to the digitising station, you can transport it with a polystyrene board, taking extra care not to damage the specimen. To view the specimen (Figure 5.11), it is advisable to consult the expert (e.g., entomologist, herpetologist, ornithologist, botanist, etc.) to look at the key morphology. In many cases, more than one view is required to document the specimen ID. Figure 5.12 shows an example of the placement of labels and scale bars on the right side and a scale bar on the top.

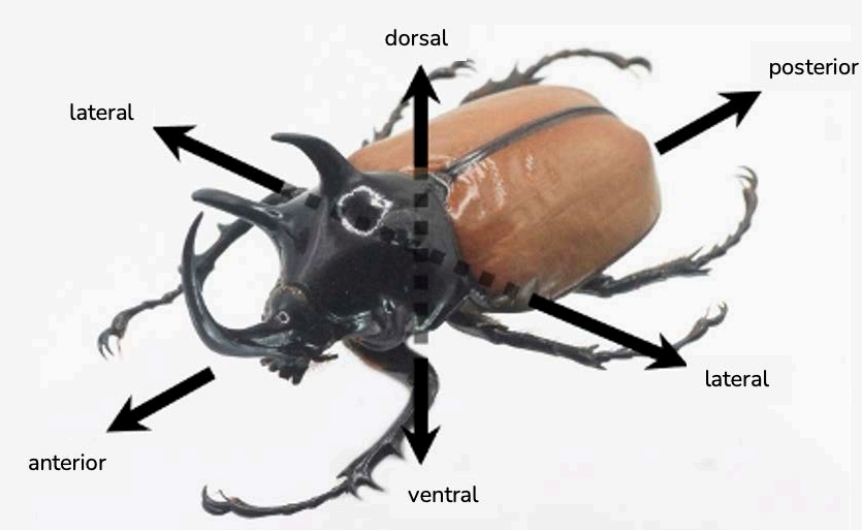



**Figure 5.11.** Specimen view, D- dorsal; V-Ventral; L-Lateral; A-Anterior; P-Posterior


**Figure 5.12.** Physical labels and scale bar placement/position

B. **Processing the images:** Step 1. Light and colour calibration – Calibrate the camera light through the lens (TTL) by switching the camera to Manual mode. Turn on the Digitising Station light source, focus the sample and adjust ISO /shutter/aperture (see step 4 for requirements). Observe that the exposure of the image (Figure 5.13) is 0 to +1. For colour calibration, use a colour chart (Figure 5.14) and set the white balance "Kelvin" to 4000K to 5000K and calibrate with the colour chart. Figures 5.13 and 5.14 show the light intensity metre on the camera screen and the colour of the calibrator, respectively.


**Figure 5.13.** Exposure of image to be 0 to +1


**Figure 5.14**. Colour meter for colour calibration

C. **Processing the images:** Step 2. Camera mode – Use either autofocus with full manual setting or aperture priority. Please note that the minimum focal length must be set manually when using macro mode.

**D. Image editing:** Step 3. ISO /shutter/aperture adjustment – The exposure triangle (Figure 5.15) is a basic concept for calibrating the light intensity or exposure of an image. If you follow the baseline of the setting listed in Figure 5.15, the result of the light intensity metre (Figure 5.13) should be between 0 and +1, and the setting is suitable for image capture.

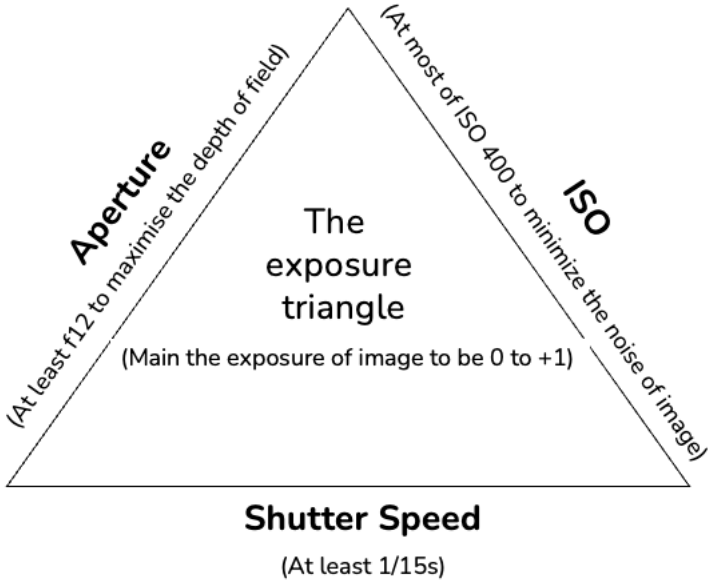



**Figure 5.15.** The exposure of the image triangle

**E. Imaging processing:** Step 4. Remote control of the camera – In this Guideline, it is strongly recommended that you use a camera remote control or camera remote control software or a self-timer (Figure 5.16, 5.17, 5.18) so that you can trigger your camera without touching the shutter button. The reason for using the remote control is two main points: Firstly, it prevents camera shake when the shutter button is pressed, which in turn results in much sharper photos, and secondly, it allows the user to operate the shutter button from a distance, which can be very convenient. It is recommended that an image quality assessment be made at the end of this phase (Appendix 7).

**Figure 5.16.** Remote control

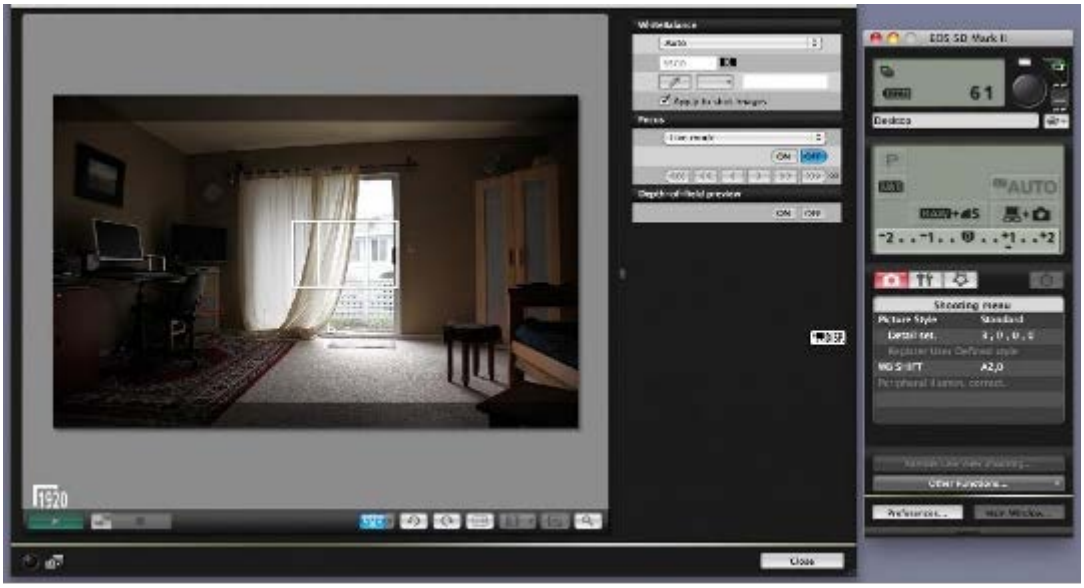


**Figure 5.17.** Remote camera control software/application

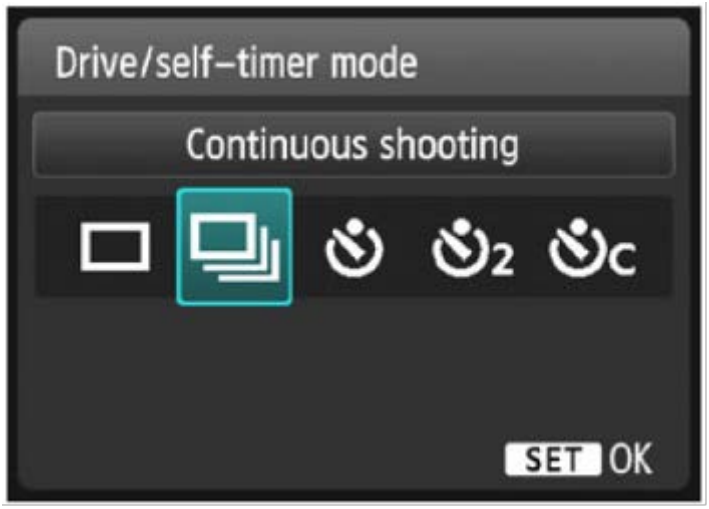



**Figure 5.18.** Timer setting

### 5.2.3 Digitisation Post-production Stage

A. **File transfer and storage:** Save the file as JPEG and name the file after the unique identifier (UID)/pattern identifier (ID) (in accordance with the step of data management/registration) or any sequence generated by the software/camera. For storage, prepare at least 1TB of hard disc space with backup software. Once the image files have been named/annotated and transferred to the hard drive, staff should duplicate/replicate the data to the institution's storage system (e.g., Google, Azure, AWS, Oracle, etc.) and an additional storage system (additional hard drive or cloud storage).

B. **Document the metadata of the image:** Document the metadata of the image according to the sample table (Appendix 5).

C. **Naming the file:** Naming the file can be done using the built-in tool of the camera remote control application and optical character recognition (OCR). The built-in tool allows the user to customise the prefix of the file by referring to the function of the folder. OCR is a type of text recognition on images. In general, it can extract and digitise the text on the image. The application mainly recommends automating the process of entering the information from the physical labels into the file name or annotation of the file. Naming files with OCR methods can be done with Google, Tesseract, and Google Colab.

# CHAPTER 6

# APPENDICES

## APPENDIX 1. METADATA QUICK REFERENCE GUIDE[97,98]

A. Collection Information Dataset

| **associatedSequences** | |
|---|---|
| Label | Associated Sequences |
| Definition | A list (concatenated and separated) of identifiers (publication, global unique identifier, URI) of genetic sequence information associated with the Occurrence. |
| Notes | |
| Example | http://www.ncbi.nlm.nih.gov/nuccore/U34853.1, http://www.ncbi.nlm.nih.gov/nuccore/GU328060 \| http://www.ncbi.nlm.nih.gov/nuccore/AF326093 |

[97] Darwin Core Standard (DwC) is a biodiversity metadata standard adopted by several main biodiversity databases such as Global Biod iversity Information Facility (GBIF), The Atlas of Living Australia (ALA), Ocean Biogeographic Information System (OBIS), FishNet2, Vernet, Encyclopedia of Life (EOL)

[98] Information on Darwin Core Terms is taken from the Darwin Core list, version 15 July 2021, published at https://dwc.tdwg.org/list/. Accessed 18 July 2022.

B. Personnel Profile Information Dataset

| firstname | lastname | nameinitial | affiliation | otherpersonalinfo |
|---|---|---|---|---|

| **firstname** | |
|---|---|
| Label | First Name |
| Definition | The first name of the individual involved in any stages of the primary biodiversity data lifecycle. |
| Notes | |
| Example | |

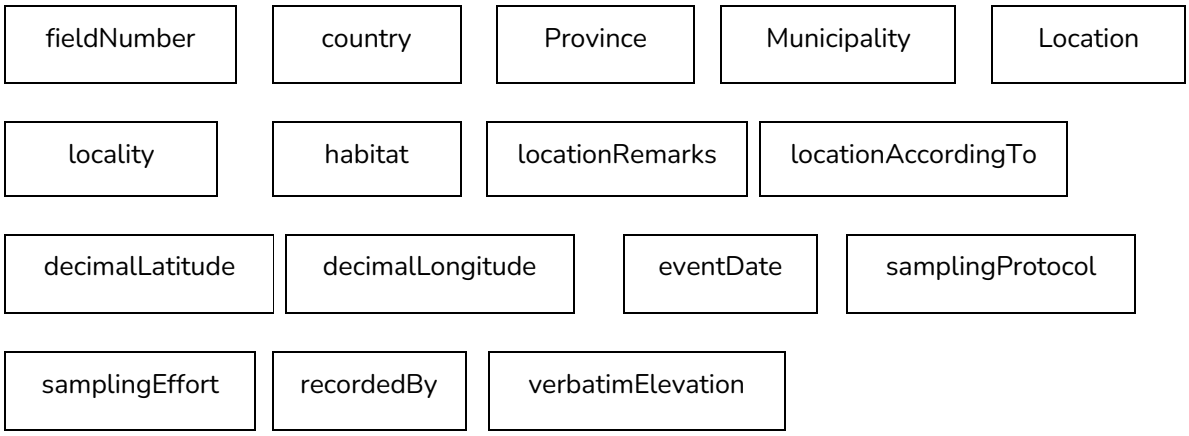
| **lastname** | |
|---|---|
| Label | Last Name |
| Definition | The last name of the individual involved in any stages of the primary biodiversity data lifecycle. |
| Notes | |
| Example | |

| **nameinitial** | |
|---|---|
| Label | Name Initial |
| Definition | The name initial (e.g. Liew, T.S.) of the individual involved in any stages of the primary biodiversity data lifecycle. |
| Notes | |
| Example | |

| **affiliation** | |
|---|---|
| Label | Affiliation |
| Definition | The affiliation of the individual involved in any stages of the primary biodiversity data lifecycle. |
| Notes | |
| Example | |

| **otherpersonalinfo** | |
|---|---|
| Label | Other Personal Information |
| Definition | The other information of the individual involved in any stages of the primary biodiversity data lifecycle. |
| Notes | |
| Example | |

C. Sampling Information Dataset




| **fieldNumber** | |
|---|---|
| Label | Field Number |
| Definition | An identifier given to the event in the field. Often serves as a link between field notes and the Event. |
| Notes | |
| Example | RV Sol 87-03-08 |

| **country** | |
|---|---|
| Label | Country |
| Definition | The name of the country or major administrative unit where the Location occurs. |
| Notes | The recommended best practice is to use a controlled vocabulary such as the Getty Thesaurus of Geographic Names. Recommended best practice is to leave this field blank if the Location spans multiple entities at this administrative level or if the Location might be in one or another of multiple possible entities at this level. Multiplicity and uncertainty of the geographic entity can be captured either in the term higherGeography or in the term locality, or both. |
| Example | Denmark, Colombia, España |

| **Province** | |
|---|---|
| Label | State Province |
| Definition | The name of the next smaller administrative region than country (state, province, canton, department, region, etc.) where the Location occurs. |

| | |
|---|---|
| Notes | The recommended best practice is to use a controlled vocabulary such as the Getty Thesaurus of Geographic Names. |
| Example | Montana, Minas Gerais, Córdoba |

| **Municipality** | |
|---|---|
| Label | Municipality |
| Definition | The full, unabbreviated name of the next smaller administrative region than county (city, municipality, etc.) where the Location occurs. Do not use this term for a nearby named place that does not contain the actual location. |
| Notes | The recommended best practice is to use a controlled vocabulary such as the Getty Thesaurus of Geographic Names. |
| Example | Holzminden, Araçatuba, Ga-Segonyana |

| **Location** | |
|---|---|
| Label | Location |
| Definition | A spatial region or named place. |
| Notes | |
| Example | The municipality of San Carlos de Bariloche, Río Negro, Argentina. The place defined by a georeference. |

| **locality** | |
|---|---|
| Label | Locality |
| Definition | The specific description of the place. |
| Notes | Less specific geographic information can be provided in other geographic terms (higherGeography, continent, country, stateProvince, county, municipality, waterBody, island, islandGroup). This term may contain information modified from the original to correct perceived errors or standardise the description. |
| Example | Bariloche, 25 km NNE via Ruta Nacional 40 (=Ruta 237), Queets Rainforest, Olympic National Park |

| **habitat** | |
|---|---|
| Label | Habitat |
| Definition | A category or description of the habitat in which the Event occurred. |
| Notes | |
| Example | oak savanna, pre-cordilleran steppe |

| **locationRemarks** | |
|---|---|
| Label | Location Remarks |
| Definition | Comments or notes about the Location. |
| Notes | |
| Example | underwater since 2005 |

| **locationAccordingTo** | |
|---|---|
| Label | Location According To |
| Definition | Information about the source of this Location information. Could be a publication (gazetteer), institution, or team of individuals. |
| Notes | |
| Example | Getty Thesaurus of Geographic Names, GADM |

| **decimalLatitude** | |
|---|---|
| Label | Decimal Latitude |
| Definition | The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic centre of a Location. Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive. |
| Notes | |
| Example | -41.0983423 |

| **decimalLongitude** | |
|---|---|
| Label | Decimal Longitude |
| Definition | The geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic centre of a Location. Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive. |
| Notes | |
| Example | -121.1761111 |

| **eventDate** | |
|---|---|
| Label | Event Date |
| Definition | The date-time or interval during which an Event occurred. For occurrences, this is the date-time when the event was recorded. Not suitable for a time in a geological context. |
| Notes | The recommended best practice is to use a date that conforms to ISO 8601-1:2019. |
| Example | 1963-03-08T14:07-0600 (8 Mar 1963 at 2:07pm in the time zone six hours earlier than UTC). 2009-02-20T08:40Z (20 February 2009 8:40am UTC). 2018-08-29T15:19 (3:19pm local time on 29 August 2018). 1809-02-12 (some time during 12 February 1809). 1906-06 (some time in June 1906). 1971 (some time in the year 1971). |

| **samplingProtocol** | |
|---|---|
| Label | Sampling Protocol |
| Definition | The names of, references to, or descriptions of the methods or protocols used during an Event. |
| Notes | The recommended best practice is to describe an Event with no more than one sampling protocol. In the case of a summary Event with multiple protocols, in which a specific protocol cannot be attributed to specific Occurrences, the recommended best practice is to separate the values in a list with space-vertical bar-space ( \| ). |
| Example | UV light trap, mist net, bottom trawl, ad hoc observation \| point count, Penguins from space: faecal stains reveal the location of emperor penguin colonies, https://doi.org/10.1111/j.1466-8238.2009.00467.x, Takats et al. 2001. Guidelines for Nocturnal Owl Monitoring in North America. Beaverhill Bird Observatory and Bird Studies Canada, Edmonton, Alberta. 32 pp., http://www.bsc-eoc.org/download/Owl.pdf |

| **samplingEffort** | |
|---|---|
| Label | Sampling Effort |
| Definition | The amount of effort expended during an Event. |
| Notes | |
| Example | 40 trap-nights, 10 observer-hours, 10 km by foot, 30 km by car |

| **recordedBy** | |
|---|---|
| Label | Recorded By |
| Definition | A list (concatenated and separated) of names of people, groups, or organisations responsible for recording the original Occurrence. The primary collector or observer, especially one who applies a personal identifier (recordNumber), should be listed first. Look up from Personal Information Table. |
| Notes | The recommended best practice is to separate the values in a list with space-vertical bar-space ( \| ). |
| Example | José E. Crespo. Oliver P. Pearson \| Anita K. Pearson (where the value in recordNumber OPP 7101 corresponds to the collector number for the specimen in the field catalog of Oliver P. Pearson). |

| **verbatimElevation** | |
|---|---|
| Label | Verbatim Elevation |
| Definition | The original description of the elevation (altitude, usually above sea level) of the Location. |
| Notes | |
| Example | 100-200 m |

D. Taxonomic Information Dataset

| | | | | |
|---|---|---|---|---|
| kingdom | phylum | class | order | family |
| subfamily | genus | subgenus | specificEpithet | infraspecificEpithet |
| scientificNameAuthorship | scientificName | taxonomicStatus | vernacularName | |
| parentNameUsageID | taxonRemarks | acceptedNameUsageID | Collection Room[99]* | |
| Collection Cabinet* | Collection Drawer* | | | |

| **kingdom** | |
|---|---|
| Label | Kingdom |
| Definition | The full scientific name of the kingdom in which the taxon is classified. |
| Notes | |
| Example | Animalia, Archaea, Bacteria, Chromista, Fungi, Plantae, Protozoa, Viruses |

| **phylum** | |
|---|---|
| Label | Phylum |
| Definition | The full scientific name of the phylum or division in which the taxon is classified. |
| Notes | |
| Example | Chordata (phylum). Bryophyta (division). |

| **class** | |
|---|---|
| Label | Class |
| Definition | The full scientific name of the class in which the taxon is classified. |
| Notes | |
| Example | Mammalia, Hepaticopsida |

[99]* Fields included in the SDBMS that are not listed in the DarwinCore (DwC) terms

| **order** | |
|---|---|
| Label | Order |
| Definition | The full scientific name of the order in which the taxon is classified. |
| Notes | |
| Example | Carnivora, Monocleales |

| **family** | |
|---|---|
| Label | Family |
| Definition | The full scientific name of the family in which the taxon is classified. |
| Notes | |
| Example | Felidae, Monocleaceae |

| **subfamily** | |
|---|---|
| Label | Subfamily |
| Definition | The full scientific name of the subfamily in which the taxon is classified. |
| Notes | |
| Example | Periptyctinae, Orchidoideae, Sphindociinae |

| **genus** | |
|---|---|
| Label | Genus |
| Definition | The full scientific name of the genus in which the taxon is classified. |
| Notes | |
| Example | Puma, Monoclea |

| **subgenus** | |
|---|---|
| Label | Subgenus |
| Definition | The full scientific name of the subgenus in which the taxon is classified. Values should include the genus to avoid homonym confusion. |
| Notes | |
| Example | Strobus, Amerigo, Pilosella |

| **specificEpithet** | |
|---|---|
| Label | Specific Epithet |
| Definition | The name of the first or species epithet of the scientificName. |
| Notes | |
| Example | concolor, gottschei |

| **infraspecificEpithet** | |
|---|---|
| Label | Infraspecific Epithet |
| Definition | The name of the lowest or terminal infraspecific epithet of the scientificName, excluding any rank designation. |
| Notes | In botany, where there can be more than one infraspecific rank, name strings may be provided in literature and in identifications that have more than two epithets. Only the last of these epithets is the infraspecificEpithet, and only |

| | |
|---|---|
| | the first and the last epithets belong to the scientificName. For example: the infraspecificEpithet in the string "Indigofera charlieriana subsp. sessilis var. scaberrima" is scaberrima and the scientificName is Indigophera charlieriana var. scaberrima. |
| Example | concolor (for scientificName "Puma concolor concolor"), oxyadenia (for scientificName "Quercus agrifolia var. oxyadenia"), laxa (for scientificName "Cheilanthes hirta f. laxa"), scaberrima (for scientificName "Indigofera charlieriana var. scaberrima"). |

| **scientificNameAuthorship** | |
|---|---|
| Label | Scientific Name Authorship |
| Definition | The authorship information for the scientificName formatted according to the conventions of the applicable nomenclaturalCode. |
| Notes | |
| Example | (Torr.) J.T. Howell, (Martinovský) Tzvelev, (Györfi, 1952) |

| **scientificName** | |
|---|---|
| Label | Scientific Name |
| Definition | The full scientific name, with authorship and date information if known. When forming part of an Identification, this should be the name in the lowest taxonomic rank that can be determined. This term should not contain identification qualifications, which should instead be supplied in the IdentificationQualifier term. |
| Notes | This term should not contain identification qualifications, which should instead be supplied in the IdentificationQualifier term. When applied to an Organism or Occurrence, this term should be used to represent the scientific name that was applied to the associated Organism in accordance with the taxon to which it was or is currently identified. |
| Example | Coleoptera (order). Vespertilionidae (family). Manis (genus). Ctenomys sociabilis (genus + specificEpithet). Ambystoma tigrinum diaboli (genus + specificEpithet + infraspecificEpithet). Roptrocerus typographi (Györfi, 1952) (genus + specificEpithet + scientificNameAuthorship), Quercus agrifolia var. oxyadenia (Torr.) J.T. Howell |

| **taxonomicStatus** | |
|---|---|
| Label | Taxonomic Status |
| Definition | The status of the use of the scientificName as a label for a taxon. Requires taxonomic opinion to define the scope of a taxon. Rules of priority then are |

| | |
|---|---|
| | used to define the taxonomic status of the nomenclature contained in that scope, combined with experts' opinions. It must be linked to a specific taxonomic reference that defines the concept. |
| Notes | The recommended best practice is to use a controlled vocabulary. |
| Example | invalid, misapplied, homotypic synonym, accepted |

| **vernacularName** | |
|---|---|
| Label | Vernacular Name |
| Definition | A common or vernacular name. |
| Notes | |
| Example | Andean Condor, Condor Andino, American Eagle, Gänsegeier |

| **parentNameUsageID** | |
|---|---|
| Label | Parent Name Usage ID |
| Definition | An identifier for the name usage (documented meaning of the name according to a source) of the direct, most proximate higher-rank parent taxon (in a classification) of the most specific element of the scientificName. |
| Notes | This term should be used for accepted names to refer to the taxonID of a Taxon record that represents the next higher taxon rank in the same taxonomic classification. For Darwin Core Archives, the related record should be present locally in the same archive. |
| Example | tsn:41074 (ITIS), urn:lsid:ipni.org:names:30001404-2 (IPNI), 2704173 (GBIF), 6T8N (COL) |

| **taxonRemarks** | |
|---|---|
| Label | Taxon Remarks |
| Definition | Comments or notes about the taxon or name. |
| Notes | |
| Example | this name is a misspelling in common use |

| **acceptedNameUsageID** | |
|---|---|
| Label | Accepted Name Usage ID |

| | |
|---|---|
| Definition | An identifier for the name usage (documented meaning of the name according to a source) of the currently valid (zoological) or accepted (botanical) taxon. |
| Notes | This term should be used for synonyms or misapplied names to refer to the taxonID of a Taxon record that represents the accepted (botanical) or valid (zoological) name. For Darwin Core Archives, the related record should be present locally in the same archive. |
| Example | tsn:41107 (ITIS), urn:lsid:ipni.org:names:320035-2 (IPNI), 2704179 (GBIF), 6W3C4 (COL) |

| **Collection Room[100]*** | |
|---|---|
| Label | Collection Room |
| Definition | The designated room in which the specimens of the taxa were kept. |
| Notes | |
| Example | |

| **Collection Cabinet*** | |
|---|---|
| Label | Collection Cabinet |
| Definition | The cabinet/compartment in the designated room where the specimens of the taxa were kept. |
| Notes | |
| Example | |

| **Collection Drawer*** | |
|---|---|
| Label | Collection Drawer |
| Definition | The drawer of the cabinet/compartment in the designated room where the specimens of the taxa were kept. |
| Notes | |
| Example | |

[100]* Fields included in the SDBMS that are not listed in the DarwinCore (DwC) terms

E. Image Data[101]*

| | | | | |
|---|---|---|---|---|
| Image File | View Metadata | Caption | Measurement | Measurement Type |
| Sample ID/UID | License Holder | License | License Contact | Photographer |

| **Image File** | |
|---|---|
| Label | Image File |
| Definition | Complete name, including extension, and identical to the image file. |
| Example | |

| **View Metadata** | |
|---|---|
| Label | View Metadata |
| Definition | Standardised term to group images depicting a specific set of features of the organisms |
| Example | Dorsal; Ventral; Lateral; Anterior; Posterior |

| **Caption** | |
|---|---|
| Label | Caption |
| Definition | Free text description of the subject with a max of 200 characters. |
| Notes | Short descriptions are recommended |
| Example | Part of the organism photographed, life stage, sex, etc. |

| **Measurement** | |
|---|---|
| Label | Measurement |
| Definition | Any single relevant measurement that was taken in metric units. |
| Example | |

| **Measurement Type** |
|---|

[101]* Fields included in the SDBMS that are not listed in the DarwinCore (DwC) terms

| | |
|---|---|
| Label | Measurement Type |
| Definition | Item or feature that was measured. |
| Example | Wingspan, Body length, Head width |

| **Sample ID/UID** | |
|---|---|
| Label | Sample ID/UID |
| Definition | Sample ID for record, must match Sample ID/UID from data management worksheet |
| Example | UMS_ITBC_ENT0078 |

| **License Holder** | |
|---|---|
| Label | License Holder |
| Definition | The primary individual holder of the license. This is less critical when using Creative Commons licenses. |
| Example | ITBC Universiti Malaysia Sabah |

| **License** | |
|---|---|
| Label | License |
| Definition | License for the use of the image, short forms are accepted |
| Example | Copyright (c)<br>No Rights Reserved (nrr)<br>CreativeCommons – Attribution (by)<br>CreativeCommons – Attribution Share-Alike (by-sa)<br>CreativeCommons – Attribution No Derivatives (by-nd)<br>CreativeCommons – Attribution Non-Commercial (by-nc)<br>CreativeCommons – Attribution Non-Commercial Share-Alike (by-nc-sa)<br>CreativeCommons – Attribution Non-Commercial No Derivatives (by-nc-nd) |

| **License Contact** | |
|---|---|
| Label | License Contact |
| Definition | Contact information for the license holder. Can be an email address, mailing address, phone number, or all of the above. |
| Example | xxx@ums.edu.my/+6088190222 |

| **Photographer** | |
|---|---|
| Label | Photographer |
| Definition | The individual or team responsible for photographing and editing the media prior to submission. |
| Example | En Azrie Allia |

## APPENDIX 2. METADATA CATEGORY REQUIREMENT

A. Collection Information

| TERM | CATEGORY REQUIREMENT OF EXISTING SPECIMEN | CATEGORY REQUIREMENT OF NEWLY COLLECTED SPECIMEN |
|---|---|---|
| Institution Code | Required | Required |
| Collection Code | Required | Required |
| Catalogue Number | Required | Required |
| Full catalogue ilformation | Required | Required |
| Other Catalogue Numbers | Required if available | Required if available |
| Type Status | Required when available | Required when available |
| Catalogued By | Strongly recommended | Strongly recommended |
| Material Citation | Strongly recommended | Strongly recommended |
| Taxonomy Information ID | Required | Required |
| Identified By | Strongly recommended | Strongly recommended |
| Identification References | Strongly recommended | Strongly recommended |
| Identification Remarks | Strongly recommended | Strongly recommended |
| Sampling Information ID | Required | Required |
| Individual Count | Required | Required |
| Basis of Record | Strongly recommended | Strongly recommended |
| Preparations | Required | Required |
| Material Sample | Required | Required |
| Prepared By | Strongly recommended | Strongly recommended |
| Disposition | Strongly recommended | Strongly recommended |
| Sex | Recommended when available | Recommended when available |
| Life Stage | Recommended when available | Recommended when available |
| Measurement or Fact | Recommended when available | Recommended when available |
| Measurement Remarks | Recommended when available | Recommended when available |
| Information Withheld | Recommended when available | Recommended when available |
| Associated Media | Strongly recommended | Strongly recommended |
| Associated Sequences | Strongly recommended | Strongly recommended |

B. Personnel Profile Information

| TERM | CATEGORY REQUIREMENT OF EXISTING SPECIMEN* | CATEGORY REQUIREMENT OF NEW SPECIMEN |
|---|---|---|
| First Name | Required when available | Required |
| Last Name | Required when available | Required |
| Name Initial | Required when available | Required |
| Affiliation | Required when available | Required |
| Other Personal Information | Strongly recommended | Strongly recommended |

C. Sampling Information

| TERM | CATEGORY REQUIREMENT OF EXISTING SPECIMEN* | CATEGORY REQUIREMENT OF NEW SPECIMEN |
|---|---|---|
| Field Number | Required when available | Required |
| Country | Required when available | Required |
| State Province | Required when available | Required |
| Municipality | Required when available | Required |
| Location | Required when available | Required |
| Locality | Required when available | Required |
| Habitat | Required when available | Required when available |
| Location Remarks | Required when available | Required |
| Location According To | Strongly recommended | Strongly recommended |
| Decimal Latitude | Required when available | Required |
| Decimal Longitude | Required when available | Required |
| Event Date | Required when available | Required |
| Sampling Protocol | Required when available | Required |
| Sampling Effort | Required when available | Required |
| Verbatim Elevation | Required when available | Required |
| Recorded By | Required when available | Required |

D. Taxonomy Information

| TERM | CATEGORY REQUIREMENT OF EXISTING SPECIMEN* | CATEGORY REQUIREMENT OF NEW SPECIMEN |
|---|---|---|
| Kingdom | Required | Required |
| Phylum | Required | Required |
| Class | Required | Required |
| Order | Required when available | Required when available |
| Family | Required when available | Required when available |
| Subfamily | Required when available | Required when available |
| Genus | Required when available | Required when available |
| Subgenus | Required when available | Required when available |
| Specific Epithet | Required when available | Required when available |
| Infraspecific Epithet | Required when available | Required when available |
| Scientific Name Authorship | Required when available | Required when available |
| Scientific Name | Required when available | Required when available |
| Taxonomic Status | Strongly recommended | Strongly recommended |
| Vernacular Name | Recommended when available | Recommended when available |
| Parent Name Usage ID | Recommended when available | Recommended when available |
| Taxon Remarks | Strongly recommended | Strongly recommended |
| Accepted Name Usage ID | Recommended when available | Recommended when available |
| Collection Room | Required | Required |
| Collection Cabinet | Required | Required |
| Collection Drawer | Required | Required |

E. Image Data Information

| TERM | CATEGORY REQUIREMENT OF EXISTING SPECIMEN* | CATEGORY REQUIREMENT OF NEW SPECIMEN |
|---|---|---|
| Image File | Required | Required |
| View Metadata | Required | Required |
| Caption | Strongly recommended | Strongly recommended |
| Measurement | Strongly recommended | Strongly recommended |
| Measurement Type | Strongly recommended | Strongly recommended |
| Sample ID/UID | Required | Required |
| License Holder | Required when available | Required when available |
| License | Required | Required |
| License Contact | Required when available | Required when available |
| Photographer | Required | Required |

## APPENDIX 3. DATA QUALITY ASSESSMENT TEMPLATE

A. SAMPLING INFORMATION (EXISTING SPECIMEN*)

Mark ( / ); Yes= 1 /No= 0. *Count as 1 mark only for any "**Yes**" ticked in that category.

| TERM | COMPLETE | | MARK |
|---|---|---|---|
| | **Yes** | **No** | |
| **Location Data** | | | |
| Field Number | *( / ) | *( ) | 1 |
| Country | *( / ) | *( ) | |
| State Province | *( / ) | *( ) | |
| Municipality | *( ) | *( / ) | |
| Location | *( / ) | *( ) | |
| Locality | *( ) | *( / ) | |
| Habitat | *( / ) | *( ) | |
| Location Remarks | *( ) | *( / ) | |
| Recorded By | *( / ) | *( ) | |
| Location According To | *( ) | *( / ) | |
| **Elevation Data** | | | |
| Verbatim Elevation | *( ) | *( / ) | 0 |
| **GPS Data** | | | |
| Decimal Latitude | *( / ) | *( ) | 1 |
| Decimal Longitude | *( / ) | *( ) | |
| **Temporal Data** | | | |
| Event Date | *( / ) | *( ) | 1 |
| **Sampling Protocol Data** | | | |
| Sampling Protocol | *( ) | *( / ) | 0 |
| Sampling Effort | *( ) | *( / ) | |
| **Total Marks** | | | 3 |

B. SAMPLING INFORMATION (NEWLY COLLECTED SPECIMEN)

Mark ( / ); Yes= 1 /No= 0. *Count as 1 mark only for any "**Yes**" ticked in that term.

| TERM | COMPLETE | | MARK |
|---|---|---|---|
| | **Yes** | **No** | |
| **Location Data** | | | |
| Field Number | ( / ) | ( ) | 1 |
| Country | ( / ) | ( ) | 1 |
| State Province | ( / ) | ( ) | 1 |
| Municipality | ( ) | ( / ) | 0 |
| Location | ( / ) | ( ) | 1 |
| Locality | ( ) | ( / ) | 0 |
| Location Remarks | ( / ) | ( ) | 1 |
| Recorded By | ( / ) | ( ) | 1 |
| Habitat | *( / ) | *( ) | 1 |
| Location According To | *( ) | *( / ) | |
| **Elevation Data** | | | |
| Verbatim Elevation | ( ) | ( / ) | 0 |
| **GPS Data** | | | |
| Decimal Latitude | ( ) | ( / ) | 0 |
| Decimal Longitude | ( / ) | ( ) | 1 |
| **Temporal Data** | | | |
| Event Date | ( / ) | ( ) | 1 |
| **Sampling Protocol Data** | | | |
| Sampling Protocol | ( / ) | ( ) | 1 |
| Sampling Effort | ( ) | ( / ) | 0 |
| **Total Marks** | | | 10 |

## APPENDIX 4. BUDGET FOR THE DIGITISATION EQUIPMENT

*****The types of equipment, level of sophistication, and price listed are reviewed by 24 June 2022.

A. Digitisation Equipment Range below RM5,000

Example (the product could be any brand/model that has similar specs)

- Canon EOS 4000D with 18mm-55mm kit lens - RM2000
- Tamron SP AF 60mm f/2 1:1 Macro Lens for Canon - RM1600
- Portable Photo Lighting Studio 40cm - RM120
- Portable Photo Lighting Studio 80cm - RM400
  (80CM 60CM 40CM 4 x LED EXTRA BRIGHTNESS LIGHTING CONTROL 3 Light Modes Portable Camera Photo Studio Photography)
- USB 2.0/3.0 Hub - RM50
  (Micro USB 2.0/3.0 hub 4/7 port high-speed USB hub with on/off switch for laptops)
- Tripod - RM700
  (Benro TAD18AIB1 Series 1 Adventure Aluminium Tripod with B1 Ball Head)
- Dry cabinet - RM90
  (Samurai Dry Box F380 Grey with Free Blue Silica Gel Bottle (500g)

B. Digitisation Equipment Range below RM10,000

Example (the product could be any brand/model that has similar specs)

- Canon EOS 90D (EF-S18-55mm f/3.5-5.6 IS STM) - RM6000
- Tamron SP AF 60mm f/2 1:1 Macro Lens for Canon - RM1600
- Portable Photo Lighting Studio 40cm - RM120
- Portable Photo Lighting Studio 80cm - RM400
  (80CM 60CM 40CM 4 x LED EXTRA BRIGHTNESS LIGHTING CONTROL 3 Light Modes Portable Camera Photo Studio Photography)
- USB 2.0/3.0 Hub - RM50
  (Micro USB 2.0/3.0 hub 4/7 port high-speed USB hub with on/off switch for laptops)
- Tripod - RM1700
  (Manfrotto MT055XPRO3 MK055XPRO3 Aluminium 3-Section MVH500AH / MHXPRO-3W / Tripod only (Original Manfrotto Malaysia)
- Dry cabinet - RM250
  (AIPO AS-25 Dry Cabinet Dry Box AS25 (25L))

C. Digitisation Equipment Range below RM20,000

Example (the product could be any brand/model that has similar specs)

- Canon EOS 6D Mark II + 24mm-105mm f4L lens - RM11,000
- Canon RF 85mm f/2 Macro IS STM Lens (MSIA) - RM2700
- Tamron SP AF 60mm f/2 1:1 Macro Lens for Canon - RM1600
- Portable Photo Lighting Studio 40cm - RM120
- Portable Photo Lighting Studio 80cm - RM400

  (80CM 60CM 40CM 4 x LED EXTRA BRIGHTNESS LIGHTING CONTROL 3 Light Modes Portable Camera Photo Studio Photography)
- USB 2.0/3.0 Hub - RM50

  (Micro USB 2.0/3.0 hub 4/7 port high-speed USB hub with on/off switch for laptops)
- Tripod - RM1700

  (Manfrotto MT055XPRO3 MK055XPRO3 Aluminium 3-Section MVH500AH / MHXPRO-3W / Tripod only (Original Manfrotto Malaysia)
- Dry cabinet - RM1100

  (AIPO AP-102EX AP102EX Dry Cabinet Dry Box (102L) AP102)

## APPENDIX 5. EXAMPLE OF IMAGE DATA SPREADSHEET

| **Image File*** | **View meta data*** | **Caption** | **Measure ment** | **Measure ment type** | **Sample ID*** | **License Holder** | **License*** | **License Contact** | **Photo grapher*** |
|---|---|---|---|---|---|---|---|---|---|
| IMG0 002 | Dorsal | Adult stage | | | UMS21 04 | | | | Azrie |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |

*Required at the point of data registry

## APPENDIX 6. PRE-PRODUCTION EVALUATION FORM TEMPLATE

| No. | Events | Yes | No |
|---|---|---|---|
| 1. | The specimen that plans to be digitised are ready at the holding site | / | |
| 2. | The specimen is free from dusk, parasites, or unwanted objects on its surface | / | |
| 3. | The spreadsheet is ready for databasing | / | |
| 4. | The timeline and workforce are planned and arranged for digitisation | / | |
| 5. | The digitisation station is set up | / | |

## APPENDIX 7. IMAGE QUALITY ASSESSMENT

| | **Criteria** | **Yes** | **No** |
|---|---|---|---|
| Composition | Does the placement of specimens and labels follow the guidelines? | | |
| Brightness | Is the image's brightness satisfying? | | |
| Colour | Does the image's colour look similar to the real specimen? | | |
| Sharpness | Is the image's sharpness, especially the key morphology, sharp enough to identify? | | |

## Acknowledgements

We would like to thank the funders of this project, ISC ROAP and ASM for their continuous support in implementing this project, project team members who have been instrumental in supporting and steering the project direction, and subject matter experts who have diligently worked and contributed their insights, time, experience and expertise to complete this important task. Credit is also due to the Institute for Tropical Biology and Conservation (ITBC) Universiti Malaysia Sabah (UMS) for providing us inputs on the Biodiversity Data Management workflow, Forest Research Institute Malaysia (FRIM) for providing us the venue and materials for video filming, the biodiversity communities for their valuable feedback to our findings, and Rimba Ilmu Universiti Malaya as the co-host for Capacity Building Activities for FAIR Biodiversity Data Stewardship.


ACADEMY OF SCIENCES MALAYSIA
www.akademisains.gov.my