# Self-assessment guidelines for biodiversity data holding institutions

## Introduction

This self-assessment tool enables data-holding institutions to carry out a high-level review of their readiness and progress in delivering digital information on the materials they hold. It is organized around a simple five-component framework for planning and implementing a sustainable content mobilization strategy. The tool is intended to assist collection managers to understand the steps involved in sharing data on their collections and to help them identify priority areas for further attention and monitor long-term progress in these tasks.

At the same time, GBIF will use the framework presented here to plan and organise training and support materials relevant to each component. The goal is to ensure that biological collections of all kinds are well-positioned to become part of a global web of interconnected information, supporting 21st century taxonomy and sustainable use of biodiversity. Even simple steps, such as sharing basic metadata about a collection, can contribute to these goals.

This self-assessment model aims to establish the status and capacity of an institution in five key areas:

1. **Strategy**: Planning how to execute the data mobilization. Key considerations include the purpose for mobilization; prioritization of what materials to digitize; labour/expertise; equipment; data standards; choice of database system; and intellectual property rights and licensing issues.

2. **Digitization**: Conversion and capture of data associated with physical specimens and artefacts into electronic formats and databases. This process may involve imaging.

3. **Publishing**: Making digital data (including associated images and multimedia) publicly available online, whether through dedicated institutional and community websites or major Internet repositories.

4. **Curation and Maintenance**: Ongoing performance of tasks including cleaning, correcting and updating data, in part based on feedback.

5. **Preservation and Archiving**: Long-term conservation of digital content and databases to ensure its integrity and persistence.

These areas are quite clearly interlinked, but data holders will not always address them in a strictly chronological sequence. In some cases, an institution will be well advanced in one of the ‘later’ areas while still needing to progress in one or more of the earlier ones. Progress in one area may also reinforce work in other areas.

## Self-assessment implementation

This self-assessment is intended for data-holding institutions wishing to develop their capacity to mobilize, manage and disseminate their data. An assessment should evaluate the status in the five core capacity areas, using a three-point scale for each measured aspect. The overall scores are to help gauge the institutional capacity at the time of assessment and at the same time give suggestions on what can be done to improve. The notes section is for writing various notes, including, comments, links, contact information, references, reminders, etc., that can help the assessor to record details about a section and make needed follow ups. Writing notes can also help an institution to understand the context under which an assessment was done especially if subsequent assessments are not carried out by the same institutional representative.

Production of this document is funded by the European Union

## Explanatory notes

**Biodiversity data-holding institutions** are entities or organizations that house or are in charge of biodiversity data in its various forms, including digitized and undigitized data. Examples include natural history museums and herbaria, culture collections, botanic gardens, arboreta, zoos, citizen science organizations, and libraries, among others.

**Biodiversity data mobilization** is the process of capturing data and information about living organisms, notably, species, or other biodiversity artefacts, in digital formats and publishing the resulting data in ways that make them available for discovery, use and reuse. Most of the data held by the world's biodiversity data-holding institutions remains digitally inaccessible. This information is suboptimal for use and, in some cases, cannot be used at all. Data-holding institutions with unmobilized data have difficulty fully accounting for the contents of their holdings.

**Digitization** involves the conversion and capture of data associated with physical specimens and artefacts (including paper records) into electronic formats and databases. This usually involves label data transcription and imaging of objects or data that is in analog format such as text from paper formats. Digitization provide multiple benefits, including digital preservation, improved collections management and dissemination through data publishing.

**Digital preservation** involves the tools and processes required to ensure the long-term persistence of digitized collections to ensure their integrity and persistence, including the creation of a back-up and/or a high-quality digital replica of the original artefact. Given that preservation is one of the core mandates of data-holding institutions, digital preservation is an important consideration in planning digitization projects. A project that involves digital preservation as one of its goals will be more demanding because the technical specifications for the outputs are much higher. The file formats generated for digital preservation are commonly large archival formats not directly suitable for web publishing that must be converted into other formats to share via the Internet.

**Database management system** is a software tool for storing and managing biodiversity data and information.


Figure 1: Self-assessment model for data holding institutions

GBIF Global Biodiversity Information Facility

Date

# BIODIVERSITY DATA HOLDER SELF-ASSESSMENT QUESTIONNAIRE

## Institution / Collection name:

## Section 1: Strategy

| | No (0) | In progress/ incomplete (1) | Yes (2) |
|---|---|---|---|
| 1. Does the collection have a defined (documented) mission which determines its role and purpose? | ☐ | ☐ | ☐ |
| 2. Has the purpose of the mobilization activity been defined? | ☐ | ☐ | ☐ |
| 3. Have the scope and extent of digitization efforts—their completeness, data elements, digital preservation, etc.—been defined? | ☐ | ☐ | ☐ |
| 4. Have curators identified or assessed any necessary or useful pre-digitization activities—sorting, selection of samples, taxonomic review, cross-checking against field notebooks, etc.? | ☐ | ☐ | ☐ |
| 5. Does the institution have a defined data policy that covers topics like intellectual property rights, custodial responsibilities, access, licensing, liability and privacy, sensitivity? | ☐ | ☐ | ☐ |
| 6. Has a database management system/software been selected? | ☐ | ☐ | ☐ |

**Section 1: Overall score**

Notes on strategy

# Section 2: Digitization

| | No (0) | In progress/ incomplete (1) | Yes (2) |
|---|---|---|---|
| 1. Has a suitable digitization workspace been identified or set up? | ☐ | ☐ | ☐ |
| 2. Has pre-digitization curation been carried out (sorting, selecting what to digitize, updating taxonomy and labels, specimen “health”/condition, assigning unique identifiers such as barcodes and general tracking)? | ☐ | ☐ | ☐ |
| 3. Have digitization processes and technologies (text capture, imaging, multimedia, other) been defined? | ☐ | ☐ | ☐ |
| 4. Have relevant data elements been mapped against important international data standards? | ☐ | ☐ | ☐ |
| 5. Are workflows, including the handling of physical materials, well-defined? | ☐ | ☐ | ☐ |
| 6. Are staff adequately and appropriately trained and equipped? | ☐ | ☐ | ☐ |
| 7. Have quality control processes and standards—including taxonomy, georeferencing, typographical errors, field mismatch, technical and metadata specifications—been clearly specified? | ☐ | ☐ | ☐ |

**Section 2: Overall score**

Notes on digitization

# Section 3: Publishing

| | No (0) | In progress/ incomplete (1) | Yes (2) |
|---|---|---|---|
| 1. Have the needs of expected end users and web publishing been fully considered in selecting appropriate data formats? | ☐ | ☐ | ☐ |
| 2. Are the selected licences machine-readable and supported by GBIF? | ☐ | ☐ | ☐ |
| 3. Has a (web) data publishing tool been selected and set up? | ☐ | ☐ | ☐ |
| 4. Does the institution have access to stable data-hosting facilities? | ☐ | ☐ | ☐ |

**Section 3: Overall score**

Notes on publishing

# Section 4: Curation and maintenance

| | No (0) | In progress/ incomplete (1) | Yes (2) |
|---|---|---|---|
| 1. Have processes been defined to accommodate updates from quality control and new information (e.g., taxonomic updates)? | ☐ | ☐ | ☐ |
| 2. Does the institution have resources and processes to handle corrections and suggestions offered by web users? | ☐ | ☐ | ☐ |
| 3.Are adequate processes in place to keep digitized data up to date with changes in curatorial information about the original materials, and vice versa? | ☐ | ◉ | ☐ |

**Section 4: Overall score**

Score 1-3: Initiate what is not started and consider actions to accelerate work in progress

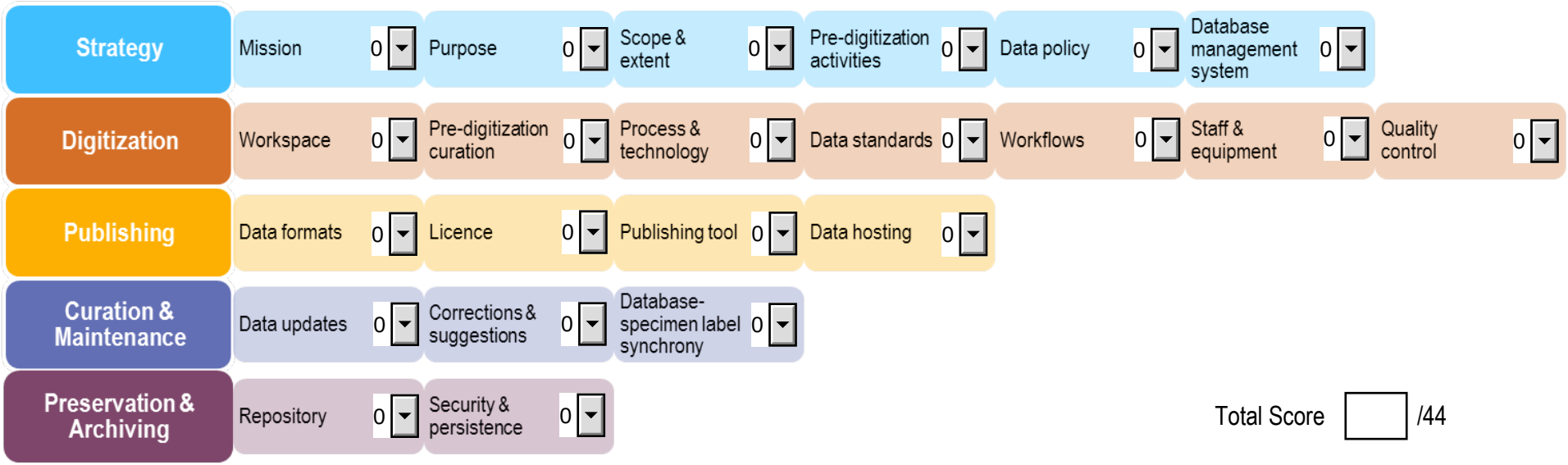
Notes on curation and maintenance

# Section 5: Preservation and archiving

| | No (0) | In progress/ incomplete (1) | Yes (2) |
|---|---|---|---|
| 1. Does the institution have access to suitable long-term data archival repositories? | ☐ | ☐ | ☐ |
| 2. Does the solution adequately ensure the security of data and safeguard against obsolescence of data formats and applications? | ☐ | ☐ | ☐ |

**Section 4: Overall score**

Additional notes on preservation and archiving


Strategy
Mission 0
Purpose 0
Scope & extent 0
Pre-digitization activities 0
Data policy 0
Database management system 0
Digitization
Workspace 0
Pre-digitization curation 0
Process & technology 0
Data standards 0
Workflows 0
Staff & equipment 0
Quality control 0
Publishing
Data formats 0
Licence 0
Publishing tool 0
Data hosting 0
Curation & Maintenance
Data updates 0
Corrections & suggestions 0
Database-specimen label synchrony 0
Preservation & Archiving
Repository 0
Security & persistence 0
Total Score
/44

## Selected support materials

**Initiating a Collection Digitisation Project** (https://www.gbif.org/document/80574/initiating-a-collection-digitisation-project)
This guide describes how to plan and execute digitization work. While focused on natural history collections, it can be applied to other situations.

**Consortium of Northeastern Herbaria** (http://neherbaria.org/digit_resource)
An annotated list of resources (with links) relevant to the digitization of biological collections. This is targeted to those initiating a biological collection digitization effort. The resources are grouped into overviews, databasing (text digitisation), georeferencing, imaging, mobilisation, standards, and tools.

**iDigBio Digitization Resources** (https://www.idigbio.org/wiki/index.php/Digitization_Resources)
This web page provides resources and information for the series of digitisation training workshops conducted by iDigBio as well as a plethora of digitization information and resources. Included is a growing list of links to documents, websites, videos, presentations, and other important information related to biological collection digitization. iDigBio also frequently hosts free online webinars open to all.

**The Atlas of Living Australia (Atlas) Digitisation Guidance (**http://www.ala.org.au/about-the-atlas/digitisation-guidance/)
Guidance on different aspects of digitisation developed by the Atlas partners. The guidance materials address the following themes: What is digitisation?, Imaging and Using volunteers for digitisation. The web page provides links to detailed documents about the three themes. The guidance materials can be reused without modifications or adapted for other non Atlas projects.

**Biodiversity Data Standards** (http://www.tdwg.org/)
Basic recommendations and documented agreements about representations, formats, and definition for primary biodiversity data. Biodiversity Information Standards, the body responsible for the standards, is commonly known by its original acronym, TDWG..

**BioSharing.org** (https://biosharing.org)
A curated, searchable portal of interrelated data standards, databases, and policies across the life, environmental and biomedical sciences.

**Towards demand driven publishing: Approaches to the prioritisation of digitization of natural history collections data** (https://journals.ku.edu/index.php/jbi/article/view/3990)
This article reviews useful high level information (metadata) about natural history collections that can be digitally captured and help to expose them even when they are not yet fully digitized at the specimen/object level. It also argues for a user demand driven approach to digitization and provides possible metrics that can be used in prioritizing collections for digitisation when there are inadequate resources to digitise the the whole collection.

**Data publishing guidance from GBIF and benefits** (https://www.gbif.org/publishing-data)
GBIF.org provides guidelines and processes on publishing data through the GBIF network, including the benefits of data publishing, publisher endorsement by nodes, and links to additional resources. For incentives and benefits of publishing open-access biodiversity data, please check the following links: https://www.gbif.org/article/11g6LFdAOylaUOM0EuCqso/quick-guide-to-publishing-data-through-gbiforg#incentives; http://bioscience.oxfordjournals.org/content/59/5/418.full

**Creative Commons licensing** (https://www.gbif.org/terms)
This resource describes the process, results and rationale for GBIF's adoption of one of three machine-readable licenses for all species occurrence datasets published through GBIF.org, namely, CC0, CC-BY and CC-BY-NC