August 2011 IAIA

Special Publication Series No. 7

Publishing EIA-Related Primary Biodiversity Data: GBIF-IAIA Best Practice Guide

International Best Practice Principles

> "PUBLISHING" BIODIVERSITY DATA MAY BE DEFINED AS MAKING BIODIVERSITY DATASETS PUBLICLY ACCESSIBLE IN A STANDARDISED FORMAT, VIA AN ONLINE ACCESS POINT (TYPICALLY A WEB ADDRESS, OR URL). THIS ACCESS POINT IS RECORDED IN A REGISTRY MANAGED BY THE GLOBAL BIODIVERSITY INFORMATION FACILITY (GBIF). PUBLISHED DATASETS CAN ALSO BE DISCOVERED AND ACCESSED VIA THE GBIF DATA PORTAL (HTTP://DATA.GBIF.ORG).

Publishing EIA-Related Primary Biodiversity Data

PURPOSE

The overall purpose of this guide is to enable EIA practitioners, consultants and other interested and affected parties to discover, capture, manage and publish to common standards, the primary biodiversity data generated during environmental impact assessment processes.

It represents a summarized version of a more comprehensive guide (ISBN: 87-92929-35) that can be accessed at http://links.gbif.org/eia_biodiversity_data_publishing_guide_en_v1

ABOUT GBIF

The Global Biodiversity Information Facility (GBIF) was established by countries as a global mega-science initiative to address one of the great challenges of the 21st century – harnessing knowledge of the Earth's biological diversity. GBIF's mission is to be the foremost global resource for biodiversity information, and to engender smart solutions for environmental and human well-being. To achieve this mission, GBIF encourages a wide variety of biodiversity data holders, generators and users across the globe to discover and publish (make discoverable) data through the GBIF network. For more information, visit http://www.gbif.org

HOW TO CITE THIS PUBLICATION

Cadman, M., Chavan, V., King, N., Willoughby, S., Rajvanshi, A., Mathur, V., Roberts, R., and Hirsch, T. (2011). Publishing EIA-Related Primary Biodiversity Data: GBIF-IAIA Best Practice Guide. IAIA Special Publication Series No. 7. August 2011. Pp. 6. Accessible at http://www.iaia.org/publications/

COLLABORATORS

This best practice guide is an outcome of the GBIF led collaboration of the following institutions:

Global Biodiversity Information Facility
South African National Biodiversity Institute
Wildlife Institute of India
International Association for Impact Assessment

INTERNATIONAL ASSOCIATION for IMPACT ASSESSMENT

- *Headquarters*
  1330 23rd Street South, Suite C
  Fargo, ND 58103-3705 USA
  Phone + 1.701.297.7908
  Fax + 1.701.297.7917
  info@iaia.org
  www.iaia.org

Introduction

The issue

Primary biodiversity data is defined as "digital text or multimedia data records detailing facts about the occurrence of an organism." Knowledge about the identity and occurrence of organisms forms the backbone of our understanding of the biological world, and is essential for monitoring the state of natural ecosystems, for developing sound environmental management policies, and making ecologically sustainable development decisions. ***Environmental Impact Assessment*** (EIA) provides opportunities for integrating biodiversity values with development, but, for a variety of reasons, biodiversity has not always been given specific or appropriate consideration in EIAs (Rajvanshi *et al*., 2007).

Ideally, ***biodiversity-inclusive*** EIA, which is promoted by the Convention on Biological Diversity, should: (a) use biodiversity information to determine the biological or ecological sensitivity of a site, and (b) generate new biodiversity records about the site. To make meaningful assessments, EIA practitioners need access to verifiable biodiversity data that are in a readily usable form and that can be accessed using standardized protocols. To date, however, there has been no easy-to-use mechanism for discovering and accessing digital biodiversity data for use in EIA, or for publishing the biodiversity data that EIA generates (King, *et al*., in prep.).

This means that EIA-related biodiversity data is, generally, unavailable for use in subsequent EIAs, or for informing research programmes, environmental planning and decision-making. This compromises the quality of the EIA, reduces the transparency of the EIA process, and, ultimately, the confidence that can be placed in decisions based on the EIA.

The solution

Through the ***Global Biodiversity Information Facility*** (GBIF), digital biodiversity data are being made freely and openly available via the Internet for scientists, researchers, authorities and the general public. GBIF provides a suite of standards and data publishing tools that can be employed to discover and publish primary biodiversity data. This best practice guide describes the tools, standards and infrastructure that are available to EIA practitioners, and explains when and how they should be used. It represents a summarized version of a more comprehensive guide (ISBN: 87-92929-35) that can be accessed at http://links.gbif.org/eia_biodiversity_data_publishing_guide_en_v1. Sources of additional assistance are also provided.

Principles and concepts underpinning data publishing

Types of biodiversity data

There are several different categories of biodiversity data, or levels at which data can be gathered and used, and it is important to distinguish between these, and to use terms about data precisely in order to avoid any confusion.

The first distinction to be drawn is between ***primary biodiversity data*** (species occurrence data), ***taxonomic data*** (information about the identity of organisms, species checklists), and ***synthesized or interpretive (secondary) data*** (a wide range of ecological information about the site and the organisms found there). Although much of the information presented in EIAs tends to be interpretive or synthesized data, this is based on large volumes of primary biodiversity data.

From a data publication perspective, GBIF makes the distinction between several terms relating to biodiversity data, including: ***data resources*** or datasets, ***data elements***, ***data values*** and ***metadata***. These terms are described in Table 1, as well as in related GBIF publications (GBIF, 2011a).

Primary biodiversity data, taxonomic data and metadata are each supported by a different data publishing option within the GBIF network.

Table 1: Data terminology

What it is called	What it is	Example
Metadata	Information about the dataset	Who collected the data, when it was collected
Dataset or data resource	A collection of data records	List of species recorded at a site
Data elements	Categories of information comprising each data record	Scientific name, latitude, longitude
Data values	These are "***the data***"—content of each data element comprising each record of occurrence	A data value for the element "Scientific name" could be *Acacia karoo*

Metadata, which are the descriptive information that accompanies a dataset, are required for all datasets published through the GBIF network (GBIF 2011b). The metadata provide the data user with a means of verifying the authenticity of the dataset, its appropriateness for the desired usage and a measure of the confidence with which it can be used.

Guiding principles of best practice

Publishing biodiversity data through the GBIF network calls for adherence to six basic principles (adapted from Chapman, 2005): accuracy, precision, fitness-for-use, effectiveness, efficiency and transparency.

Accuracy: refers to how correct the data are. For example, is the organism correctly identified or is the correct locality supplied? If the data are correct, then they are accurate.

Precision or resolution: refers to the exactness or level of detail of the data. In the case of occurrence data, if only the broad area of occurrence is given, the precision of the data is low. If exact geographic coordinates are supplied, then the precision of the data is high.

Quality, or "fitness for use": in the context of this guide, data are described as "fit for use" or "potential use" (Chapman, 2005), if they are suitable for the intended use in EIA and subsequent decision-making about development. GBIF strives to publish only high quality data that are maximally fit-for-use. Data of low accuracy and low precision are poor quality data that will, generally, not be fit-for-use. High quality data are both accurate and precise, as well as being comprehensive, complete, up to date, easy to access and interpret and consistent with other sources.

Effectiveness: this is the likelihood that the data, or a method, might have of achieving the intended outcomes.

Efficiency: relates to the ratio of output to input.

Transparency: relates to how complete, accurate and precise the information is that describes the dataset (i.e., the metadata). Transparency enhances accessibility and also the fitness-for-use of the data.

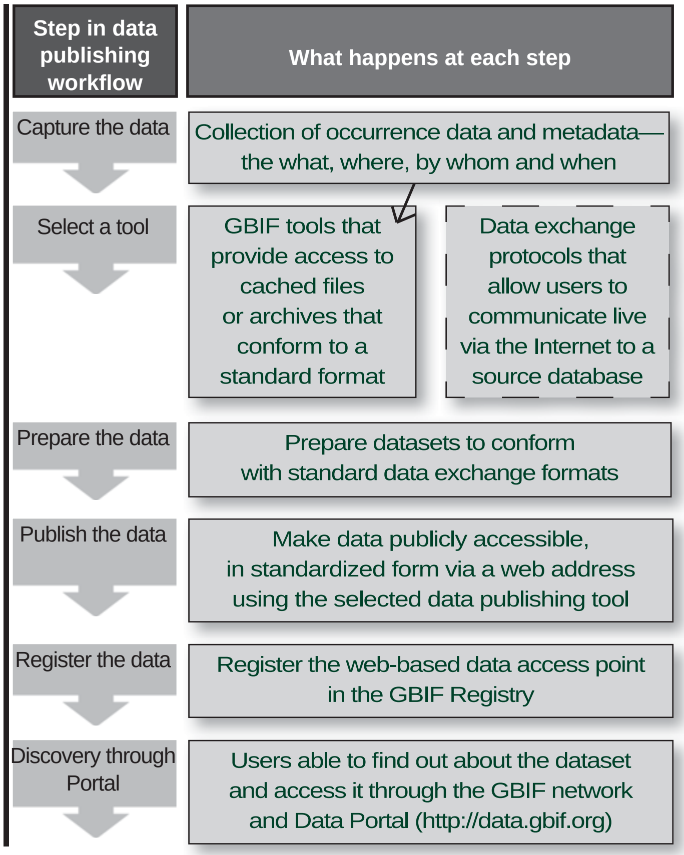
Each of these principles can be applied to the primary biodiversity data themselves, and to the tools, protocols and practices that are employed at each step of the data publishing workflow.

Operating principles: *Steps in the publishing process*

GBIF provides a means of sharing biodiversity data, through the process known as "**publishing**," that makes it universally accessible over the Internet, using simple tools and following standard procedures and protocols. Data publishing through the GBIF network follows a series of clear steps, shown in Figure 1. Each of these steps is described in more detail in the subsequent sections of this document.

This guide will help environmental assessment practitioners, consultants and other interested and affected parties to choose the most suitable option or tool for publishing the primary biodiversity data they have gathered, as an integral part of the EIA process.

Figure 1: The data publishing workflow


Step in data publishing workflow
What happens at each step
Capture the data
Collection of occurrence data and metadata—the what, where, by whom and when
Select a tool
GBIF tools that provide access to cached files or archives that conform to a standard format
Data exchange protocols that allow users to communicate live via the Internet to a source database
Prepare the data
Prepare datasets to conform with standard data exchange formats
Publish the data
Make data publicly accessible, in standardized form via a web address using the selected data publishing tool
Register the data
Register the web-based data access point in the GBIF Registry
Discovery through Portal
Users able to find out about the dataset and access it through the GBIF network and Data Portal (http://data.gbif.org)

Step 1: Capturing the data

GBIF provides a set of pre-configured Excel spreadsheets that serve as templates for capturing occurrence data (primary biodiversity data), metadata and simple species checklists. These spreadsheets are simple tools that provide a common format and standard for collecting data, using consistent terminology. This makes it easier for data to be exchanged between users, compared between sites, and integrated into national and global biodiversity databases.

Use of the GBIF Excel data capture templates (a) makes it easier for EIA practitioners to collect and manage primary biodiversity data; (b) improves the consistency and utility of data collection, and (c) ensures that the data are collected in a form that is suitable for publishing using GBIF infrastructure.

There are three GBIF spreadsheet templates available (for occurrence data, taxonomic data and metadata); those that will be of greatest use to EIA practitioners are the metadata and occurrence spreadsheets. These spreadsheets, which can be downloaded from the GBIF web site (http://tools.gbif.org/spreadsheet-processor/), are easy to use and include online help, which is accessed by hovering the cursor over spreadsheet cells with red upper-right corners. Each spreadsheet includes a large number of possible data fields (or data elements), into which data (or data values) can be captured. These data fields are described using a standardised set of terms referred to as the ***Darwin Core***. Although it is recommended that as many fields as possible are used in order to maximise the quality of the data, there is a minimum set of six compulsory fields that must be filled in.

A number of GBIF User Guides (GBIF 2011b, GBIF 2011c) provide step-by-step assistance for use of the Excel spreadsheets.

Steps 2 – 4: Selecting a tool to prepare data for publishing

GBIF provides a rich array of support and tools for customizing data formats and for publishing primary biodiversity data in compliance with global standards.

To be published via the GBIF network, primary biodiversity datasets must first be converted to a standardised format, known as a ***Darwin Core Archive file*** (DwC-A). EIA data publishers do not have to generate Darwin Core Archive files themselves, unless they choose to do so.

GBIF tools that are currently available for transforming the data into a Darwin Core Archive are:

- The GBIF Spreadsheet Processor
- The GBIF Integrated Publishing Toolkit (GBIF IPT)
- The Darwin Core Archive Assistant (DwCA-Assistant).

For EIA practitioners, the simplest, quickest and most effective route would be to use the GBIF Spreadsheet Processor. This is also the only tool that can be used if the data are not already digitised.

Using the Spreadsheet Processor: The Spreadsheet Processor is a web based application that transforms pre-configured Excel spreadsheet files for occurrence data or metadata into GBIF-supported formats (GBIF 2011c). The Spreadsheet Processor accepts the completed Excel spreadsheet templates as a web form or as an email attachment. It then performs a series of data checking (validation) and transformation steps, and then returns a validated Darwin Core Archive file to the user, suitable for publishing via GBIF (or other biodiversity networks that support this format). The Spreadsheet Processor is hosted at http://tools.gbif.org/spreadsheet-processor/.

If the data are already digitized, or are already in Darwin Core Archive format, then the GBIF IPT or the Darwin Core Archive Assistant are options:

Using the GBIF IPT: The Integrated Publishing Toolkit (IPT) is a software platform developed by GBIF to facilitate easy and efficient publishing of biodiversity data on the Internet. To use the IPT, data must already be digitised as existing Darwin Core Archives or as any delimited text files (e.g. text files using comma or tab-separated values). The IPT also supports automatic registration of the dataset.

Currently, data publishers wishing to use the GBIF IPT need to install and host a local version of the IPT at their home institution. In future, it will be possible to access the IPT via a GBIF-endorsed Data Hosting Centre*, and this will be the easiest option for EIA practitioners to use.

* *Note: Data Hosting Centres are currently being developed by GBIF. They will serve as a "one-stop-shop" through which EIA practitioners will be able to capture, prepare, publish, register, archive and discover primary biodiversity data.*

Using the Darwin Core Archive Assistant: This facility can be used when data are already digitised or in a relational database. It would be suited to those users who have access to high levels of data management and IT capacity. It is not recommended for EIA practitioners.

Getting help for publishing: As a first step towards publishing biodiversity data, EIA practitioners can seek assistance from the wide network of GBIF country and organization Participant Nodes. A list of the national Participant Nodes is accessible at www.gbif.org/participation/participant-nodes/who-we-are/countries/, and the regional and thematic Participant Nodes are listed at www.gbif.org/participation/participant-nodes/who-we-are/other-associate-participants/. A majority of these nodes encourage, coordinate and assist in biodiversity data publishing activities within their respective jurisdictions and domains.

Steps 5 and 6: Registering the data with GBIF

Registration is the final step in the data publication process using Darwin Core Archive files. An entry for the dataset URL is made in the GBIF registry (http://data.gbif.org) that serves to make the Internet location of the dataset freely and openly available.

There are three options for registration of datasets:

(1) Using the GBIF Integrated Publishing Toolkit
(2) Using the Spreadsheet Processor
(3) Using other tools

The GBIF IPT supports automatic registration in the GBIF network (see the online manual for the IPT). Using the Spreadsheet Processor or other tools there is no automatic registration. An email must be sent to helpdesk@gbif.org with the following information:

1. Dataset title
2. Dataset description
3. Technical contact (the person to be contacted in matters regarding technical availability or resource configuration issues on the side of the dataset or data publisher)
4. Administrative contact (the person to be contacted in all matters regarding scientific data content and usage of a specific dataset or data publisher)
5. Institution name
6. Your relation to this Institution
7. The name of the GBIF Participant Node (the agency that coordinates data publishing in your country/region) that can endorse the publishing institution
8. The dataset URL: either the access point URL (if you are publishing using one of the provider softwares), or the DwC-Archive URL (if you are publishing via a zipped DwC-Archive)
9. The metadata document URL.

The GBIF Helpdesk will attend to your registration request as quickly as possible.

Once endorsement has been received and the registration is completed, the registered dataset can be found on the GBIF Registry website, through searching by institution name or dataset title.

Following registration, the GBIF Helpdesk will queue the newly registered dataset for indexing. Depending on the size of the dataset, indexing can take anywhere from minutes to weeks. If problems are encountered during indexing, the GBIF Helpdesk will work with you to resolve them as quickly as possible.

When indexing is successful, the new dataset will become publicly available (or discoverable) in the GBIF Data Portal (http://data.gbif.org).

Summary of steps to be followed:

Figure 2, below, provides a simple summary of what needs to be done at each step of the data publishing process. For EIA practitioners, the "route to undigitized data" (middle column of boxes) will be the simplest and most effective and efficient to follow.

Figure 2: What EIA practitioners should do

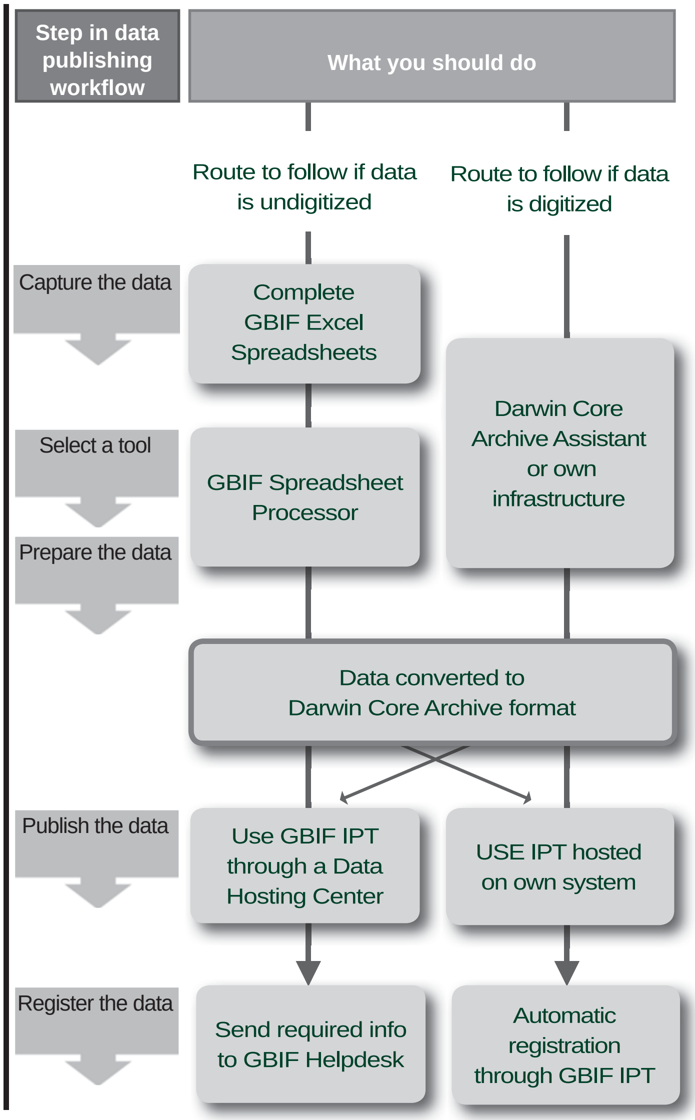
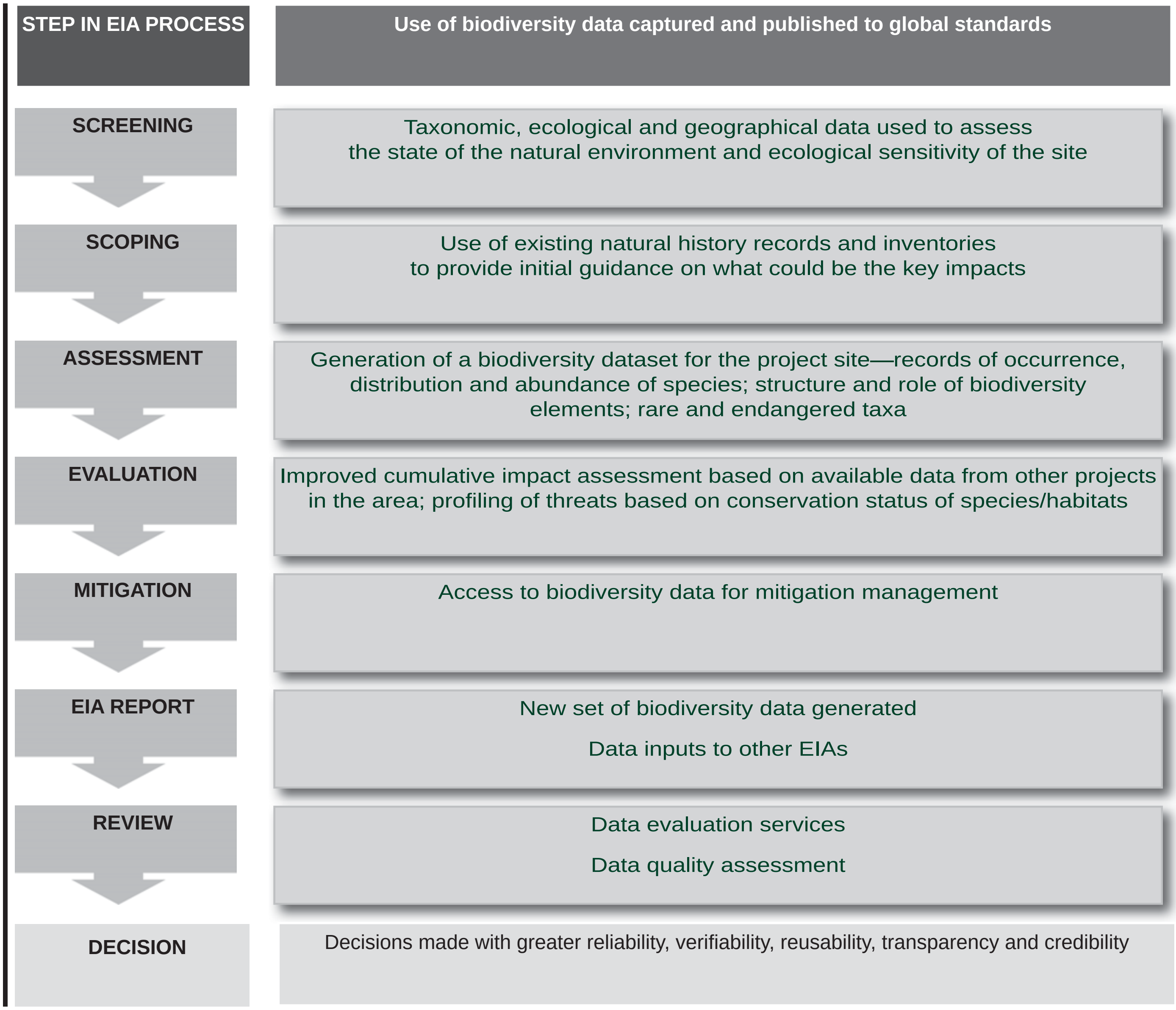

Step in data publishing workflow
What you should do
Route to follow if data is undigitized
Route to follow if data is digitized
Capture the data
Complete GBIF Excel Spreadsheets
Darwin Core Archive Assistant or own infrastructure
Select a tool
GBIF Spreadsheet Processor
Prepare the data
Data converted to Darwin Core Archive format
Publish the data
Use GBIF IPT through a Data Hosting Center
USE IPT hosted on own system
Register the data
Send required info to GBIF Helpdesk
Automatic registration through GBIF IPT

Figure 3: Steps in the EIA process, showing benefits of using verified, published primary biodiversity data



STEP IN EIA PROCESS	Use of biodiversity data captured and published to global standards
SCREENING	Taxonomic, ecological and geographical data used to assess the state of the natural environment and ecological sensitivity of the site
SCOPING	Use of existing natural history records and inventories to provide initial guidance on what could be the key impacts
ASSESSMENT	Generation of a biodiversity dataset for the project site—records of occurrence, distribution and abundance of species; structure and role of biodiversity elements; rare and endangered taxa
EVALUATION	Improved cumulative impact assessment based on available data from other projects in the area; profiling of threats based on conservation status of species/habitats
MITIGATION	Access to biodiversity data for mitigation management
EIA REPORT	New set of biodiversity data generated Data inputs to other EIAs
REVIEW	Data evaluation services Data quality assessment
DECISION	Decisions made with greater reliability, verifiability, reusability, transparency and credibility

Benefits of publishing EIA-related primary biodiversity data

This best practice guide describes a suite of simple, inexpensive tools and procedures that can be used by the impact assessment community to capture, publish and discover EIA-related primary biodiversity data. Publishing these data using consistent, internationally standardised formats is a relatively quick and easy procedure that can be easily adopted as an integral part or step of the EIA process (see Figure 3). Uptake of the tools and processes described in this best practice guide will:

- Enable free and open access to the biodiversity data which is essential for biodiversity-inclusive environmental assessments.
- Facilitate the ongoing expansion and improvement of the local, national and global biodiversity databases on which EIAs and other areas of scientific work and land-use management frequently rely, improving baseline knowledge of the ecosystems of a particular site, region or country.
- Help EIA practitioners to gain recognition for their work by enabling them to be cited in future uses of their data.
- Enhance the quality, predictive value, verifiability and transparency of EIAs, thus improving the land-use decisions that they inform and the confidence civil society can place in these decisions.

Glossary

Biodiversity: the variability amongst living organisms from all sources including, *inter alia*, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems

Data publishing: a process through which biodiversity datasets are made freely and openly available in standardised formats, via an Internet access point that is indexed in the GBIF Registry

Darwin Core: an internationally standardised set of terms for describing the identity and occurrence of organisms

Darwin Core Archive: a standardised format in which data must be presented in order to publish it through the GBIF infrastructure

Fitness-for-use (describing data): the suitability, effectiveness or usefulness of GBIF-mediated data in delivering accurate, authenticated, replicable and scientifically valid data for analysis and forecasting in conservation and management of natural resources

Metadata: information (data) about a dataset

Primary biodiversity data: digital text or multimedia data records of the occurrence of organisms

Methods and tools web sites: Sources of additional assistance

Getting started: overview of data publishing in the GBIF network
http://links.gbif.org/getting_started_publishing_en_v1

Publishing and Registering data with GBIF
http://links.gbif.org/dwc-a_publishing_guide_en_v1

GBIF Spreadsheet templates: User Guide
http://www.gbif.org/orc/?doc_id=2823

GBIF Metadata Profile: Reference Guide
http://www.gbif.org/orc/?doc_id=2820

GBIF Metadata Profile: How-to-Guide
http://www.gbif.org/orc/?doc_id=2821

Darwin Core Quick Reference Guide
http://links.gbif.org/gbif_dwc-a_guide_en_v1.1

Darwin Core Archive: How-to-Guide
http://links.gbif.org/gbif_dwc-a_how_to_guide_en_v1

GBIF
www.gbif.org

References

Chapman, A. (2005). *Principles of Data Quality, version 1.0. Copenhagen: Global Biodiversity Information Facility*. 58 pp. ISBN: 87-92020-03-8. Accessible at http://www.gbif.org/orc/?doc_id=1229&l=en

GBIF (2011a). *Getting started: An overview of data publishing in the GBIF network* (contributed by Remsen, D., Ko, B., Chavan, V., Raymond, M.). Copenhagen: Global Biodiversity Information Facility, 16 pp. ISBN: 87-92020-28-3. Accessible at http://links.gbif.org/getting_started_publishing_en_v1

GBIF (2011b). GBIF Spreadsheet templates: User Guide, version 1.0. (contributed by Remsen, D., Doring, M., Robertson, T.), Copenhagen: Global Biodiversity Information Facility, 20 pp., IBSN: 87-920-27-5. Accessible at http://www.gbif.org/orc/?doc_id=2823

GBIF (2011c). Publishing Species Checklists: Best Practices, version 1.0 (contributed by Remsen, D., Doring, M., Robertson, T.), Copenhagen: Global Biodiversity Information Facility, 20 pp., ISBN: 87-92020-26-7. Accessible at http://www.gbif.org/orc/?doc_id=2814&l=en

King *et al*. Improving access to biodiversity for, and from, EIAs – a data publishing framework built to global standards. (In prep).

Rajvanshi, A., Mathur, V., Iftikhar, U.A. (2007). *Best-practice guidance for biodiversity inclusive impact assessment: a manual for practitioners and reviewers in South East Asia.* CBBIA-IAIA Guidance Series, Capacity Building in Biodiversity and Impact Assessment (CBBIA) Project, International Association of Impact Assessment, North Dakota, USA.

A comprehensive set of references is provided in the following GBIF publication:

GBIF (2011). *Improving EIA Practice: Best Practice Guide for publishing primary biodiversity data* (contributed by Cadman, M., Chavan, V., King, N., Willoughby, S., Rajvanshi, A., Mathur, V.B., Roberts, R., and Hirsch, T.). Copenhagen: Global Biodiversity Information Facility, 51 pp. ISBN:.87-92020-35-6. Accessible at http://links.gbif.org/eia_biodiversity_data_publishing_guide_en_v1