GBIF-ICLEI Best Practice Guide for Biodiversity Data Publishing by Local Governments: *Concise Summary*

Version 1.0

May 2012

GBIF-ICLEI Best Practice Guide for Biodiversity Data Publishing by Local Governments: Summary

'Publishing' biodiversity data is defined as making biodiversity datasets publicly accessible in a standardized format via an online access point (typically a web address). This access point is recorded in a registry managed by the Global Biodiversity Information Facility (GBIF) and can be accessed (or 'discovered') via the GBIF Data Portal.

This best practice guide serves to enable local governments, their consultants and other interested and affected parties to discover, capture, manage and publish the biodiversity data generated during local government planning, decision-making, assessment and reporting processes. It represents a summarized version of a more comprehensive guide published by GBIF and its project collaborators (ICLEI and the CBD):

GBIF (2012). GBIF-ICLEI-CBD Best Practice Guide for Publishing Biodiversity Data by the Local Governments, (contributed by Cadman, M.J.; Chavan, V.; Patrickson, S.; Galt, R.; Mader, A.; Sood, R.; Hirsch, T.) Copenhagen: Global Biodiversity Information Facility, Pp. 62, ISBN: 87-92020-37-2, http://links.gbif.org/gbif_best_practice_guide_data_publishing_by_local_governments_en_v1

1. Introduction

Local governments as users and generators of biodiversity data

Primary biodiversity data are the digital text or multimedia records detailing facts about the occurrence of organisms. 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, developing sound environmental management policies, and making ecologically sound, sustainable development decisions.

Local governments are becoming increasingly important as managers and users of biodiversity assets and ecological infrastructure. They are responsible for environmental management and planning, regulation of land-use through planning and decision-making, and supporting the implementation of global, national and sub-national policies and strategies relating to biodiversity and sustainable development. Biodiversity management options arise in almost all traditional fields of activity for which local governments are responsible, such as integrated development planning, service provision and management of urban green spaces. They are, therefore, both important users and generators of biodiversity data.

For a variety of reasons, local governments face many challenges when it comes to dealing with biodiversity, and access to readily usable and verifiable biodiversity data is often problematical. Furthermore, much of the biodiversity data collected as part of local government planning processes is either lost after completion of the report, or is collected in inconsistent formats that cannot be easily archived or shared. This has largely been due

to the lack of awareness of the tools and protocols suitable for capturing, sharing, archiving and accessing primary biodiversity data, or lack of knowledge about how to use these tools.

The purpose of this best practice guide

This best practice guide explains important principles that underlie the data publishing process; describes the tools, standards and infrastructures that are available to practitioners in local government for publishing biodiversity data, and explains when and how these tools should be used. It also sets out to:

- Make local governments aware of the benefits to them of being able to access biodiversity data via the GBIF network, and highlights the important role they can play as contributors of biodiversity data; and,
- Explain how data publishing can be incorporated into planning, policy development and decision-making processes in local government.

Publishing biodiversity data

Through the Global Biodiversity Information Facility (GBIF), digital biodiversity data are being made freely and openly available via the Internet for scientists, researchers, national and local authorities and the general public. GBIF provides a suite of standards and tools that can be employed to discover and publish primary biodiversity data. 'Publishing' is the process through which biodiversity datasets are made publicly accessible in a standardized format, via an online access point (typically a web address, or URL). This access point is recorded in a registry managed by GBIF. Published datasets can then be discovered and accessed via the GBIF Data Portal.

2. 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 ***synthesised 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 local government reports tends to be interpretive or synthesised 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.

Metadata, which is 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.

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*

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 co-ordinates 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' (Crisman, 1983), or 'potential use' (Chapman, 2005a), 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 (fit-for-use data) to input (investment of time in data capture and publishing).

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 of best practice

The data publishing workflow

'Publishing' makes datasets 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, as follows:

- **Capturing the data** in a consistent, exchangeable format;
- **Preparing the data for publication** (i.e. using GBIF tools for converting it into a standardized format known as a *Darwin Core Archive* that can be accepted on the GBIF network);
- **Publishing the dataset** (i.e. making it publicly accessible via a web address using the GBIF tools); and,
- **Registering the data** in the web-based data access point in the GBIF registry.

Once the data have been published and registered with GBIF, then they are freely and openly accessible (or '**discoverable'**) through the GBIF network and the GBIF data portal (http://data.gbif.org).

Tools and protocols for data publishing

At each step in the data publishing workflow there are simple ***tools and protocols*** available for data publishers to use. The option selected should be matched with the technological and data management capacity of the user. This guide will help practitioners in local government, their 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 local government planning, reporting and decision-making processes.

The tools include:

- pre-configured **Excel Spreadsheets** for capturing data in a standardized way;
- the **Spreadsheet Processor** or **Darwin Core Archive Assistant** for generating the Darwin Core Archive file, and
- the **Integrated Publishing Toolkit** or other manual tools for publishing the Darwin Core Archive file through the GBIF network.

The simplest route for publishing biodiversity data would be to use the pre-configured GBIF Excel Spreadsheet templates, prepare them for publishing using the GBIF Spreadsheet Processor and then publish the datasets using the GBIF Integrated Publishing Toolkit (IPT) or through a Data Hosting Centre* (see note, below), if one is available. This workflow is summarized in the flow chart shown in Figure 1.

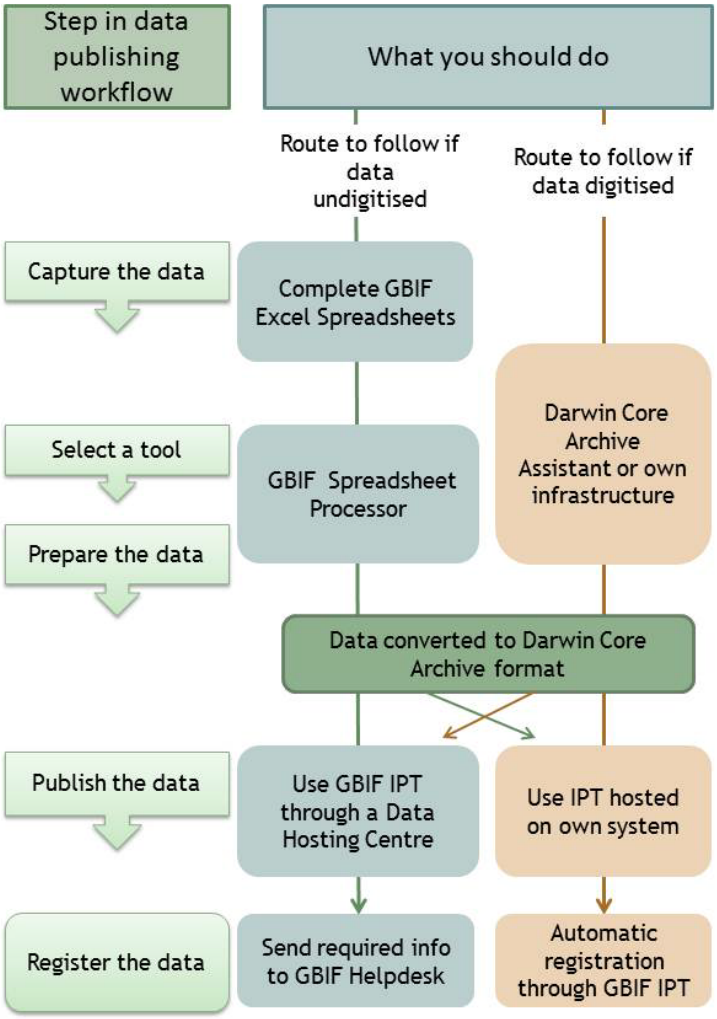

Figure 1: Steps in the data publishing workflow and tools available for use at each step

Step-by-step guide to data publishing using GBIF tools and protocols

Step 1: Capturing the data

There are three GBIF spreadsheet templates available (for occurrence data, simple checklists and metadata); those that will be of greatest use to local government practitioners are the metadata and occurrence spreadsheets. These spreadsheets are simple tools that provide a common format and standard for collecting data, using consistent terminology.

Use of the GBIF Excel data capture templates (a) makes it easier for publishers to collect, manage and share 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.

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.

There are a number of GBIF User Guides (GBIF 2011b, GBIF 2011c) that provide step-by-step assistance for use of the Excel spreadsheets. Box 1, below, provides a quick summary of how to use these spreadsheets.

Box 1: GBIF Excel spreadsheets - how to use these tools

Step 1: Access the spreadsheets by logging onto the GBIF website and downloading the appropriate template (i.e. occurrence or metadata templates). Visit http://www.gbif.org

Step 2: Populate the spreadsheet with your data using at least the compulsory fields; make use of the inline help if you need to, by hovering the cursor over the cells marked with red upper corners.

Step 3: Upload the completed spreadsheet to the Spreadsheet Processor so that it can be turned into a format suitable for publishing.

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

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)***. 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).

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 digitized.

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).

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.

* *Note: Data Hosting Centres are currently being developed by GBIF. They will serve as a "one-stop-shop" through which 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 digitized 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.

Use of the GBIF Spreadsheet Processor and GBIF IPT is explained in Box 2, below.

Box 2: The GBIF Spreadsheet Processor and Integrated Publishing Toolkit- how to use these tools

Step 1: Access the Spreadsheet Processor at http://tools.gbif.org/spreadsheet-processor/ and upload your completed Excel spreadsheet by following the instructions provided.

Step 2: Once Spreadsheet Processor has checked and transformed the data, a Darwin Core Archive file will be returned to you, saved in the same folder as your original spreadsheet.

Step 3: Publish the DwC-A file in one of the following ways:

(a) do it yourself by posting it on a web server and registering the URL with GBIF through a Participant Node (note: registering of datasets is explained in further detail in section 4.5. of this best practice guide);

(b) send it by FTP or email to a Data Hosting Centre for publication via the GBIF IPT; or

(c) use the IPT yourself to publish the file.

Currently, data publishers wishing to use the GBIF IPT need to install and host a local version of the IPT at their home institution. Information on installing and operating the IPT can be found in the IPT user manual or on the IPT website, at http://code.google.com/p/gbif-providertoolkit/

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 local government practitioners to use.

Why using the IPT is recommended for local governments:

- It can be used to manage and publish primary occurrence data, taxonomic checklists and metadata.
- It can be used for data that have already been converted to Darwin Core Archives (using the Excel spreadsheets and the Spreadsheet Processor) or it can accept any delimited text files (e.g. text files using comma or tab-separated values).
- The IPT supports automatic registration of the dataset (see below).
- The IPT can be used to author the metadata files, and can be used to create Data Papers (See full best practice guide, Section 4.3)
- The same version of the IPT can be used by many different data publishers. For example, if ICLEI were to host a version of the IPT on its systems, then any number of local governments could use it to publish their data, while keeping clear attribution and a distinct identity for the datasets.

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:

(a) Using the GBIF Integrated Publishing Toolkit;
(b) Using the Spreadsheet Processor; and,
(c) 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 information contained in Box 3.

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[1], through searching by institution name or dataset title.

Box 3: Information required by GBIF for registration of datasets

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 co-ordinates 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.

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.

3. Biodiversity data publishing by local government

It is difficult to make a 'one-size-fits-all' set of recommendations to local governments as to which data publishing option they should follow, as these institutions vary widely in respect of capacity and resources. However, some general rules of thumb that could be applied include:

- Use the GBIF Excel spreadsheet templates to capture all biodiversity data that are gathered as part of local government planning processes - this will mean that your data will automatically be suitable for publishing via the GBIF network.
- Build the use of these tools, and this best practice guide, into the Terms of Reference for all consultants that are contracted to do assessments or prepare biodiversity reports or plans for any relevant local government process. You should also recommend to consultants that they make use of the GBIF network to source appropriate data when embarking on a new study.
- Make data discovery and publishing a part of your Local Biodiversity Strategy and Action Plan.
- Become part of the GBIF network and benefit from the knowledge exchange and support it offers.

Benefits to local governments of publishing biodiversity data

There are a number of compelling reasons for local governments to publish biodiversity data using the GBIF tools and network. Chief amongst these are that data publishing will:

- enable free and open access to biodiversity data, which is essential for biodiversity-inclusive planning and development at local government level;
- facilitate the ongoing expansion and improvement of the local, national and global biodiversity databases on which environmental planning, EIAs, land-use management, policy development and areas of scientific work frequently rely, improving baseline knowledge of the ecosystems of a particular site, region or country;
- help practitioners who do specialist work for local governments 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 local government planning processes, thus improving the land-use decisions that they inform and the confidence civil society can place in these decisions.

Where to find further assistance:

The principles, tools and processes described in this Executive Summary are explained in greater detail in the full-length GBIF/ICLEI/CBD best practice guide to data publishing by local governments. There are also numerous other GBIF User Guides that are available online to assist with the publication of primary biodiversity data and their associated

metadata, using the GBIF tools.These guides provide detailed, step-by-step instructions in the use of all the key tools used at different steps in the data publishing process. The table below summarises the key documents that can provide assistance at each step of the data publsihing pathway.

Should further assistance be required, you can obtain assistance by contacting your local GBIF Participant Node. GBIF works with a large number of Participants that can be countries or institutions. A Participant Node is the regional or national committee that co-ordinates data sharing activities within its domain, and helps link the region into the global GBIF network. You can find out who your Participant Node is by looking it up on the GBIF website at: http://www.gbif.org/particpation/participant-nodes/who-we-are/

Useful references and websites:

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://links.gbif.org/dwca-spreadsheet-processor-guide
GBIF Metadata Profile: Reference Guide - http://links.gbif.org/gbif_metadata_profile_how-to_en_v1
GBIF Metadata Profile: How-to-Guide - http://links.gbif.org/gbif_metadata_profile_how-to_en_v1
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

Publishing EIA-related biodiversity data: GBIF (2011). *Promoting biodiversity data inclusive EIA: 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.D.) 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

Glossary of frequently used terms:

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