Getting Started

Overview of data publishing in the GBIF Network

Version 1.1

April 2015

Suggested citation:
GBIF (2015). Getting Started: An overview of data publishing in the GBIF network, version 1.1 (contributed by Chavan, V., González-Talaván, A., Ko, B., Raymond, M. & Remsen, D.). Copenhagen: Global Biodiversity Information Facility, 17 pp. ISBN: 87-92020-28-3 (for version 1.0). Accessible at http://links.gbif.org/getting_started_publishing_en

ISBN: 87-92020-28-3 (for version 1.0)
Persistent URI: http://links.gbif.org/getting_started_publishing_en_v1.1
Language: English

Document Control:

Version	Description	Date of release	Author(s)
1.0	Review, edits and final styling	5 April 2011	DR, BK, VC, MR
1.1	Update of critical elements	24 April 2015	AGT

Cover Art Credit: *Gregory Basco*
Brown Pelican, Pelecanus occidentalis

About GBIF

The Global Biodiversity Information Facility (GBIF) was established 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 envisions 'a world in which biodiversity information is freely and universally available for science, society, and a sustainable future'. GBIF's mission is to be the foremost global resource for biodiversity information, and engender smart solutions for environmental and human well-being[1]. To achieve this mission, GBIF encourages a wide variety of data publishers across the globe to discover and publish data through its network.

[1] GBIF (2011). GBIF Strategic Plan 2012-16: Seizing the future. Copenhagen: Global Biodiversity Information Facility. 7pp. ISBN: 87-92020-18-6. Accessible at http://links.gbif.org/sp2012_2016.pdf

Table of Contents

1. Introduction

This document provides an overview of biodiversity data publishing through the GBIF network. The word, "publish", in this sense, refers to making biodiversity datasets publicly accessible, in a standardised form, via an access point, typically a web address (a URL). This access point is recorded in the GBIF central services, which serves to make the dataset locations globally discoverable. GBIF also maintains the website GBIF.org, which provides discovery and access services to data indexed from datasets published through GBIF. This data index accessible through GBIF.org is updated constantly.

GBIF provides a ***means*** to share biodiversity data. The data being shared remain at the location from which they are being shared and under the control of the data publisher. The index GBIF maintains represents a cached set of data that is regularly refreshed. There are two different means by which data are commonly shared which will be discussed in this guide:

1. A connection to a "live" database from which a copy of data is extracted and transferred to a user on demand.
2. Access to a data file in a standard format representing a copy of a source dataset that has been extracted and is served as a complete file through a web server URL.

This guide provides a high-level overview of the biodiversity data types that can be published through the GBIF network. It presents the scope of the core data types currently supported by GBIF and publishing options for each. Its major objective is to help potential data publishers choose the most suitable option and/or tool to achieve the goal of publishing biodiversity data through the GBIF network. The guide itself does not provide specific details on each publishing option. Instead, it provides an overview of these options with links to more detailed documentation and online resources.

2. Scope

From a data-publication perspective, GBIF makes the following distinctions:

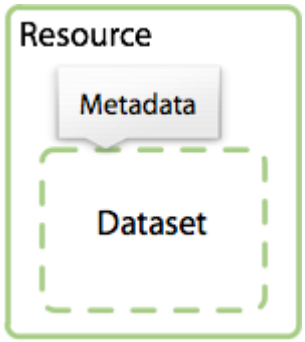

Figure 1. Scope

- Biodiversity data published through GBIF are organised into ***datasets*** or ***data resources***.
- A dataset is a collection of ***data records***.
- Datasets are described by ***metadata***. In the context of GBIF, metadata provide information about the suppliers of biodiversity data and about the origins, purpose and nature of those data.
- A data record is a collection of ***record elements*** or properties. An example data record may describe a museum specimen. One of the data elements would almost certainly be a ***scientific name*** element.
- A record element contains the ***data values*** (i.e., the data). An example value in a ***scientific name*** record element would be *Limulus polyphemus*.

2.1. Three Core Data Types

The GBIF data-publishing platform supports the publication of three primary classes of data.

1. Primary Biodiversity Data[2] or Occurrence Data

This category of information refers to data or information relating to a specific instance of a taxon, usually a species, in nature, in a collection or in a dataset. An example dataset would be a collection of bird observation data records where a data record provides details of a particular bird sighting. Another example would be a collection of specimen data records from a natural history museum. A single taxon may be the subject of many records in a single occurrence dataset. The occurrence of biological species in spatial and temporal terms is the fundamental data unit on which services and analytical workflows are based.

2. Taxonomic Data

This category of information refers to data or information relating to a taxon and ***not*** necessarily to a specific instance (occurrence) of an individual within that

[2] Primary biodiversity data is defined as: digital text or multimedia data record detailing facts about the instance of occurrence of an organism, i.e. on the what, where, when, how and by whom of the occurrence and the recording.

taxon. An example dataset would be an annotated checklist of bird species where a data record provides information about a single species. A single taxon is generally the subject of only a single record in a taxonomic dataset.

3. Resource (or Dataset) Metadata

Metadata are data records that provide descriptive information about datasets. In the context of GBIF, metadata provide information about the suppliers of biodiversity data and about the origins, purpose and nature of those data together with the statement of their 'fitness-for-use'. GBIF supports both the authoring and publishing of metadata that conform to a GBIF Metadata Profile (GMP)[3]. Metadata are required for all datasets published through the GBIF network. Metadata are important to improve dataset discovery and to provide potential users with details on the 'fitness-for-use' of the data they describe. Metadata can describe both digital and non-digital data sets: data publishers can also publish metadata about datasets that are yet ready to be published.

Each of these three data classes is supported by different data-publishing options within the GBIF data-publishing platform and will be detailed in order.

2.2. Data Publishing Workflow

Publishing data through GBIF network is achieved by following certain steps. Figure 2, depicts a model data-publishing workflow. Major steps leading to the discovery and accessibility of the biodiversity data through the GBIF network include; (a) the selection of appropriate data-publishing tools (or options) on the basis of data-type, technical skill-sets, and available technical capacity, (b) preparing your dataset to conform with the standard data-exchange format, (c) publishing datasets employing the appropriate data-publishing tool, and (d) registering the data access-point in the GBIF central services. Once these steps are accomplished, your data will be discoverable and accessible through the GBIF network and GBIF.org.

Figure 2. Data Publishing Workflow

[3] The GBIF Metadata Profile - http://rs.gbif.org/schema/eml-gbif-profile/

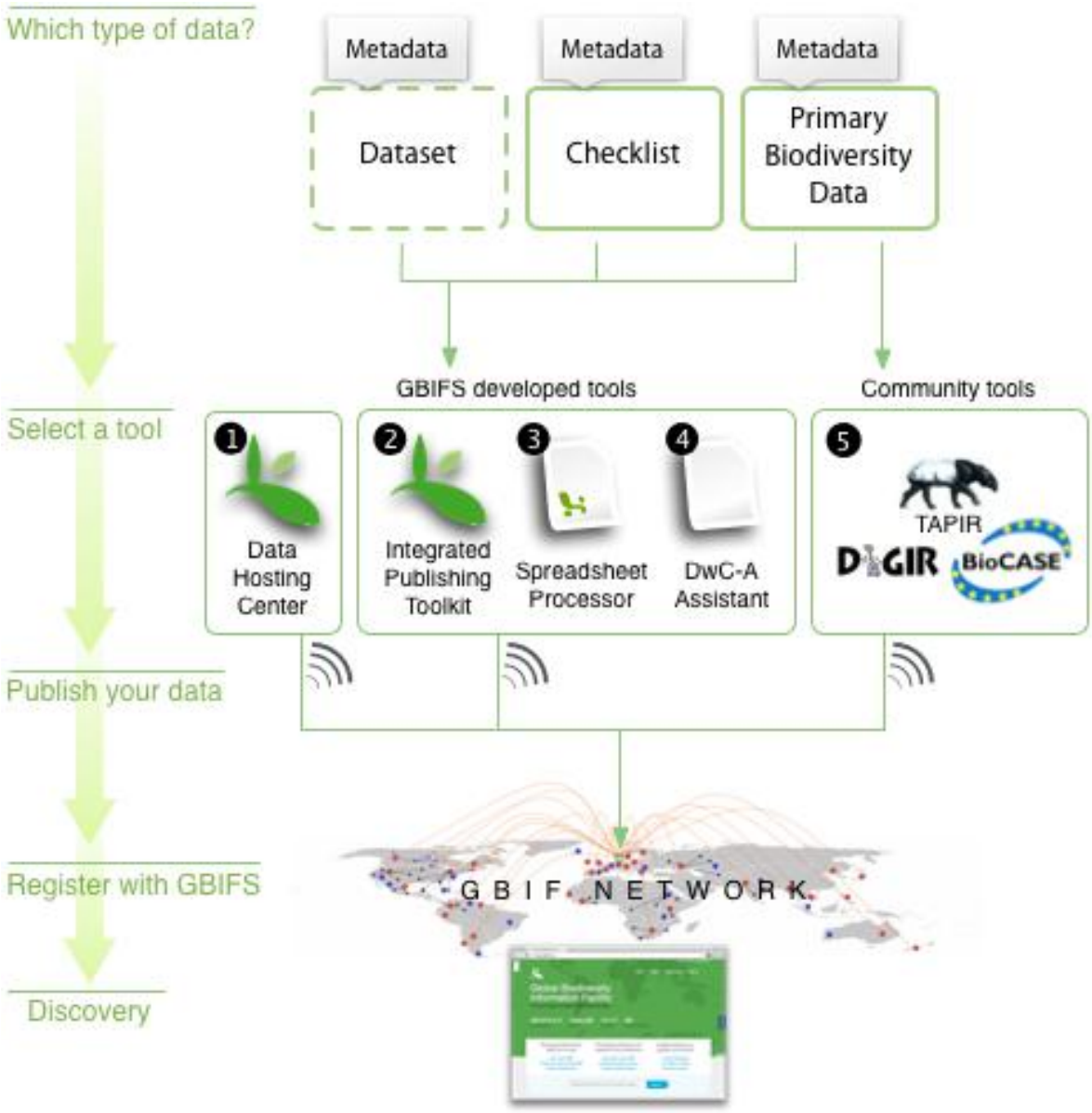

Figure 3. An overview of data publishing options in the GBIF Network

3. Publishing Primary biodiversity data or Occurrence Data

Occurrence data may be published through GBIF via two different methods:

1. Via access to complete or partial datasets provided as cached data files or ***archives*** that conform to a standard format. ***This is the preferred approach for new data publishers to GBIF***.
2. Through the use of biodiversity ***data exchange protocols*** that allow users to communicate "live" via the internet to a source database. This is the traditional way data has been published through GBIF and remains an option today.

3.1. The Darwin Core Archive Format

The preferred approach to publishing occurrence and taxonomic data to the GBIF network, both for new data publishers as well as an evolutionary migration path for existing data publishers, is through the use of Darwin Core Archives.

The **Darwin Core Archive (DwC-A)** format is an internationally recognised and formally ratified biodiversity informatics data standard. It simplifies the publication of biodiversity data by combining the use of a stable and internationally ratified glossary of terms, the Darwin Core, with the ease and readability of standard Comma-Separated-Values (CSV)-style text files. An archive is a collection of files that conform to the described standard and are compressed into a single file. Darwin Core Archives do not require the installation of dedicated software by the data publisher and can easily be produced and published with nothing other than a web server to host the published archive (note that data hosting services are also available for data publishers without access to a web server).

GBIF provides a rich array of support and tools for publishing Darwin Core Archives and for customising the format to include new data types for even more flexibility. Go to Darwin Core Archives How-to Guide at http://links.gbif.org/gbif_dwc-a_how_to_guide_en_v1.

In addition to data files, the Darwin Core Archive requires the inclusion of a resource metadata document (See Publishing Resource Metadata below).

3.2. Data Exchange Protocols

There are three biodiversity data exchange protocols that GBIF can accept, ***but which are no longer the preferred method for publishing data to the GBIF network***. Each of the protocols defines a particular process for interacting with a database. Users can follow these processes to make specific queries to the source database and retrieve subsets of the data in response. Data is returned to the user in a standardised response format. Supporting these protocols and responses requires data publishers to install and configure

software "wrappers" that connect to their database and support a "live" connection to users.

3.2.1. Biological Collections Access Service (BioCASe)

BioCASe[4] refers to the Biological Collections Access Service and is currently the best supported and maintained of the protocols. The acronym refers both to the protocol (BioCASe) and to the project (BioCASE), which focuses on European natural history collections. The BioCASe protocol returns data records in an XML format called Access to Biological Collections Data (ABCD[5]), capable of expressing rich, deeply nested data. The primary software implementation for BioCASe is BioCASE Provider Software[6] and is actively maintained by BioCASE.

Go to the BioCASe website for more information. http://www.biocase.org/

3.2.2. TDWG Access Protocol for Information Retrieval (TAPIR)

TAPIR is an open standard developed under the auspices of Biodiversity Information Standards (TDWG). TAPIR is currently used to serve biodiversity data to the GBIF network using an older version of Darwin Core, which is not up-to-date with the ratified Darwin Core standard. Different server and client implementations exist, but none are directly supported by GBIF.

TAPIR Protocol - http://www.tdwg.org/standards/449/

TAPIR software - http://wiki.tdwg.org/twiki/bin/view/TAPIR/TapirSoftware

3.3.3. Distributed Generic Information Retrieval (DiGIR)

DiGIR was the pioneer implementation of the biodiversity data exchange protocols and is still in use within the GBIF network. DiGIR software is currently not under active development and is not actively supported by GBIF.

Go to the DiGIR website for more information. http://digir.sourceforge.net/

3.3. To Publish Occurrence Data

There are two mechanisms for publishing occurrence data through GBIF (described above): using Darwin Core Archives or using data exchange protocols. The first decision a data publisher must take is which of these mechanisms is most appropriate for them.

[4] BioCASe - http://www.biocase.org/
[5] ABCD Schema - http://www.bgbm.org/TDWG/CODATA/Schema/
[6] BioCASE Provider Software - http://www.biocase.org/products/provider_software/index.shtml

3.3.1 Darwin Core Archive publishing

This is the preferred mechanism for publishing through GBIF and several tools are now available to support Darwin Core Archive publishing. These tools are designed to provide a broad range of data publishing workflows for different publishers, including those seeking to publish data using simple spreadsheet tools, those wishing to make use of data hosting services, those able to create their own Darwin Core Archives from existing databases, and those wishing to install a data-publishing tool on a dedicated server with a permanent internet connection. The 'Darwin Core Archive How to Guide' explains the options available and how to select the most appropriate tool.

Workflow for publishing occurrence data using Darwin Core Archives:

1. To publish the metadata associated to your dataset, see the section To Publish Resource Metadata (below)
2. Refer to the following manuals
    a. Darwin Core Archive How to Guide[7]
    b. Reference Guide to Darwin Core Terms
3. These will guide users to select a publishing solution from the following:

Publishing Solution	Data Format	User Guide
Integrated Publishing Toolkit	Darwin Core Archive	http://links.gbif.org/ipt_user_manual
Spreadsheet Templates	Darwin Core Archive	http://links.gbif.org/xls
Make your own DwC-A	Darwin Core Archive	http://links.gbif.org/dwc-a_own

4. See Publishing and Registering Data with GBIF[8]

3.3.2. Data Exchange Protocols

These protocols can still be used to publish occurrence data through the GBIF network. The use of protocols to publish data requires the installation and configuration of dedicated software (called a 'wrapper') connected to a 'live' database. The links below provide more information about them.

Workflow for publishing occurrence data using Data Exchange Protocols:

1. Select a publishing solution from the following table:

Publishing Solution	Data Format	User Guide
BIOCASE	ABCD	http://links.gbif.org/biocase
TapirLink	Darwin Core XML	http://links.gbif.org/tapirlink
DiGIR	Darwin Core XML	http://links.gbif.org/digir

[7] http://links.gbif.org/gbif_dwc-a_how_to_guide_en_v1
[8] http://links.gbif.org/dwc-a_publishing_guide_en_v1

2. Refer to the user guides in the table above.
3. Publishing and registration are built into the wrapper tools and set up at the time of configuration.

4. Publishing Taxonomic Data

Darwin Core Archives are the ***only*** format that GBIF supports for publishing species data through GBIF. Note that documenting the provenance and scope of datasets is ***required*** in order to publish data through the GBIF network. In addition to data files, the Darwin Core Archive requires the inclusion of a resource metadata document (See 5. Publishing Resource Metadata below).

The capacity to publish species data in a standard manner is not restricted to simple species checklists. The extensibility of the Darwin Core Archive format supports the sharing of:

- Taxonomic catalogues and monographic data
- Species descriptions such as might appear on a website "species page"
- Images and other multimedia
- Distribution details
- Measurements and Facts
- And more…

4.1. To Publish Taxonomic Data

GBIF has developed a number of tools to assist with the creation and publication of Darwin Core Archives. These are described in detail in the Darwin Core Archive How to Guide.

Workflow for publishing taxonomic data using Darwin Core Archives:

1. To publish the metadata associated to your dataset, see To Publish Resource Metadata (below)
2. Refer to the following Manuals
    a. Darwin Core Archive How to Guide[9]
    b. Best Practices in Publishing Species Checklists[10]
    c. GBIF GNA Profile: Reference Guide[11]
3. Select a publishing solution from the following table:

Publishing Solution	Data Format	User Guide
Integrated Publishing Toolkit	Darwin Core Archive	http://links.gbif.org/ipt_manual

[9] http://links.gbif.org/gbif_dwc-a_how_to_guide_en_v1
[10] http://links.gbif.org/checklist_best_practices
[11] http://links.gbif.org/gbif_gna_profile_reference_guide

Spreadsheet Templates	Darwin Core Archive	http://links.gbif.org/xls
Make your own DwC-A	Darwin Core Archive	http://links.gbif.org/dwc-a_own

4. See Publishing and Registering Data with GBIF

5. Publishing Resource Metadata

Metadata are literally 'data about data'. They provide information on such aspects as the 'who, what, where and when' of data and can be considered from the perspective of both the data producer and the data user.

GBIF supports the publication and exchange of metadata documents that describe the properties of biodiversity datasets, particularly occurrence datasets such as natural history collections data, as well as taxonomic and species-level datasets such as taxonomic catalogues.

For the producer, metadata are used to document data in order to inform prospective users of their characteristics, while for the user, metadata are used to both discover data and to assess their appropriateness for particular needs – their 'fitness for use'. Metadata thus complement the two core classes of biodiversity data supported by the GBIF data-publishing platform: occurrence datasets and taxonomic/species datasets. A metadata document may also be used to describe a dataset with no accessible data service, such as an un-digitised natural history collection, or a dataset that contains data that are in a format not easily publishable through the normal GBIF infrastructure, but could nonetheless be manually accessed and extracted by an interested user.

GBIF has developed a specific resource metadata description profile that is based on the internationally recognised Ecological Metadata Language (EML) standard. Other metadata standards can be accepted but full authoring support for these is not currently available using GBIF tools.

5.1. To Publish Resource Metadata

It is a requirement that resource metadata are published to accompany all occurrence or taxonomic datasets published through the GBIF network. The GBIF tools that support the publication of Darwin Core Archives also assist data publishers in the creation and publication of resource metadata.

Workflow for publishing resource metadata using the GBIF Metadata Profile:

1. Refer to the following manuals
    a. GBIF Metadata Profile: How-to Guide[12]
    b. GBIF Metadata Profile: Reference Guide[13]

[12] http://links.gbif.org/gbif_metadata_profile_how-to_en_v1
[13] http://links.gbif.org/gbif_metadata_profile_guide_en_v1

2. These will guide users to select a publishing solution from the following:

Publishing Solution	Metadata Format	User Guide
Integrated Publishing Toolkit	GBIF EML Profile	http://links.gbif.org/ipt_manual
Spreadsheet Templates	GBIF EML Profile	http://links.gbif.org/xls
Make your own EML	GBIF EML Profile	http://links.gbif.org/dwc-a_asst

6. Additional resources

Darwin Core Archive – Reference Guide to the XML Descriptor File (for technical users)

For any help on data publishing please contact helpdesk@gbif.org.