Getting Started Using the IPT

About GBIF

The Global Biodiversity Information Facility (GBIF) was established as a global megascience 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-being1. To achieve this mission, GBIF encourages a wide variety of data publishers across the globe to discover and publish data through its network.

Introduction

This document provides an step by step guide for biodiversity data publishing through the GBIF network using the IPT (Integrated Publishing Tookit). 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 Registry, which serves to make the dataset locations globally discoverable.
GBIF also maintains a Data Portal. The data portal provides discovery and access services to data indexed from datasets published through GBIF, which 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 within the Data Portal represents a cached set of data that is regularly refreshed.

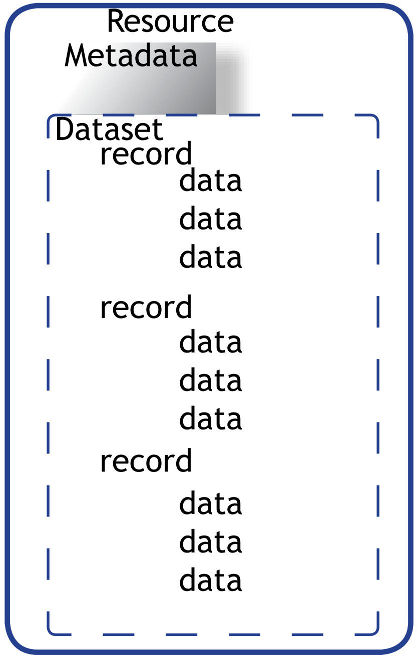
This guide provides a high-level overview of

1. the biodiversity data types that can be published through the GBIF network.
2. It presents the scope of the core data types currently supported by the IPT
3. Its major objective is to help potential data publishers learn how to use the IPT hosted by AntaBIF in order to achieve the goal of publishing biodiversity data through the GBIF network.
4. The guide itself Provides a detailed step by step overview, that wil guide you through the data publication process..

Data Portal : http://data.gbif.org

Scope

From a data-publication perspective, GBIF makes the following distinctions:
• Biodiversity data published through GBIF are organized 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 and purpose 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.



Three Core Data Types

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

1. Primary Biodiversity Data 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.
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.

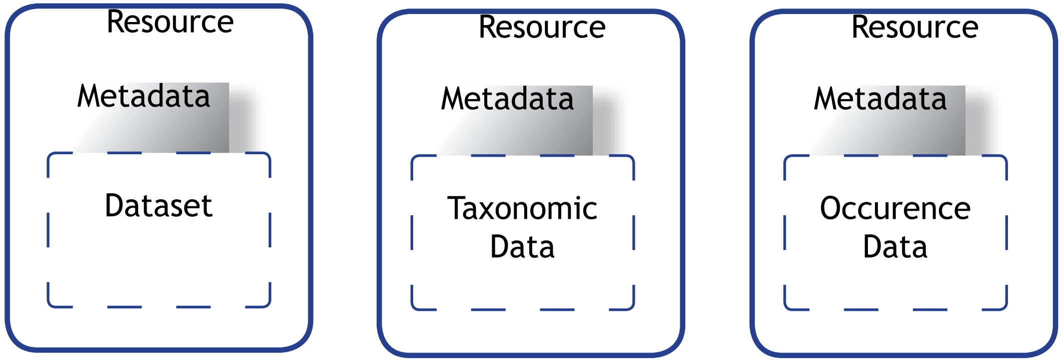
2. Taxonomic Data

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 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 and purpose 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).
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.

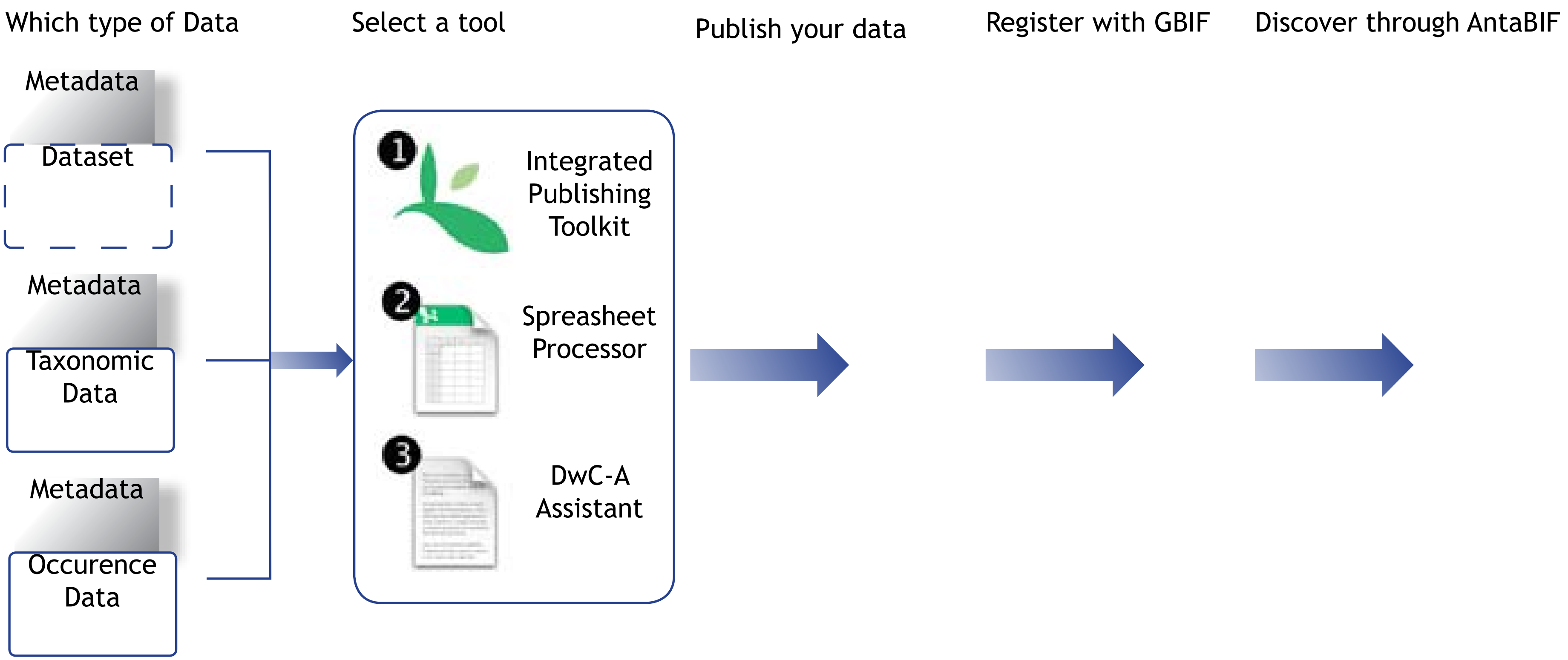
Figure 1. Three Core Data Types



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) preparing your dataset to conform with the standard data-exchange format, (b) publishing datasets employing the Integrated Publishing toolbox, and (c) registering the data access-point in the GBIF Registry.

Once these steps are accomplished, your data will be discoverable and accessible through the GBIF network and Data Portal (http://data.gbif.org).



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 datapublishing 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.

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
    b. GBIF Metadata Profile: Reference Guide

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

Publishing Primary biodiversity data or Occurrence Data

Occurrence data may be published through GBIF 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.

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).
Darwin Core Archive publishing is the preferred mechanism for publishing through GBIF and several GBIF 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, from 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, or those wishing to install a data-publishing tool on a dedicated server with a permanent internet connection.

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 (above)

2. Refer to the following manuals
   a. Darwin Core Archive How to Guide
   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_manual
Spreadsheet Templates	Darwin Core Archive	http://links.gbif.org/xls
Make your own EML	Darwin Core Archive	http://links.gbif.org/dwc-a_own

4. See Publishing and Registering Data with GBIF

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 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…

To Publish Taxonomic Data

The only way to publish taxonomic data through the GBIF network is using Darwin Core Archives. 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 (above)

2. Refer to the following Manuals
    a. Darwin Core Archive How to Guide
    b. Best Practices in Publishing Species Checklists
    c. GBIF GNA Profile: Reference Guide

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

4. See Publishing and Registering Data with GBIF