

GBIF Metadata Profile: How-To Guide

Version 1.0



April 2011

This document is a how-to guide for creating a metadata document that conforms to the GBIF Metadata Profile.

Suggested citation: GBIF (2011). GBIF Metadata Profile - How-to Guide, (contributed by Ó Tuama, Eamonn, Braak, K. Remsen, D.), Copenhagen: Global Biodiversity Information Facility, 11 pp, accessible online at http://links.gbif.org/gbif_metadata_profile_how-to_en_v1

Persistent URI: http://links.gbif.org/gbif_metadata_profile_how-to_en_v1

ISBN: 87-92020-24-0

Language: English

Copyright © Global Biodiversity Information Facility, 2010

License:



This document is licensed under a [Creative Commons Attribution 3.0 Unsupported License](http://creativecommons.org/licenses/by/3.0/)

Document Control:

Version	Description	Date of release	Author(s)
1.0	Checked consistencies across relevant documents, updated links to production sites, updated text and pictures to reflect current functionalities.	1 Mar 2011	Kyle Braak, Burke Chih-Jen Ko, Melianie Raymond

Cover Art Credit: John Giez (<http://www.flickr.com/photos/johngiez-/4453912650/>)

Maidenhair fern sporophyte, (*Adiantum* sp.)

This document is also part of the 'GBIF Data Publishing Manual version 1.0, ISBN 87-92020-31-3, available at http://links.gbif.org/data_publishing_manual.

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-being¹. 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

About GBIF	iii
Introduction	1
Metadata Authoring: What to Use?	2
Integrated Publishing Toolkit (IPT)	2
GBIF Darwin Core Spreadsheet Templates	4
Modify an Existing Sample Document	7
Appendix	8
<i>Required Elements</i>	8
Best practices guide to debugging/validating XML.....	9

List of Figures

Figure No.	Caption of the Figure	Page
1	Figure 1 - Screenshot of a help dialog for the term “Study Area Description”	6
2	Figure 2 - Screenshot of the field validation message displayed when an email field is submitted with an irregular email address.	6
3	Figure 3 - Screenshot of the Published Release section of the Manage Resource page of the IPT	7
4	Figure 4 - Screenshot of part of the Basic Metadata section of the Dataset Description page in Spreadsheet mode	7
5	Figure 5 - Screenshot of part of the People and Organisations section of the Dataset Description page showing the hover-over help in Spreadsheet mode	8

Introduction

Documenting the provenance and scope of datasets is required in order to publish data through the GBIF network. Dataset documentation is referred to as ‘resource metadata’ that enable users to evaluate the fitness-for-use of a dataset.

There are various ways to write a metadata document conforming to the GBIF Metadata Profile (GMP). This How-To Guide will go through the most common ways, such as using the [GBIF Integrated Publishing Toolkit \(IPT\)](#)² metadata editor, the GBIF Darwin Core Spreadsheet templates metadata forms³, or simply taking a sample metadata document and replacing fields of relevance with the source data. The guide extends the [Reference Guide to the GBIF Metadata Profile](#)⁴.

If metadata describing a dataset are also being published using Darwin Core Archives (DwC-A), the metadata file will be included in the DwC-A file that bundles it together with the data (based on the Darwin Core terms) that it describes. For help with making the complete DwC-A, refer to the [Darwin Core Archive: How-To Guide](#).⁵

Once the metadata document has been written and validated, it is ready to be published. For help publishing metadata, users can refer to the [Publishing and Registering Data with GBIF](#)⁶

Ultimately, the goal in publishing the metadata is that the data resource described therein can be fully documented and registered in the GBIF Registry. In so doing, the data resource becomes globally discoverable.

³ <http://tools.gbif.org/spreadsheet-processor/>

⁴ http://gbif-metadata-profile-guide_en_v1

⁵ http://links.gbif.org/gbif-dwc-a-how-to-guide_en_v1

⁶ http://links.gbif.org/dwc-a-publishing-guide_en_v1

Metadata Authoring: What to Use?

If occurrence data or taxonomic (checklist) data are being published using the [GBIF Integrated Publishing Toolkit \(IPT\)](#)⁷ or the [GBIF Darwin Core Spreadsheet Templates](#)⁸, there is a built-in metadata authoring functionality that can be used to write an accompanying metadata document conforming to the GMP. It may be convenient to use either of these tools for authoring metadata even if no data is being published. Although the IPT is software that requires installation, it might save a user time in the long run if there is a need to author and manage several metadata documents. On the other hand, if only a few metadata documents are needed, it might be easiest to make use of a Spreadsheet Template. Another option is to modify a [sample document](#)⁹. Below is a description of each of these methodologies.

Integrated Publishing Toolkit (IPT)

The IPT is a GBIF tool for publishing occurrence and taxonomic data, as well as resource metadata, using Darwin Core Archives. The process for installing the IPT on a web server is described in detail in the [IPT User Manual](#)¹⁰.

The IPT contains a user-friendly interface that makes authoring metadata easy. Once a user has inputted and saved the minimum required metadata, they can return to it at any time to add to or modify the metadata.

In total, there are 12 different forms in the IPT that logically organise metadata entry. These are:

1. Basic Metadata
2. Geographic Coverage
3. Taxonomic Coverage
4. Temporal Coverage
5. Other Keywords
6. Associated Parties
7. Project Data

⁷ <http://code.google.com/p/gbif-providertoolkit/>

⁸ <http://tools.gbif.org/spreadsheet-processor/>

⁹ <http://tools.gbif.org/eml-gbif-sample.xml>

¹⁰ http://code.google.com/p/gbif-providertoolkit/downloads/detail?name=GBIF_IPT_User_Manual_1.0.pdf

8. Sampling Methods
9. Citations
10. Collection Data
11. Physical Data
12. Additional Metadata

The [IPT User Manual](#)¹¹ goes through each form and its respective fields in some depth. Occasionally, the form will provide help dialogs to aid the user in understanding what an element means (Figure 1).

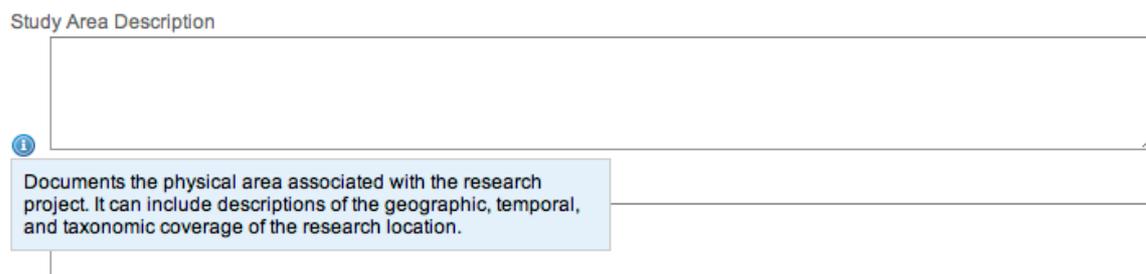


Figure 1 - Screenshot of a help dialog for the term “Study Area Description”

To ensure suitable data are entered, the fields are validated and informative messages displayed back to the user to assist them in filling out the forms (Figure 2).



Figure 2 - Screenshot of the field validation message displayed when an email field is submitted with an irregular email address.

For further reference, the [GBIF Metadata Profile Reference Guide](#)¹² includes a description of each element with an accompanying example.

Once installed on a web server, the IPT conveniently publishes the metadata document making it available on a publicly accessible URL and registering the URL with the GBIF

¹¹ http://links.gbif.org/ipt_user_manual, for metadata fields, please go to [http://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes - Metadata](http://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes_-_Metadata)

¹² http://links.gbif.org/gbif_metadata_profile_guide_en_v1

Registry. The IPT ensures that it is validated against the GBIF Metadata Profile so the user does not have to worry about validation.

If at any time the metadata are modified, the user only has to update the document and click the “Publish” button on the Manage Resource page (Figure 3).



Figure 3 - Screenshot of the Published Release section of the Manage Resource page of the IPT.

GBIF Darwin Core Spreadsheet Templates

The GBIF Darwin Core Spreadsheet Templates are simple tools to assist in the publication of either occurrence or taxonomic data through the GBIF network. Once data have been entered into the templates, the [GBIF Spreadsheet Processor](#)¹³ can be used to generate a Darwin Core Archive file containing both the dataset and its associated metadata. The Darwin Core Archive will then need to be made publically accessible on a web server and the URL registered with GBIF. This process is described in [Publishing and Registering Data with GBIF](#)¹⁴.

Companion [User Guides](#)¹⁵ for the GBIF Darwin Core Spreadsheet Templates (either the [Species & Vernacular Names Spreadsheet Template](#)¹⁶, the [Species Checklist Spreadsheet Template](#)¹⁷ or the [Occurrence Spreadsheet Template](#)¹⁸) provide thorough assistance on how to author metadata.

¹³ http://tools.gbif.org/spreadsheet_processor

¹⁴ http://links.gbif.org/dwc-a_publishing_guide_en_v1

¹⁵ <http://links.gbif.org/dwc-a-spreadsheet-processor-guide>

¹⁶ http://tools.gbif.org/spreadsheet-processor/templates/checklist/checklist-3_v0.9.xls

¹⁷ http://tools.gbif.org/spreadsheet-processor/templates/checklist/checklist-1_v0.9.xls

¹⁸ http://tools.gbif.org/spreadsheet-processor/templates/occurrence/occurrence-1_v0.9.xls

Basic Metadata		<i>Required fields are listed with an asterisk (*)</i>
Universally Unique Identifier (UUID)		b7998a69-1cb5-4cfe-b912-aaa066ba09a5
* Title		The Catalogue of Life in Taiwan
* Publication Date (MM/DD/YYYY)		12/31/11

Figure 4 - Screenshot of part of the Basic Metadata section of the Dataset Description page in Spreadsheet mode

There is hover-over help for fields that have a red triangle at the top right (Figure 5).

The spreadsheet templates make use of controlled lists (vocabularies) for some fields, such as country and basis of record. In addition, the required fields are all clearly marked with asterisks (Figure 5).

People and Organizations		
<i>Note: select agent roles from pulldown list (in rows 19-30)</i>		
* Resource Creator	First Name	Last Name
* Metadata Author	REQUIRED - The creator	
Author	is the person or	
Author	organisation	
Author	responsible for the	
	data in the dataset	
	associated with this	

Figure 5 - Screenshot of part of the People and Organisations section of the Dataset Description page showing the hover-over help in Spreadsheet mode

Where there is doubt about what a field means, the [GBIF Metadata Profile Reference Guide](http://links.gbif.org/gbif_metadata_profile_guide_en_v1)¹⁹ can be used to look up the description of its corresponding element with an accompanying example.

Once the metadata have been filled out, they need to be converted into a metadata document conforming to the GMP. To do so, the spreadsheet is submitted to the GBIF Spreadsheet Processor that accepts the spreadsheet file, validates it, and transforms it into a zipped Darwin-Core Archive that contains the metadata document. Note: other data-related sheets found in the Spreadsheet Template do not have to be completed and can be left with the default data. Inside the returned archive is an eml.xml file - this

¹⁹ http://links.gbif.org/gbif_metadata_profile_guide_en_v1

represents the metadata document. Other files packaged inside the archive can be ignored and correspond to the data-related sheets. For help filling in the data-related sheets could refer to the [Darwin Core Archive How-to](#)²⁰ document or the companion [User Guides for the Spreadsheet Templates](#).²¹

²⁰ http://links.gbif.org/gbif_dwc-a_how_to_guide_en_v1

²¹ http://links.gbif.org/dwca_spreadsheet_processor_guide

Modify an Existing Sample Document

Another quick and easy way to author a metadata document conforming to the GMP is to take the [reference sample²²](#) document and edit it. Refer to the following list to ensure it is completed properly:

1. Fill in all required elements (Refer to the Appendix: Required Elements).
2. Remove example placeholder information from all other fields that do not pertain to the data resource. Be careful to not remove fields (columns) as this could break the XML validity of the document. Nor should extra elements be added, unless it is to the “additionalMetadata” section.
3. Set the “packageId” attribute inside the <eml:eml>. Remember, the packageId should be any globally unique ID fixed for that document. Whenever the document changes, it must be assigned a new packageId. For example: packageId='619a4b95-1a82-4006-be6a-7dbe3c9b33c5/eml-1.xml' for the 1st version of the document, packageId='619a4b95-1a82-4006-be6a-7dbe3c9b33c5/eml-2.xml' for the 2nd version, and so on.

To ensure that the new document is valid, both as an XML document and in conforming to the GMP, it is essential to validate the new document against the GML schema. To do so, use the online validation service provided by GBIF: <http://tools.gbif.org/dwca-validator/eml.do>. For help with debugging the XML document, please refer to the Appendix: Best practices guide to debugging/validating XML.

²² <http://tools.gbif.org/eml-gbif-sample.xml>

Appendix

Required Elements

It is recommended that as many elements are completed as possible to ensure that the metadata are as descriptive and complete as possible.

Nevertheless, the GMP has a minimum set of five mandatory elements. These mandatory elements are designed to allow a prospective end-user of the data to discover the resource's name, description, contact person, and rights management information.

These elements are:

1. title
2. metadataProvider, creator, and contact with as much of the following as possible:
 - a. givenName - part of individualName (Required)
 - b. surname - part of individualName (Required)
 - c. organizationName
 - d. positionName
 - e. role
 - f. deliveryPoint - part of address
 - g. city - part of address
 - h. administrativeArea - part of address
 - i. postalCode - part of address
 - j. country - part of address
 - k. phone
 - l. electronicMailAddress (Required)
 - m. onlineUrl
3. abstract
4. pubDate
5. language (of metadata document)

For each element's detailed description, you can refer to the [GBIF Metadata Profile Reference Guide](#)²³

²³ http://links.gbif.org/gbif_metadata_profile_guide_en_v1

Best practices guide to debugging/validating XML

- Q: I receive the following error: “Eml validation error org.xml.sax.SAXParseException: The processing instruction target matching “[xX][mM][lL]” is not allowed.” - what does this mean?
- A: Ensure there is no whitespace before `<?xml version="1.0" encoding="utf-8" ?>`
- Q: I receive something similar to the following error: “org.xml.sax.SAXParseException: An invalid XML character (Unicode: 0x1c) was found in the element content of the document - what does it mean?”
- A: XML has a special set of characters that cannot be used in XML strings. These need to be replaced with their escaped equivalent:
 1. `&` should be replaced by `&`;
 2. `<` should be replaced by `<`;
 3. `>` should be replaced by `>`;
 4. `"` should be replaced by `"`;
 5. `'` should be replaced by `'`;