World Flora Online

General Guidelines for Data Contributors

Prepared by the World Flora Online Taxonomy and Technology Working Groups

Version 2.01 – DRAFT November 26, 2017

Updated by Mark Watson Feb 2018

Updated by William Ulate Apr 2018

Version 2.02 – Updated by Technical Working Group June 2018

Version 2.03 – Updated by WFO Council and Chuck Miller June 2018

Version 2.05 – Updated by Walter Berendsohn, Mark Watson and Marc Sosef July 5, 2018

Version 2.06 – Updated by Colin Pendry and Walter Berendsohn up to November 18, 2018

Version 2.07 – Updated by Chuck Miller and the Technical Working Group, Istanbul, April 26, 2019

Version history: Based on the first part of the "Guidelines for Backbone Data Contributors" v. 1.0. distributed by W. Berendsohn March 31, 2017. Changes were discussed by the WFO's Taxonomic Working Group in their meeting in Paris and documented by Marc Sosef. The technical part of the original document was separated as the "Guidelines for Backbone Data Contributors" v. 2.0.

NB: The following documents relate to Data Contributions:
1. WFO General Guidelines for Data Contributors (this document)
2. WFO Guidelines for Taxonomic Backbone Contributors
3. WFO Guidelines for Content Contributors

Table of Contents

1. Definition of Terms Used in the WFO Guidelines Documents

CCO waiver - used to opt out of copyright and database protection, and the exclusive rights automatically granted to creators. See https://creativecommons.org/share-your-work/public-domain/cc0

Content – Data relating to the description of plants and their properties, including text descriptions, images, altitude ranges, distribution, ecology, conservation assessments, etc. All *Content* is linked to names in the *Taxonomic Backbone*, and displayed under the name currently accepted by WFO[1].

Content Provider – A *Data Provider,* being an individual or group of monographic or floristic experts providing *Content* data to WFO. Note that nomenclatural and classification data from *Content Providers* are not included in the *WFO Taxonomic Backbone*[1].

Coordinating Taxonomist(s) - Person(s) designated by the *WFO Council* to administer and manage *Data Providers*, and liaise with the *WFO Gatekeeper* to upload data into the WFO system.

Darwin Core (DwC) - A body of internationally agreed standards, including a glossary of terms, intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries. See http://rs.tdwg.org/dwc

Darwin Core Archive (DwCA) - A biodiversity informatics data standard that makes use of the Darwin Core terms to produce a single, self-contained dataset for species occurrence or checklist data. Essentially it is ZIP file containing a set of tables (text files, usually CSV or Tab-separated) with a simple descriptor file (meta.xml) to inform others how your files are organized. See http://tools.gbif.org/dwca-assistant and the descriptions in the WFO Guidelines for Contributors to the *Taxonomic Backbone* and *Content*.

Data Provider - The administrative unit used by the WFO system to assign accreditation to those who provide data. A *Data Provider* maybe an individual, a team, a consortium, an institution, etc. When an individual or group contributes data on several distinct taxonomic groups (generally families), they will act as several distinct *Data Providers* within the WFO system, e.g. Caryophyllales Group providing data to the *Taxonomic Backbone* as separate *Data Providers* for Caryophyllaceae, Nepenthaceae, etc. (as the *Taxonomic Backbone* is sliced primarily at the rank of family).

Taxonomic Expert Network (TEN) – A *Data Provider* authorised by the WFO to take responsibility for providing updated information for one or more taxonomic groups, usually families, in the WFO *Taxonomic Backbone*.

WFO Council – The governing body of the World Flora Online Consortium.

WFO Gatekeeper – Person(s) designated by the *WFO Council* to technically administer and manage WFO data, and liaise with accredited *Data Providers* to upload their data into the WFO system.

WFO-ID – Unique identifier originated by WFO for all scientific names of plants, composed of 'WFO-' and 10 digits. This is used to unambiguously link data within WFO that is associated with a scientific name.

WFO Taxonomic Backbone – The updatable list of scientific names and their classification that contains all Effectively Published plant names and differentiates between accepted names and synonyms. *Content* data are displayed according to the WFO consensus classification in the *Taxonomic Backbone*. Updates to the *Taxonomic Backbone* are made by expert networks (*TENs*) who are assigned taxonomic groups (usually families).

WFO Taxonomic Working Group – The committee entrusted by the *WFO Council* with the oversight and organisation of the taxonomic contributions to the WFO.

WFO Technology Working Group – The committee entrusted by the *WFO Council* with the oversight and organisation of the technical setup of the WFO.

---

[1] Content that has been provided under a name considered a synonym in the WFO Taxonomic Backbone will currently be displayed on the synonym's name page in the WFO portal.

2. Introduction

The World Flora Online (WFO) is a database and Internet portal which brings together authoritative information on all known species of plants: including bryophytes, ferns and their allies, gymnosperms, and flowering plants (www.worldfloraonline.org). The WFO was created in response to the Global Strategy for Plant Conservation 2020 Target 1 (an online Flora of all known plants), and aims to primarily be a resource for conservationists, especially those working towards GSPC Targets and those involved in other CDB activities (CITES, Ramsar, etc.).

WFO will include information on all plant names, their classification and history of publication, and descriptions, distributions, conservation status and images. Biodiversity data from published floristic, monographic and other sources will be included within WFO and presented using a single, consensus classification. This classification of accepted names and their synonyms (the *Taxonomic Backbone*) is managed and updated by networks of taxonomic experts to reflect current global knowledge of plant species and their relationships.

WFO will:
- be a free, open access, online web-based resource, enabling access to floristic data on a global scale;
- be a global overview of the diversity of plant species on the planet;
- be comprehensive and authoritative, and include vascular plants and bryophytes;
- be a synoptic Flora with a defined data set on all the world's plant species;
- publish static versions of a global plant checklist;
- largely contain pre-existing data from national and regional Floras, and monographs.

The global community of plant systematists are invited to participate in WFO by either contributing to the curation of data in the *Taxonomic Backbone* and/or providing *Content* information such as descriptions, distributions, etc. to be displayed on taxon pages.[2]

This document describes the general process of contributing data to the World Flora Online and the way that information is handled in the WFO portal and databases. Technical details such as the format of such data contributions are set out in separate documents, the "WFO Guidelines to Taxonomic Backbone Contributors" and the "WFO Guidelines to Content Providers". Terms italicised in the text are defined in the glossary given above.

3. General Process for Data Contribution

WFO uses a derivative of the eMonocot (www.emonocot.org) software system to store and display information in the WFO portal. Some of the data, as defined in the contributor guidelines, is stored, but not currently displayed in the portal for technical reasons. These will be displayed in future versions of the WFO portal as resources permit.

Data Providers will be accredited as the source of the dataset. Individual contributors can be acknowledged under the general information provided by the *Data Provider* or by using the taxon-Remarks field to provide a bibliographic reference for the correct citation of the authority for the data record.

Data Providers, of both *Content* and *Taxonomic Backbone* information, are expected to contribute data electronically using a *Darwin Core Archive* file (see below). Those who do not have their data in this format, or are not able to easily convert it to such, are requested to contact the *WFO Gatekeeper* for support.

---

[2] At present content is stored and displayed in WFO with the scientific plant name in the *WFO Taxonomic Backbone* matching the scientific plant name given by the content data provider.

4. Taxonomic Backbone

At the core of WFO is an updatable *Taxonomic Backbone* of scientific names and their classification that contains all Effectively Published plant names (as defined in the International Code of Nomenclature for Algae, Fungi and Plants), and differentiates between accepted names and synonyms. Data in the *Taxonomic Backbone* will be in the Public Domain, under a *CC0 waiver*.

The first version of the *WFO Taxonomic Backbone* was created based on *The Plant List* v1.1 as this was the most comprehensive single information resource covering all plants. *The Plant List* is not perfect - ca. 20% of names are unresolved and not all the taxonomic decisions within it are derived from a peer reviewed, curated, authoritative source. Also, many infraspecific taxa are not covered by the *The Plant List.* Data within the *WFO Taxonomic Backbone* will be actively curated by taxonomic experts and so become increasingly authoritative over time.

For each taxonomic group (typically a Family) within the *WFO Taxonomic Backbone*, the data are managed and updated by a *Taxonomic Expert Network* (*TEN*) authorised by the *WFO Council.* A *TEN* has responsibility delegated by the WFO Council for providing reviewed nomenclatural and classification information for <u>all</u> Effectively Published names within their taxonomic group within a given timeframe. This responsibility may be revoked if the conditions and timeframes are not met, and the taxonomic group be re-assigned.

Taxonomists working in networks with a regional focus (e.g. Flora projects) are also encouraged to participate in the curation of data in the *WFO Taxonomic Backbone.* This could be done by contributing classification and nomenclature data to an established *TEN*, or by applying to become a *TEN* when one does not exist for important families within their geographical area. Such a *Data Provider* can for example apply to become the *TEN* for families which are (sub)endemic to their region. Likewise, for families where there is substantial coverage in their area they can apply to be a 'caretaker' *TEN* with the remit to curate data relevant to their area, and the expectation that this will be handed over to a regular *TEN* in the future when one is appointed.

The *WFO Taxonomic Working Group* manages the taxonomic hierarchy for Families and higher ranks, (currently based on APG III for angiosperms and on equivalent, internationally supported systems for non-angiosperm plants)[3]. In exceptional circumstances an alternative classification can be adopted by WFO if proposed by a *TEN* and agreed by the *WFO Taxonomic Working Group.*

5. Authorization and Registration of Contributors

Contributors to the World Flora Online must be authorised by the *WFO Taxonomic Working Group* as a *Data Provider*. Send an email to taxwg@worldfloraonline.org to request authorization to become a *Data Provider* to WFO. Authorizations may be reviewed by the *WFO Council*. A *Data Provider* may contribute data to the *Taxonomic Backbone*, or *Content*, or both. After authorization, the *Taxonomic Working Group* will notify the requestor, the *Coordinating Taxonomist* and the *WFO Gatekeeper* of the authorization.

The next step will be for the *Data Provider* to register with the *WFO Gatekeeper*. To register, the *Data Provider* must supply the information in the table below as a text document to the *WFO Gatekeeper* before other data are uploaded [by email](mailto:) – this information can be modified later by contacting the *WFO Gatekeeper*.

The text file should be in the form of a two-column table with the term in the first column and the value in the second. Required fields are in **BOLD.** See below for an example.

---

[3] *WFO Taxonomic Working Group* intends to adopt APG IV with modifications as required.

Fields are limited in length to 255 characters (except the dataset description [field 'abstract'], which is practically unlimited). If you want to have longer text, use a full URL (http://...) to refer to an external page containing the information.

After registration, refer to the Data Ingestion Process section of either the Guidelines for Taxonomic Backbone Contributors or the Guidelines for Content Data Contributors for the next steps to contribute data to the WFO.

WFO *Data Provider* Registration File Content

Data needed for both *Taxonomic Backbone* and *Content Providers*

(Fields in **bold** are mandatory)

Field[4]	Simplified Description
title [A]	Name of the data provider
emailCommentsTo [B]	Email used to send comment alerts
contactEmail [B]	Email of the contact for dataset
onlineUrl [A]	Weblink to the data provider's homepage
logoImageUrl [B]	Weblink to contributor's logo graphic
abstract [A]	Description of the dataset
citation [A]	Text showing how the dataset should be cited, shown after 'Reference' in the portal.
createdBy [B]	Statement giving accreditation to those who created the dataset.
publisherName [B]	Text
publisherEmail [B]	Email address
keywords [B]	A list of terms, separated by commas, describing the dataset

EXAMPLE DATA PROVIDER FILE

Field	Value
title	GrassBase - The Online World Grass Flora
emailCommentsTo	enquieries@about.e-monocot.org
contact email	m.vorontsova@kew.org
onlineUrl	http://www.kew.org/data/grasses-db
logoImageUrl	[NULL]
abstract	GrassBase will ultimately provide an integrated, online view of the World Grass Species databases which have historically been held in two separate downloadable databases. The first step towards this integration has been the generation of nearly 11,000 species descriptions from the DELTA format that they're encoded in. In addition to this the synonymy/nomenclature database now contains links to these species descriptions integrated with searches for the accepted name and synonyms for just over 60,000 grass names.
citation	Clayton, W.D., Vorontsova, M.S., Harman, K.T. and Williamson, H. (2006 onwards). GrassBase - The Online World Grass Flora. http://www.kew.org/data/grasses-db.html [accessed 24 Jan 2012; 14:30 GMT]
createdBy	Maria Vorontsova
publisherName	Science Applications Support
publisherEmail	apps-support@kew.org

---

[4] Field titles derived from (A) EML (Ecological Markup Language) or (B) eMonocot

Field	Value
keywords	e-Taxonomy, GrassBase, Royal Botanic Gardens Kew, Poaceae, Graminae, Grasses

6. Creating a Darwin Core Archive File

Data for the WFO *Taxonomic Backbone* or *Content* are ingested into the WFO system as a file in *Darwin Core Archive (DwCA)* format. A *Darwin Core Archive* is a ZIP-compressed file containing several text files and an XML metadata file. The text files include elements/fields based on the *Darwin Core (DwC,* http://rs.tdwg.org/dwc/) standard, as defined by the Biodiversity Information Standards (TDWG) organization. *Darwin Core* defines many terms (data types / column headings). The WFO portal only recognizes those terms outlined in the table specifications in the Guidelines for Backbone and Content Contributors. There are also some additional data elements specific to the WFO.

The *DwCA* (ZIP compressed) file contains one or more comma-delimited (e.g. CSV) text files, and an XML file. The text files contain the data in tabular form, the XML "metafile" defines the columns in these tables and holds additional technical information (e.g. the data type of the columns). The tables in a *Darwin Core Archive* have a star-like structure: a single central record (in WFO, a name in the Classification or Taxon table) can have multiple "extensions", each one with records of a different type linked to it — like descriptions, references, distributions, etc..

Darwin Core Archives can be produced by the GBIF Integrated Publishing Tool (IPT) system (https://www.gbif.org/ipt)Scratchpads, BioCASe, Botalista, the EDIT Platform for Cybertaxonomy, or manually. GBIF provides a tool called Darwin Core Archive Assistant (http://tools.gbif.org/dwca-assistant/) to help creating the XML "metafile", which serves to describe the physical format and arrangement of published files within a *Darwin Core Archive* file. For more information on the *DwCA* format, please consult the Darwin Core Archive Assistant Help tab and its User Guide[5].

For instructions on the specific data values, formatting and restrictions that apply to WFO Darwin Core Archives please refer to either the Guidelines for *Taxonomic Backbone* Contributors or the Guidelines for *Content* Data Contributors. Please note that the References.txt table for backbone contributions is not a formal DwC extension and thus does not need to be specified in the respective *DwCA* metadata.

7. Testing the DwCA File

The DwCA file can be tested with the GBIF DwC-A Validator service under https://www.gbif.org/tools/data-validator. Another tool provided by GBIF called the Darwin Core Archive Validator (http://tools.gbif.org/dwca-validator/) also serves to check for some of the most common problems when generating a DwCA file. However, please note that the non-DwC elements in the DwCA will be marked as errors, as well as the link (parentNameUsageID) of the family to the higher taxon ID (either because the ID is missing or because the taxon referred to - e.g. an Order name- is not present in your file). Also, the test will let some terms pass that are not accepted by WFO (e.g. "nom. nov.", "nom. nud." etc. as values for nomenclatural status) – please refer to the respective lists of permitted values in the Guides for Backbone and Content providers. Please also note the Typical Issues section below.

---

[5] GBIF (2010). Darwin Core Archive Assistant, version 1.1, released on 1 April 2011, (contributed by Remsen, D, Sood, R.), Copenhagen: Global Biodiversity Information Facility, 23 pp. Available at: https://www.gbif.org/document/80637/

8. Typical Issues with Data Contributions

8.1. When registering a Data Provider:

1. Keywords must be separated by a comma.
2. Creation date format must be YYYY-MM-DD or YYYYMMDD

8.2. When creating the DwCA file:

1. Verify that the encoding of each file within the DwCA file is correctly defined in the meta.xml file (for example: *encoding="UTF-8"* or *encoding="ANSI"*) and that all data conforms to such encoding.
2. Verify that all text files (tables) included in a Darwin Core Archive (ZIP file) include the *WFO-ID* of the name in the taxonID or '0' field (with the exeption of the References.txt table in a *Taxonomic Backbone* data contribution).
3. Verify that for each one of the files included in the DwCA file, the first line is ignored if it contains headers for the fields (for example: *ignoreHeaderLines="1"*). If no headers are used in a file, indicate this in the corresponding entry of the meta.xml file (with *ignoreHeaderLines="0"*).
4. Make sure all lines in each one of the files in the DwCA file are appropriately terminated (for example: *linesTerminatedBy="\r\n"*). The line endings in the file may be converted when saved in certain editing programs (like Microsoft Excel) or a computer with particular Operating Systems (like Linux or UNIX based OS).
5. Make sure all fields in all lines of the files in the DwCA file are properly terminated with the delimiter defined in meta.xml (for example: *fieldsTerminatedBy="\t"* or *fieldsTerminatedBy=","*).
6. If CSV files are created, verify fields are quote-delimited when commas are contained in the field and that the correct number of commas are included, one for each column in the file and that the corresponding definition in the meta.xml file is the correct one for the field separation (with *fieldsTerminatedBy=","*) and the field delimitation (it could be double quotes: *fieldsEnclosedBy="""* or single quotes: *fieldsEnclosedBy="'"*). One of such fields is the *Source* field in the *description* file which could contain several references separated by commas (that is, any comma inside this field will produce a new record in the references table), so it's recommended to define the field as quote-delimited.
7. Check that the type of record defined in the meta.xml file corresponds to the content of the files (for example: *rowType="http://rs.tdwg.org/dwc/terms/Taxon"* or *rowType="http://rs.gbif.org/terms/1.0/Reference"*).
8. Verify that the name and extension of each file included in the DwCA file corresponds to the appropriate entry in the meta.xml (for example: *<files><location>taxon.txt</location></files>*).
9. Do not include a time into date fields, use YYYY-MM-DD or YYYYMMDD only.
10. Use only the permitted values for the nomenclatural status and taxonomic status.
11. Pay attention to the size of the text fields provided, undetermined errors might occur when loading the content if a string longer than expected is provided in the file (for example, the namePublishedIn field must have fewer than 255 characters).
12. Separate description source field references by commas (any comma in the field will produce a new record in references). Make sure there's no space left before or after the comma as it will be considered as part of the reference.
13. Make sure no records are created when no actual *Content* data are provided. For example: do not include a record in description file with only the ID when the text is empty, as this will create an empty record for the description and an empty text block with reference might also be created in the portal.