DATA 4 NATURE

SHARE TO PROTECT

Practical recommendations guide for publishing primary biodiversity data

AN INITIATIVE TO ENCOURAGE THE SHARING OF BIODIVERSITY DATA GENERATED BY DEVELOPMENT PROJECTS, VIA GBIF.

1 Preamble

Starting in 2021, the Agence Française de Développement (the AFD) is requiring project owners and service-providers to publish any primary biodiversity data generated within the framework of funded projects, through the global network of the GBIF (Global Biodiversity Information Facility). Primary biodiversity data refers to the observation of fauna or flora situated in time and space. This type of data, also known as taxon occurrence data, provides evidence that a species (or some other taxon) is present in a particular location on a given date.

In addition to the requirements relating to the publication of the biodiversity data set out in the funding agreements and terms of reference, this document provides practical recommendations for consultants and other stakeholders in charge of the publication of the data.

The application of AFD contracting clause relating to the publication of the biodiversity data requires the involvement of at least three or four different players:

- The AFD ;
- Its client, usually a project owner ;
- One or more consultancy firms specializing in the collection and analysis of biodiversity data ;
- The GBIF.

As the following figure shows, each of the players has a specific role in the process of publication of the data. The practical implementation of data publishing is generally delegated to the consultant(s) (or other players) in charge of data collection. Those consultants must, as a result, carry out several successive stages of work before proceeding to publish the data as such.

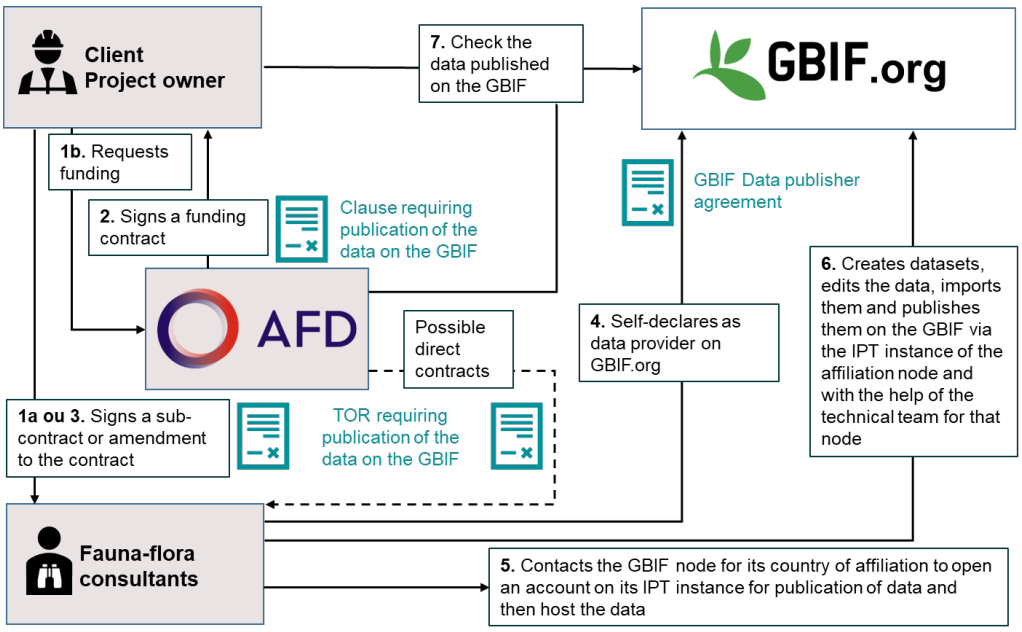

Figure 1: Diagram of the process for publication of primary biodiversity data generated with the help of AFD funding.

2 How to become a GBIF data publisher

2.1 Acceptance of the GBIF's Data Publisher Agreement

If you are not yet a GBIF data publisher, the first step towards publishing data on the GBIF network consists of registering yourself as a data provider. To do so, you must log onto the GBIF web portal https://www.gbif.org/become-a-publisher in order to accept the terms and conditions of the data publisher contract, then complete a short form describing your organisation, the type of data that you will provide and how the data will be made available.

Only organisations can be endorsed as GBIF data publishers. Endorsement as an individual is not possible.

Table 1 : Recommendations for completing the request form for registration as a GBIF data provider.

Headings	Recommendations
Information about the organisation	For your organisation's contact details, provide those of the head office, as appropriate. Similarly, show your organisation's head office on the location map.
Endorsing node	Keep the default setting active unless you intend to supply marine data exclusively.
GBIF projects	Keep the setting "No" proposed as the default.
Contacts	Provide, at the very least, the contact information for your organisation, usually the person in charge of managing biodiversity data.
What and how	Check, at the very least, "Occurrence-only data". However, you can select other types of data if you wish. Then provide a short description of all the types of data that you are likely to publish. If you do not have a computer server through which you will give direct access to your data to the GBIF, we recommend that you ask for help in publishing your data. The staff of the GBIF node to which you belong will contact you to define the methods for publishing your data.

As soon as you register the form, it is submitted to the GBIF node to which you depend. That node will assess your request before accepting it.

If your organisation is based in a country that is not a member of GBIF, endorsement will be made by the GBIF's Nodes Steering Group.

*Further information on the endorsement of an organisation as a GBIF data publisher is available on the following web page: **https://www.gbif.org/fr/endorsement-guidelines***

As soon as your request for registration as a GBIF data publisher is accepted, a dedicated web page will automatically be created for you on the GBIF.org site with the information that you have supplied in the form. From that page, you will also have access to the datasets that you have published and to metrics for those data.

2.2 Opening an IPT instance for publishing data

The publication of your data on the GBIF platform requires:

- A computer server which you will use both to store your primary biodiversity data and to give access to that thanks to the GBIF website ;
- Data publication software, such as the public and free "IPT" (Integrated Publishing Toolkit) software made available by the GBIF.

If you do not have a computer server or if it is not possible for you to share your data from your own servers, the GBIF node to which you depend is able to host your data and open an IPT instance for publishing your data on their own servers. In that case, your GBIF node will give you the URL of the IPT instance to use, together with your user log-in. Where you are in a country with no GBIF node, the GBIF Secretariat can provide you with alternative data hosting solutions.

*You can find the contact details for the GBIF node to which you belong on the following address: **https://www.gbif.org/the-gbif-network***
*If the country where your organisation is based does not have a GBIF node, contact the GBIF Secretariat in Copenhagen at **helpdesk@gbif.org** to provide alternative solutions.*

In every case, we recommend you to reach out to the GBIF node to which you belong in order to set up and configure your data publication tools.

3 Recommendations for data collection and formatting

3.1 Data collection

The publication of primary biodiversity data on the GBIF does not require the collection of additional information in the field. It only involves formatting your data in accordance with data standards.

It is therefore recommended to anticipate that formatting as soon as data is collected in the field in order to facilitate the publication of the data afterwards. That anticipation of the publication of the data can result in a number of good practices which will also contribute to improving the quality of your data:

- Use of data-entry sheets for observations in the field which enable, at the very least, specification of the type of observation, the scientific name of the taxon, its location and the date of observation (an example of a field datasheet is attached) ;
- Mapping of the most interesting surveys and observations with the help of a GPS ;
- Centralisation and digitisation of the observations in a table structured in a similar way as the standard defined for the publication of the species occurrences on the GBIF (see Section 3.2 below in this document).

Supplying information on the location of the observations requires a basic knowledge of cartography. It is, further, strongly recommended to use a geographic information system and/or GPS position management software to complete the geographic coordinates of the observations.

Mapping the observations with the help of a GPS is not essential. Each observation must nevertheless be associated with a geographic position and be published with the same location accuracy as that collected in the field, except for the species considered as sensitive (See Section 3.3 on page 6 of this document).

For observations that are not precisely georeferenced, the geographic position may relate to the coordinates of a point located at the centre of the study area. In that case, the radius of the smallest circle encompassing the study area as from that central point will be evaluated and incorporated into the data (field *coordinateUncertainty* to be completed, see Table 2 on page 5).

3.2 Formatting data in accordance with the GBIF standards

The terms of reference for procedures in the appraisal phase and project/programme funding agreements drawn up by the AFD imply that primary biodiversity data collected are published, at the very least, under the form of **species occurrences** to provide sufficient information on the location of the observations in time and space.

Those species occurrences must be aggregated and published by datasets for which metadata will be provided, i.e., additional information detailing the conditions of data acquisition and use. Section 4.4 on page 8 gives details of the metadata to enter.

In the absence of additional specific requirements by the AFD for the publication of the data, **each project funded will result in, at least, the publication of one dataset containing all the observations made in the field**.

If you wish to do so, you are free to publish data in several datasets in order to supply more specific details on the inventory protocols used for each data record.

In order to prepare for the publication of your dataset, it is necessary to list your observations in a spreadsheet completed with the detailed information set out in the table below. You can use the model Excel file made available by the GBIF here: https://github.com/gbif/ipt/wiki/occurrenceData. You may need to add some fields manually to the given model. In that case, it is important to use the exact field names listed in the following table.

Table 2 : Description of "species occurrence" standard used by the GBIF and expected in the publication of biodiversity data acquired through AFD funding.

	Field name	Description
Fields always required	**occurrenceID**	Unique identifier for the taxon occurrence. You are free to define that unique identifier yourself. At the very least, it must be unique within the dataset. It may be a numerical or alphanumeric identifier (e.g. 2354 or Obs0015). Best practice is, however, to use a unique and permanent identifier of a UUID (Universally Unique Identifier) type or possibly a computer-generated URI (Uniform Resource Identifier).
	basisOfRecord	Nature of the observation that was the source of the occurrence. It may be, for example, a human observation or a recording made by a machine (e.g. ultrasound recorder, camera trap, Argos buoy, radar, etc.).
	scientificName	Full scientific name on the lowest possible taxonomic level, including the author of the description of the taxon and the date of publication of the name (e.g. *Canis lupus* Linnaeus, 1758). The taxon's scientific name may be looked up on the GBIF site through the following page https://www.gbif.org/species/search or using the online tool Species matching tool provided by the GBIF (see Part 3.4 on page 6 of this document for more information on that tool).
	eventDate	Date when, or period during which, the taxon was observed. The date must be formatted according to the ISO 8601 standard (e.g. for 18th November 2020, we must write 2020-11-18)
	decimalLatitude	Geographic coordinates of the occurrence in decimal degrees. E.g. decimalLatitude = 48.861037 and decimalLongitude = 2.335860.
	decimalLongitude	
	geodeticDatum	Geographic coordinate system used to indicate the latitude and the longitude of the occurrences. For information, the benchmark geographic coordinate system used by the GBIF is the WGS84 (World Geodetic System 1984).
Fields conditionally required	**coordinateUncertainty**	Specification of the location of the taxon occurrence in metres. In other words, this is the radius of the smallest circle containing the observation. **Condition:** the coordinates provided correspond to the centre of an area within which the occurrence was observed.
	dataGeneralizations	Description of a possible intentional degradation of the data (for example, a blurring of the geographic position). Enables the user to know if a more precise version of the published data exists. **Condition:** observation of a species considered to be sensitive.
Recommended fields	**individualCount**	The *individualCount* field enables the number of individuals observed for the taxon in question to be specified. It is also possible to quantify the taxon occurrence with other measurement systems, for example, abundance-dominance coefficients, a density, an abundance indicator, etc. In that case, the fields *organismQuantity* and *organismQuantityType* are to be used to show the quantity and the measurement unit used to quantify the taxon occurrence.
	organismQuantity	
	organismQuantityType	
	countryCode	Two-letter code showing the country in which the taxon was observed. That code must follow the standard ISO 3166-1-alpha-2
	taxonRank	Taxonomic rank of the scientific name given for the taxon occurrence in Latin or in English. (Examples: kingdom, phylum, class, order, family, genus, species)
	kingdom	Full scientific name, in Latin, of the reign to which the observed taxon belongs. The possible values for this field are therefore: Animalia, Archaea, Bacteria, Chromista, Fungi, Plantae, Protozoa or Viruses.
	informationWithheld	In the case of intentional degradation of the data, it is useful to complete the field *informationWithheld* in addition to the field *dataGeneralizations* in order to show that more precise information exists but that it has intentionally not been given.

More information on the fields associated with the description of the species occurrences may be accessed on the GBIF web portal: https://www.gbif.org/fr/data-quality-requirements-occurrences

The data in each data record may be entered in the language of your choice. The language used to describe the data and metadata must, however, be specified in the metadata of each dataset. That information is declared at Step 4 of the process for publication of the data (see Part 4.4 on page 8 of this document).

The GBIF also allows you to group your species occurrences together in sampling events. Your observation lists can then be associated with different surveys for which you can publish specific additional information. Although these groupings are not required by the AFD, they further increase the value and usability of your data records. More information on sampling events can be accessed on the GBIF site here: https://www.gbif.org/fr/sampling-event-data.

3.3 Degradation of the geographic location of sensitive data

In order not to compromise the survival of local populations of species threatened by intentional removal or destruction of individuals (poaching, for example), you may purposely degrade the accuracy of the location of the observations of species called sensitive.

The simplest method of degrading the location of observations consists of rounding up the recorded geographic coordinates. That rounding can be adapted depending on the sensitivity level of the species. Thus, for the most sensitive species, you could, for example, round up the geographic coordinates to the nearest whole numbers. For species that are a little less sensitive, you could round the coordinates to one or two digits after the decimal point.

Beyond the polar circles, blurring of the location of sensitive species by degradation of the geographic coordinates becomes less useful due to the fact that the meridians draw closer together.

Table 3 Relation between the number of decimals used for the geographic coordinates and the level of degradation of the location of the data at the equator

Number of digits after the decimal point to round the geographic coordinates	Value in decimal degrees	Equivalence in metres in the field / margin of error in the location of the observations
0	1.0	111.32 km
1	0.1	11.13 km
2	0.01	1.11 km

When you intentionally degrade the location accuracy of an observation, it is necessary to describe the nature of the degradation in the *dataGeneralizations* field. You can, for example, write in that field: “Rounding of the geographic coordinates to the nearest whole numbers”. In addition to the field *dataGeneralizations,* it is also useful to complete the field *informationWithheld* in order to show that more precise information about the location exists but that it has purposely not been given.

For more information on degradation of the location of sensitive species, you may refer to the best practice guide made available by the GBIF (CHAPMAN, 2020) which can be downloaded here: https://www.gbif.org/document/80512/guide-to-best-practices-for-generalising-sensitive-species-occurrence-data

3.4 Checking the validity of the datasets

For your datasets to be published correctly, the observations lists files that you import into the IPT data publication software must comply with the standards used by the GBIF.

In order to check that your Excel files are properly formatted, we recommend that you use successively the two following tools made available by the GBIF. These tools enable you to highlight any possible mistakes that you might have made in your files in order to correct them more easily before publishing them.

Table 4 : Principal tools for checking the validity of your datasets

Step	Tool name and link to the tool	Function	Remark
1	GBIF Species matching tool	Checking the validity of the taxon’s scientific name and seeking the taxon names accepted by the GBIF	Your Excel file must first have been converted into CSV format
2	GBIF Data validator	Overall checking of the compatibility of your datasets with the data standards used by the GBIF	Pre-version of the application

4 Data publishing on the GBIF

The publication of datasets on the GBIF web portal is done through the IPT web application (Integrated Publishing Toolkit). When you use an IPT instance made available by the GBIF node to which you belong, you access the application using an URL, the user account and the password attributed to you by your GBIF node.

When you use the IPT application, you have constant access to a user manual by clicking on the dedicated link at the foot of each screen.

In order to correctly use the IPT application, we recommend you consult the page https://www.gbif.org/fr/ipt and to watch the demonstration video made available by the GBIF. You can also request a presentation of the application or training from your GBIF node.

4.1 Step 1: Creation of a data record

Once you have checked and confirmed that your data comply with the GBIF standards, you are ready to publish them through the IPT application. All you have to do is to create a new resource of an "Occurrence" type or a "Sampling event" type from the IPT "Manage Resources" tab.

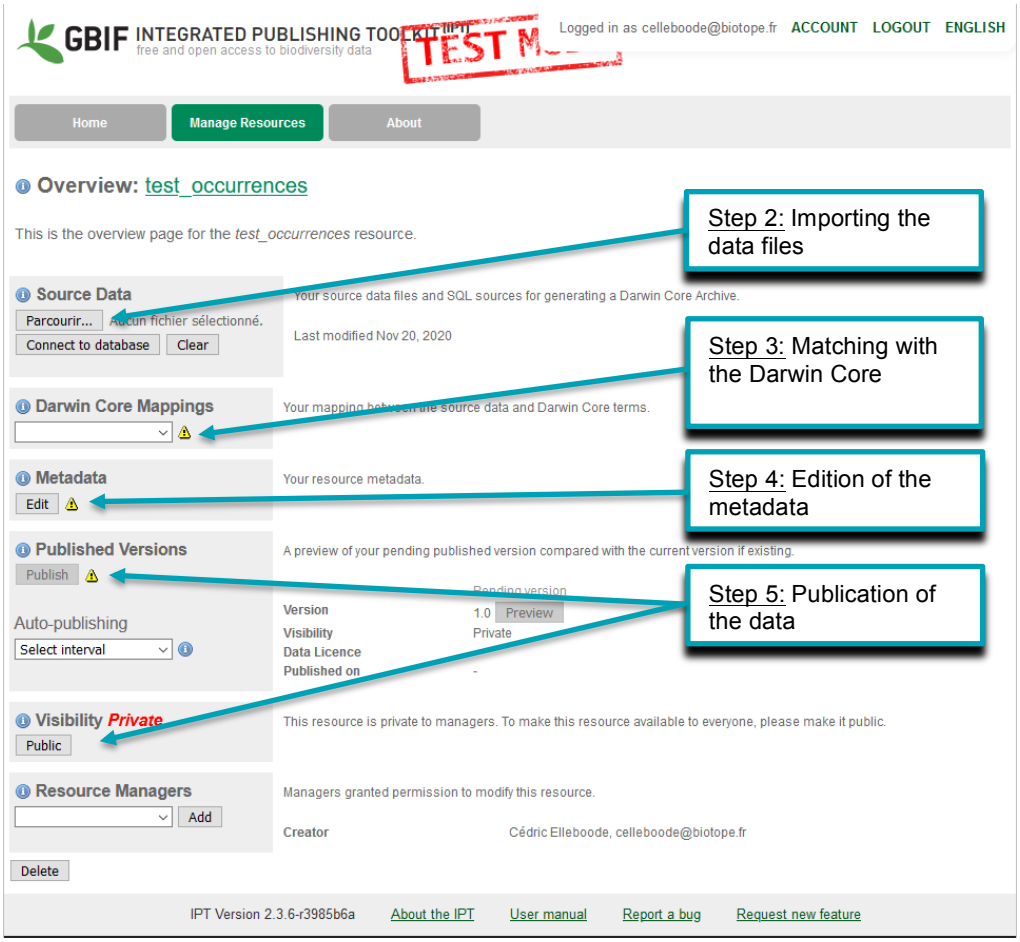

Figure 2 Graphic interface of the management page of a dataset in the IPT application

4.2 Step 2: Importing data in the dataset

When you have created a dataset on the IPT, you can import the Excel spreadsheet containing your data from the section "Source Data". When you import your data, an intermediate window opens so that you can edit the format of your source data if necessary. At that stage, we recommend you to preview the data that you have imported in order to check that the formats used have been correctly integrated.

To preview the imported data, click on the eye to the right of the button "Analyze" in the edition window for your data.

4.3 Step 3: Matching imported data with the Darwin Core standard

The Darwin Core is the format of the data file standard used by the GBIF. Thus, this step is intended to ensure that the imported data files are indeed compatible with the Darwin Core. If some fields of data are missing or are not automatically recognised, it is possible to add fields with pre-defined values for the dataset or to match some fields manually with the Darwin Core standard.

In order to simplify this stage of the work, we recommend you to make maximum use of the Excel file made available by the GBIF to format the list of your species occurrences and to respect the format of the data fields listed in Part 3.2 on page 4 of this document.

To learn more about the Darwin Core data standard used by the GBIF, you may consult the following web site: https://dwc.tdwg.org

4.4 Step 4: Editing the metadata

The metadata enable you to specifically describe the conditions of production and use of the data record. The importance of that information should not be overlooked because the usability of the data record depends heavily upon it. Entering certain metadata such as the name of the data record, its description and the contact details for the producer of the data is required by the GBIF and it is a prior condition for publication of the data.

Furthermore, the funding agreements and terms of reference of the AFD require certain metadata to be entered according to pre-defined rules.

This is the case, in particular, for the conditions of use of the data. Thus, in the section “Basic Metadata”, you are obliged to opt for one of the two least restrictive levels of rights for the field “Data Licence”, namely the “Public Domain (CC0)” licence or the “Creative Commons Attribution (CC-BY)” licence.

This is also the case for the information on the Project within the framework of which you have collected the data. In the section “Project Data”, it is essential to give the name of the project and also a short description of that, followed by the name of the project owner. In the field “Funding”, you are obliged to give the names of those funding the project, including the Agence Française de Développement (AFD).

4.5 Step 5: Publication of the data

By default, the data that you publish on the GBIF are private and accessible only by the persons to whom you have attributed the role of “Manager” of your datasets. Consequently, before publishing your datasets, we recommend you to make them accessible to the public by clicking on the button “Public” in the section “Visibility” of the IPT “Manage Resources” tab.

Once your dataset has “Public” status, you can go to the last step of the publication process by clicking on “Publish” in the section “Published version”.

The IPT application allows you to amend or complete the data record that you have already published, if necessary. In that case, it is useful to specify the amendments made at the time of publication of a new version of a dataset.

5 Further information

The GBIF can support you throughout the process described in this recommendations guide. In addition to the wealth of information supplied on the GBIF web portal (https://www.gbif.org/), it is possible for you to send your questions directly to private-sector@gbif.org.

The GBIF regularly organises training for data publishers and provides you with access to an e-learning platform and freely downloadable training materials. https://www.gbif.org/article/2lE7tH4dlcik1BnmnilPAc/training-and-e-learning

Appendix: example of field data-entry sheet facilitating publication of the data on the GBIF

Project name	
Name of survey	
Date of survey	*The date and, potentially, the time of the survey*
Location of survey	*Description, GPS position and/or cross-reference with a location map*
Inventory protocol	*Name or description of the protocol or, failing that, of the method used to gather the data*
Conditions of survey	*Important information for the interpretation of the survey (example: meteorological conditions for a bird inventory)*

Species observed	Number of individuals	Additional comments	GPS point identifier
		Additional information, especially for species threatened with extinction (example: reproduction sign)	*In particular for observations of threatened species*

CONTACTS

GBIF Secretariat in Copenhagen (Danemark)

Email : helpdesk@gbif.org

Phone : +45 35 32 14 70

Agence Française de Développement

Email : data4nature@afd.fr

Illustrator : Sébastein Pelon