



Pacific Invasive Species Battler Series



SHARE PACIFIC INVASIVE SPECIES DATA USING THE GLOBAL BIODIVERSITY INFORMATION FACILITY





SPREP Library Cataloguing-in-Publication Data

Share Pacific invasive species data using the Global Biodiversity Information Facility.
Apia, Samoa : SPREP, 2018.

28 p. 29 cm.

ISBN: 978-982-04-0785-5 (print)
978-982-04-0786-2 (ecopy)

1. Invasive species – Databases – Oceania.
2. Non-indigenous pests – Control – Data processing – Oceania.
3. Introduced organisms – Control – Data processing – Pacific Ocean.
4. Biological invasions – Information management – Oceania.
- I. Pacific Regional Environment Programme (SPREP).
- II. Global Biodiversity Information Facility (GBIF).
- III. Biodiversity Information for Development (BID).
- IV. European Union (EU).
- V. Title.
- VI. Series.

632.90995



As part of SPREP's commitment to the environment, this item is printed on recycled paper.

Secretariat of the Pacific Regional Environment Programme (SPREP)

PO Box 240
Apia, Samoa
sprep@sprep.org

www.sprep.org

Our vision: A resilient Pacific environment sustaining our livelihoods and natural heritage in harmony with our cultures

Copyright © Secretariat of the Pacific Regional Environment Programme (SPREP), 2016. Reproduction for educational or other non-commercial purposes is authorised without prior written permission from the copyright holder provided that the source is fully acknowledged. Reproduction of this publication for resale or other commercial purposes is prohibited without prior written consent of the copyright owner.

Cover photo: Stuart Chape

TABLE OF CONTENTS

Dear Invasive Species Battler	2
About This Guide	3
Why Share Species Data?	3
What is GBIF?	4
How do I use GBIF to find data?	5
Who owns the data?	8
What are people allowed to do with the data I share?	10
How do I publish data on GBIF?	11
Pathways to data publishing	12
How do I register as a data publisher on GBIF?	13
Georeferencing basics	14
How do I prepare data for publishing?	15
Preparing data using the Integrated Publishing Toolkit	16
Mapping to Darwin Core Terms	17
Constructing a unique identifier	19
Cleaning data	19
Republishing data	20
How do I plan for data publication?	22
Glossary of definitions	23
For More Information	24
Ongoing Support	24
Websites	24

Dear Invasive Species Battler,

We are a diverse bunch of people in the Pacific region, which spans about one third of the earth's surface and encompasses about half of the global sea surface. We have ~2,000 different languages and ~30,000 islands. The Pacific is so diverse that its ecosystems make up one of the world's biodiversity hotspots, with a large number of species found only in the Pacific and nowhere else. In fact, there are 2,189 single-country endemic species recorded to date. Of these species, 5.8 per cent are already extinct or exist only in captivity. A further 45 per cent are at risk of extinction. We face some of the highest extinction rates in the world.

The largest cause of extinction of single-country endemic species in the Pacific is the impact of invasive species. Invasives also severely impact our economies, ability to trade, sustainable development, health, ecosystem services, and the resilience of our ecosystems to respond to natural disasters.

Fortunately, we can do something about it.

Even in our diverse region, we share many things in common. We are island people, we are self-reliant, and we rely heavily on our environment to support our livelihoods. We also share many common invasive species issues as we are ultimately connected. Sharing what we learn regionally makes us and our families benefit economically, culturally, and in our daily lives.

The "Invasive Species Battler" series has been developed to share what we have learned about common invasive species issues in the region. They are not intended to cover each issue in depth but to provide information and case-studies that can assist you to make a decision about what to do next or where to go for further information.

The SPREP Invasive Species Programme aims to provide technical, institutional, and financial support to regional invasive species programmes in coordination with other regional bodies. We coordinate the Pacific Invasive Learning Network (PILN), a network for invasive species practitioners battling invasive species in Pacific countries and territories, and the Pacific Invasives Partnership (PIP), the umbrella regional coordinating body for agencies working on invasive species in more than one Pacific country.

For knowledge resources, outreach tools, and more information on SPREP, the Invasive Species Programme, PILN, and PIP, please visit the SPREP website: www.sprep.org

Thank you for your efforts,

SPREP Invasive Species Team



About This Guide

Data about tropical island species can be difficult to obtain and store safely, but these data are valuable to Pacific communities, environmental managers, and the global research community. A Pacific solution to efficiently managing and sharing data about invasive species can contribute to significant biodiversity outcomes. The purpose of this guide is to assist the practitioner in publishing invasive species data and training others to prepare and publish to GBIF—the Global Biodiversity Information Facility—using consistent terminology that allows data to be used many times.

This guide was produced as part of the 'National and Regional Alien and Invasive Species Data and Information Mobilization and Capacity Building in the Pacific' Project, funded by the European Union through GBIF's Biodiversity Information for Development (BID) programme and operated by SPREP on behalf of the Pacific island countries and territories.

Why Share Species Data?

Data are valuable only when used, and the value of data grows when the amount of data and the links to other datasets increase. Sharing information reciprocally can raise the value of the information.

It can be a challenge to manage the biosecurity information required by quarantine and biosecurity staff, particularly when surveillance and monitoring of pest species is often done by different groups with different reporting systems. Biosecurity officers often benefit from data from other countries to enable early detection and response within national borders. Establishing routines for sharing data can also improve the consistency in data collection and analysis and can help grow trust in data sharing across relevant sectors and ministries.

Informed decision-making depends on quality information. Data-sharing routines, systems, and trusting relationships increase the availability of high-quality data and make those data easier to use. Quality data that are easier to use make it easier and faster for national staff to analyse and interpret data into knowledge products that address national priorities.

There are many reasons to share data:

- We share data to support science research, education, and decision-making globally.
- Sharing data allows us to make ourselves known to the biodiversity research and education community. Sharing makes the statement that we are an active member of the community, that we want to help and want to be engaged within the field of biodiversity. By sharing, you gain recognition for your data collection and sharing efforts through data citations in research, demonstrating the value you are adding to global knowledge about our biodiversity.
- Storing data in a shared repository keeps the data safe from server crashes, computer loss, and file corruption. The metadata and explanatory information required for data sharing also help keep the data useful after staff turnover.
- Sharing makes our data as useful as possible – to have a greater impact! The datasets we publish and share can be used for research, training, educational curricula, biodiversity protection, and more, and invite the world to gather in one place to collaborate and communicate through shared interest in the various biodiversity data collections.
- Sharing invites interested individuals to come together and work with other scientists, researchers, students, and so on, and it sends a strong message that we want to be there, we want our work to be recognised, and we want to help, as well as be open to the expertise that is available from other data publishers.
- The data we publish can be made even better by opening up communication with a global network of people who have a shared interest in our datasets. Assistance from others may be able to resolve any limitations that we may have with our published dataset, essentially improving the quality of information and/or making it complete.
- For a small island nation, publishing our data collections from our own home environment celebrates development in areas of science and research on a global stage.
- In some instances, donors who may have funded our organisation to facilitate the collection of the invasive species data may require that the data be made publicly available. One avenue for this public sharing is through GBIF – the Global Biodiversity Information Facility.



Regional Invasive Species Data and Resources

There are several ways to obtain and share data and information relevant for Pacific invasive species management and decision-making. In addition to the GBIF data collection, Pacific data and information can be found at the [Battler Resource Base](#) and the [Pacific Islands Protected Area Portal](#).

What is GBIF?

GBIF – the Global Biodiversity Information Facility – is an open-data research tool funded by various governments from around the world, which aims to provide free and open access to biodiversity data to anyone, anywhere.

The GBIF website at www.gbif.org contains over a billion biodiversity records freely and openly available for users. The site provides access to these data through a variety of search mechanisms.

The GBIF system of data sharing allows for consistency across countries through the use of a fixed vocabulary: the [Darwin Core standard](#) for biodiversity. Using these standard terms is part of the process of making sure the data are robust and high-quality. The Darwin Core standards are maintained by [Biodiversity Information Standards \(TDWG\)](#), also known as the Taxonomic Databases Working Group, a not for profit scientific and educational association focusing on the development of standards for the exchange of biological/biodiversity data.

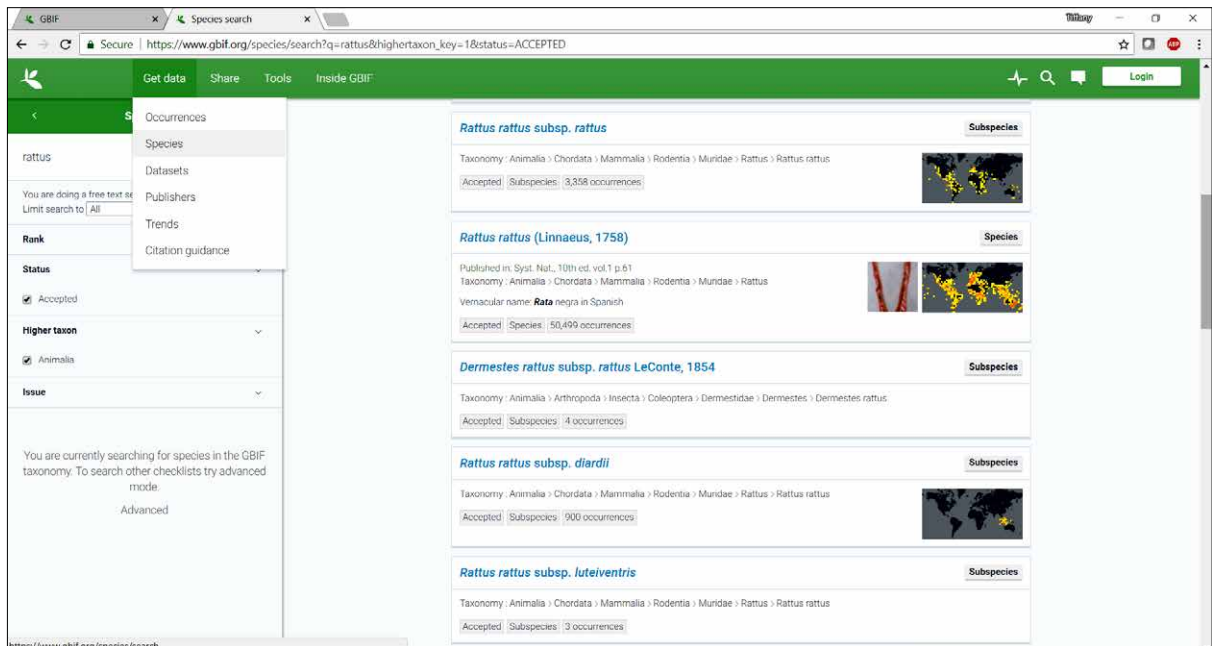
Interactions with the GBIF infrastructure are facilitated through Participant Nodes. In the Pacific, SPREP provides a regional-level Participant Node, the coordinating team designated to establish and strengthen GBIF-related activities of its member countries and partner organisations. Participant nodes are also considered knowledge hubs for both biodiversity data, guiding stakeholders to relevant sources of biodiversity information and data, as well as their own expertise on biodiversity and data management.

- As a Participant Node, SPREP can endorse biodiversity Data Publishers. National ministries, agencies, institutes, or organisations who wish to publish their data can apply to be registered in GBIF. SPREP's Node Manager will receive an email alert to begin the verification process when an appropriate partner country, organisation, or institute requests to be registered on GBIF. Once endorsed, they may publish as often as they wish, with new or updated datasets.
- SPREP's goal as a Participant Node is to help set up agencies, institutes, and organisations as data publishers and to assist them with the data publishing process.
- Note: A key criterion for GBIF is that a 'data publisher' must be an institution rather than an individual. Individuals wishing to publish datasets should work to get their affiliated organisations endorsed as a data publisher. See <https://www.gbif.org/become-a-publisher>

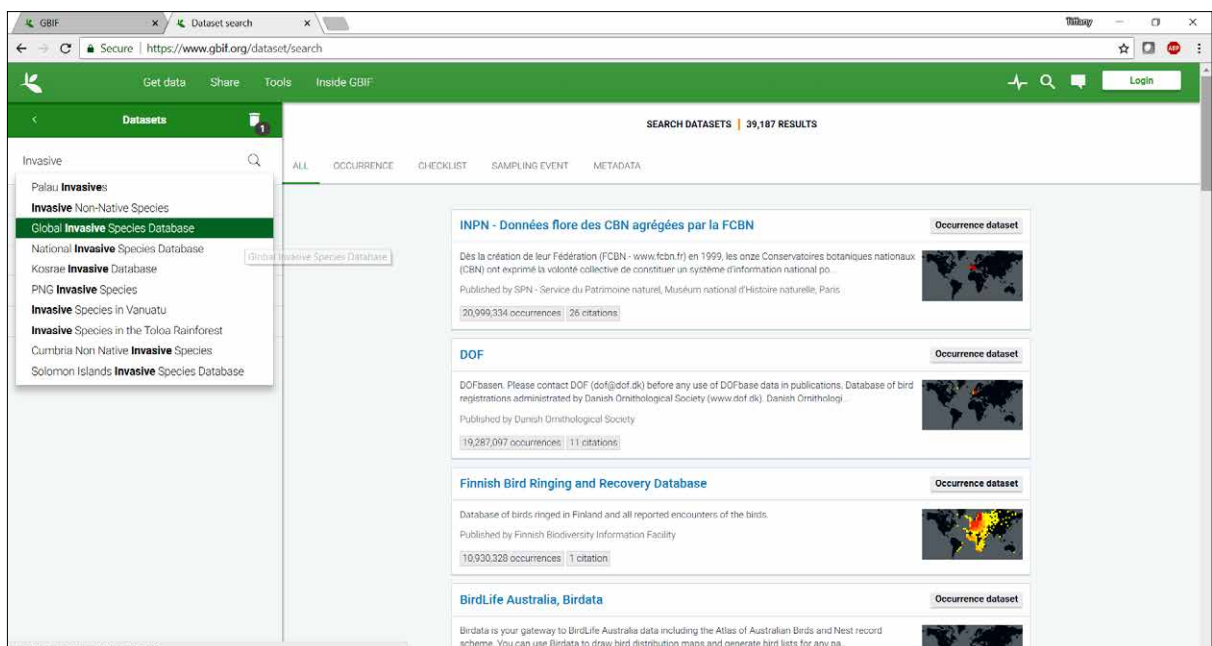
How do I use GBIF to find data?

The site www.gbif.org contains over one billion records of species, specimens, observations, and samples. Each record will have information about how to download and cite the data. Try exploring the site!

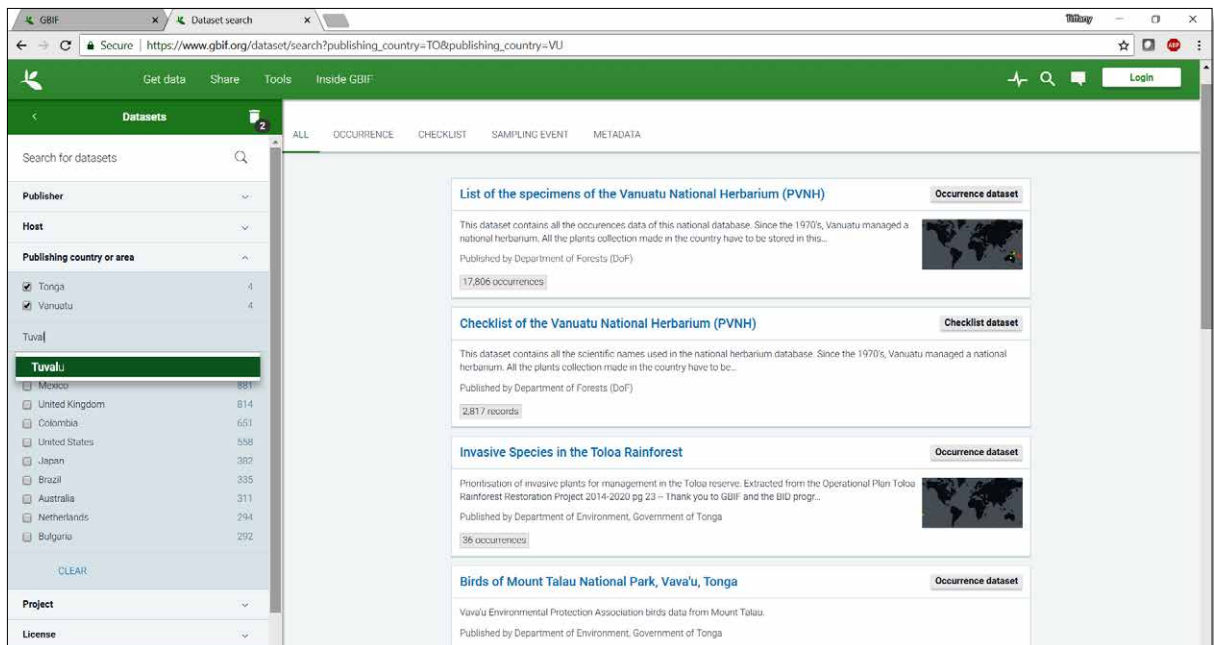
Start by searching from the homepage or clicking on “Get Data” in the top panel. You can look for occurrence data, search by species, look for whole datasets, and more.



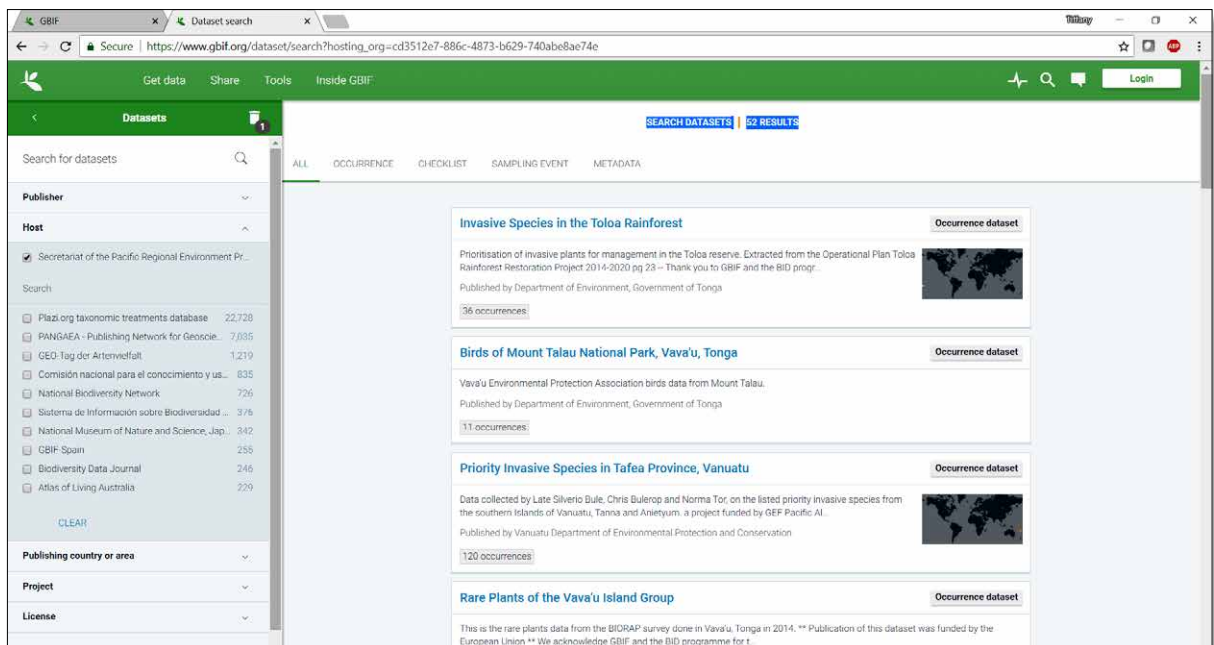
You can search for whole datasets using keywords. In this example, we’re looking for datasets with titles that include “invasive”. (Note that you might not find all the data available about invasive species this way. You will only find datasets that have been labelled using that exact word.)



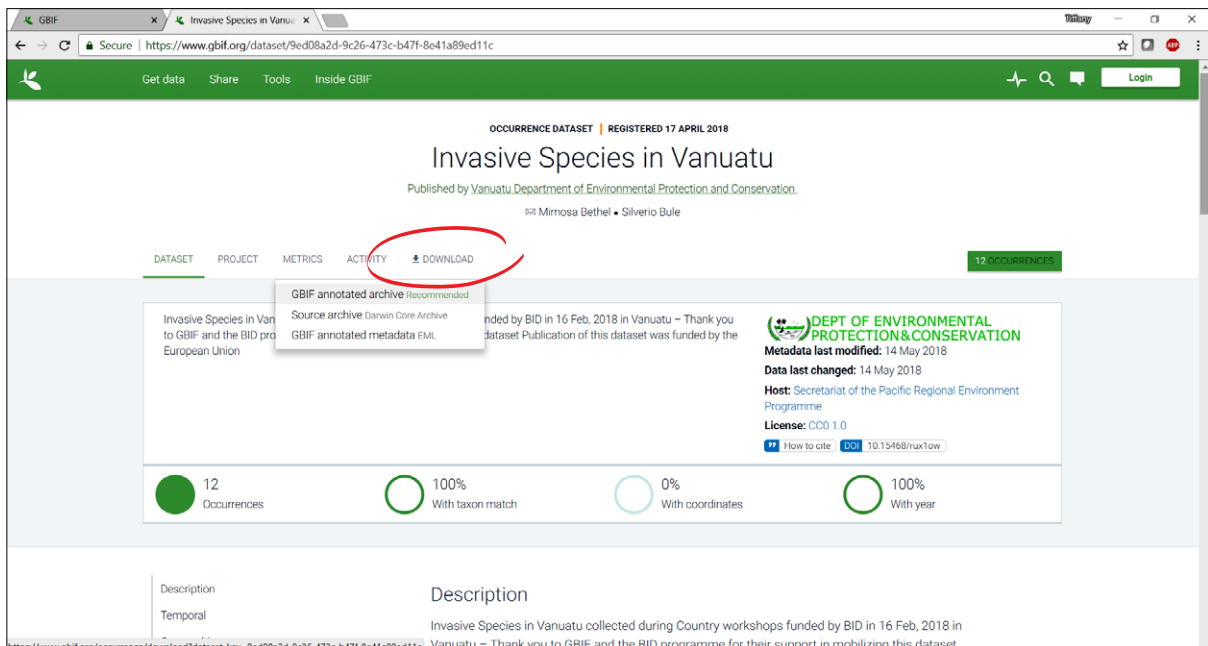
You can search for datasets by region, called “Publishing country or area” in the left panel. You can search for multiple countries at a time and combine with other filters: each checked box will be included in your search.



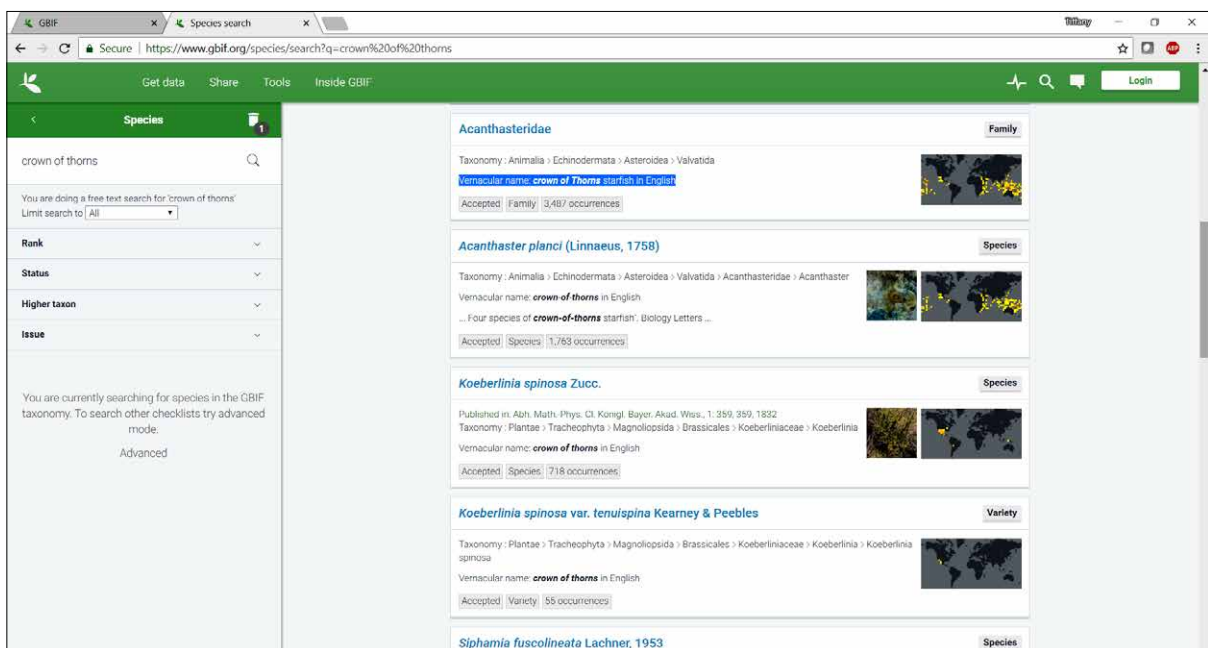
For those of us in the Pacific, a useful way to find related datasets from our region is to look at the datasets hosted by the SPREP Node. Here, we searched using “Secretariat of the Pacific Regional Environment Programme” as the Host. There were 52 results at the time, and this number is growing as you and your colleagues publish more data!



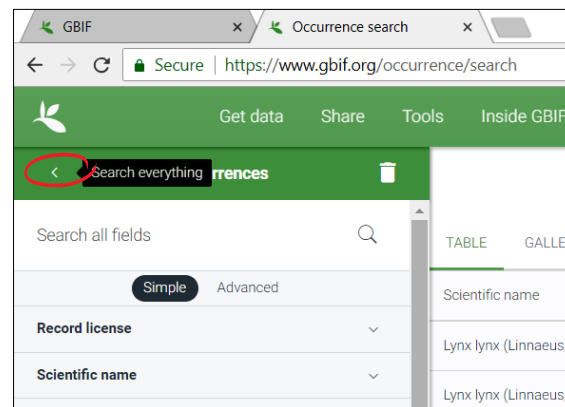
Click on a dataset name to find out more. You can also download the whole dataset if you want, in a few different formats:



There are many ways to search, using filters, keywords, and more. One of the most common ways to search is for a single species, and you can use either the scientific name (Latin name) or the common name (vernacular) to search. However, it is important to remember that the vernacular name might be different in different places, or the same name might be used for different species. For example, here is a search for “crown of thorns”. The search results include the sea star or starfish called “crown of thorns” but also include a plant, a fish, an insect, and more. For best results, you may want to search a few times using different names.



You can also choose to search “everything” (occurrences, species, whole datasets, and more) by entering a search, then clicking on the top left arrow here:



Credit where credit is due

Always cite your references! Each dataset is given a unique digital object identifier (DOI) reference by GBIF (for example: <https://doi.org/10.15468/suj1kg>). When you use these data to create a report, you would also state the citation given in their metadata and/or that DOI to acknowledge who provided the data. When you use your GBIF search results, you may be downloading and using data from many different datasets shared by different institutions—for example, all occurrences of a particular species observed or collected in a country. In this case, GBIF will provide a single, unique DOI for the download itself, which when cited will give credit to each of the contributing data publishers (including direct links to the published citation from the dataset page). Just follow the recommended citation provided when you order the download!

Who owns the data?

The primary way that users access data on GBIF is by searching for attributes rather than datasets. For example, you could use GBIF to quickly find out the global distribution of a particular species. This means that the data are going to be extracted from the dataset more often than not. Rather than picking up one whole dataset, a user is more likely to find pieces of data from many datasets and will then have to figure out where each piece came from. To make this possible, each piece of data on GBIF is accompanied by linked metadata.


This data about the data, or “metadata”, should include who owns the data and who collected it using which methodology. It is important to give credit where credit is due, and to show people who they would need to contact if they want to ask about using the data.

All occurrence records published through GBIF must be shared with an open, machine-readable licence or waiver using the [Creative Commons](#) categories CC0, CC-BY or CC-BY-NC. The standard licensing requirement enables the terms of use to be applied to the download as a whole.

Regarding rights to the data, a data registry like GBIF does not own the data or have access to your database computer where the data was entered. The data still belong to the data publisher. The data publisher is the one who extends the rights for others to use the data and under what conditions (according to the Creative Commons category), while the registry or index, like GBIF, simply provides the infrastructure to share the data. The responsibility to use the data as licensed falls with the data user. There are publications on how users of data can be the best possible citizens by following community norms of attribution (e.g., <http://vertnet.org/resources/norms.html>).

When you share data, the data still belong to you but will be accessible to others seeking data or information. Sharing data provides an opportunity for others on the global stage to suggest improvements to make the data better (with your approval of changes, of course), add to the data, or combine the data in new ways to gain greater understanding of global biodiversity.

In the GBIF system, GBIF users only have access to the data that the data publisher has agreed to share online. You have the option to publish data or metadata only. You can also choose to share only selected elements of the data, such as an approximate rather than precise location, or withhold certain details in contextual fields for reasons of cultural sensitivity or privacy.



Did you know?

Pieces of **Data** are raw numbers or facts, without interpretation or analysis. By themselves, data can be hard to use or even understand.

Information is a useful combination of data, with context and interpretation. Information is the valuable result we seek to build on a foundation of good data, analysis, and understanding.



What are metadata?

Metadata are a summary of the basic information of a dataset and how the data have been managed. It shows the “who, what, when and where” of the data. GBIF has its own requirements for data and metadata. See <https://www.gbif.org/data-quality-requirements> SPREP as the Pacific Participant Node requests the following metadata for each dataset published to GBIF:

Metadata fields requested by SPREP	
Title of dataset	
Description of dataset	
Link to original source, if available	
Associated Parties	
	name
	position
	organisation
	contact information



What are others allowed to do with the data I share?

Controlled data sharing means that *access* to the data is controlled, using mechanisms like paid access or password-protected access, or that the *usage* of the data is controlled, using mechanisms like data licenses.

Controlling access, using mechanisms like paywalls, passwords, or restricted copies of a dataset on only one or a few computers, requires long-term capacity for data storage, data management, and responding to requests. It is possible to publish only information about the data (metadata) and make people ask for access to the original datasets, although this again requires resources to monitor those requests.

The GBIF system relies on open access to the data and [Creative Commons](#) licensing. The users are licensed to use the data in a specified way or purpose. This system places responsibility on the data user to understand and abide by the sharing rules.

All occurrence records published through GBIF must be shared with an open, machine-readable licence or waiver using the [Creative Commons](#) categories CC0, CC-BY or CC-BY-NC, which in general provide data free of charge for “fair” uses. The license CC0 (public domain) is recommended wherever possible. CC0 means the dataset comes under a public domain, and people are free to use the data in any way. However, it is uncommon in the community for people to pull out data and use it for monetary gain. In addition, it is unlikely the data publisher will have all the resources to follow up on people who use the data incorrectly. It is important to check your local laws and confirm the data licensing with those who collected the data. Additional information can be found at www.gbif.org

For more information about open-access categories, knowledge rights, and data management for environmental information in the Pacific, we encourage you to engage with the [Inform Project](http://www.sprep.org/inform/home) (www.sprep.org/inform/home) and the [Access and Benefit Sharing \(ABS\) Project](http://www.sprep.org/abs/about) (www.sprep.org/abs/about) facilitating knowledge management under the Nagoya Protocol, both of which are active partnerships with the Pacific island countries and territories.



How do I publish data on GBIF?

First, you must register your organisation to GBIF, indicating that you'll be using the SPREP Integrated Publishing Tool (see below). After you have registered as a data publisher, you can follow this step-by-step process to publish data on GBIF:

- 1 Enter metadata for your dataset on the SPREP IPT.
- 2 Have your source dataset Darwin Core ready (see below: "How do I prepare data for publishing").
- 3 Upload dataset to the IPT.
- 4 Map your dataset to Darwin Core Terms, either as a Checklist Dataset or Occurrence Dataset (as appropriate for your data)
- 5 Publish.
- 6 Validate the dataset with the GBIF Data Validator, prior to registering the dataset (<https://www.gbif.org/tools/data-validator>).
- 7 Register the dataset with GBIF (first time-only).
- 8 Receive feedback from GBIF and users.
- 9 Update your source data.
- 10 Repeat from step 4. Note that the responsibility for maintaining datasets falls on the data publisher.



What if my data were published by another institution?

Datasets have unique Identifiers. If the dataset you want to share is already on GBIF, it is recommended that you don't re-publish because GBIF will identify the dataset as redundant. A better solution, as a publisher, would be to link to that data already uploaded by the other institution.

Remember, the publisher is not necessarily the only or original owner of the dataset, as defined in the metadata, they are just the one who did the work to publish it through GBIF. If another institution has already done that for you, go ahead and use it as an asset!

Pathways to data publishing

How do you decide what to publish? First, consider the type of species data that you most need and use. Sharing that data can help provide you a stable, recorded copy, and those data are also likely to be useful to others.

Consider the goals of your active projects. In addition to new data, are there older data you would need to compare with any new findings? During project development, did someone have to dig up old files to try to identify the scope of the problem? Those data may be worth sharing.

While data about endemic species or new areas are always interesting, it's also true that data about species found in common areas are helpful to identify and plan for potential species invasions.

Data that are already in tables or spreadsheets are ideal, but data in any format can be extracted and put into columns and rows. For example, you may have a report that verbally describes encounters with certain species in a particular area or areas. You can convert that into a dataset with the species name matched to the area, year of observation, and so on.

You can decide whether it is occurrence data, checklist data, or sample event data. Each of these types requires different amounts and details of data. For more information, see <http://rs.tdwg.org/dwc/terms/> and www.gbif.org/data-quality-requirements.



Did you know?

Occurrence data are evidence for the existence of an organism at a specified place and time.

Checklist datasets are a list of taxa. In Darwin Core terms, these data only include taxonID (UUID) and scientificName.

Sampling event datasets provide greater detail about a species occurring at a given location and date, including the methods, events and relative abundance of species recorded in a sample.



Case Study: Kosrae

A dataset was available that had many variables, listed as scientificName, Vernacular name, Ecosystem, Pathway, Status (which included 'introduced'), Abundance estimates, and more. In this case, the dataset had more data variables than required for a typical checklist or occurrence dataset on GBIF.

If the goal was to publish these data as a checklist, the data clean-up would involve extracting only the values under "scientificName", adding the metadata about the organisation and dataset, and publishing.

To publish these data as an occurrence dataset, the data clean-up was a bit more involved. Some terms had to be standardised: for example, the label "Ecosystem" was changed to the Darwin Core Term "habitat" because the values indicated where a species was found: the data included "terrestrial", "marine", "stagnant water" and "wetland/terrestrial". The term Status actually indicated whether the species was native to Kosrae or introduced, so it was changed to the Darwin Core Term establishmentMeans. You can view the result here: <https://doi.org/10.15468/ypzayn>

How do I register as a data publisher on GBIF?

The first step to sharing data with GBIF is to register, identifying your organisation as one that has quality data to share. Once you are registered, you can publish as many datasets as you like.

The GBIF requirement for registration is that you have access to a node – in the Pacific islands, this means access to the SPREP node and Integrated Publishing Toolkit (IPT). You can register to become a publisher at any time, you do not need to have data ready. A publisher can also be registered to provide access to metadata, even if they are not ready to publish a dataset.

A publisher may be a ministry, institute, organisation, and so on. For simplicity, we will use the term “institution” throughout these instructions.

Step 1 Access the GBIF website at www.gbif.org. Click on *Share tab – Become a Publisher* and search for your institution to confirm whether or not your institution is already registered. If your institution is not registered, proceed to Step 2.



Step 2 There is a data publisher agreement to consider. Agree to continue.

Step 3 Fill in the details requested (e.g., name of institution, primary contact, website address, description of organisation, map location, etc.).

Remember to select that the endorsing node, as well as access to an IPT, will be through the SPREP node.

The site will ask whether you need help in publishing – the answer here is “No” because there is already assistance available to you via the SPREP Node Manager.

Contact Information: Provide your contact details within your institution. It is a good idea to also provide a general contact email for the institution (i.e. info@.....) that will last even if a staff member leaves. The Technical Point of Contact will require the contact information of the Node Manager at SPREP. Note that if contact information changes, you may contact the GBIF Help Desk directly to update the contact information.

Provide a short description of your institution.

Select “No” to the question “Are you planning to install and run publishing software (such as the [Integrated Publishing Toolkit – IPT](#)) to publish your data directly to GBIF.org?”

Step 4 Submit

Step 5 A notification email will be sent to the contact person (Data Publisher) and a notification email will be sent to SPREP as the GBIF Participant Node.

Step 6 Once the Participant Node has endorsed the institution for GBIF, the registration will be complete and now has a new publisher added to the GBIF site.

Step 7 GBIF will send the contact person a password for the newly registered institution. This password will be required by the Node Manager to assist with registering new datasets to GBIF in the IPT.

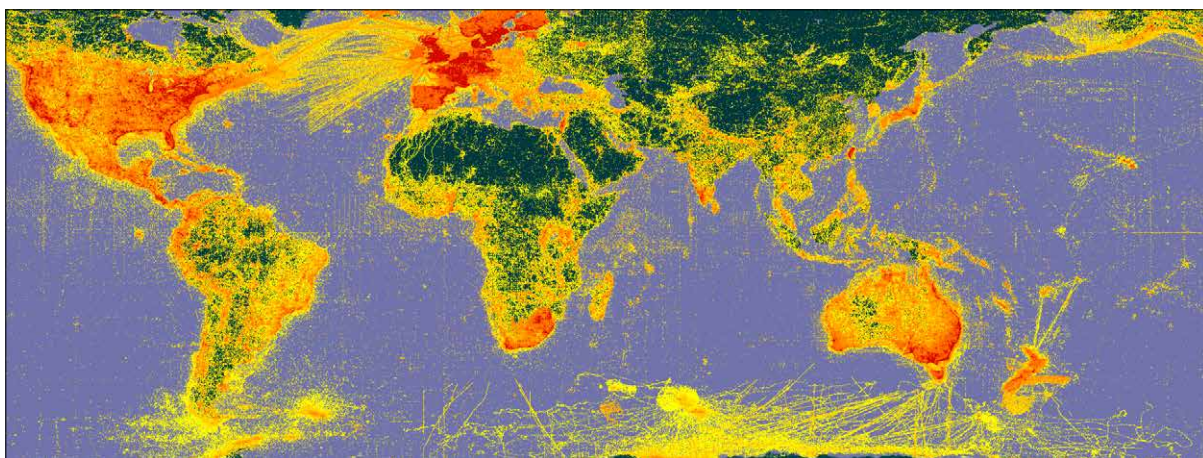


Georeferencing

Georeferencing provides the 'where' for the dataset, allowing the records to be placed on a map and to be used in spatial analysis. Georeferencing means attaching a location tag, or reference, to each piece of data. For example, in a bird survey, we would report bird sightings by species name but also by location, using coordinates.

- It is important to add latitude and longitude values to give spatial context to the data.
- Best practices for georeferencing promote a minimum of four critical Darwin Core fields to be included: decimalLatitude, decimalLongitude, geodeticDatum, and coordinateUncertaintyInMeters.
- The geodetic datum is a mathematical model of the surface of the globe upon which coordinates are based. There are more than 200 geodetic datums: some are used only locally; some, such as the popular WGS84 used by Google Maps and others, are global in coverage.
- Uncertainty in Metres describes the size of a described place. The coordinateUncertaintyInMeters is the maximum distance you could go from the location of the given coordinates and still be within the place that is being described. It shows data users, in one variable, how specific the location is in a way that applies anywhere on the planet. The uncertainty can be easily determined using Google Maps by using the measuring tool (right-click on the map) to draw a line across the widest cross-section of the place. Half of that distance is the coordinateUncertaintyInMeters, and the midpoint of that line is where the coordinates should be.
- Georeferencing can be a complex process. Please refer to the [Georeferencing Quick Reference Guide](#) to understand this fascinating sub-discipline of biodiversity informatics.

There are a variety of tools to assist with georeferencing – see www.georeferencing.org.



How do I prepare data for publishing?

Data preparation involves pulling the information out of different sources, creating the occurrence and checklist datasets, and mapping to Darwin Core Terms so that the data can be found and understood.

There are 3 basic types of datasets that GBIF can accept: Occurrence, Event, and Taxon or checklist. A Taxon dataset type is primarily a list of taxa. The metadata would describe the scope of the list. The existence of a taxon at a place and time—an Occurrence—is an important concept because it provides a richer description of the existence of taxa in the world. An Event-based dataset is meant for monitoring, where sites are sampled or observed repeatedly over time, with information such as abundance and measurements or facts about the events.

Remember that checklist datasets require a unique identifier (TaxonID) and the species name (ScientificName), while occurrence datasets require the who/what/when/where information. You can find the data requirements for the different dataset types at: www.gbif.org/data-quality-requirements

The first step is to make sure the data are in a format ready for IPT, meaning a delimited file such as an Excel file or a database connection.

Palau Case Study

It is often possible to pull data from a text or combine data from several different reports to make the desired dataset for publication. In this case, a published scientific article contains data that we want to add to GBIF. The paper has a unique identifier, the article DOI, so we will list that in the associatedReferences cell.

Table 4. Detected introduced, cryptogenic and potentially introduced species during the preliminary Palau introduced species survey. Species are listed with an indication of the sampling locations/sites (wharves, moorings, vessel hulls, or “pristine” locations without commercial activity). A “•” indicates presence upon a substrate. Please note that scientific names and taxonomic authorities (for species names) were verified using the WoRMS database (<http://www.marinespecies.org>).

Phyla	Species	Status	Wharves	Vessels	Mooring	Pristine
Porifera	<i>Haliclona caerulea?</i> (Hechtel, 1965)	Potential	•		•	•
	<i>Mycale</i> sp. (orange sponge)	Potential	•			•
Hydroida	<i>Eudendrium carneum</i> Clarke, 1882	Introduced	•			
	<i>Obelia</i> sp.	Cryptogenic	•	•	•	
	<i>Thyrosocyphus fruticosus</i> (Esper, 1793)	Introduced	•		•	
Polychaeta	<i>Sabellastarte</i> sp.	Potential	•			
	Serpulididae	Potential	•	•		
Cirripedia	<i>Amphibalanus amphitrite</i> (Darwin, 1854)	Cryptogenic (cosmopolitan)	•	•		
	<i>Chthamalus proteus</i> Dando and Southward, 1980	Introduced			•	•

Source: Campbell et al. (2016) Marine pests in paradise [...]. Management of Biological Invasions 7:351-363 doi:10.3391/mbi.2016.7.4.05 Courtesy of the authors..

Some of the data that we want to include are already in table form. Each species record that we pull from this table will have the associated References cell stating the citation of this particular article. In the metadata, the data ownership will also be stated with appropriate citations for each data source.

Tip

Get into the practice of attaching an identifier (occurrenceID for occurrence datasets and taxonID for taxon datasets) to each record of your datasets when you are preparing your data. These identifiers will be required when mapping your data through the IPT. You can generate these yourself, or you can use a Universal Unique Identifier (UUID) site (e.g., www.uuidgenerator.net) to generate them for you.

Preparing data using the Integrated Publishing Toolkit

The IPT is a software tool installed on a server. The IPT we use in this instance is the SPREP IPT. The SPREP Node Manager is committed to assist with the initial publishing of datasets. In addition to this guide, there will be on-going technical support from the SPREP Node Manager, as well as assistance from GBIF and BID mentors as necessary.

- After an institution is successfully registered, the SPREP Node Manager will add the institution as one of those that can publish datasets on their server, using the password that was sent to the contact of the Data Publisher.
- The publisher will convert their datasets into a Checklist or Occurrence dataset using Darwin Core Terms. A quick reference guide for Darwin Core terms can be found here: <http://rs.tdwg.org/dwc/terms/>

Once you have prepared your dataset and uploaded it to the IPT, it will be possible to do the field to field mapping to Darwin Core Terms and publish the data. Note that your source data in its original form will also be archived on the IPT.

GBIF has an IPT wiki for documentation and tutorials. There is also a video demonstration of creating an occurrence resource with extensions found on the main GBIF IPT page: www.gbif.org/ipt.



Recommended Darwin Core fields for a simple occurrence dataset

Darwin Core Terms and Definitions:		http://rs.tdwg.org/dwc/terms/	
What – taxon	Where – location	When – event	Other
scientificName	decimalLatitude	year	occurrenceID (UUID)
	decimalLongitude	month	basisOfRecord
	geodeticDatum	day	organismQuantity
	coordinateUncertaintyInMeters	eventDate	organismQuantityType
	country	verbatimEventDate	establishmentMeans
	countryCode		dynamicProperties
	island		
	islandGroup		
	stateProvince		
	waterBody		

Mapping to Darwin Core Terms

Mapping is the matching of data in source data fields to fields in the Darwin Core Standard, so that it can move from the original dataset into a format that can be used by GBIF.

Darwin Core Terms are a list of terms or fields that are of interest in the field of Biodiversity. Using standardised terms allows for comparison across datasets from different places, times, and countries, often done using automated searches. A quick reference guide for Darwin Core terms can be found at <http://rs.tdwg.org/dwc/terms/>. This guide includes definitions and examples of each term. The terms are organised by categories (in bold) in the index. The categories correspond to Darwin Core terms that are classes (terms that have other terms to describe them). The terms that describe a given class (the class properties) appear in the list immediately below the name of the category in the index. The index provides links to the term descriptions in the table below the index.

In the IPT, one can either map a source field to a Darwin Core field or set the value of a field to a constant. For information that does not have a corresponding Darwin Core term, the information can always be added in the Darwin Core term dynamicProperties but might also exist in and be possible to add in a Darwin Core extension.



Darwin Core Terms for Invasive Species

There are some Darwin Core Terms that are especially relevant for describing invasive species, but it is important to note that the Darwin Core Terms do not include every aspect of a species and a case may be made to review and update the Terms and associated vocabulary.

Managers want to know...	Darwin Core Term	Suggested vocabulary
How did it get there?	establishmentMeans	native, introduced, naturalised, invasive, managed, or uncertain
Where does it live?	occurrenceStatus	present; absent; common; irregular; rare; doubtful; or excluded (GBIF vocabulary only). [However, in practice, some people include conservation status in occurrenceStatus.]
Is it native?	<i>Proposed term:</i> Origin	<i>Proposed:</i> native, reintroduced, introduced (potentially subdivided into before or during the modern era), vagrant, or unknown
How well established is it?	<i>Proposed term:</i> degreeofEstablishment	<i>Proposed:</i> native, captive, cultivated, transported, released, casual, reproducing, established, dispersed, colonizing, invasive

For further discussion, see <https://github.com/tdwg/dwc-qa/wiki/Webinars#chapter4>

Here are some examples of mapping Darwin Core terms. Definitions are found in the [quick reference guide](#):

- ‘Species’ name becomes **scientificName**
- ‘Tongan name’ becomes **vernacularName** (note that this becomes a loss of information about what language the name is, and this would need to be specified perhaps in dynamicProperties for Occurrence datasets or in a Vernacular Names extension for a Taxon dataset).
- **dynamicProperties**: other information that does not fit into Darwin Core can be placed here. It is a way to store fields of information, structured data, without losing the context. Invasive Species Variables can be placed here in this field. Links to associated references can be placed here, such as a link to download the pdf source dataset.
- **identificationRemarks** – any remarks about the process of assigning a Taxon to an observed or collected entity.
- **previousID** – any scientific name that was previously applied to an entity, where the scientificName is the currently accepted or valid name.

“We are creating an action plan and working with agencies to make data available. I really want to do this to raise the profile of invasive species issues in Solomon Islands.”

– Josef Hurutarau, Solomon Islands



Case Study: Solomon Islands

In this case, all the right data were there but the labels did not match Darwin Core Terms. The terms you use in data collection, or the labels provided by individual researchers, are often not standardised for global comparison. In this case, the required data clean-up was as simple as relabelling the columns.

Serial number	country/territory	accepted_species_name_OK	accepted_species_authority	synonym used b	synonym species	Kingdom	environment system	provenance /origin	Invasiveness	establishment means	Invasiveness	year	occurrence
53467	Solomon Islands	Abelmoschus moschatus	Medik.			Plantae	terrestrial	alien	not specified	alien	not specified	2016	
53468	Solomon Islands	Acacia auriculiformis	Benth.			Plantae	terrestrial	alien	invasive	alien	invasive	2016	
53469	Solomon Islands	Acacia farnesiana	Willd. (L.)			Plantae	terrestrial	alien	invasive	alien	invasive	2016	
53470	Solomon Islands	Acacia sp.	Mill.			Plantae	terrestrial	alien	not specified	alien	not specified	2016	
53471	Solomon Islands	Achatina fulica	(Férussac, 1821)			Animalia	terrestrial	alien	invasive	alien	invasive	2016	
53472	Solomon Islands	Achatina fulica	(Férussac, 1821)			Animalia	terrestrial	alien	not specified	alien	not specified	2016	
53473	Solomon Islands	Achyranthes aspera	L.			Plantae	terrestrial	alien	not specified	alien	not specified	2016	
53474	Solomon Islands	Acmella uiginosa	Cass. (Sw.)			Plantae	terrestrial	alien	not specified	alien	not specified	2016	
53475	Solomon Islands	Acridotheres tristis	(Linnaeus, 1766)			Animalia	terrestrial	alien	invasive	alien	not specified	2016	

Constructing a unique identifier

Each occurrence data record requires a unique identifier. It is highly recommended that your source dataset has globally unique and persistent identifiers (such as UUIDs) for the values of occurrenceID and that these identifiers remain with the record from then on. If you are using Excel, it is recommended good practice to start attaching an ID to each of your occurrence records (i.e. to each field, or individual row); this ID will go into the IPT under the Darwin Core Term occurrenceID. This is crucial in GBIF for people searching to identify a specific record within your dataset among the records in the global dataset once it is published. Having global unique identifiers that are UUIDs will allow references to specific records to be made easily, and these will unambiguously point to specific records any time in the future.

- The SPREP Node Manager can assist by generating a global UUID using IDs you may already have attached to your records and sending the global UUID back for incorporation in the source data. It is very important that these identifiers accompany the records at the source and that they do not change.
- Similarly, records in a Checklist dataset require unique identifiers for the taxonID field. These can be generated using a UUID site, like www.uuidgenerator.net, and captured in the Taxon records.
- If you are using a database other than Excel, it may be able to generate a UUID for you automatically. Excel can also be programmed to generate a UUID (see www.idigbio.org/wiki/images/0/03/GUIDgeneration.pdf).
- Your dataset as a whole will be given a DOI when registered and published with GBIF (e.g., Tonga Toloa Rainforest = doi:10.15468/suj1kg).

Cleaning data

Cleaning up a dataset may range from simply checking that all the information is included in a standardised way through to filling gaps in the data or defining new pieces that you want to include. There are many ways to clean data just using Microsoft Excel. (You can find a list of helpful formulae here: www.sprep.org/attachments/Publications/IOE/gbif-sprep-invasive-species-dataset-templates-resources.xlsx . Open Refine is an excellent data-cleaning tool. See the *Basic Use Guide to Open Refine for Biodiversity Data Cleaning* at <https://tinyurl.com/ORforBiod>.

The basic steps are standardising and updating.

- **Standardising** means taking fields like country codes (e.g. TO/TON for Tonga) and standardising to just one code (in this case, TO).
- **Updating** makes a dataset more complete, e.g. if you only have a scientific name for a species in a dataset, you could make it more complete by adding kingdom and phylum, for example, if you only had genus and species.

Georeferencing may also be done during the cleaning and updating step – cleaning can complete all levels of geography. People search for all kinds of terms, so you may use several to describe one location, such as South Pacific, Tonga, Island, and so on.

As you clean each part of the data, check against standard terms used elsewhere by known experts.

Republishing data

Once the source dataset is mapped through IPT, published and registered with GBIF, you can view the published dataset on GBIF. You may then find that some of the dataset as displayed on GBIF does not include all the information that you wanted, that you need to add more information in the georeferencing, that you may have used the wrong Darwin core term, or that GBIF's data quality checks have identified potential issues. You will then have the option to go through the process of addressing the issues in the data and republishing them.

We will go through an [example dataset](#) from the Toloa Rainforest in Tonga. You may download the Excel file here to follow along: www.sprep.org/attachments/Publications/IOE/invasive-species-toloa-rainforest.xlsx

- The Tongan Toloa Rainforest published on GBIF had a few issues that needed to be resolved: 3 potential Darwin Core mismatches. Therefore, it was chosen to republish the data with improvements.
 - The data publisher tech support had incorrectly assumed that the field 'habit' was Darwin Core Term 'habitat'. This was corrected.
 - One alert was an incorrect scientific name – possibly a simple spelling error red-flagged by GBIF. To correct this, we can use the Global names resolver site (<http://resolver.globalnames.org/>). Always confirm first with the data custodian if it is OK to make these improvements.
 - The value "Tongan Name" needed to be in the Darwin Core field vernacularName.
- **Updating the Tongan Toloa Rainforest Data:**
 - Log into the IPT
 - Go to the Manage Resources tab
 - Select the dataset (called resource in the IPT) to update
 - In the Source Data section, Choose File and upload it, replacing the previous version of the file
 - The Source Data review screen will allow you to set the characteristics of the uploaded file and select specific worksheets. You can view the data to see if they are being interpreted as expected
 - Click on Save, and the data are now in the IPT and ready for field mapping
 - In the Darwin Core Mappings sections, edit the mappings to see which fields have been mapped and which have not – for example, the following corrections made:
 - Species should be mapped to scientificName
 - Habit should be mapped within dynamicproperties (this is a step that had to be done prior to upload)
 - Tongan Name should be mapped to vernacularName
 - Status should be mapped to establishmentMeans
 - 2014 dominance should be mapped within dynamicProperties (this is a step that had to be done prior to upload)

- Invasive should be mapped within dynamicProperties (this is a step that had to be done prior to upload)
- PreviousID should be mapped to previousIdentification
- ID Remarks should be mapped to identificationRemarks
- Newly added fields Latitude and Longitude, Datum, and coordinateUncertainty should be mapped to decimalLatitude, decimalLongitude, geodeticDatum, and coordinateUncertaintyInMeters
- Save and return to the Overview page

Now that everything is mapped, you can publish by clicking on the Publish button in the Published Versions section. It is recommended you add a short summary of what has been done in the republishing process; in this case, something like “added new fields for georeferences and dynamicProperties. Two taxonomic updates based on issues discovered by GBIF ingestion process.”

Your data are now republished. In this Tonga example, we updated an existing dataset.

The Data Validator tool (<https://www.gbif.org/tools/data-validator>) is a way of checking for quality issues prior to publication. Once the dataset is published, the ‘metrics’ tab gives a summary of the contents of the dataset including e.g. taxonomic distribution and quality issues; for example, see <https://www.gbif.org/dataset/7c0cd863-8b81-4937-84f9-2f596fd3fa79/metrics>.

Further feedback may come via GBIF either from its data quality reports or from people trying to use your data, sometimes including improvements. After you go through this feedback, you can make associated changes and republish again if needed.



“What we have accomplished together is our legacy for the people of Tonga. It’s a love of nature that has driven me to work and help protect my island home through managing invasive species.”
– Lisa Fenua, Tonga

How do I plan for data publication?

It is strongly recommended that you include data management as part of the operational plan for your project to maximise the likelihood that you can mobilise data through digitisation, management, and online publishing. Your plan helps ensure that you collect all of the information you will need for publication at the time of sampling and that you account for the time required to prepare, clean, and publish data. The operational plan should establish a timeline for the project, the resources that will be required, and how the project will take place. Seek expert advice when writing your operational plan.

Requirements for the provision and management of data may be necessary to include in consultancy contracts and memoranda of understanding (MOU), if part of your invasive species management is done through the use of external consultants and/or partner organisations.

The templates that you use for data collection or data processing can make it easier to standardise data for later publishing. You will not have to change all your systems of data planning and collection, to have it GBIF ready. However, if you plan to publish using GBIF, there are certain steps you can take to make things easier, such as attaching an occurrence data identifier to each occurrence when collecting.

Although it is not necessary, it might be useful to prepare your datasets in a standardised format that meets data requirements or uses a common vocabulary like the Darwin Core terms, which will make the mapping process easier through IPT to GBIF. Using Darwin Core terms gives you the advantage of being able to refer to consistent definitions used throughout the biodiversity community.

A data template is available here: www.sprep.org/attachments/Publications/IOE/Alien%26Invasive-Species-template.xls

Using standard terms is also important for efficient data management. A set of standard definitions and codes useful for GBIF publishing is available here: www.sprep.org/attachments/Publications/IOE/All-lookups.xls.

You now have the tools to prepare, publish, and use shared biodiversity data. We hope this guide encourages you to explore the options available with GBIF to support Pacific biological diversity.

Glossary of definitions

Data Publisher – any organisation / institution that shares data via GBIF.

Darwin Core Terms – Darwin Core Terms are a list of terms or fields that are of interest in the field of Biodiversity. A quick reference guide for Darwin Core terms can be found here: <http://rs.tdwg.org/dwc/terms/>, including definitions and examples of each term. A 2012 paper describing the Darwin Core Standard is here: <https://doi.org/10.1371/journal.pone.0029715>

Integrated Publishing Toolkit (IPT) – A software tool developed by GBIF to carry out dataset publishing, installed on a local or regional server. Participant Nodes often have one or more IPT instances for Data Publishers to use. The IPT we use in this instance is the SPREP IPT.

Metadata – data about the data, including the “who, what, where and when”. For example, if the piece of data that you want to publish is the occurrence of a species in a certain forest in May, your metadata may include who you are (name of your organisation, etc.), the purpose of data collection, methodology, range of sampling dates and geographic scope.

Node Manager – a person who manages the Participant Node. The Node Manager either is or directs the technical point of contact for data publishers.

OccurrenceID – An identifier for the Occurrence (as opposed to a particular digital record of the occurrence). In the absence of a persistent global unique identifier, construct one from a combination of identifiers in the record that will most closely make the occurrenceID globally unique.

Participant Node – a coordinating team (within a country or organisation) that works to establish and strengthen GBIF-related activities.

TaxonID – An identifier for the set of taxon information (data associated with the Taxon class). Maybe a global unique identifier or an identifier specific to the data set.



For More Information

The Battler Resource Base contains information materials and resources for battling invasive species: www.sprep.org/piln/resource-base A basic [template for datasets](#) and a [list of common codes and categories](#) for GBIF use are available from the Battler Resource Base.

You can contact the Invasive Species Programme through the SPREP website: www.sprep.org/Invasive-Species/bem-invasive-species or by e-mailing sprep@sprep.org

Ongoing Support

Tech support – Pacific Node Manager, SPREP, Apia, Samoa: gbifnodemanager@sprep.org

SPREP Invasive Species Team: sprep@sprep.org

GBIF Help Desk: helpdesk@gbif.org

Websites

GBIF homepage	www.gbif.org	Access, explore and publish biodiversity data
Invasive species project dataset list	www.gbif.org/dataset/search?project_id=bid-pa2016-0005-reg	Section of GBIF with invasive species data from the Pacific islands region
Darwin Core Quick Reference	http://rs.tdwg.org/dwc/terms/	Web page with definitions for Darwin Core terms
Darwin Core Hour	https://github.com/tdwg/dwc-qa/wiki/Webinars	Webinars and presentations on Darwin Core
Improving Darwin Core for Invasive Species Research	https://github.com/tdwg/dwc-qa/wiki/Webinars#chapter4	Webinar on what might be done about limitations of Darwin Core terms occurrenceStatus and establishmentMeans when talking about invasive species
Date Parser	http://data.canadensys.net/tools/dates	Tool to determine dates in format yyyy-mm-dd from various other formats
Integrated Publishing Toolkit (IPT)	www.gbif.org/ipt	Description of the Integrated Publishing Toolkit for getting data into GBIF
IPT User Manual	https://github.com/gbif/ipt/wiki/IPT2ManualNotes.wiki	How-to guide to using the Integrated Publishing Toolkit
Universal Unique Identifier (UUID) Generator	www.uuidgenerator.net/	Tool to create UUID global unique identifiers
Copyrights for Data	http://vertnet.org/resources/datalicensingguide.html	Description of the choice of license for published data

Data Use Norms	http://vertnet.org/resources/norms.html	Description of the expected behaviour of a good data-using citizen
Open Refine Basic Use	http://tinyurl.com/ORForBiod	Guide on how to do commonly-needed data cleaning on biodiversity data
Export PDF to Excel	https://acrobat.adobe.com/es/es/acrobat/how-to/pdf-to-excel-xlsx-converter.html	If you do not have a full version of Adobe that can do this, there is a paid service online to do it, with a free trial
Georeferencing Best Practices	www.gbif.org/document/80536/biogeomancer-guide-to-best-practices-in-georeferencing	Description of methods and reasons for georeferencing
Georeferencing Quick Reference	http://manisnet.org/GeoreferencingQuickReferenceGuide.pdf	Description of exactly how to georeference distinct types of location descriptions
InfoXY	http://splink.cria.org.br/infoxy?criaLANG=en	Tool to determine geography from decimalLatitude and decimalLongitude
Coordinate Converter	http://data.canadensys.net/tools/coordinates	Tool to determine decimalLatitude and decimalLongitude from latitude and longitude in other formats
Georeferencing	http://georeferencing.org/index.html	Site with further information about georeferencing





Join the Fight

Protect our islands from invasive species



Håfa Adâi

Aloha

Mogetin

Rahn Anim

Iokwe

Alii

Kaselehlie Len Wo

Mauri

Ekawomir Omo

Mālō te ma'uli

Halo

Tālofa nī

Halo

Tālofa

Halo

Tālofa

Ni sa Bula Fakaalofa lahi atu

Bonjour

Mālō e lelei

Kia Orana

Ia Orana
Bonjour

Hello

Kia Ora

