Manaaki Whenua
Landcare Research

Potential for regional councils to use GBIF to access and share species occurrence data

Envirolink Grant: 2340-ORC006

Prepared for: Otago Regional Council

November 2023

Potential for regional councils to use GBIF to access and share species occurrence data

Contract Report: LC4381

Aaron D. Wilton, Ursula Jewell, Becky Goodsell

Manaaki Whenua – Landcare Research

Reviewed by:	*Approved for release by:*
Dr Jerry Cooper Mycologist and former GBIF-NZ Node Manager Manaaki Whenua – Landcare Research	Dr Geoff Ridley Portfolio Leader – Biota Manaaki Whenua – Landcare Research

Contents

Summary

The Global Biodiversity Information Facility (GBIF) is an international network established to provide open access to biodiversity data from around the world. The vision for GBIF is 'A world in which the best possible biodiversity data underpins research, policy and decisions.' Established in 2001, GBIF now delivers over 2.5 billion species occurrence records globally, including over 13 million records for New Zealand.

GBIF provides a rich, standards-based infrastructure for mobilising and accessing species occurrence data. This infrastructure includes, for example:

- harvesting processes that integrate and validate data from over 90,000 data sets globally
- web services that can be used to access and publish data
- web-based software (the Internet Publishing Toolkit – IPT) to assist data holders publish their data (e.g. through assisting with metadata creation and data mapping)
- online training and learning resources, manuals, and guidelines
- hosted solutions for establishing national or community portals (hosted portals and living atlases) and data publishing (hosted-IPT).

New Zealand has been a participant in GBIF since 2001 but only recently established a web portal (www.gbif.org.nz) and a hosted-IPT instance to assist New Zealand-based data holders (ipt.gbif.org.nz).

Regional council species occurrence data

Twenty-seven exemplar species occurrence data sets were provided by regional council staff to assess for compatibility with GBIF and the requisite data standards (e.g. Darwin Core, Ecological Metadata Language). All the data sets were found to be compatible with the data standards utilised by GBIF and would be appropriate to be published to GBIF. Two generic issues, which could affect the long-term integrity of data, were found across multiple data sets: the lack of persistent unique identifiers, and reliance on vernacular names for recording taxon identifications.

A survey of regional council staff identified several frequently encountered difficulties for these staff associated with species occurrence data. These included issues associated with the discovery and accessing of existing data, and reformatting and integrating data sets, through to sharing data sets with other parties. Nearly all these difficulties could be addressed by integrating GBIF with regional council information systems or processes.

Key recommendation

Regional councils should adopt GBIF as a primary means of preparing, sharing, and accessing publicly available species occurrence data.

1 Introduction

Regional councils collect biodiversity and biosecurity data under the Resource Management Act 1991 and Biosecurity Act 1993. This role for regional councils has expanded under recently released national policy statements for freshwater management and indigenous biodiversity. In the future, biodiversity and biosecurity management will require regional councils to collaborate closely with other organisations and groups in the development of standardised surveillance and monitoring methods, data management, and data sharing.

Increasingly, the need for biodiversity and biosecurity data to be collected, managed, and accessible in standardised ways has been recognised to ensure its quality and allow it to be federated to inform national policy development and state of the environment monitoring (e.g. see Goals 4.1 & 4.2, Te Mana o te Taio – Aotearoa New Zealand Biodiversity Strategy). The Parliamentary Commissioner for the Environment recognises these needs in their reports on the national state of the environment reporting programme and pest plants, as do the regional councils, as evidenced in a Te Uru Kahika | Regional and Unitary Councils Aotearoa think piece on biodiversity and the role of regional councils.[1] Work is underway to build consensus among central and regional government agencies to develop indicators and national scale datasets that will inform progress towards multiple environmental and social outcomes, and focus investment on monitoring and data collection.

New Zealand is not unique in the need to access to biological data in a timely, coordinated and standardised manner. Internationally this has seen the development of standards bodies (e.g., Biodiversity Information Standards[2]) and various initiatives to federate data at different regional scales (e.g., the Atlas of Living Australia[3]). More recently, the Global Biodiversity Information Facility (GBIF) has emerged as a global biodiversity data infrastructure that is supported by many of the world's governments – including New Zealand. GBIF provides a data infrastructure that is networked internationally and aims to ensure 'the best possible biodiversity data underpins research, policy and decisions'. GBIF utilises a federated model with some centralised elements which permits local flexibility and autonomy for data holders whilst providing data holders and users with data aggregation services based around common tools and standards, data integration and quality services, a registry of data holders and their direct data access points, and data access via a common webservice.

Here we present our findings from investigating how the GBIF network and tools could be utilised to enable regional councils to meet their requirements for access to species occurrence to fulfil their biodiversity and biosecurity mandates.

---

[1] G. Willis 2017. Addressing New Zealand's biodiversity challenge: a regional council think piece on the future of biodiversity management in New Zealand. Enfocus.

[2] Biodiversity Information Standards - https://www.tdwg.org/

[3] Atlas of Living Australia (https://www.ala.org.au/)

1.1 Object and scope

The overall goal of the project was to investigate how GBIF could be utilised by regional councils to ensure species occupancy data is accessible in standardised ways to meet biodiversity and biosecurity mandates and enable a federated approach to inform national policy development and state of the environment monitoring. The scope of this reports includes:

- advice for regional councils, based on sample data sets, survey questions, and other interactions with regional council staff
- expertise and first-hand experience with biological data, standards, and GBIF.

After this introduction, section 2 describes the GBIF network and provides a foundation for the analysis in section 3. Within section 2 a list of additional resources is provided at the end of each subsection to assist readers seeking additional detail.

Section 3 covers:

- the results from testing the suitability of sample species occurrence data sets that were received from regional councils
- insights into the needs regional councils have regarding occurrence data, and where they encounter difficulties in obtaining, analysing, sharing, or reporting on the data.

Section 4 provides the conclusions of the report, and the final section outlines key recommendations and potential actions in the form of a draft road map.

2 About GBIF

The Global Biodiversity Information Facility (GBIF) is an international network and data infrastructure that aims to provide anyone, anywhere, with open access to data about Earth's biodiversity.

GBIF arose from a recommendation[4] of the Biodiversity Informatics Subgroup of the OECD's Megascience Forum. The recommendation was to create a mechanism to make biodiversity data more accessible globally, and it was endorsed by the science ministers of the OECD member states. In 2001 GBIF was officially established through a memorandum of understanding[5] between participating governments.

GBIF is funded by the world's governments and is coordinated through its Secretariat, located in Copenhagen. The GBIF network consists of participating countries and organisations that work through participant nodes (e.g. GBIF-NZ). Via the participant nodes, the Secretariat provides data-holding institutions around the world with common standards, best practices, and open-source tools that enable them to share information about where and when species have been recorded; i.e. 'species occurrences'.

The next following summarises some of the key aspects of GBIF.

2.1 Scope of data in GBIF

The core data in GBIF are species occurrences: the occurrence of a species in place and time established through an observation obtained by various methods, or through material evidence (e.g. natural history specimens). GBIF harvests

The GBIF vision

'A world in which the best possible biodiversity data underpins research, policy and decisions.'

The GBIF mission

'To mobilize the data, skills and technologies needed to make comprehensive biodiversity information freely available for science and decisions addressing biodiversity loss and sustainable development.'

https://www.gbif.org/what-is-gbif

Key statistics

Global

107 participants (including NZ)

2,186 publishing institutes

90,161 data sets

2,579,347,923 occurrence records

https://www.gbif.org/

New Zealand

Member since 2001

429 publishers of NZ occurrences

10 publishers within NZ

13,176,012 NZ occurrences

1,605 data sets that include NZ occurrences

https://www.gbif.org.nz/

[4] http://www.oecd.org/science/inno/2105199.pdf

[5] https://www.gbif.org/document/80661

these data from the publishers, integrates the data into a central data structure, then makes the data available via websites, web services, and data downloads.

To support the vision of open global access to these data, GBIF accepts species occurrence data published under three Creative Commons licences:

- CC0: data are made available for any use without restriction
- CC BY: data are made available for any use provided attribution is appropriately given for the sources of data used, in the manner specified by the owner
- CC BY-NC: data are made available for any use provided attribution is appropriately given and provided the use is not for commercial purposes.

GBIF[6] and Creative Commons[7] recommend using the latest version of CC licensing (version 4.0). This aligns with the New Zealand Government Open Access and Licensing (NZGOAL) framework's recommendations[8] for releasing public domain material for reuse by others.

To meet the increasing needs of the GBIF community, GBIF has a work programme that will expand the level of detail that can be included through the development of a new data model.[9] This model is expected to allow publishers to include even richer information alongside their species occurrences. The model is being expanded to support a wider array of the data capture methods (e.g. eDNA and camera traps) used for recording biotic interactions and absence data.

Data sets (often also referred to as 'resources') within GBIF fall into four classes: metadata-only, checklist, occurrence, and sampling event.

- **Metadata only**: resources describe a species occurrence data set that is either undigitised or has yet to be published fully to GBIF. Although not providing the full occurrence data, metadata are a valuable resource for showing that the data set already exists and may be accessible upon request to the data holder, and may also be useful for prioritising data sets for digitisation and/or publication. The metadata standard used for these metadata-only resources is also applied to the other three data set classes.
- **Checklist data set**: this provides a list of the names of organisms for a specific context. The context of each checklist is usually defined by factors such as taxonomic group, geographical extent, and ecological context, but can also include factors such as management or threat status. For example, one checklist might cover the indigenous wetland plants of Canterbury; another might list the bird species in Rotokare Scenic Reserve.
- **Occurrence data set**: these are constructed with a 'core' of occurrence records to which additional information can be linked (see Darwin Core Archive below). Each record details one occurrence, containing multiple data fields that cover (at least)

[6] https://ipt.gbif.org/manual/en/ipt/latest/applying-license
[7] https://wiki.creativecommons.org/wiki/License_Versions#License_Versioning_History
[8] https://www.data.govt.nz/assets/Uploads/nzgoal-version-2-december-2014.pdf
[9] https://www.gbif.org/composition/HjlTr705BctcnaZkcjRJq/gbif-new-data-model

occurrence, identification, locality, and event data. Occurrence data sets are the most frequent data set class in GBIF, and they are particularly suited to mobilising data based on natural history specimens, field observations, and automated camera traps.

- **Sampling-event data set**: these are constructed with a core of sampling events to which species occurrences are linked. Each core record provides details of one sampling event and location. Species observations are linked to these events to provide the occurrence and identification data. Sampling-event data sets are particularly suited to occurrence data obtained through structured ecological investigations or monitoring programmes that are using standard data collection protocols.

It should be noted that occurrence and sampling data sets both utilise Darwin Core fields but differ in the arrangement, or structure, of the data. As a consequence, they have different required and recommended fields.

2.1.1 Additional resources

- NZ Government Open Access Licensing (NZGOAL): https://www.data.govt.nz/toolkit/policies/nzgoal/
- Creative Commons: https://creativecommons.org/
- GBIF Terms of Use: https://www.gbif.org/terms
- GBIF Data Use Agreement: https://www.gbif.org/terms/data-user
- GBIF Data Publisher Agreement: https://www.gbif.org/terms/data-publisher

2.2 Data standards and formats

GBIF utilises a standards-based approach to enable the harvesting and integration of occurrence data sets of varied and variable origins. There are three standards that are most frequently used within the GBIF network: Darwin Core, Ecological Markup Language (EML), and the Darwin Core Archive.

2.2.1 Darwin Core

Darwin Core,[10] sometimes abbreviated as DwC, is a data standard that has been developed by Biodiversity Information Standards (TDWG),[11] an open, international, not-for-profit organisation established to develop and promote the use of standards for recording and sharing data about organisms. Darwin Core was formally ratified by TDWG in 2009 and provides the dictionary of terms that enable sharing information about organisms, their occurrence, and related information. It includes terms (along with their definition and examples) covering multiple aspects of species occurrence data, such as record-level metadata, location information, details of occurrence and observation events,

[10] https://www.tdwg.org/standards/dwc/
[11] https://www.tdwg.org/

identification of the organism, and more (e.g. Figure 1). Darwin Core is being actively maintained and extended by the TDWG community.

GBIF uses Darwin Core as a 'stable, straightforward and flexible framework for compiling biodiversity data'.[12] GBIF has published several vocabularies to support the use of Darwin Core (see http://rs.gbif.org/vocabulary/gbif/).

recordedBy	
Identifier	http://rs.tdwg.org/dwc/terms/recordedBy
Definition	A list (concatenated and separated) of names of people, groups, or organizations responsible for recording the original dwc:Occurrence. The primary collector or observer, especially one who applies a personal identifier (dwc:recordNumber), should be listed first.
Comments	Recommended best practice is to separate the values in a list with space vertical bar space (|). This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.
Examples	`José E. Crespo` `Oliver P. Pearson | Anita K. Pearson` (where the value in recordNumber `OPP 7101` corresponds to the collector number for the specimen in the field catalog of Oliver P. Pearson)

Figure 1. The term 'recordedBy' from the Darwin Core Quick Reference Guide. (Source: TDWG, https://dwc.tdwg.org/terms/#dwc:recordedBy, licensed under CC BY 4.0)

2.2.2 Ecological Metadata Language (EML)

Ecological Metadata Language (EML)[13] is a metadata standard developed for recording information about ecological data sets in a series of modular and extensible XML document types. EML is an open-source standard that is administered and maintained by the Knowledge Network for Biocomplexity.[14] The EML modules allow the description of multiple facets of a data set, including, for example, the scope or extent of the data, the methods and protocols used to collect and analysis the data, any associated resources, and parties associated with the data.

GBIF utilises EML to describe all data sets within the network, and each Darwin Core Archive (see below) includes an EML file as one of its components.

2.2.3 Darwin Core Archive

Darwin Core Archive (sometimes abbreviated as DwC-A) is the preferred format for publishing data in the GBIF network. DwC-A is a GBIF specification for a self-contained

[12] https://www.gbif.org/standards
[13] https://eml.ecoinformatics.org/
[14] https://knb.ecoinformatics.org/

data set consisting of the metadata and data files, which are arranged using a star-schema approach (Figure 2). The four types of file in the archive are as follows.

- **Core data file**: the main or central data file, containing sampling-event, occurrence or checklist data. This file is formatted as a comma- (CSV) or tab-separated value (TSV) text file, with each record on a new row and consisting of Darwin Core terms that are separated using commas or tabs respectively. For examples, see Figures 8, 10 and 12.
- **Extension files:** optional data files that contain additional data that link to the records in the core file. These are also CSV or TSV files, which consist of data mapped to Darwin Core or other data standards (e.g. Audiovisual Core Multimedia Resources Metadata Schema[15]). The list of extensions available is maintained in the GBIF Extension Repository.[16]
- **Metafile** (meta.xml in Figure 2): an XML-formatted file that describes the other files in the archive. For each file it maps the data columns in the core and extension files to a Darwin Core or Extension term.
- **Resource metadata** (EML.xml in Figure 2): an XML file that records a description of the data set using EML (see above).

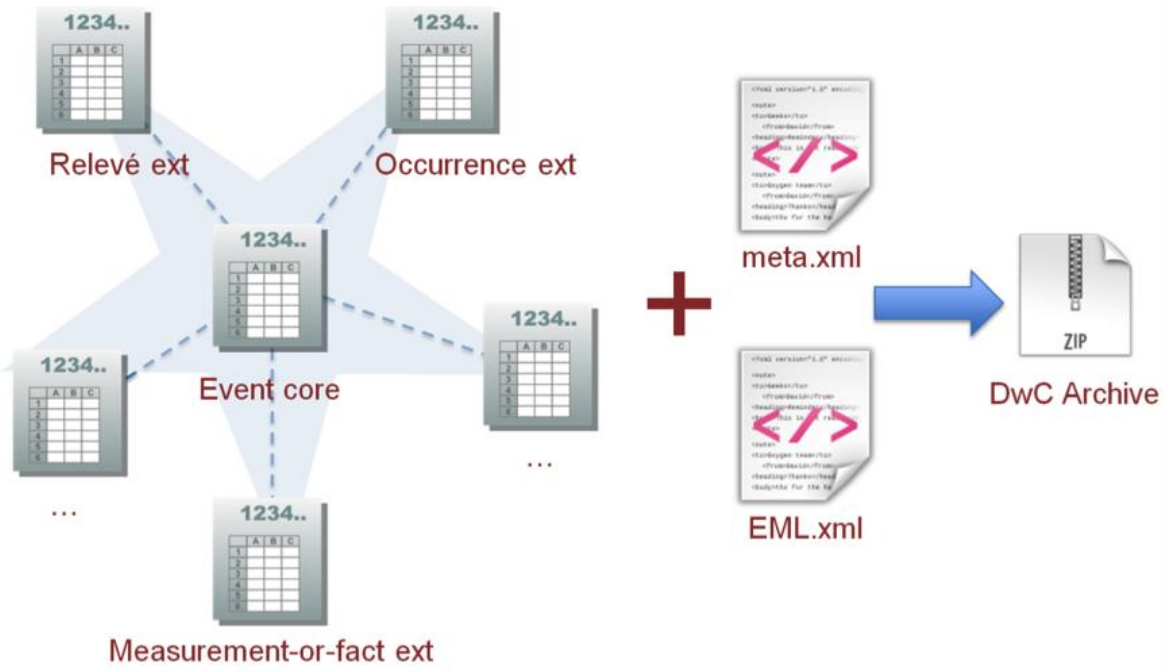


Figure 2. Structure and typical contents of a Darwin Core Archive.
(Source: GBIF IPT Manual, https://ipt.gbif.org/manual/en/ipt/latest/dwca-guide, CC-BY 4.0)

[15] https://www.tdwg.org/standards/ac/
[16] https://rs.gbif.org/extensions.html

2.2.4 Additional resources

- What is Darwin Core and why does it matter? (https://www.gbif.org/darwin-core)
- GBIF vocabularies: http://rs.gbif.org/vocabulary/, particularly http://rs.gbif.org/vocabulary/gbif/
- digital object identifier (DOI) https://www.doi.org/.

2.3 Publishing data to GBIF

The most common method of publishing data is as Darwin Core Archive files generated using an Integrated Publishing Toolkit (IPT). It is also possible to publish data to GBIF using other methods, such as the GBIF API (Figure 4), or by creating Darwin Core Archives using other processes.

2.3.1 Integrated Publishing Toolkit (IPT)

The Integrated Publishing Toolkit (usually called IPT) is a free toolkit that data holders can use to organise and share their data about biological organisms. IPT is a web-based tool that has been created, and is maintained, by the GBIF Secretariat.

IPT helps data holders to document (i.e. add metadata) and structure their data, then publish the data as a Darwin Core Archive. It provides a series of interfaces that leads a **resource manager** through the process of creating a resource and associating it with a publishing organisation, adding metadata, linking to the data sources (which may be based on file or database sources) for the resource, and then mapping the data onto the selected IPT data core and extensions.

The interfaces also allow the user to preview the raw and mapped data, create a Darwin Core Archive, and publish and register the resource with GBIF. While a Darwin Core Archive is being created, IPT validates the resource and provides information on any issues encountered. Until resources are set to public and published, they are only accessible to the resource author, the IPT instance administrator, and any registered users the resource author has added to that particular resource.

Resource managers may be configured with or without publication rights, allowing multiple people without publication rights to collaborate to prepare a data set while restricting the publication privilege to nominated resource managers. In some circumstances it may be necessary (e.g. security policy, hosting arrangements) or more convenient (e.g. to restructure data) to export data from an internal system before it is added to an IPT resource.

Each IPT installation has at least one person in an **administrator** role. The administrator has responsibility for creating and managing user accounts and for configuring the IPT instance. Each IPT installation can be configured to support multiple publishing organisations and retain a specified number of versions for each resource. The administrator also manages the IPT data cores and extensions that are available on that IPT installation.

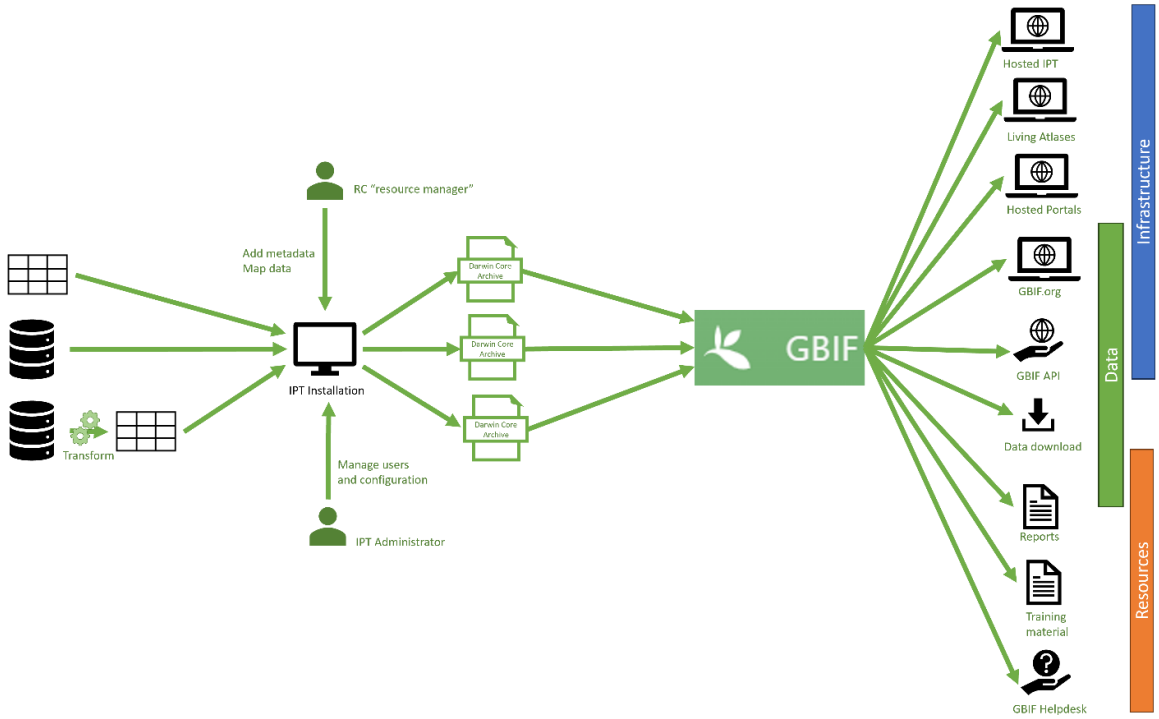


Figure 3. A conceptual overview of the GBIF network, showing publication using IPT, through to data, resources, and infrastructure provided by GBIF.

IPT is well documented, with a comprehensive manual and associated tools (see 'Additional information' below).

2.3.2 IPT deployment

IPT can be utilised and deployed in different ways depending on the ability or desire of an organisation to install and maintain it. A publisher with good levels of technical support may choose to stand up their own installation of IPT (**self-hosted** in Figure 4). Those with lower levels of technical support (which may incur high IT costs) or who are at the start of the process of becoming data publishers may choose to temporarily or permanently utilise a hosted IPT installation. These installations can be hosted by another data publisher (**hosted installation** in Figure 4) or a participant node (**node-hosted** in Figure 4).

During 2023, GBIF-NZ has worked with the Secretariat to establish a node-hosted instance of IPT for New Zealand.[17] This instance is administered by GBIF-NZ while being hosted in the GBIF infrastructure and receiving technical support (e.g. software updates) from the Secretariat. As at the end of September 2023 this installation is available to New Zealand-based publishers.

[17] https://ipt.gbif.org.nz/

It should be noted that resources published using one installation of IPT can be transferred to a different installation if this becomes necessary, or is desired by the data publisher, at a later time.

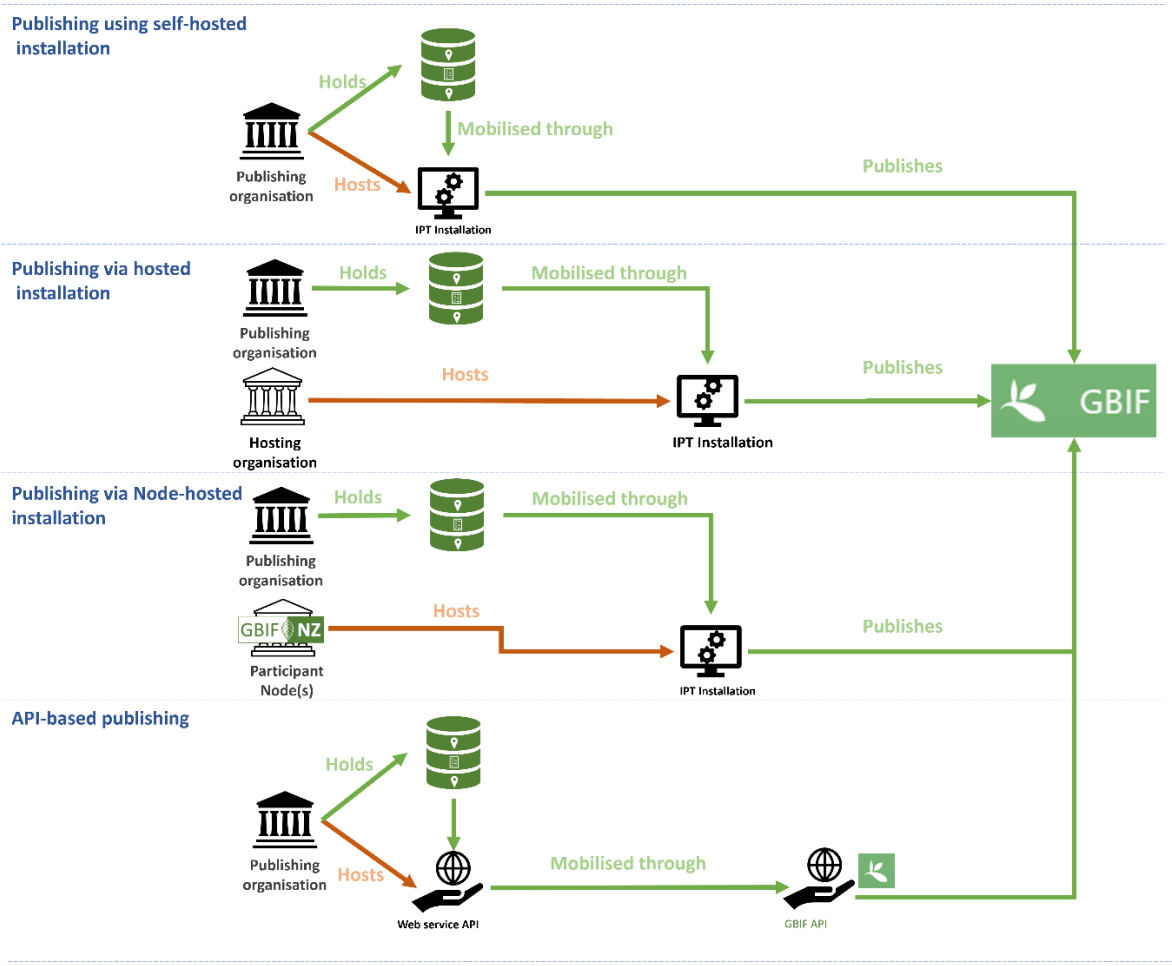


Figure 4. Summary of different approaches to publishing data to GBIF using IPT or the GBIF API.

2.3.3 Becoming a publisher

Publication of data is open to any organisation that meets a simple set of requirements (e.g. a stable arrangement for data hosting) and receives endorsement from the relevant node (i.e. GBIF-NZ for New Zealand organisations)[18] and agrees to the GBIF Data Publisher Agreement.[19] Application to become a publisher is made using a simple online process.[18]

[18] https://www.gbif.org/become-a-publisher
[19] https://www.gbif.org/terms/data-publisher

2.3.4 Additional information

- IPT Manual: https://ipt.gbif.org/manual/en/ipt/latest/. This manual extends beyond IPT and includes, for example, links to templates and exemplar data sets (see the section 'How to publish biodiversity data through GBIF.org' (https://ipt.gbif.org/manual/en/ipt/latest/how-to-publish) .
- Data quality requirements:
  - checklist data set: https://www.gbif.org/data-quality-requirements-checklists
  - occurrence data set: https://www.gbif.org/data-quality-requirements-occurrences
  - sampling-event data set: https://www.gbif.org/data-quality-requirements-sampling-events
- Online DwC-A validator: https://www.gbif.org/tools/data-validator
- GBIF API: https://www.gbif.org/developer/summary
- GBIF Terms of Use: https://www.gbif.org/terms

2.4 Infrastructure and services

In addition to the infrastructure described above, GBIF provides other tools and services. These are briefly outlined below.

- **Hosted portals**[20]**:** GBIF has developed, maintains, and hosts a web-portal infrastructure that provides a simple way for participant nodes, or other communities, to establish a website for their node that delivers species occurrence data, alongside supporting content and branding created by the node participants for their community. This infrastructure has been adopted by multiple countries and groups, including GBIF-NZ.[21]
- **IPT Hosting:** GBIF offers cloud-hosted instances of IPT for participants unable to access another hosting solution or who lack the infrastructure to host their own IPT instance. GBIF-NZ has a hosted IPT[22] instance that is available to New Zealand-based data holders to publish their data.
- **Training and learning:** The GBIF Secretariat manages a wealth of training and learning materials developed by GBIF staff in collaboration with the GBIF community.
- **Global Registry of Scientific Collections** (GRSciColl)[23]: This 'is a comprehensive and community-curated clearing house of information about scientific collections in the GBIF registry'.[24]

[20] https://www.gbif.org/hosted-portals
[21] https://www.gbif.org.nz
[22] https://ipt.gbif.org.nz
[23] https://scientific-collections.gbif.org/
[24] https://scientific-collections.gbif.org/about

- **Data access tools:** GBIF maintains a list of tools that facilitate data access and analysis.[25] These include, for example, an R library (rgbif[26]) and a python library (pygbif[27]) for accessing data from the GBIF API.

2.4.1 Additional resources

- Data standards: https://www.gbif.org/standards
- IPT manual: https://ipt.gbif.org/manual/en
- GBIF metadata overview: https://ipt.gbif.org/manual/en/ipt/latest/gbif-metadata-profile
- Derived data sets: https://data-blog.gbif.org/post/derived-datasets/

2.5 GBIF in New Zealand

New Zealand has been a participant in GBIF since 2001 and established a national node, GBIF-NZ, in 2002. GBIF-NZ supports the mobilisation of species occurrence data held by New Zealand organisations and the use of GBIF-mediated biodiversity data about New Zealand's biota.

Funding for New Zealand's membership of GBIF is provided through the Strategic Science Investment Fund, administered by the Ministry for Business, Innovation & Employment (MBIE). MBIE is also responsible for appointing the Head of Delegation and Node Manager, which are the formal roles required for New Zealand to participate in the GBIF network.

In 2021 GBIF-NZ participated in GBIF's hosted portals[28] initiative, resulting in the development and publication of the GBIF-NZ portal.[29] GBIF-NZ hopes this portal, which is hosted on GBIF infrastructure, will raise awareness and use of the biodiversity data that are being mobilised, help stimulate the development of a community of biodiversity data users and publishers, and act as a stepping-stone to establishing a Living Atlas[30] for New Zealand.

GBIF-NZ has worked with the GBIF Secretariat to establish a national hosted IPT installation.[31] This installation is administrated by GBIF-NZ, on infrastructure that is provided and maintained by the GBIF Secretariat. This instance enables New Zealand-based organisations to mobilise data using IPT without having to set up and maintain an IPT instance themselves. GBIF-NZ hopes this will remove a key barrier to any New Zealand-based organisations seeking to mobilise their biodiversity data.

---

[25] https://www.gbif.org/resource/search?contentType=tool
[26] https://www.gbif.org/tool/81747/rgbif
[27] https://www.gbif.org/tool/OlyoYyRbKCSCkMKIi4oIT/pygbif-gbif-python-client
[28] https://www.gbif.org/composition/3kQFinjwHbCGZeLb5OhwN2/gbif-hosted-portals
[29] https://www.gbif.org.nz
[30] https://living-atlases.gbif.org/
[31] https://ipt.gbif.org.nz/

2.5.1 New Zealand data publishers

As noted early, the majority of New Zealand species occurrences records available via GBIF are sourced from New Zealand based data holders (Figure 5).[32] These providers are currently Crown Research Institutes, Museums, and community initiatives (Figure 6). However, this composition is expected to change significantly over the next few years. For example, GBIF-NZ has recently approved two new data publishers – Antarctica New Zealand and wildlife.ai – who are working towards publishing their first data sets. In addition, GBIF-NZ has had discussions with data holders from other sectors and organisations, including MPI (contact: Michael Berardozzi) and DOC (contact: Meredith McKay), investigating the potential to use of GBIF to publish and/or access species occurrence data.

Figure 5. The number of New Zealand species occurrence records available via GBIF according to the publishing country. (Data accessed: 11 November 2023, https://doi.org/10.15468/dl.he43sa)

[32] The definition of New Zealand providers is based on the country of publication provided by the data holder when publishing the data set, even if the underpinning information infrastructure resides overseas (e.g. eBird).

Figure 6. The number of records contributed by New Zealand based data providers.[33] (Data accessed: 11 November 2023, https://doi.org/10.15468/dl.he43sa)

3 Analysis

3.1 Compatibility of regional council data with GBIF

The primary goal in assessing the ability of regional councils to mobilise their species occurrence data using the GBIF network is to ascertain the alignment of these data with GBIF data structure, standards, and tools, as well as the GBIF goals. Regional council staff were invited to submit exemplar data sets so that we could evaluate their compatibility with the GBIF network.

Twenty-seven data sets were received from seven regional councils (Appendix 2), ranging from structured monitoring and survey data, to trap data, pest surveillance, and species lists. The exemplar data sets included presence, presence and absence and quantitative occurrence based records. The data sets provided were all found to be compatible with the Darwin Core standard used by GBIF, and our evaluation found that 19 (70%) of the data sets would be suitable to be published in the GBIF network.

Six (23%) were judged likely to be fully compatible with GBIF, but additional data or metadata are required to enable a more complete assessment. One of these data sets contained only a few columns but is likely to be suitable to be published as a checklist for a specified area.

Two (7%) of the data sets, while they could be mapped to Darwin Core, consisted of highly derived data aggregated from multiple sources, including sources outside of the council's

[33] New Zealand based providers follows the country information included in the GBIF dataset registration, even if the underpinning information infrastructure resides overseas (e.g. eBird).

region. Before being published as a primary resource to GBIF, these data sets would require additional considerations, in particular:

- appropriate permissions – do the data licences and/or permissions from the original data holder(s) permit the data to be published to GBIF under a Creative Commons licence?
- derived data – publication of derived data to GBIF requires careful consideration as it may introduce data duplication and other undesirable data artifacts (e.g. degradation of data through gridding of data); and if the original data are already available in GBIF, should the derived data sets also be published as a primary data set? Are there subsets of the data that may not otherwise be published to GBIF?

This second consideration (derived data sets) does not prohibit making the data set accessible via GBIF in other ways. For example, this could be achieved by:

- registering and uploading the data sets as derived data[34] within the GBIF network – derived data sets enable the sources of information that were used to create the data set to be acknowledged, and for users of the derived data set to correctly acknowledge and cite the derived data set.
- creating a Darwin Core Archive using IPT, but (given that it is not primary data) not publishing that resource to GBIF, and instead publishing a metadata-only resource to advertise the availability of the data set, which would ensure the data set is compatible with, and easily usable alongside, other GBIF-sourced data.

Three of the exemplar data sets that were representative of the biodiversity and biosecurity data sets provided were selected to be imported into an isolated test instance of IPT as part of our evaluation and to demonstrate the output of that process:

- eDNA data set from Otago Regional Council (Figures 8 and 9);
- ungulate monitoring data set from Environment Canterbury (Figures 10 and 11);
- bird monitoring data set from Environment Canterbury (Figures 12 and 13).

The first two data sets were mapped to a (sampling) event core, while the bird data were mapped to an occurrence core (Figure 7). To map the data into an IPT, some simple transformations of the data sets was required using spreadsheet functions, including:

- mapping column names to Darwin Core
- stacking or unstacking data
- adding scientific names
- adding unique identifiers (see below).

By using cell formulas these transformations could be templated, and rapidly applied and reused for other data sets with the same structure. Within the exemplar data sets we noted some common issues, which are outlined below.

[34] https://www.gbif.org/derived-dataset/about

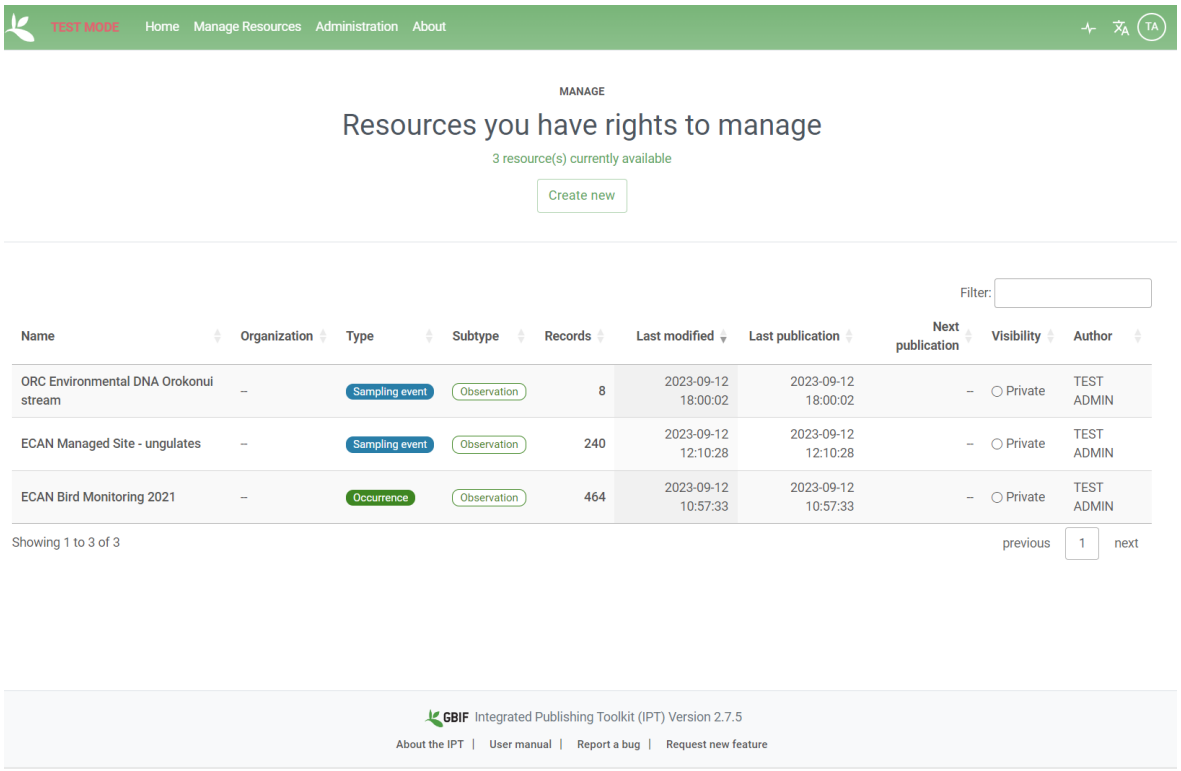


Figure 7. The resource management page on the test ITP installation showing the three resources based on one Otago Regional Council (ORC) and two Environment Canterbury (CAN) data sets.

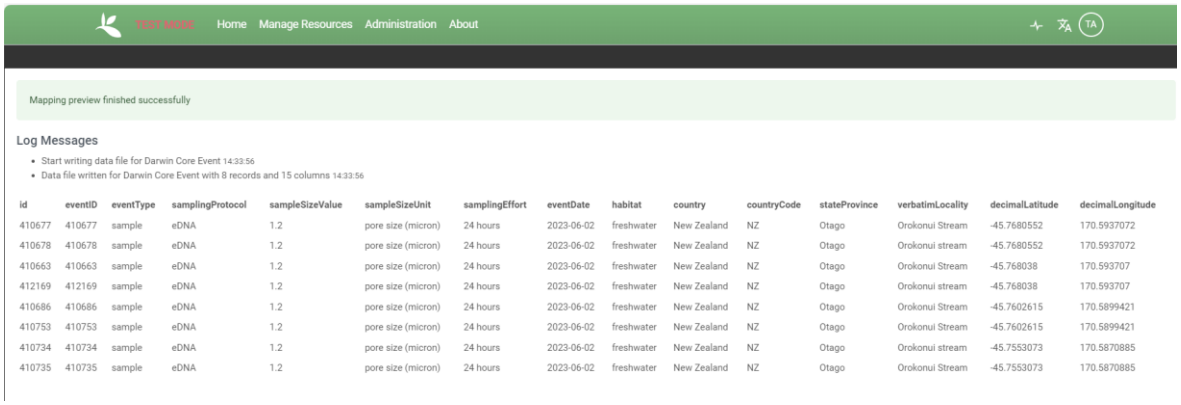
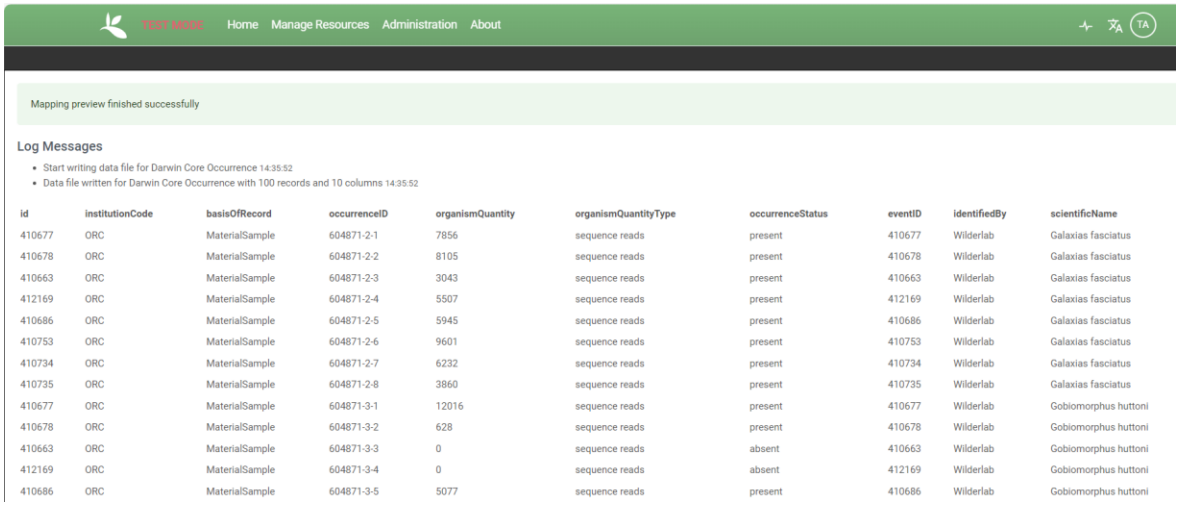


Figure 8. IPT event data preview of the Otago Regional Council eDNA data set into Darwin Core.



TEST MODE Home Manage Resources Administration About

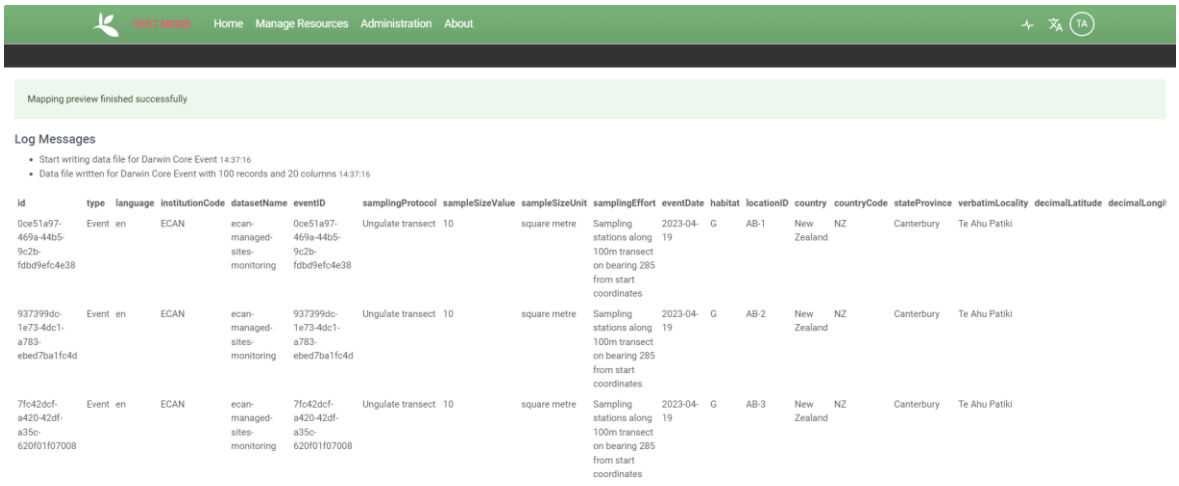
Mapping preview finished successfully

Log Messages

- Start writing data file for Darwin Core Occurrence 14:35:52
- Data file written for Darwin Core Occurrence with 100 records and 10 columns 14:35:52

id	institutionCode	basisOfRecord	occurrenceID	organismQuantity	organismQuantityType	occurrenceStatus	eventID	identifiedBy	scientificName
410677	ORC	MaterialSample	604871-2-1	7856	sequence reads	present	410677	Wilderlab	Galaxias fasciatus
410678	ORC	MaterialSample	604871-2-2	8105	sequence reads	present	410678	Wilderlab	Galaxias fasciatus
410663	ORC	MaterialSample	604871-2-3	3043	sequence reads	present	410663	Wilderlab	Galaxias fasciatus
412169	ORC	MaterialSample	604871-2-4	5507	sequence reads	present	412169	Wilderlab	Galaxias fasciatus
410686	ORC	MaterialSample	604871-2-5	5945	sequence reads	present	410686	Wilderlab	Galaxias fasciatus
410753	ORC	MaterialSample	604871-2-6	9601	sequence reads	present	410753	Wilderlab	Galaxias fasciatus
410734	ORC	MaterialSample	604871-2-7	6232	sequence reads	present	410734	Wilderlab	Galaxias fasciatus
410735	ORC	MaterialSample	604871-2-8	3860	sequence reads	present	410735	Wilderlab	Galaxias fasciatus
410677	ORC	MaterialSample	604871-3-1	12016	sequence reads	present	410677	Wilderlab	Gobiomorphus huttoni
410678	ORC	MaterialSample	604871-3-2	628	sequence reads	present	410678	Wilderlab	Gobiomorphus huttoni
410663	ORC	MaterialSample	604871-3-3	0	sequence reads	absent	410663	Wilderlab	Gobiomorphus huttoni
412169	ORC	MaterialSample	604871-3-4	0	sequence reads	absent	412169	Wilderlab	Gobiomorphus huttoni
410686	ORC	MaterialSample	604871-3-5	5077	sequence reads	present	410686	Wilderlab	Gobiomorphus huttoni

Figure 9. IPT occurrence data preview of the Otago Regional Council eDNA data set into Darwin Core using IPT.



TEST MODE Home Manage Resources Administration About

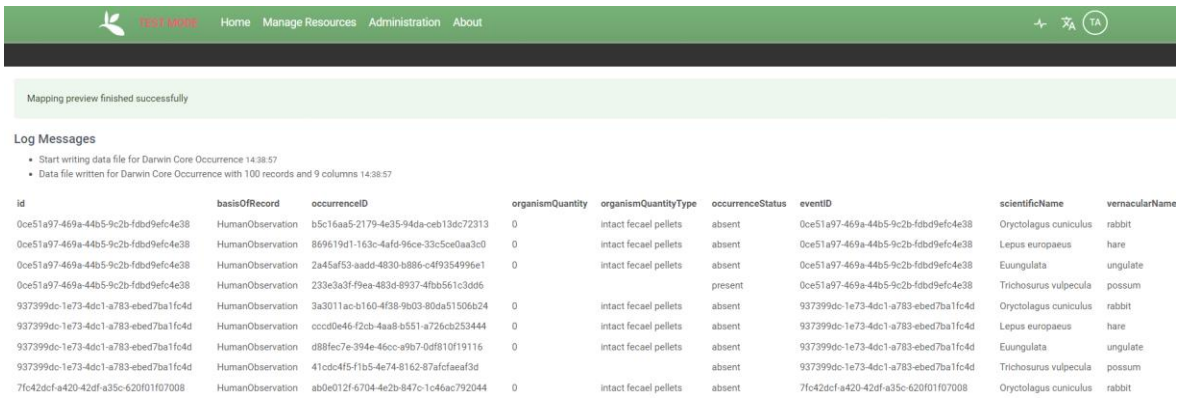
Mapping preview finished successfully

Log Messages

- Start writing data file for Darwin Core Event 14:37:16
- Data file written for Darwin Core Event with 100 records and 20 columns 14:37:16

id	type	language	institutionCode	datasetName	eventID	samplingProtocol	sampleSizeValue	sampleSizeUnit	samplingEffort	eventDate	habitat	locationID	country	countryCode	stateProvince	verbatimLocality	decimalLatitude	decimalLongi
0ce51a97-469a-44b5-9c2b-fdbd9efc4e38	Event	en	ECAN	ecan-managed-sites-monitoring	0ce51a97-469a-44b5-9c2b-fdbd9efc4e38	Ungulate transect	10	square metre	Sampling stations along 100m transect on bearing 285 from start coordinates	2023-04-19	G	AB-1	New Zealand	NZ	Canterbury	Te Ahu Patiki		
937399dc-1e73-4dc1-a783-ebed7ba1fc4d	Event	en	ECAN	ecan-managed-sites-monitoring	937399dc-1e73-4dc1-a783-ebed7ba1fc4d	Ungulate transect	10	square metre	Sampling stations along 100m transect on bearing 285 from start coordinates	2023-04-19	G	AB-2	New Zealand	NZ	Canterbury	Te Ahu Patiki		
7fc42dcf-a420-42df-a35c-620f01f07008	Event	en	ECAN	ecan-managed-sites-monitoring	7fc42dcf-a420-42df-a35c-620f01f07008	Ungulate transect	10	square metre	Sampling stations along 100m transect on bearing 285 from start coordinates	2023-04-19	G	AB-3	New Zealand	NZ	Canterbury	Te Ahu Patiki		

Figure 10. IPT event data preview of the Environment Canterbury managed sites ungulates monitoring data.



TEST MODE Home Manage Resources Administration About

Mapping preview finished successfully

Log Messages

- Start writing data file for Darwin Core Occurrence 14:38:57
- Data file written for Darwin Core Occurrence with 100 records and 9 columns 14:38:57

id	basisOfRecord	occurrenceID	organismQuantity	organismQuantityType	occurrenceStatus	eventID	scientificName	vernacularName
0ce51a97-469a-44b5-9c2b-fdbd9efc4e38	HumanObservation	b5c16aa5-2179-4e35-94da-ceb13dc72313	0	intact fecael pellets	absent	0ce51a97-469a-44b5-9c2b-fdbd9efc4e38	Oryctolagus cuniculus	rabbit
0ce51a97-469a-44b5-9c2b-fdbd9efc4e38	HumanObservation	869619d1-163c-4afd-96ce-33c5ce0aa3c0	0	intact fecael pellets	absent	0ce51a97-469a-44b5-9c2b-fdbd9efc4e38	Lepus europaeus	hare
0ce51a97-469a-44b5-9c2b-fdbd9efc4e38	HumanObservation	2a45af53-aadd-4830-b886-c4f9354996e1	0	intact fecael pellets	absent	0ce51a97-469a-44b5-9c2b-fdbd9efc4e38	Euungulata	ungulate
0ce51a97-469a-44b5-9c2b-fdbd9efc4e38	HumanObservation	233e3a3f-f9ea-483d-8937-4fbb561c3dd6			present	0ce51a97-469a-44b5-9c2b-fdbd9efc4e38	Trichosurus vulpecula	possum
937399dc-1e73-4dc1-a783-ebed7ba1fc4d	HumanObservation	3a3011ac-b160-4f38-9b03-80da51506b24	0	intact fecael pellets	absent	937399dc-1e73-4dc1-a783-ebed7ba1fc4d	Oryctolagus cuniculus	rabbit
937399dc-1e73-4dc1-a783-ebed7ba1fc4d	HumanObservation	cccd0e46-f2cb-4aa8-b551-a726cb253444	0	intact fecael pellets	absent	937399dc-1e73-4dc1-a783-ebed7ba1fc4d	Lepus europaeus	hare
937399dc-1e73-4dc1-a783-ebed7ba1fc4d	HumanObservation	d88fec7e-394e-46cc-a9b7-0df810f19116	0	intact fecael pellets	absent	937399dc-1e73-4dc1-a783-ebed7ba1fc4d	Euungulata	ungulate
937399dc-1e73-4dc1-a783-ebed7ba1fc4d	HumanObservation	41cdc4f5-f1b5-4e74-8162-87afcfaeaf3d			absent	937399dc-1e73-4dc1-a783-ebed7ba1fc4d	Trichosurus vulpecula	possum
7fc42dcf-a420-42df-a35c-620f01f07008	HumanObservation	ab0e012f-6704-4e2b-847c-1c46ac792044	0	intact fecael pellets	absent	7fc42dcf-a420-42df-a35c-620f01f07008	Oryctolagus cuniculus	rabbit

Figure 11. IPT occurrence data preview of the Environment Canterbury managed sites ungulates monitoring data.

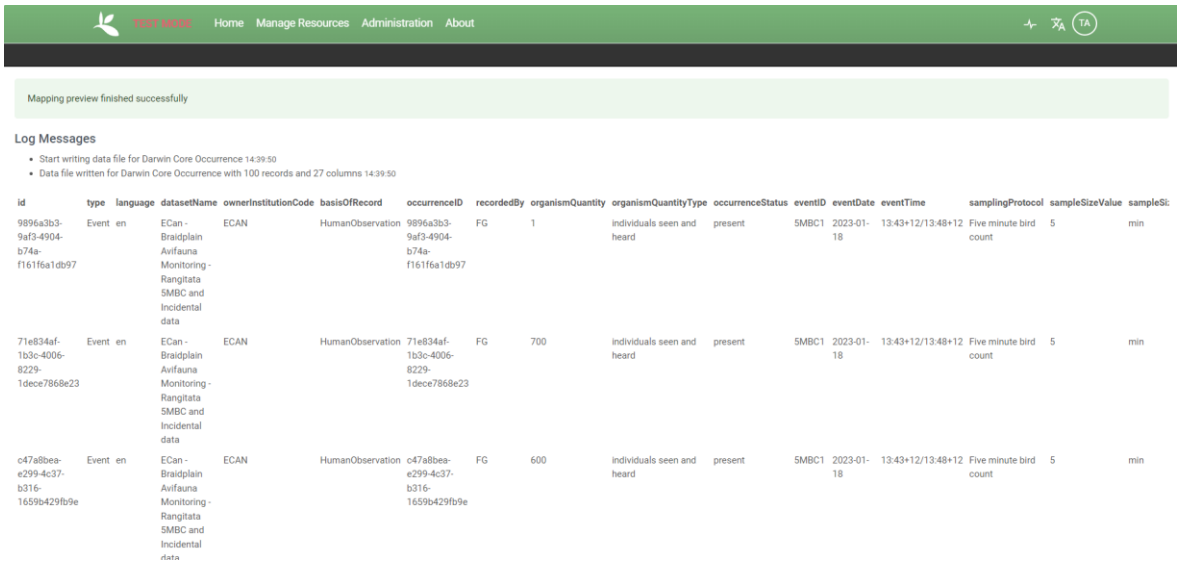


Figure 12. IPT occurrence data preview of the Environment Canterbury bird monitoring sample data.

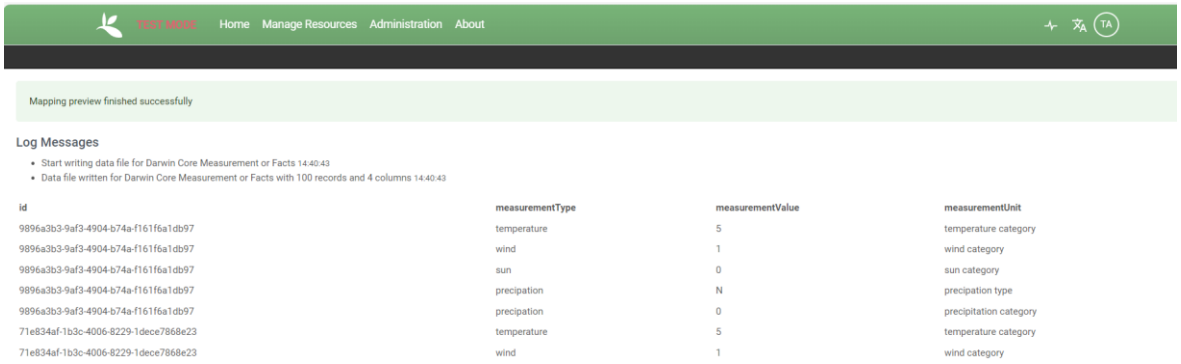


Figure 13. IPT preview of the environmental observations associated with the Environment Canterbury bird monitoring.

3.1.1 Lack of persistent unique identifiers

Many of the data sets lacked any form of persistent unique identifier for records and digital objects. Data published within the GBIF network will ideally include persistent global unique identifiers for data objects. These identifiers are essential for activities such as re-indexing, linkage of associated data, and citation of data. They can also be useful for enabling detection of duplication and tracking the provenance of data. As a minimum, identifiers need to be unique within a single data set, and persistent (i.e. an identifier stays with, and refers to, the same record).

Persistent unique identifiers can be formed in many different ways, but a common recommendation is to use a system that relies on universal unique identifiers (UUIDs, e.g. c821a27f-8ff8-4dd2-9597-8a8dcB80fd7d is the persistent UUID assigned to a specimen at the Allan Herbarium with catalogue number CHR 92742). A key aspect for consideration is

that once assigned, the identifiers are stable and maintained with the digital object/record (ideally in the primary data repository).

To incorporate the three exemplar data sets into an IPT, UUIDs were added using the Excel formulas provided in Appendix 3. However, the formula results in a dynamic UUID that changes frequently based on events in the spreadsheet; this was made persistent by copying and pasting the UUID back into the same cells as a value.

3.1.2 Georeference coordinate data

The exemplar data sets examined nearly all contained high-precision georeference coordinates recorded as New Zealand Transverse Mercator (NZTM). NZTM, and georeferences using other coordinate systems, can be mapped into the Darwin Core standard (e.g. using verbatim coordinate fields). However, given that GBIF is a global resource, it is recommended that whenever georeference data are available they should also be included as decimal latitude and longitude values with a stated datum (e.g. WGS84) and (optional) an uncertainty measure.

Provision of georeference data as decimal latitude and longitude enables easier use of data, because users do not need to transform values from a various region/national specific projections, or from historical coordinate systems (e.g. NZMS1 and NZMG within New Zealand). Further, GBIF utilises these decimal latitude and longitude coordinates to undertake various data quality checks as part of the aggregation process.

3.1.3 Taxon names

Within the exemplar data sets approximately half captured the taxon identification using only vernacular names (also referred to as common names)[35] or taxon codes. To publish data sets to GBIF, the vernacular names or codes need to be supplemented with scientific names at the applicable taxonomic rank (i.e., in some cases this may be a species bionomical, but in others only the name of a genus or other higher rank).

More generally, although vernacular names and codes may provide a convenient handle for capturing data, their use as the only method for permanently recording species identifications can be problematic for the reuse, integration, and long-term storage of data. Vernacular names can be problematic because the application of a vernacular name is frequently ambiguous, for a number of reasons.

- A taxon may have more than one vernacular name. For example, *Acaena anserinifolia* (J.R.Forst. & G.Forst.) J.B.Armstr. has been recorded as having variously been assigned 10 vernacular names: bidibid, huruhuru-o-hine-nui-te-pō, hutiwai, kaiā, kaiārurerure, kaikaiā, kaikaiārure, pirikahuk piripiri, piriwhetau.
- A single vernacular name may be used for more than one taxon. For example, puka has variously been applied to *Brassica oleracea* L., *Syzygium maire* (A.Cunn.) Sykes &

[35] 'Vernacular name' is used here to refer to any informal name, in any language, used for a taxon.

Garn.-Jones, *Meryta sinclairii* (Hook.f.) Seem., *Muehlenbeckia australis* (G.Forst.) Meisn., and *Elaeocarpus hookerianus* Raoul.

- Use of vernacular names is highly dependent on the context of the space, time, and culture of a particular community.
- Vernacular names, and their spelling and application, are not governed by a formal code – these being the remit of the community using the name.

The potential ambiguity of vernacular names means that the integration of data based on them is likely to be problematic, particularly when integrating data sets of differing age and provenance.

Scientific names also change over time as result of systematic research and nomenclatural process. However, they are governed by formal codes that result in a link between the names being documented, providing significantly less ambiguity in comparison to vernacular names.

In addition to the issues noted above, vernacular and scientific names both suffer from high rates of transcription error, often requiring complex or manual processing to integrate data fully. The use of taxon dictionaries, particularly those that assign permanent unique identifiers to names (e.g. NZOR[36]), is highly recommended.

3.2 Pain points

Understanding the difficulties – so called 'pain points' – that regional council staff are encountering when working with species occurrence data is another important aspect of understanding the utility of the GBIF network to regional council staff. Some pain points were experienced by a high proportion of the regional council staff responding to the survey (Figure 14). These ranged equally across the categories of access and sharing of data, preparing data for use, and the presence of adequate metadata.

We believe a widespread use of the GBIF network would directly address all of these pain points, with one exception: access to technical assistance to wrangle, analyse, or visualise data would not be directly addressed by use of GBIF. However, using GBIF would create opportunities for assistance from peers, reduce handling of data (because it will become available in a common format and standards), and create the potential to use, or develop, shared tools for analysis, visualisation or training (e.g. the Atlas of Living Australia spatial portal guides[37]).

The pain points are covered further in the remainder of this section.

[36] NZOR, the New Zealand Organisms Register (https://nzor.org.nz/), is an initiative to provide an integrated source of the names and taxonomy of the organisms found in, or otherwise relevant to, New Zealand.
[37] How to guides : ALA Support (https://support.ala.org.au/support/solutions/folders/6000234079)

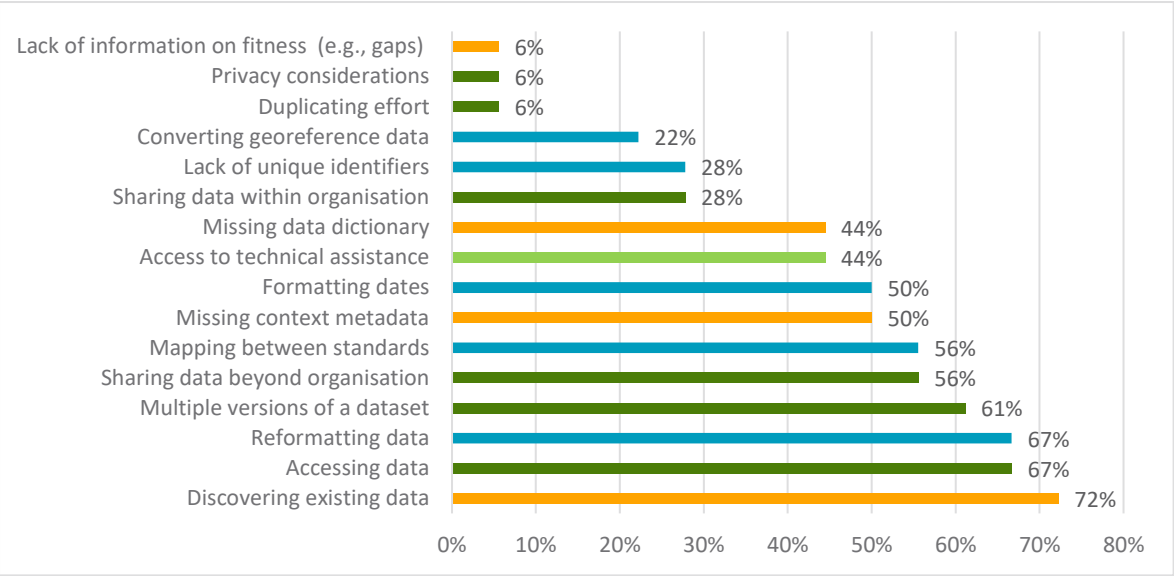


Figure 14. Percentage of surveyed regional council staff experiencing each pain point associated with species occurrence data, categorised by data access and sharing (dark green), data preparation and handling (blue), data discovery (orange), and capability (light green).

3.3 Discovering species occurrence data

Discovering existing species occurrence data was the most frequent pain point encountered by regional council staff: 72% of survey respondents identified this as a pain point they had experienced (Figure 14). The survey results also indicate that the level of awareness or visibility of data is an issue across a range of organisational contexts (Figure 15).

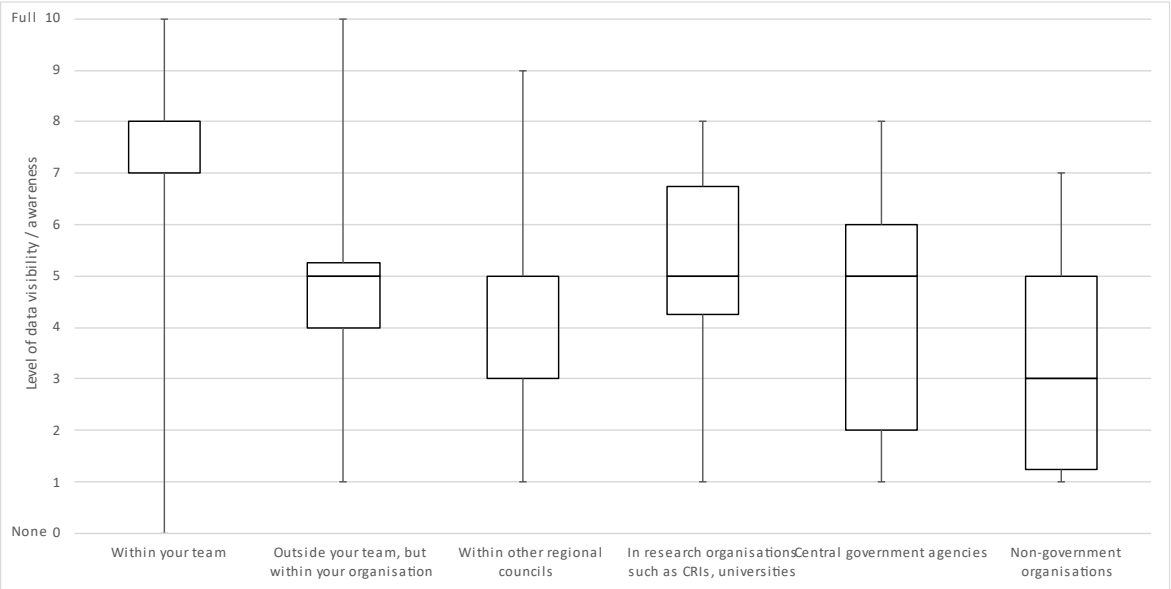


Figure 15. Summary of survey responses to the perceived level of awareness of species occurrence data sets in differing organisational contexts. Respondents rated their perceived awareness from no visibility (0) through to full visibility or awareness (10).

The inability to discover existing species occurrence data is likely to be a key contributor to the 6% of respondents indicating duplication of data collection as a significant pain point.

The survey respondents also indicated a need for data across different spatial and temporal scales. Over 20% of respondents required data at a global scale (Figure 16), while over 60% required data with an age spanning from the current year through to 10 or more years old (Figure 17).

Figure 16. The percentage of survey respondents requiring data at different spatial scales.

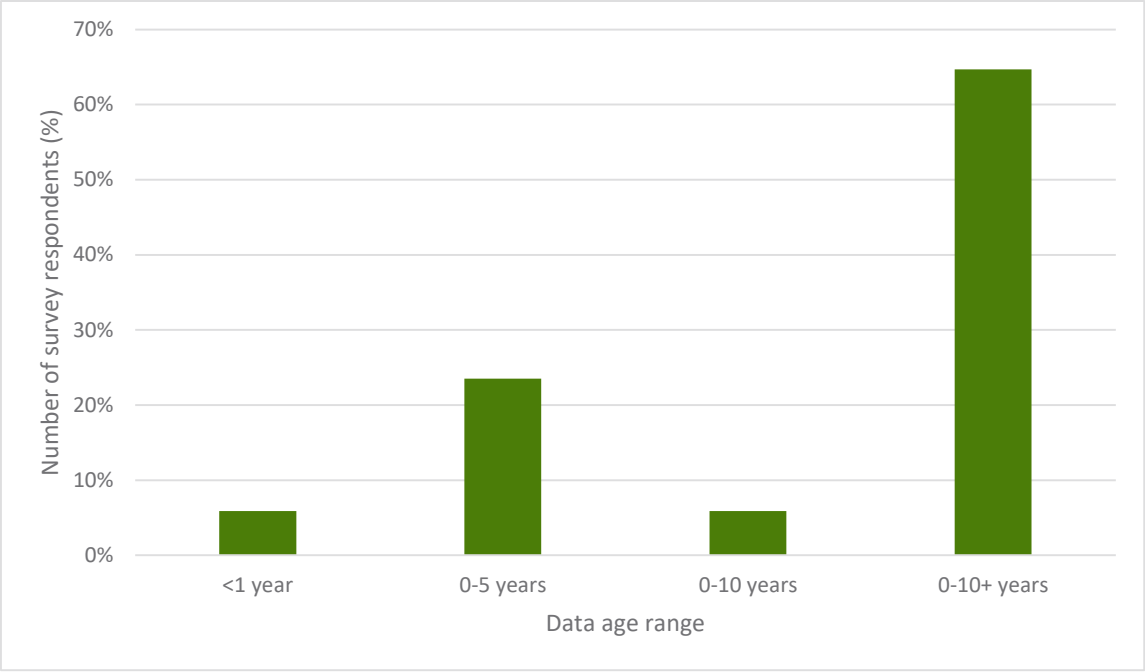


Figure 17. The number (%) of survey respondents requiring data with differing temporal scopes. Note that time periods overlap.

Only 50% of survey respondents had used GBIF, yet an unrefined search of GBIF[38] for "otago", limited to records using New Zealand as country filter, showed that a significant number of resources are already potentially available via GBIF (Table 1). Note that these numbers represent an unrefined total and will vary when the records are filtered (for example, using the taxonomic group of interest or various data quality metrics), depending on the specificity and required application of the data.

Table 1. Counts of the resources available in GBIF resulting from a search for "otago" and filtered to New Zealand (as at 10 November 2023)

Resource type		
Occurrences	Number of records	873,479
	Date range of the occurrences	1800 to 2023
Data sets		479
Publishers		233

GBIF is not limited to fully digitised occurrence data sets. Metadata-only resources can also be published to GBIF to document the existence of species occurrence data sets that are not yet able to be published in full. This could increase the awareness of, for example, 'analogue' data sets that are yet to be digitised, data sets where there are insufficient resources to publish fully, or data sets with high sensitivity or restrictions that means they cannot be published in full in GBIF.

With widespread adoption GBIF could become the go-to system for species occurrence data. It can provide a means to discover and access species occurrence data across differing temporal and spatial scales, and differing data sources from other regional councils in New Zealand, other New Zealand data publishers, and other data publishers around the world.

[38] https://www.gbif.org/occurrence/search?q=otago&country=NZ

Perspective: Use of GBIF for invasive species mapping

Morgan Shields, Biosecurity Advisor for invasive species, Environment Canterbury

'As part of my role, I use GBIF almost daily, to indicate the distributions of potentially invasive species as this is a critical component of prioritising species for surveillance and future management. This will enable the regional council to strategically respond to potentially high impact invasive species before they become the next gorse or wilding conifer.

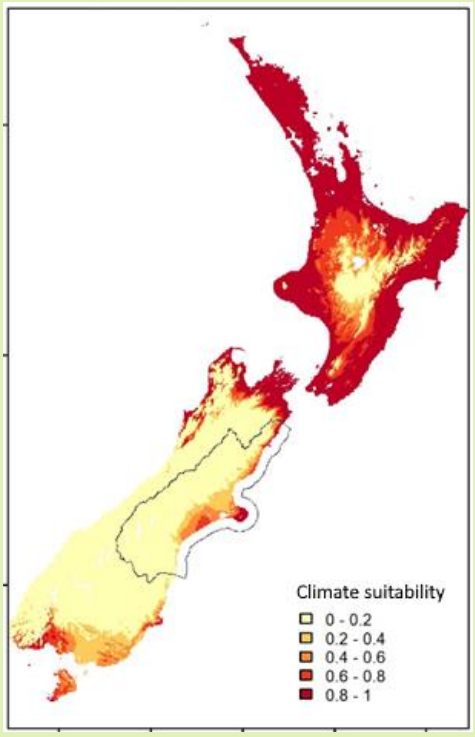


Figure 18. Palm grass (*Setaria palmifolia*) predicted 2070 climate range using GBIF occurrence data. (0.8-1 optimal suitability). Image: Environment Canterbury & Manaaki Whenua Landcare Research 2023.

The strength of GBIF is that it provides reliable coordinates and species identification at local, national and global scale from a multitude of sources ... regardless of the type of organism. This means a picture of a species distribution can often be formed [using] only one platform with a multitude of data filtering utility. GBIF [occurrence] data has also proved essential for climate niche modelling to suggest where exotic and native species could survive now and in future climate scenarios in Aotearoa/New Zealand so different regional councils can react accordingly. Furthermore, the ease of access to historical observations on GBIF has enabled surveillance of invasive species sites that Environment Canterbury would not have otherwise been aware of.

The data sharing challenge was highlighted in the 2021 Parliamentary Commission for the Environment 'space invaders' report on environmental weeds by Simon Upton and while arguments have been made that regional councils do not benefit from sharing their data such as on GBIF, this is incorrect. Regional councils are already directly or indirectly benefiting from ecological research and applications that use GBIF data e.g., weed biological control programmes. However, GBIF-derived benefits for regional councils and the nation could be greatly enhanced by sharing some of the observational data those councils are already recording. These benefits underpin both biodiversity and biosecurity management decisions and do not have to be restricted by territorial borders. They may include accurate species distribution data at varying spatial scales indicating temporal changes, greatly enhanced predictive power of climate change effects on fauna and flora populations and more targeted surveillance and pathway management opportunities.

While there [are] substantial challenges for regional councils to overcome, by preparing for future integration of GBIF with council databases, where appropriate, these challenges can be overcome."

3.4 Accessing species occurrence data

Accessing existing species occurrence data was a pain point frequently encountered by regional council staff: 67% of survey respondents identified this as a pain point (Figure 14). Respondents indicated that they needed to use a variety of means to access data (Table 2), and the need for more timely responses to data requests was raised multiple times during online discussions. Almost half the respondents (47%) lodge a formal request with the data holder to access data, and 72% rely on their peer network to gain data. Both methods come with an administrative burden and are time-consuming.

Table 2. Frequency of methods used by survey respondents to obtain species occurrence data

Method of obtaining data	Percentage of respondents
My team or I collect it directly	94
Peer network	72
Internal information system, file system or data catalogue	72
Download via a website or web service	61
Formal request to the data holder(s)	44
Other methods	22

An additional difficulty encountered by respondents when accessing data is the existence of multiple versions of a data set: just over 60% of respondents encountered this issue.

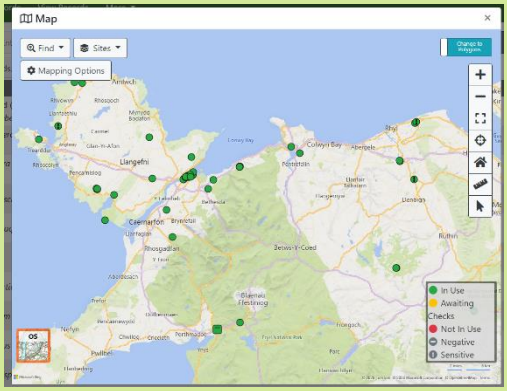
GBIF provides free and open access to biodiversity data via data downloads and the GBIF API. It is also possible to directly access data sets from IPT instances when that method has been used to publish the data by the data holder. Thus, GBIF offers immediacy of access to the incorporated data: there is no need to use peer networks or make formal requests. When this immediacy is combined with the expectation that data holders will maintain their data sets (as specified in the publisher agreement), GBIF becomes a primary access point and reduces the risk of data being stored in multiple places and in multiple versions.

In addition, GBIF issues a unique identifier (in the form of a digital object identifier, or doi) to all data set downloads and allows recover of the data set for up to 6 months by using that doi. After 6 months GBIF retains the query used, the number of records, the data sets they came from, the licence, and the EML metadata. This service means it is possible for a data user – or indeed other parties wanting to use or verify the same data – to later retrieve the data set or reuse the query parameters.

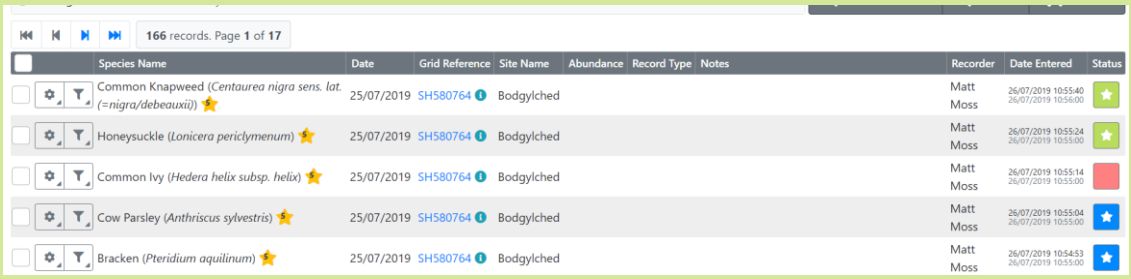
Perspective: occurrence data for ecological assessments
Matt Moss, Ecologist, Tasman District Council

"Any ecological assessment requires a desktop survey to see what protected sites/species have been previously recorded in the area. Ideally, ecologists would search local and national datasets as part of this process and discuss the results accordingly in their report. This then grants ecologists reason to undertake specialist surveys for species or require specific mitigation for the works as a precautionary measure – for example, threatened or protected species.

I previously worked as an Ecological Consultant in the UK, where Local Data Centres hold biological information and provide it on request. They often charge a small fee but the data is well maintained and validated by local experts. You'd receive a shapefile and excel sheet of data (see images). Many of these centres work in combination with national providers to ensure the data is shared accordingly."



In New Zealand, "I've found access to data requires using multiple resources (e.g. eBird and iNaturalist) and there are often issues of data validation. Councils, DOC, CRIs hold valuable data but may not be contacted either because the ecologist lacks knowledge of the data resource or perceives resistance to share the data. Coming from my previous consultancy experience, I realise that unless you know who to ask you may not be able to get appropriate data. Having all this data in one place would ensure ecological assessments can be done to a high standard. It also means [an] ecologist could refer to a centralised location for data rather than asking multiple providers for this – or be unsure of who to ask."


166 records. Page 1 of 17

Species Name	Date	Grid Reference	Site Name	Abundance	Record Type	Notes	Recorder	Date Entered	Status
Common Knapweed (*Centaurea nigra sens. lat. (=nigra/debeauxii)*)	25/07/2019	SH580764	Bodgylched				Matt Moss	26/07/2019 10:55:40 26/07/2019 10:56:00	
Honeysuckle (*Lonicera periclymenum*)	25/07/2019	SH580764	Bodgylched				Matt Moss	26/07/2019 10:55:24 26/07/2019 10:55:00	
Common Ivy (*Hedera helix subsp. helix*)	25/07/2019	SH580764	Bodgylched				Matt Moss	26/07/2019 10:55:14 26/07/2019 10:55:00	
Cow Parsley (*Anthriscus sylvestris*)	25/07/2019	SH580764	Bodgylched				Matt Moss	26/07/2019 10:55:04 26/07/2019 10:55:00	
Bracken (*Pteridium aquilinum*)	25/07/2019	SH580764	Bodgylched				Matt Moss	26/07/2019 10:54:53 26/07/2019 10:55:00	

3.5 Data integration and use

The ability to combine data sets is important to regional councils: the survey showed that half the respondents need to combine different data sets, and often need to undertake different data preparation steps (Table 3). Despite this, only 22% of respondents indicated that they were receiving data sets to a known data standard, and 83% indicated that

currently metadata is only 'sometimes' included with a supplied data set. Survey respondents encountered difficulties due to missing metadata (50% of respondents) or missing data dictionaries (44% of respondents).

Table 3. The percentage of survey respondents undertaking different data preparation steps

Process	Percentage of respondents
Combine it without transformation with other data	44
Summarise it and then combine it with other data	50
Re-format the data to match another structure	50
Map or transform the data into another data standard	39
Do data validations	28

Survey respondents were asked about three specific aspects of species occurrence tasks: formatting dates, converting georeference coordinates, and the lack of unique identifiers. These three difficulties were reported as issues by 50%, 22%, and 28% of respondents, respectively.

GBIF's primary service is the aggregation and integration of species occurrence data, and subsequent delivery of the data in known formats and data standards. The ability to access data to a consistent format and standard should facilitate the use of species occurrence data. Here are some of the benefits.

- Regional council occurrence data obtained via GBIF would all be to a consistent standard and format, including date and georeferenced coordinate fields, and will have been assigned unique identifiers.
- A consistent format enables repeatable data processing steps to be developed and applied to the data with confidence.
- Consistent availability of associated metadata enables the data user to understand the provenance of the data, and therefore its suitability for their particular use.

3.6 Data provision

Sharing data for which they are responsible was a pain point encountered by survey respondents when sharing the data either within or outside their organisation (28% and 56% of respondents, respectively). Several common concerns were identified by respondents when needing to share data (Table 4).

Table 4. Frequency of concerns reported by survey respondents when publishing or providing their data set to another party.

Concern	Percentage of respondents
Tracking and reporting on the use of the data	33
Suitably formatting the data set	39
Needing to add metadata	39
Quality of my data set(s)	50
Time to prepare the data set	50
Time to explain or document the data set	56
Authorisation to provide the data	56
IP, privacy or other legal considerations	56

GBIF provides a variety of tools, training materials, and guidelines to support data holders to publish their species occurrence data. These have the potential to at least reduce the impact of these publishing concerns, as follows.

- GBIF supports the tracking of published data sets based on the DOIs issued when data sets are downloaded and subsequently cited in papers or reports. It is also possible to manually add data set usage when a particular use may not be detected by GBIF.
- GBIF supports the preparation of data sets through the provision of well-documented and consistent standards for data and metadata, and tools. The use of IPT also provides an easy-to-use interface to write and update metadata, and to map data to the appropriate fields.
- GBIF's tools can help data holders to identify potential data quality issues before publication using the online validator[39] (Figure 19), and to explore and address quality issues after publication by viewing the details in the data set metrics or on the occurrence records themselves. Data quality tests are also included in the Darwin Core Archive downloads.
- Using an IPT, a data holder is quickly able to create a new resource using the metadata from an existing data set as a starting point, avoiding the need to re-enter metadata when creating several similar resources.

[39] https://www.gbif.org/tools/data-validator

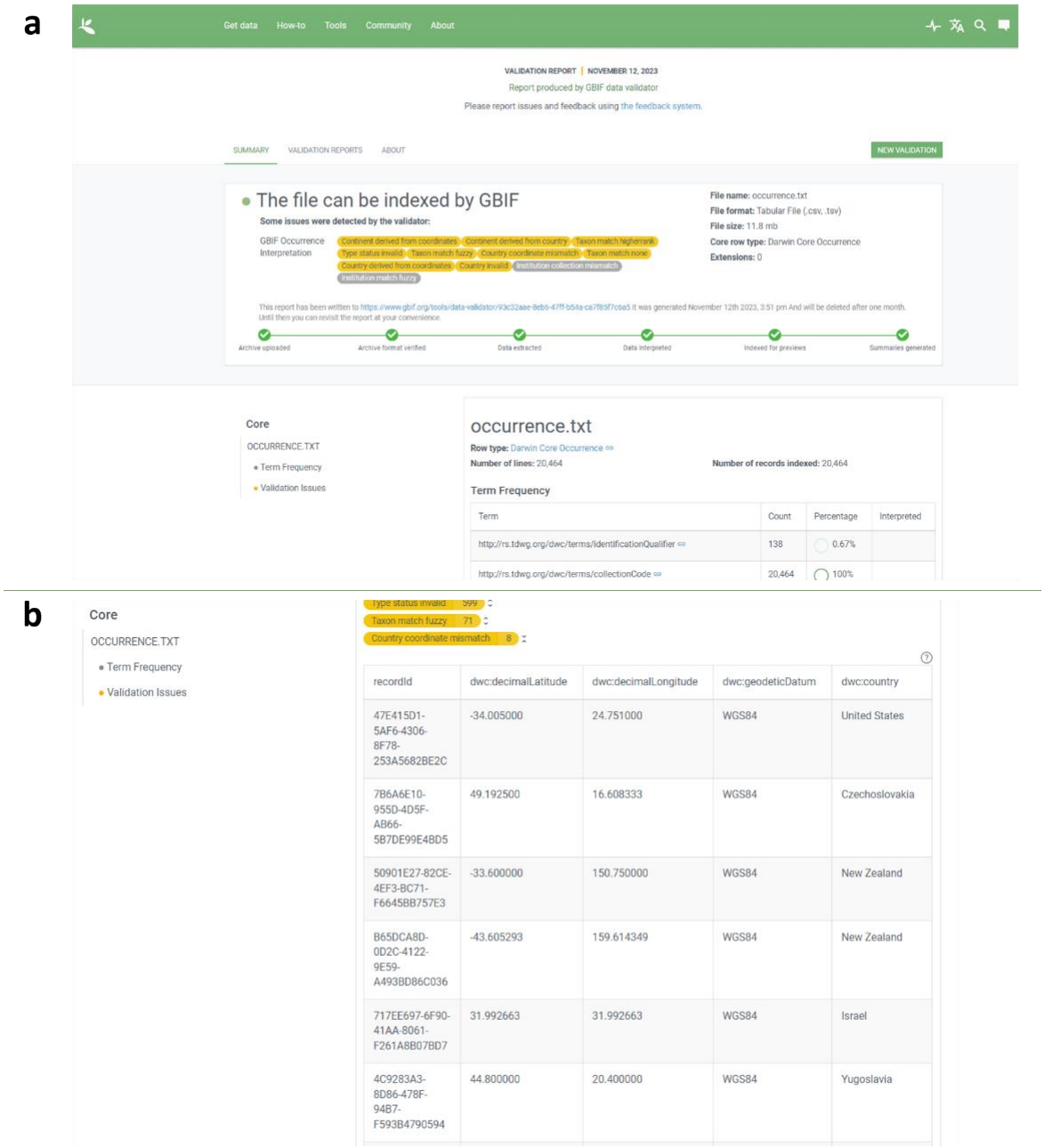


Figure 19. An example report from the GBIF online data validator showing (a) the report summary and (b) details for a specific test (country coordinate mismatch).

3.7 Sensitive data

Sensitivity of species occurrence records may result, for example, from the particular taxa being recorded (e.g. observations of rare and threatened species, species of biosecurity concern, taonga species), the process of collecting the data (e.g. privacy of the observer), or the location of the observation (e.g. private land or land with other restrictions). These and other data sensitivities were invariably mentioned during the discussions with regional council staff. This high degree of awareness is reflected in the survey, with 56% of respondents indicating concern about data sensitivities when publishing data (Table 4).

Since its establishment, GBIF has been concerned about the unprotected distribution of sensitive species occurrence data. In 2006 GBIF initiated a work programme on sensitive data based on taxon sensitivity. This resulted in the publication of a best practice guide for generalising data,[40] which has recently been revised.[41] Although focused on taxon-based sensitivities, many of the considerations – particularly the methods for generalising data – can be applied to other contexts. GBIF encourages the publication of species occurrence data as openly as possible, yet at the same time 'respect[s] the wishes of data providers to restrict information on sensitive taxa'.[42]

Within the GBIF network the federated approach to data publishing ensure that data holders retain the primary copy of the data, and can selectively publish a version of the data, with appropriate stakeholder engagement, to the GBIF network. The data sets that are published are made available under Creative Commons licensing, which allows for data to be shared and reused under CC0 1.0, CC BY 4.0, or CC BY-NC 4.0. The data standards used within the GBIF network allow for omission or generalisation or removal of data and provide ways of recording these actions at both the data set and record level.

Fine-grained obfuscation of sensitive data can be applied by modifying fields across the whole data set (Figure 20 top), an entire record (Figure 20 bottom), or at the level of field plus record (Figure 20 middle).

At the data set level there are several potential approaches to safeguarding sensitive data, aside from full withholding of the data set from the GBIF network (Figure 21). These include:

- publishing the data set as a metadata-only record, to permit discovery and consultation
- using an IPT to create a Darwin Core Archive, then downloading the archive to secure storage and removing the resource from the IPT (the metadata file from this archive could be used to create a metadata-only record, and the archive used to provide data in a known and consistent format when requested for legitimate uses)
- creating private resources in an IPT that require a login and permissions to access or establish a second IPT instance that also requires web-server authentication to access.

It is also possible to mix these strategies; for example, by creating subsets of a single species occurrence data set based on data sensitivities for publication. This would allow non-sensitive data to be published fully to GBIF while safeguarding sensitive data.

[40] Guide to Best Practices for Generalising Sensitive Species-Occurrence Data 2008
[41] Current Best Practices for Generalizing Sensitive Species Occurrence Data 2023
[42] https://docs.gbif.org/sensitive-species-best-practices/master/en/#introduction

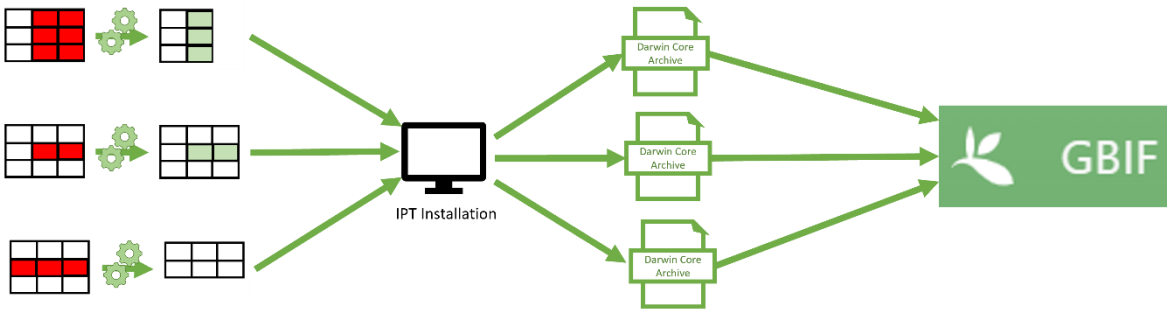


Figure 20. Potential approaches to publishing data containing sensitive fields, fields within some records, and records.

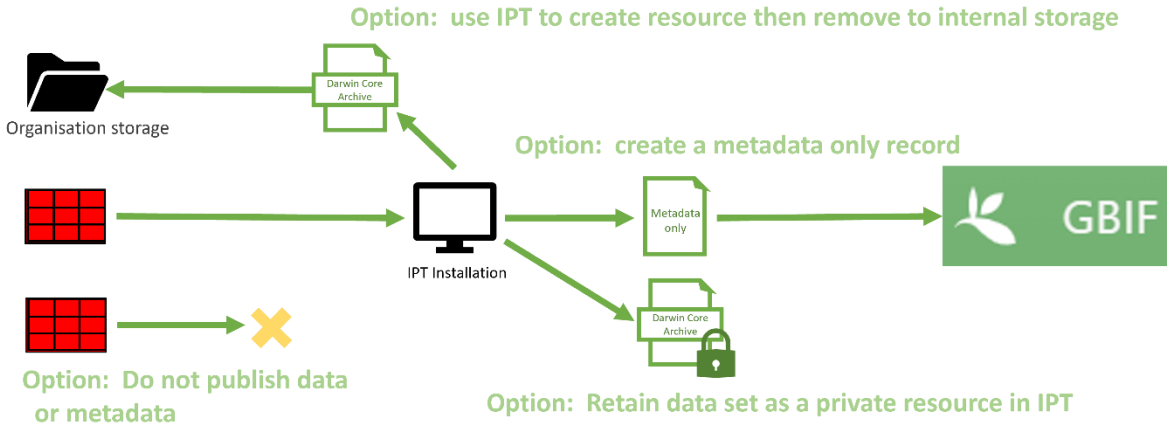


Figure 21. Potential approaches for sensitive data sets.

3.8 Indigenous data sovereignty and governance

Indigenous data sovereignty and governance were raised during discussions and regional council staff emphasized the importance of their relationships with mana whenua. Comprehensive coverage of this topic is beyond the scope of this report, however some key points in the context of GBIF are provided below.

A key aspect of GBIF in this regard is the federated architecture applied to data publishing. This architecture ensures that data holders have full local autonomy and flexibility as to what data they publish. This will enable councils to respect and reflect the interests of iwi and hapu in the data they publish, both in terms of data sets published, data included and in the data set metadata (see Sensitive data above for examples approaches that could be applied). In addition, the primary version of data and the intellectual property is retained by the data holder with only a Creative Commons license to permission use applied. Further, the ability to create metadata only and private resources using IPT offers a mechanism that could potentially be used to provide iwi and hapu with information about existing data sets as part of an ongoing partnership.

GBIF is supportive of indigenous data governance initiatives with work items on the implementation of the CARE principles[43] included in the work plan for 2023/24[44], and active discussions within the GBIF community of ensuring indigenous interests in data are maintained when data are published and aggregated. These community discussions are currently focussed on the Traditional Knowledge and Biocultural Labels and Notices that have been developed by Local Contexts.[45] These are arguably one of the first practical methods that can be used to allow indigenous communities to assert their cultural authority in physical collections and data. Manaaki Whenua is part of this GBIF community discussion and recently partnered with Local Contexts and four iwi to test the application of Biocultural Labels to the biological collections held at Manaaki Whenua. This partnership resulted in the successful addition of Biocultural Labels to records collected within from the rohe of each iwi (example shown in Figure 22), and we are now working to ensure the Local Contexts Labels and Notices are retained when data are published to GBIF and any in any other context. While this information can already be included in published data sets, this is dependent on using generalised fields (e.g., dynamicProperties[46] in the Darwin Core standard) so a more specific and nuanced methodology is being sought.

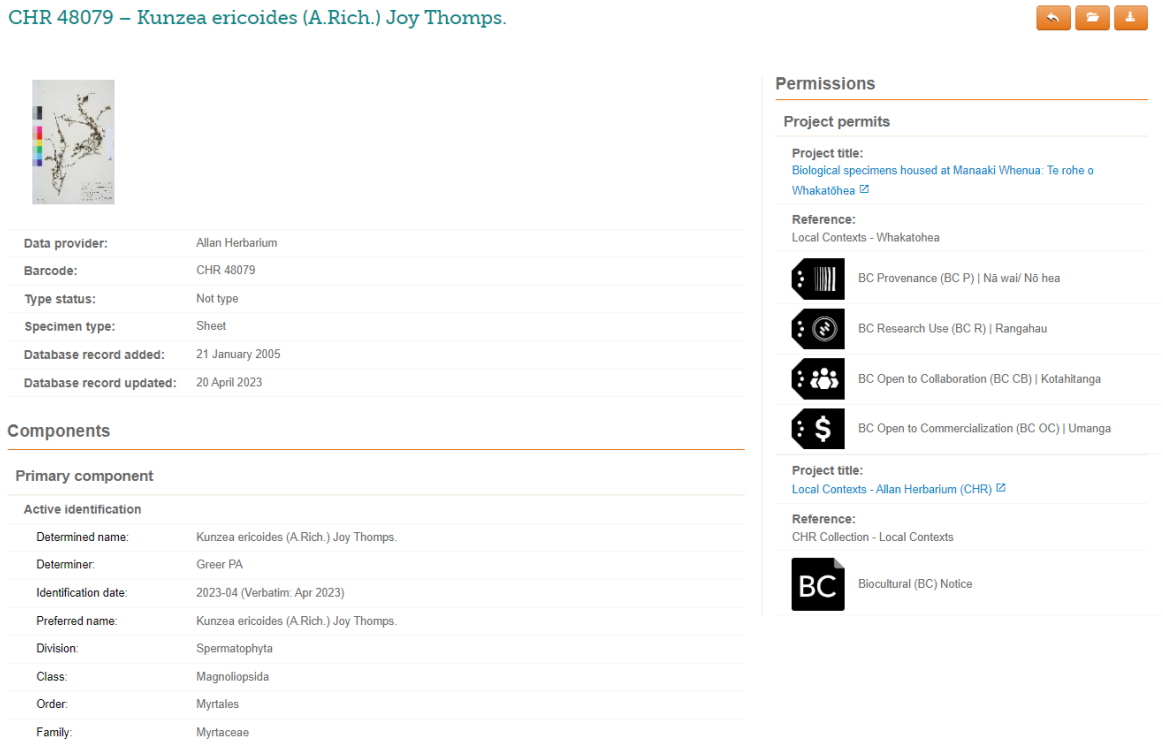


Figure 22. Partial screenshot of a specimen record from Manaaki Whenua – Landcare Research's collections data web site displaying Biocultural Labels and Notices on the right-hand side. (Source: Manaaki Whenua (2023) Systematics Collections Data. https://scd.landcareresearch.co.nz/Specimen/CHR%2048079)

[43] Global Indigenous Data Alliance (GIDA) – CARE principles - https://www.gida-global.org/care
[44] The 2023/24 GBIF work plan is currently in draft form and will be published after consultation with Participants.
[45] Local Contexts - https://localcontexts.org/
[46] Darwin Core Guide – dynamicProperties definition - https://dwc.tdwg.org/terms/#dwc:dynamicProperties

3.9 Data quality

Data quality was only identified as a difficulty by 6% of survey respondents when accessing data, yet 50% of respondents were concerned about the quality of their own data when publishing or providing it to another person. As noted above ('3.6 Data provision'), GBIF provides various tools and guides to assist data holders to determine the quality of their data. It also applies a number of data quality tests to aggregated data, resulting in tags on occurrence records that are available to filter the data.

Users of GBIF should note that the provenance of data contributed to GBIF ranges from citizen science through natural history collections to machine observations. Provenance is well documented, and records are tagged with metadata and other fields that assist with the filtering of data.

Although GBIF provides extensive data, there might still be gaps in certain regions or species affecting the comprehensiveness of information for local decision-making. It can be expected that these gaps will be addressed, at least in part, as the scope of data holders publishing to GBIF increases. A widespread adoption of GBIF would also allow any meta-analysis undertaken to more accurately identify gaps and prioritise them for additional data gathering.

3.10 Capability and capacity

Forty-four percent of survey respondents indicated that access to technical assistance to support their use of species occurrence data is a frequent difficulty. As noted above, GBIF does not directly address the lack of capability and capacity. However, utilising GBIF can connect regional councils with national and international experts, facilitating knowledge exchange and collaboration on biodiversity and biosecurity challenges.

There are also other aspects of the GBIF network that support access to, or development of, capability and capacity.

- Within the organisation model developed by GBIF it is intended that each participant node, if sufficiently resourced, will provide support to data publishers and users within their country.
- GBIF provides a range of training and learning material.
- The use of a common infrastructure with standards and processes across a variety of sectors creates opportunities to source expertise and resources from other organisations.

3.11 Summary

This report aimed to investigate how GBIF could be utilised by regional councils to ensure species occupancy data are accessible in standardised and to enable a federated approach to inform national policy development and state of the environment monitoring. In particularly, this report focussed on the applicability of GBIF to mobile species occurrence data and address difficulties identified with council staff.

Overall, we found that the exemplar data sets were compatible with GBIF, and that GBIF was able to directly, or indirectly (for capability and capacity), assist with the difficulties experienced by regional council staff. Some of these benefits are collated in the summary below.

The data access and use benefits that can be obtained by adoption of GBIF include:

- GBIF provides centralised services that enable discovery of species occurrence records and data sources supporting the need for species occurrence data to underpin biodiversity and biosecurity policy, measurement and management decisions.
- Data access is provided via the GBIF API and from the GBIF website, including as data downloads. In additional data can also be obtained directly via data holders IPT publishing sites or via a hosted-portal or living atlas site
- Data downloads are available in Darwin Core Archives ensuring that metadata accompanies each download
- Data downloads are issued with doi's providing ability to declare the data that was used to support research, policy, or management
- Data are made accessible in consistent and well-supported data standards which should not only reduce the handling difficulties experienced by staff (i.e. they would have a reduced number of formats etc to process), but would enable the (ideally collaborative) development of stable data processes to support activities such as analysis and visualisation, and integration with other types of council data
- Data are accessible in both "raw" form and integrated form and are accompanied by data quality tests. These enable rapid filtering of data and independent verification of the data (e.g. to access the accuracy of the integration result or suitability for a particular purpose)
- GBIF provides hosted-portal infrastructure that can be used to rapidly develop a website to provide access to GBIF-mediate data for a specific community

Data sharing benefits include:

- Data publishing within GBIF uses a federated model which ensures local autonomy and flexibility enabling data holders, when publishing data, to meet requirements of legislation, partners, and other stakeholders.
- GBIF provides a free and open-source tools to help prepare then publish species occurrence data to a consistent and standards-based format.
- GBIF provides guides, manuals, exemplar data sets and training material to support data holders become publishers.

- The primary publication tool, IPT, can be extended to permit the publication of data associated with occurrences that is not already covered by existing standards and extensions.
- Data publishers can use the GBIF validators and data quality tests to identify potential data quality issues enabling them to address issues that may impact the long term integrity and reuse of data.

4 Conclusions

Regional councils can leverage the GBIF network as a tool to help meet their legislated mandates, for example to meet their mandate to develop objectives, policies or methods to maintain viable populations of species of conservation concern (e.g. NPS-IB[47] clause 3.20.3 and 3.20.4 for highly mobile fauna) and to support integrated approaches that cross administrative boundaries (e.g. NPS-IB clause 3.4.1(b)). GBIF has been established with the vision to ensure the "best possible biodiversity data underpins research, policy and decisions". GBIF provides an established and mature information system whereby various biodiversity and biosecurity data sets are aggregated and made available to users in a standardised way whilst ensuring local autonomy and flexibility. GBIF provides a means for councils to access and discover existing data, efficiently share their data, and creates an opportunity to save time and resources that councils might otherwise spend on collecting data themselves, aggregating reformatting and integrating data sets, and even on the development of independent federated data infrastructure for species occurrence data. It would be reasonable to expect the utility to councils of GBIF to as the number and breadth of New Zealand data sets increases – particularly if there is adoption by central and regional governance agencies.

The species occurrence data sets provided by regional council staff were found to be compatible with the data standards utilised by GBIF and would be appropriate to be published to GBIF. Three common issues were noted across several of the exemplar data sets. Most notable are two generic issues that could affect the long-term integrity of data: the lack of persistent unique identifiers, and reliance on vernacular names for recording taxon identifications.

The key strengths of GBIF correspond to the top five most common pain points experienced across the regional councils:

1 finding existing species occurrence data

2 combining data sets with different formats or standards

3 accessing data

4 providing a common access point for species occurrence data

5 sharing (or publishing) data.

[47] National Policy Statement for Indigenous Biodiversity - https://environment.govt.nz/publications/national-policy-statement-for-indigenous-biodiversity/

Furthermore, the GBIF network provides a variety of resources to support data holders and users. Perhaps more importantly, the use of a common infrastructure will create real opportunities for collaboration between regional councils and other data holders within New Zealand.

The GBIF network can serve as a valuable resource for regional councils in New Zealand to access, utilise, and share species occupancy data. By integrating GBIF data into their decision-making processes, councils can enhance their ability to meet biodiversity and biosecurity mandates effectively.

5 Recommendations

5.1 Recommendations for adopting GBIF

Regional councils should adopt GBIF as a primary means of preparing, sharing, and accessing species occurrence data. More specifically, regional councils should look to:

- use a staged approach to adopting GBIF, allowing time for any necessary policy work, engagement with stakeholders, and training to occur in a managed way (an indicative road map is provided below)
- publish as many species occurrence data to GBIF as possible within legislative, licensing, policy, and resource constraints
- adopt policy settings, training, and technical support, and encourage staff to publish species occurrences to GBIF
- encourage staff to utilise GBIF to obtain species occurrence data
- once data are published to GBIF, encourage staff to utilise GBIF as a means to fulfil requests for data sets they steward
- use metadata-only data resources to advertise the presence of species occurrence data that cannot be published in full
- use metadata-only records to facilitate stakeholder engagement and prioritisation
- collaborate with other councils, NZ-based data publishers and GBIF-NZ to provide training and capacity building
- reinforce good data etiquette with regard to citing and attributing data, providing constructive feedback to publishers, and discouraging the 'banking' or sharing of third-party data sets
- encourage the deposition of derived data sets in appropriate repositories
- collaborate with other councils and GBIF participants to develop common analytical and reporting tools based on GBIF services
- collaborate with other councils and appropriate GBIF participants to identify areas that may need to be expanded (e.g. IPT extensions) to support regional council data requirements for other species occurrence dimensions or sources
- collaborate with other councils and GBIF participants to develop guidelines and, where necessary, vocabularies to support the publication and use of different types of data collected by councils.

5.2 General recommendations

- Ensure persistent unique identifiers are implemented within (environmental and biological) information systems and data sets, ideally as globally unique and persistent identifiers.
    - Ensure persistent unique identifiers are included routinely in data products, particularly any exports.
- Review the practice of using vernacular names to record the identification of a taxon, particularly for the long-term storage or archiving of data.

5.3 Indicative road map

The following road map is predicated on using IPT as the primary method of publishing data to GBIF. It is also possible to use API or other processes to publish to GBIF, but these are beyond the scope of what is possible within the current report. The roadmap primarily reflects adoption by a single organisational, however there would be significant benefit from a coordinated approach across multiple councils, and the roadmap could easily be interpreted in that wider context.

There are six phases in the indicative road map: three phases of data-mobilisation activity, each preceded by an approval or review phase. The width of phase is not indicative of duration, but the thickness of arrows is intended to indicate the relative amount of effort that may be expected.

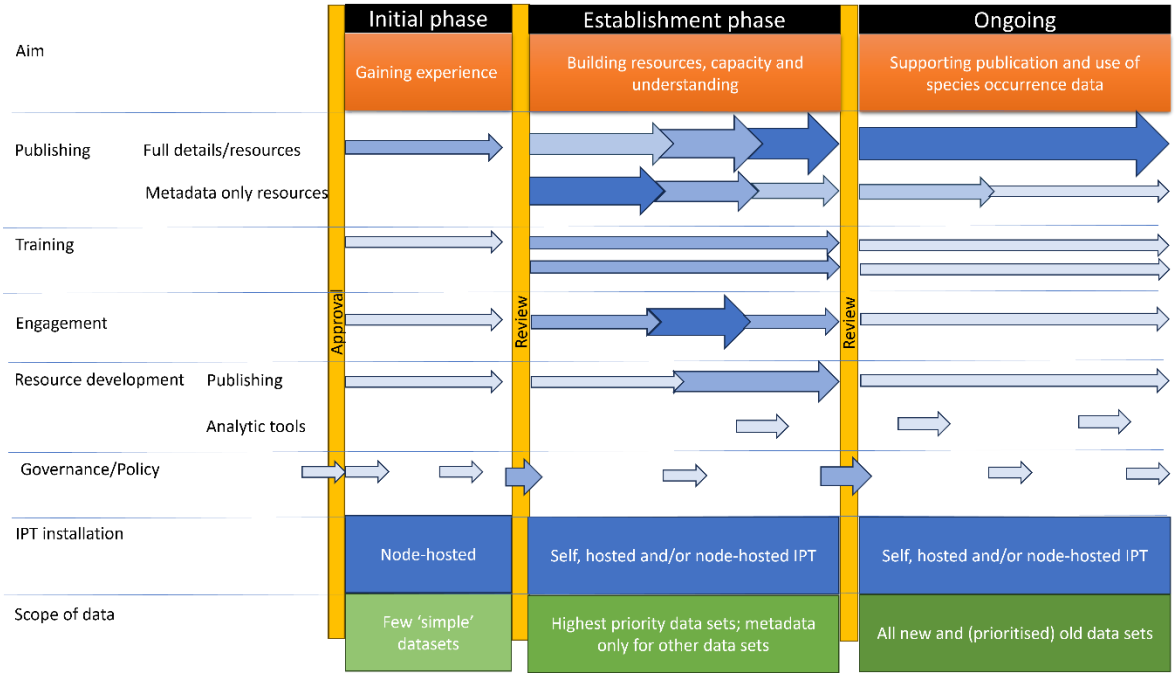


Figure 23. Overview of stages of an indicative roadmap.

Phase 1 – Approval

Purpose

Gain the necessary approvals to initiate work to publish species occurrence data to GBIF.

Key activities

These include:

- reaching agreement within council on the scope of data to be published during the initial phase
- identification of key questions or concerns to be addressed in the initial phase
- approval for the publishing method (e.g. if using IPT, will this be via a self-publishing or hosted approach?)
- identification and allocation of necessary resources and approvals.

Phase 2 – Initial phase

Purpose

Publish an agreed number of data sets to gain initial experience of mobilising to the GBIF network. This should increase understanding of GBIF within the regional council and inform subsequent phases.

Key activities

These include:

- identifying a small number of data to publish to GBIF (we recommend that these data sets have few or no legal barriers to publication, and represent a few classes of species occurrence data that are selected for their ease of publication/manipulation or to engage different business units)
- registering the council as a GBIF data publisher
- publishing data sets via the GBIF-NZ hosted IPT instance
- as necessary, undertaking training and collaborative workshops for staff engaged with this phase
- as necessary, seeking support from the GBIF-NZ node
- if necessary, initiating or continuing engagement with iwi partners and other stakeholders
- communicating findings within and between councils and other key stakeholders.

Phase 3 – Review

Purpose

Act as a go/no go gateway between phases, and ensure the findings from the previous phase are taken into consideration when determining the next steps.

Key activities

These include:

- identifying and addressing any barriers to progress (e.g. policy, resourcing or technical constraints)
- reaching agreement within council on the scope of data to be published during the next phase.

Phase 4 – Establishment phase

Purpose

Mobilise the highest-priority data sets and create a metadata-only record for other data sets.

Key activities

These include:

- mobilising priority data sets using hosted IPT, including more complex data sets
- creating metadata-only resources for other data sets, as resources permit
- using metadata-only records for prioritising data sets for full publication, identifying potential issues, and as a tool to engage iwi and other stakeholders
- identifying opportunities for collaborative development of transformation tools and processes, as well as analytical and visualisation tools
- as necessary, undertaking more detailed analysis of species occurrence data-holdings data (more complex data sets may require the development of supporting resources, such as best practice guidelines and specific vocabularies).

Phase 5 – Review

As for Phase 3.

Phase 6 – Ongoing

Purpose:

Support the routine publication of species occurrence to GBIF as the primary means for sharing and accessing these data.

Key activities

These include:

- reviewing publication policies and processes
- routine publication to GBIF of new data sets and high-priority existing data sets
- analysing emerging and new data set types, reviewing dictionaries, and identifying the need for extensions to the GBIF infrastructure to support ongoing benefits for regional councils
- reviewing data publishing processing, including the IPT hosting setting

- training new staff
- providing feedback on any aspect of the network to GBIF via GBIF-NZ
- enabling staff to engage in the development and maintenance of data standards
- responding to feedback on potential data quality issues for published data.

5.4 Future work

- Kōrero with iwi and hapū to ascertain whether GBIF satisfies Māori aspirations regarding our natural taonga and data sovereignty, and in collaboration with other New Zealand stakeholders.
- Investigate the potential for GBIF-mediated data to contribute to regional sector natural environmental modelling standards.
- Investigate the potential for GBIF infrastructure to provide thematic portals to meet New Zealand's needs (using either the hosted portals or living portal infrastructure), including the collaborative development and maintenance of reporting dashboards.
- Investigate whether the ability to recover GBIF data downloads and/or query definitions would be an acceptable mechanism for providing species occurrence data when fulfilling official information (LGOIMA) requests.

6 Acknowledgements

We would like to thank those regional council staff who provided exemplar data sets and contributed through the survey and online discussions. In addition, we would like to thank Morgan Shields (ECAN) and Matt Moss (TDC) for contributing their use cases which are included as perspectives.

We are grateful for permission from Scott Jarvie (ORC), Shaun Wilkinson (Wilderlab), and Miles Burford (ECAN) to include the IPT screen shorts of their exemplar data.

We are gratefully for the feedback provided by Scott Jarvie (ORC), Roger Urys (GWRC), Michael Berardozzi (MPI), and Jerry Cooper (MWLR).

Appendix 1 – Glossary of selected terms and abbreviations

API	Acronym for application programming interface – a software interface that allows information systems to communicate.
CSV	Comma-separated values is a text-based file in which records are separated by new lines and fields are separated by commas.
DwC	Acronym for the Darwin Core data standard maintained by Biodiversity Information Standards (TDWG)
DwC-A	Acronym for Darwin Core Archive – a self-contained data archive format defined by GBIF that contains metadata describing the provenance and structure of the data as well as the biodiversity data.
eBird	eBird is a citizen science platform maintained by the Cornell Lab of Ornithology, it is represented in New Zealand by New Zealand eBird (https://ebird.org/newzealand/home)
EML	Ecological Metadata Language is a metadata specification maintained by ecoinformatics.org (http://ecoinformatics.org/) for describing environmental/biodiversity data.
Hosted publishing	An installation of ITP on infrastructure maintained by another organisation that a data holder uses to publish their data to GBIF.
IPT	Integrated Publishing Toolkit – a web-based application developed and maintained by GBIF.
Living atlas	Living atlases refers to the open-source platform that has been developed by the Atlas of Living Australia. This platform has now been adopted by other GBIF Nodes which are part of the Living Atlases community (https://living-atlases.gbif.org/)
Node	In the GBIF network a node is the focus point for coordination and activity within a participating country.
Node-hosted	An installation of ITP provided by the participant GBIF node that a data holder uses to publish their data to GBIF.
Occurrence	Evidence of a species in time and space observed or recorded by any method.
Publisher	An organisation that is publishing their data holdings to the GBIF network.
Self-hosted (publishing)	An installation of ITP on infrastructure maintained by the data holder that they use to publish their data.
Species	In this report 'species' is used as shorthand for any organism or group of organisms irrespective of their taxonomic rank.
TDWG	Acronym for Biodiversity Information Standards. The acronym is based on the original name and scope of the organisation – Taxonomic Database Working Group.
TSV	Tab-separated values is a text-based file in which records are separated by new lines and fields are separated by tabs
UUID	A universal unique identifier is an identifier used in many information systems to uniquely label data. UUIDs can be assigned without reference to a central registration authority and yet, for practical purposes, are consider to be unique.
Vernacular name	An informal name, in any language, assigned to taxon by use within a community. Also referred to as common names.
XML	The acronym for Extensible Markup Language. XML is a hardware and software independent specification for storing and transmitting data. It is maintained by the World Wide Web Consortium (W3C)

Appendix 2 – Summary of regional council engagement

As part of this work we engaged with council staff through an online survey, online forums, and through the provision of exemplar data sets. The table below summarises the type of engagement with one or more staff from each council.

Table 5. Overview of the types of engagement with staff from each council.

	Survey response(s)	Exemplar data set(s)	Online forum
Auckland Council	✓	✓	✓
Bay of Plenty Regional Council			✓
Environment Canterbury	✓	✓	✓
Environment Southland	✓		✓
Greater Wellington Regional Council	✓	✓	✓
Hawke's Bay Regional Council	✓		✓
Horizons Region Council			✓
Nelson City Council			✓
Northland Regional Council			✓
Otago Regional Council	✓	✓	✓
Taranaki Regional Council	✓	✓	✓
Tasman District Council	✓	✓	✓
Waikato Regional Council	✓	✓	✓

Summarised roles held by survey respondents

- Bio-information Analyst
- Biodiversity Advisor
- Biosecurity Advisor
- Biosecurity Officer
- Biodiversity Officer
- Data administrator – Biodiversity
- Ecologist
- Policy Analyst
- Team Leader

Appendix 3 – Excel function to generate a GUID

Publishing data to GBIF requires the use of a unique identifier for particular data objects (e.g. occurrences and events). Ideally these would be assigned and stored in the primary data management system. However, when the primary data do not contain such an identifier, it is possible to add one to the data set during preparation.

The following are MS Excel formulae that can be used to calculate a unique identifier in the form of a UUID. Note that this is recalculated every time cells are refreshed in Excel, therefore not meeting the requirement of a permanent identifier. To avoid this, once the formula has been copied to every row down a column, the column should then be selected, copied and pasted as values back into the column so that the unique identifier becomes a fixed value.

```
=LOWER(
  CONCATENATE(
    DEC2HEX(RANDBETWEEN(0, POWER(16,8)),8), "-",
    DEC2HEX(RANDBETWEEN(0, POWER(16,4)),4),"-","4",
    DEC2HEX(RANDBETWEEN(0, POWER(16,3)),3),"-",
    DEC2HEX(RANDBETWEEN(8,11)),
    DEC2HEX(RANDBETWEEN(0, POWER(16,3)),3),"-",
    DEC2HEX(RANDBETWEEN(0, POWER(16,8)),8),
    DEC2HEX(RANDBETWEEN(0, POWER(16,4)),4)
  )
)
```