Manaaki Whenua
Landcare Research

Darwin Core and GBIF readiness assessment

February 2025

Aaron D. Wilton

Manaaki Whenua – Landcare Research

David Dale

Department of Conservation

Manaaki Whenua Contract Report: LC4583

Prepared for: Department of Conservation

Reviewed by:

Ursula Jewell
Informatics Team Leader (Acting)

Kate Boardman
Database Administrator (Allan Herbarium)

Manaaki Whenua – Landcare Research

Approved for release by:

Gary Houliston
Portfolio Leader – Plant Biodiversity & Biosecurity
Manaaki Whenua – Landcare Research

Contents

Summary

Background: about GBIF

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.' GBIF provides a rich, standards-based infrastructure for mobilising and accessing species occurrence data.

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

The key strengths of GBIF address the common pain points experienced in agencies with respect to biodiversity data. These strengths are:

- discovering existing biodiversity (species occurrence) data
- accessing the data
- integrating data of different provenance into a common standard and format
- sharing and responding to requests for data
- providing tools and information to help prepare and use data.

The project

Thirteen example species occurrence data sets were provided to Manaaki Whenua – Landcare Research (MWLR) by the Department of Conservation (DOC) to assess for compatibility with GBIF and the requisite data standards (e.g. Darwin Core, Ecological Metadata Language).

Key findings

- Biodiversity data held by DOC are a good fit with Darwin Core[1] and related standards.
- DOC's biodiversity data are suitable to be published to the GBIF network and would constitute an extremely valuable addition for New Zealand.
- Some areas for consideration are identified in this report, along with a series of recommendations, but none of these prevent DOC's immediate adoption of Darwin Core and publishing to the GBIF network.

[1] This document frequently uses the term 'Darwin Core' for brevity and because Darwin Core is the best known of a set of related standards. However, the document also covers other related standards, where applicable, including Audiovisual Core, Humboldt Extension, Taxonomic Name Core, etc.

Conclusions and recommendations

- All 13 example species occurrence data sets provided by DOC to assess for compatibility with GBIF and the requisite data standards (e.g. Darwin Core, Ecological Metadata Language) were found to be compatible with the data standards utilised by the GBIF network.
- Once transformed to the appropriate standards and structures, all the data would be suitable for publication to the GBIF network and would constitute a highly valuable addition for New Zealand. All but two of the 13 data sets could be published as original data to the GBIF network. The remaining two data sets are highly summarised and would need to be published as derived data sets.
- Two generic issues, which could affect the long-term integrity of data, were found across the data sets: the lack of persistent unique identifiers, and reliance on vernacular names (also referred to as common names) for recording taxon identifications. However, these issues do not block the publication of data using Darwin Core or to the GBIF network.
- Data collected as part of DOC Tier 1 programmes reside in three information systems: one within DOC (nine of the thirteen data sets) and two at Manaaki Whenua – Landcare Research. While it is currently possible to integrate data from these sources, we recommend that the custodians collaborate to strengthen the ability to accurately integrate these data.

1 Introduction

The Department of Conservation (DOC) recognises the critical importance of biodiversity and biosecurity data collection, management, and accessibility to achieve conservation objectives under the Conservation Act 1987 and other relevant legislation. This role has gained greater prominence with the introduction of national policy statements on freshwater management and indigenous biodiversity, which mandate a more integrated and standardised approach to managing these vital resources.

In the future, effective biodiversity and biosecurity management will require DOC to collaborate closely with regional councils, iwi, hapū, and other organisations. This collaboration will focus on developing and implementing standardised methods for surveillance, monitoring, data management, and data sharing. Such standardisation is essential to ensure data quality and usability at regional, national, and international levels, enabling its integration into broader environmental policy and monitoring frameworks.

The increasing need for high-quality, standardised biodiversity and biosecurity data is highlighted in Te Mana o te Taiao – Aotearoa New Zealand Biodiversity Strategy 2020, particularly in Goals 4.1 and 4.2, which emphasise improving data accessibility and reporting. The Parliamentary Commissioner for the Environment has also underscored the importance of such data in reports on national state of the environment reporting and pest management. Furthermore, regional councils, through Te Uru Kahika – Regional and Unitary Councils Aotearoa, have identified the necessity of coordinated efforts to address biodiversity challenges effectively.

Work is underway to foster consensus among central and regional government agencies, iwi, and other stakeholders to develop robust indicators and national-scale data sets. These initiatives aim to measure progress toward multiple environmental and social outcomes, streamline monitoring and data collection efforts, and optimise investments in these areas. DOC plays a pivotal role in supporting and aligning with these efforts to ensure the long-term protection and restoration of New Zealand's unique ecosystems and species.

New Zealand is not unique in its need to access biological data in a timely, coordinated, and standardised way. Internationally this has seen the development of standards bodies (e.g. Biodiversity Information Standards[2]), and various initiatives to federate data at differing 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, which is supported by many of the world's governments, including New Zealand's.

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, while providing data holders and users with data aggregation services based on common tools and standards, data integration and quality services, a

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

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

registry of data holders and their direct data access points, and data access via a common web service.

DOC has decided to take a staged approach to publishing its data to GBIF, with the first step being implementation of the Living Atlas software as an internal portal before making the data publicly available. The work needed to host data in this internal portal would need to be done anyway for GBIF publication. A key feature of the Living Atlas is that it stores data in the sector-best-practice format of Darwin Core and associated bioinformatics standards, which are also used in GBIF.

Here we present our findings from evaluating 13 selected DOC data sets as an indicator of Darwin Core suitability and GBIF readiness, and assessing the ability to re-integrate biodiversity data resulting from the same monitoring programmes but that are resident in different information systems owned and managed by different organisations.

1.1 Scope

The scope of this assessment is the data exported from data sources identified by DOC rather than the source systems that created these data. The report focuses on the ability to map these data to the Darwin Core and related standards.

Where adjacent issues were identified as part of this work (e.g. arising from source systems, based on the authors' previous experience), they have been noted but are not addressed fully in this assessment.

1.2 Assumptions

At the outset of the assessment the following points were established to assist the focus and scope of the work.

1 DOC is committed to participating in GBIF and using global biodiversity data standards to deliver data to staff and external stakeholders, so the report focuses on the *suitability* of DOC data to GBIF and these standards rather than *whether* to use them. However, a summary of GBIF and the advantages of using it are included for context.

2 Given the first assumption, DOC is necessarily committed to transforming its data to Darwin Core as part of maturing its approach to biodiversity data, and enabling a move to an internal data portal (perhaps using Living Atlas) and delivering to GBIF. The report therefore focuses, again, not on *whether* to use Darwin Core and related standards, but on *how* to use them and *which data sets* are suitable.

3 The 13 data sets selected for this report are representative of the types of data sets collected by DOC.

1.3 Structure of this report

The report consists of three major parts. This arrangement has been adopted to allow the document to be either used as a whole, or for different sections to be more easily utilised as separate parts.

The order of content in the report is:

1 an assessment of DOC readiness

2 general introductory information about GBIF

3 appendices, including initial mappings of the sample data sets to Darwin Core and related standards.

2 Analysis

2.1 Compatibility of DOC data with Darwin Core and GBIF

2.1.1 Data sets

The primary goal in assessing the DOC data sets was to determine their suitability for mapping to Darwin Core and related standards, and the readiness to mobilise the species occurrence data using the GBIF network.

The data sets (Appendix 2) include structured monitoring and survey data, *ad hoc* observations, and data arising from an endangered species management programme. The example data sets include presence, absence, and quantitative occurrence records. The majority of the data sets provided were found to be compatible with the Darwin Core standard, and our evaluation found that they would be suitable to be published in the GBIF network as original data – with two possible exceptions.

The exceptions are the Riverbird Count Summary and the Twizel Kaki Hide data sets. These represent highly summarised data, and, where possible, preference should be given to publishing original observations. Also, based on prior knowledge of one of the authors (AW), there is a possibility the Riverbird Count Summary data set includes data from other data owners.

These data sets therefore serve as a good example of factors that should be taken into account for all data.

- Data ownership: does the data set include third party data? Do the data licences and/or permissions from the original data holder(s) allow the data to be published to GBIF under a Creative Commons licence?
- Duplication: would publishing to GBIF result in duplicate data that is already in, or that will be published separately to, GBIF?
- Derived data: publication of derived data[4] to GBIF requires careful consideration because it may introduce data duplication and other undesirable data artifacts (e.g. degradation of data through gridding). 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 third consideration (relating to derived data sets) does not prohibit making the data set accessible within DOC or to GBIF in other ways. For example, this could be achieved by:

- registering and uploading the data sets as derived data within the GBIF network – derived data sets enable the sources of information that were used to create the

[4] https://www.gbif.org/derived-dataset/about. A derived data set is one that has been generated from original or primary data through processes such as aggregation, enrichment, and/or other analyses.

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 an Integrated Publishing Toolkit, 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-ready and sourced data
- creating a Darwin Core Archive for use within DOC.

Recommendation 1. DOC should give priority to publishing original data to the GBIF network rather than highly summarised and aggregated data.

Recommendation 2. DOC should identify data sets containing third party data for which additional permissions may be required prior to publication to the GBIF network.

Recommendation 3. DOC should publish highly summarised data sets as derived data sets rather than as original data sets, unless it is likely the original data will not be published to the GBIF network.

Darwin Core Archives

Darwin Core Archives comprise a zipped folder of metadata and data clustered in a star-schema around a central core file (*core* files are the central file in the start-schema to which other data files, so called *extensions*, are linked. In GBIF the cores may represent occurrences, sampling events or taxa. (See section 5.2.3).

All the data sets assessed could be published as Darwin Core Archives using the sampling event or occurrence cores – the choice is potentially influenced by the underlying data structures and presence (or absence) of critical identifiers for the object classes (especially a persistent identifier for the sampling event).

For most of the data sets assessed, the use of extensions, particularly the measurement or fact extension, is necessary to map the data fully. While some of these data could be passed using alternative approaches (e.g. a field called dynamicProperties), the extension provides the most structured and robust method for including 'non-standard' data fields; see, for example, the use of extensions in the Kaki Master egg check data set (Appendix 3.14).

It should be noted that the measurement or fact extension has two variants: the first version links only to the Core file; the extended version, allows linkage to more than one file in the Archive, thus supporting use of this extension in a sampling core for values that need to link to occurrence records.

It was noted in several conversations with DOC staff while preparing this assessment that the Darwin Core Archives would provide a good method for documenting and archiving data sets, particularly historical data, in a consistent way that enables long-term use of the data.

Recommendation 4. DOC should give preference to the use of extensions, especially the Measurement or Fact extension, over concatenation of values in simpler fields (e.g. dynamicProperties).

Recommendation 5. DOC should consider using Darwin Core Archives as a method for creating self-documenting data sets for publishing, sharing, and archiving biodiversity data.

Data segmentation

The data set samples provided for this assessment represented a subset of each of the data sets. Each subset represented data obtained using a specific methodological as well as a specific temporal and/or spatial focus. This type of segmentation of the data held by DOC is appropriate for the GBIF network, because:

- the creation and interpretation of metadata are simpler (i.e. the metadata does not have to cover multiple methodologies)
- segmentation supports approval processes for publication
- common patterns of transformation/mapping can be established based on an internal source and the methodology.

Recommendation 6. DOC should develop an approach to segmenting data sets to facilitate the data publishing processes; that is, a pattern for publishing data both internally and externally based on key facets, such as methodology and season.

2.1.2 Data standards

The GBIF network utilises several biodiversity data standards – most notably Darwin Core and Ecological Metadata Language (see section 5.2.2). There is ongoing work to review and extend these standards and the types of data that can be published (more fully) to the GBIF network[5], such as camera-trap data.

For the DOC data sets assessed there is a good fit with Darwin Core

Most of the fields in the sample data sets can be mapped 1:1 to fields in Darwin Core and associated standards. When a direct 1:1 mapping is not possible, there are other structures within the standards that can be used to represent the data. For example, in the Weeds data set there are measurements for different life stages of the weed that can be mapped with the measurement or fact extension.

To provide flexibility for mapping to different types of data set, very few fields are mandatory in Darwin Core. However, there is value in populating as many optional fields as possible, and making the data as complete as possible, for each data set. Two key reasons for this are that:

[5] https://www.gbif.org/new-data-model

- the uses of data cannot be known in advance (as illustrated in the GBIF Science Reviews[6]), so decisions to not populate an element cannot be based only on currently known use cases
- some of the elements that might appear to be unnecessary may be used in quality checks, so populating them might reduce the number of error/warning flags and consequently increase uptake of the data.

Given the flexibility of the data standards, DOC, in collaboration with other key stakeholders, should produce guidelines that help the mapping and publication of data sets using the Darwin Core and related standards. Such guidelines should be flexible and kept under active review, but the absence of guidelines would create the risk of inconsistent approaches across different DOC data sets. In some cases the guidelines may not need to be prescriptive, but instead could consist of a list of factors to consider when making the decision for an individual data set.

Recommendation 7. DOC should make use of Darwin Core and related standards to enable consistent integration of biodiversity data.

Recommendation 8. DOC should develop guidelines to assist the consistent publication of its biodiversity data.

2.1.3 Significant considerations

Three key aspects of the data were identified during this assessment, which are discussed below. *None of these prevent DOC using Darwin Core and related standards or publishing to the GBIF network*, but they could result in additional manual processing until they are addressed.

Lack of persistent unique identifiers

Persistent unique identifiers enhance data governance and management by ensuring consistency, traceability, and interoperability across systems. They enable accurate data integration, and prevent and detect duplication, ultimately improving data quality, compliance, and operational efficiency.

Data published within the GBIF network will ideally include persistent, globally unique identifiers for data objects. These identifiers are essential for activities such as re-indexing, linking associated data, and citing data. They can also be useful for enabling the 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).

The most significant issue apparent within the DOC data sets assessed is the lack of persistent, globally unique identifiers for many of the records and objects. It is also unclear if the local identifiers within the data are persistent or whether they represent temporary

[6] https://www.gbif.org/science-review

assignments that will be unstable over time. For example, many of the DOCMON data sets have 'ResultMasterID', which is persistent and unique with DOCMON but is not unique in the context of mapping data to Darwin Core because it represents multiple observations (i.e. it is unique at a level of granularity higher than the individual event and occurrence records). DOCMON also has 'ID', which is unique at the row level, but not persistent.

Persistent unique identifiers can be formed in many ways. A common method is to use universal unique identifiers (UUIDs). For example the UIDD c821a27f-8ff8-4dd2-9597-8a8dcB80fd7d is the persistent UUID assigned to a specimen at the Allan Herbarium with catalogue number CHR 92742.

Another method for assigning persistent identifiers is to calculate the ID using a concatenation of text and selected stable and permanent components in the data. Given the assumption that the data in the selected fields is stable, these IDs can be created in the data integration environment or at the time of export. Care is required when selecting the fields to be used to ensure they are stable and the combinations will be at least locally unique. It is also recommended that some consideration be given to the opacity of the ID; for example, inclusion of a field such as a taxon name would not be appropriate because it is likely to change over time.

There is also the potential for a user to find the taxon name in the ID to be a convenient short-hand so they rely on that string, with the likely result of misinterpretation of the data over time. Example concatenations are included for events in some of the data set assessments. These examples use concatenates of season + location identifier, then concatenate GPS identification identifier and survey methods, as required for each level of events.

For example, monitoring events during the 2023/24 season at location CO94 could have a parent event with identifier *2023-24-C094*, with a subsequent nested event at a bird station receiving the identifier *2023-24-CO94-AA.* Additional suffixes will be required to cope with repeated measures or greater levels of hierarchy. This pattern is similar to the GPS labels specified in the Tier 1 field protocols[7]. The GPS labels mandated in the field protocols could be adopted as a unique identifier for locations (locationID in Darwin Core) for Tier 1 data.

Key entities for which persistent unique IDs need to be used include:

- observations
- (survey and sampling) events
- locations
- scientific names.

[7] DOCDM-826779 – Field protocols for Tier 1 monitoring – invasive mammal, bird, bat, RECCE surveys. Version 17. Page 16.

A key aspect for consideration is that, once assigned, the identifiers must be stable and maintained with the digital object/record (ideally in the primary data repository), and that a process and policy are in place governing these IDs.

Recommendation 9. DOC should ensure persistent unique identifiers are maintained and available for data sets, and have been assigned at the correct levels and documented.

Recommendation 10. Persistent identifiers should ideally be opaque and should *not* include fields that might be changed by data management/curation processes (e.g. scientific name).

Geospatial data

The data sets assessed contained georeference coordinates recorded as New Zealand Transverse Mercator (NZTM). NZTM, and georeferences using other coordinate systems, can be mapped into the Darwin Core standard 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 (WGS84 is recommended) and (optionally) 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 various regional/national specific projections, or from historical coordinate systems (e.g. NZMS1 and NZMG in New Zealand). Further, GBIF utilises these decimal latitude and longitude coordinates to undertake various data quality checks as part of the aggregation process.

The original spatial data and referencing system should be recorded in the data sets using the verbatim fields. This allows for verification of the conversion processes and, within DOC, will provide the NZTM coordinates for consistency with other DOC spatial data.

Some biodiversity data within DOC contains spatial data that are managed as polygons. These polygons can be provided in Well-Known Text (WKT) fields, but we strongly recommend calculating a point representation (most likely the centroid) and including it in the data in the decimal latitude/longitude fields, because this will support the data classification, integration, analysis and visualisation processes used across the GBIF network and reduce complexity for less complex spatial analyses.

One limitation encountered in the current GBIF implementation of Darwin Core is a field-length limit for the FootprintWKT, which is used to provide a WKT representation of a spatial feature. For some DOC data sets the source data set includes complex polygons consisting of a significant number of nodes that describe the spatial geometry. When these polygons are converted to WKT, the resulting text exceeds the currently allowed length of this field (note that this issue has been logged with GBIF). In these cases, the polygons need to be simplified before being converted to WKT. When this simplification occurs, this should be noted in the data and/or record metadata, as appropriate, and (ideally) quantified.

There are a number of functions for which specialist geospatial processing may be required, including:

- calculating the geospatial centroids of the footprintWKT polygon and the original polygon,[8] and the pointRadiusSpatialFit of the footprintWKT
- determining the 'higher geographies' (e.g. country, province, and place names of the point or polygon, subject to confirmation of which higher geography elements will be used)
- identifying any occurrence of polygons that straddle multiple higher-geography units so that a decision can be made on how to handle these (note that the higher geographic Darwin Core fields are not 'list' fields and should be left blank when multiple values are correct)
- ensuring the WKT text follows the established conventions for handedness to ensure correct interpretation as enclosing polygons.

Recommendation 11. DOC should convert all georeference coordinate data into decimal latitude and longitude (WGS84) for publication to the GBIF network, and optionally include the original coordinates and footprint in the relevant fields.

Recommendation 12. DOC biodiversity data with spatial data stored as polygons should have a centroid calculated when publishing externally to enable integration and visualisation with other data sets.

Taxon names

Within the assessed data sets some (those related to introduced mammals) captured the taxon identification using only vernacular names.[9] We believe this practice is more widespread in other DOC biodiversity data sets that were not assessed in this report. To publish data sets to GBIF, the vernacular names or codes need to be supplemented with scientific names at the applicable taxonomic rank; in some cases this might be a species binomial, 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 is problematic for the reuse, integration, and long-term storage of data. Vernacular names are problematic because the application of a vernacular name is frequently ambiguous, for several reasons, including the following.

- 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, and piriwhetau.

[8] The centroid of the FootprintWKT is recommended because, while possibly being less accurate in some cases, it is consistent with use of the footprintWKT to define the polygon, so is less likely to generate unnecessary user feedback. If used at all, the centroid of the original precise polygon could be mapped to verbatimLatitude and verbatimLongitude, with a note to explain what they represent.

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

- 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 & 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, instead being determined by the community using the name.

The potential ambiguity of vernacular names means that the integration of data based on them will be problematic, particularly when integrating data sets of differing age and provenance. While scientific names may also change over time, this occurs only as the result of systematic research and a nomenclatural process. Scientific names 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.

Recommendation 13. DOC should ensure that scientific names are included in all biodiversity data. Where vernacular names are used for data capture, they should be supported by documentation or data that maps each vernacular name to the scientific name *as it is being applied by the team gathering that data*. These mappings should be used to add the scientific names when they are permanently stored.

Recommendation 14. DOC should establish a service and/or process to assist with the accurate integration and mapping of the taxonomic and nomenclatural data, and to enable records to be supplemented with additional taxonomic data (e.g. higher classifications). This would draw on the information collected following Recommendation 1413, as well as additional sources (e.g. NZOR[10] and the GBIF taxonomy).

Recommendation 15. DOC should establish or adopt data validation processes that allow staff to submit data sets for validation to identify any erroneous, new and/or ambiguous taxonomic data.

2.2 Integration of specimen, vegetation, and animal data

Data from the Tier 1 monitoring programme are held in three separate systems:

- DOCMON (animal data; DOC)
- the Allan Herbarium (plant voucher specimens; MWLR)
- the National Vegetation Survey Databank (vegetation data; MWLR).

[10] 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.

The purpose of this section is to indicate whether these data could be confidently and accurately integrated once mobilised to GBIF, and if so, how. The section also makes recommendations for custodians of each of these systems, either independently or in collaboration, that could strengthen this integration.

2.2.1 Allan Herbarium (CHR)

The Allan Herbarium accessions specimens from a variety of sources, including vouchers collected by the Tier 1 teams as well as *ad hoc* collections from other DOC staff. Data management at the Allan Herbarium (CHR) is done in a customised information system – the Collection Information System (CIS).

Allan Herbarium and DOC staff have developed a well-defined process for accessioning Tier 1 specimens that includes capture of the Plot ID and sample number and the consistent provision of agreed data fields. It also includes the creation of a 'standard locality' in the gazetteer within the CIS to which the Tier 1 specimens are linked. More recently, Allan Herbarium staff have started creating projects within CIS for each season for Tier 1 specimens; these are constructs that enable the creation of a virtual set of specimens.

In the context of connecting these data sources, CIS stores the following fields:

- plot ID (as part of the locality strings)
- collection date
- collector
- geospatial coordinates (stored as original and converted values: decimal latitude/longitude, WGS84)
- programme name (as part of the note fields)
- sample number.

CIS has the ability to link to external resources where an API[11] is made available.

It is important to note that catalogue (aka accession) numbers are not guaranteed to be permanent or to resolve to the same specimen. There are rare occasions when an accession number needs to be changed. Therefore, relying solely on catalogue numbers as the linkage point between information systems is not recommended. To address this issue, CIS assigns a permanent unique identifier to specimen records and collection events, as well as to several other objects in the data.

CIS has the ability for Allan Herbarium staff to set an external (i.e. outside of CIS) access level for the different parts of each specimen record (e.g. whole specimen, collection event, georeferences). This is used by CHR to control the data that are made available to external

[11] An API, or Application Programming Interface, is a mechanisms that enable two software components to communicate with each other using a set of definitions and protocols. For example, a webservice is a specific type of API that allows this communication via the internet.

sources, including GBIF, only where necessary (though the data are public by default). CHR publishes to GBIF, on a weekly schedule, data that are tagged for public access.

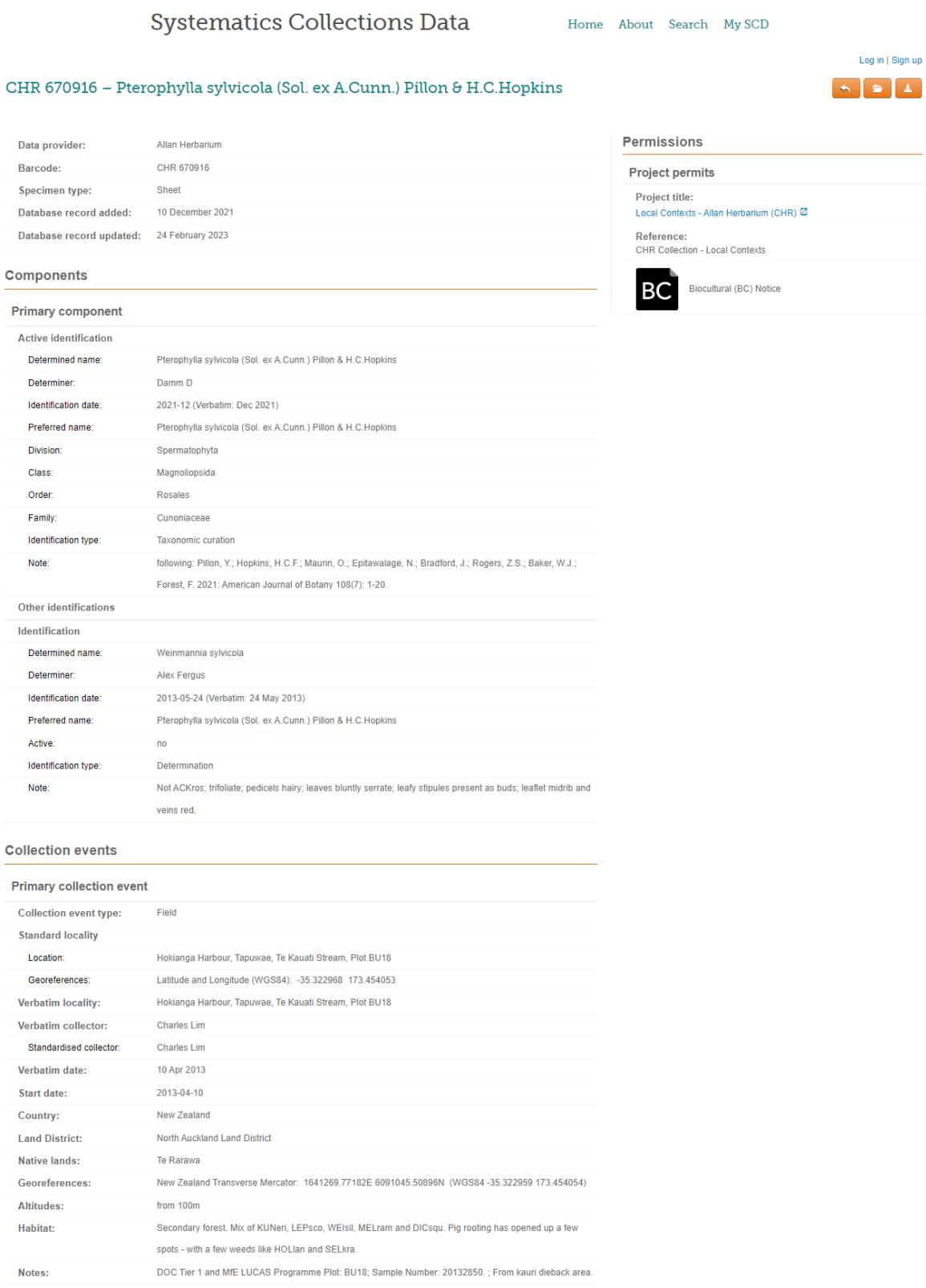

Figure 1. The Allan Herbarium data record for a Tier 1 voucher specimen. Note the plot and sample number in the Notes field. (https://scd.landcareresearch.co.nz/Specimen/CHR%20670916**)**

Recommendations for the Allan Herbarium

- Include the Plot ID in a more structured field to enable it to be used to create connections between records and external services, once available, especially the National Vegetation Survey (NVS). Consideration should also be given to whether these should be associated with the list of 'standard localities' (e.g. in Locality.Code, within CIS).
- Continue to capture the sample number as a way for Tier 1 staff at DOC to find specific specimens as needed. If sample number is added to DOC sources that are published to GBIF, it may be appropriate to move sample number from the 'Notes' fields (e.g. to external links to enable these values to be used as connectors).
- Establish a service that enables systems to create connections to specimen records:
  - External connections should be maintained using the specimen UUID.
  - To establish the connections, it is likely the service will need to be able to respond to requests based on at least sample number, plot ID, date, collector, and catalogue number, and would need to return a list of matches.

2.2.2 National Vegetation Survey

The NVS contains plot-based vegetation data from multiple sources, in addition to Tier 1 records. These data are managed across multiple access levels, with agreement from the data owners.

In the context of connecting these data sources, NVS stores the following fields:

- plot name (this is equivalent to Place in DOCMON)
- observation date
- observer
- geospatial coordinates
- metadata on the context of the observations (e.g. data owner, programme).

NVS does not have a facility to capture specimen numbers/links, because this would (1) introduce complexity in the data structures due to the many linkages between an observation event and any associated samples/specimens; and (2) create an additional and significant management burden because the accessioning of specimens is normally delayed compared to NVS data entry processes.

NVS currently publishes Level 1 (public) data to GBIF as a single resource representing species occurrences, but these are highly generalised and anonymised; for example:

- event date is a truncated to year and month (YYYY-MM)
- observer is not included
- Plot Name (Plot ID in CHR, Place in DOCMON) is not included
- latitude/longitude coordinates are generalised, with values truncated to 1 to 4 decimal places (equivalent to c. 7.5 m to 110 km) for each record.

It should be noted that plot name strings are not unique. NVS has many examples where the same string has been used as an identifier for plots in different survey systems and programs (Figure 2).

Project	PlotName	PlotID	PlotObsID	PlotObsStar
RANGIPO VOLCANIC DUNES 2009-2011	R150	789974	1162571	2009
NZ Biodiversity Monitoring System: Public Land 2014-2015	R150	1230994	1638814	2015-03-20
RANGIPO VOLCANIC DUNES 2009-2011	R151	789975	1162572	2009
NZ Biodiversity Monitoring System: Public Land 2016-2017	R151	1653095	1875633	2017-04-10
NZ Biodiversity Monitoring System: Public Land 2021-2022	R151	1653095	2239035	2022-03-24
NZ Indigenous Carbon Monitoring System: Main 2002-2007	R152	1144811	1524221	2004-02-16
RANGIPO VOLCANIC DUNES 2009-2011	R152	789976	1162573	2009
NZ Biodiversity Monitoring System: Public Land 2011-2012	R152	1144811	1455434	2011
NZ Indigenous Carbon Monitoring System: Main 2009-2014	R152	1144811	1522590	2012-01-27
NZ Biodiversity Monitoring System: Public Land 2015-2016	R152	1144811	1795726	2015-12-08
NZ Biodiversity & Indigenous Carbon Monitoring System: Public Land 2020-2021	R152	1144811	2143150	2020-12-14

Figure 2. Screen shot from the NVS data management tool showing three examples of duplicated Plot Names (R150, R151 and R152). PlotID and PlotObsID are unique identifiers for the Plot and Plot survey event NVS (and can be supported/supplemented by a UUID field).

Recommendations for NVS

- In collaboration with other stakeholders, consider establishing a service for resolving plots and plot observations to enable the connection of other data records to NVS (i.e. formal survey places).
  - Plot: Plot UUID, plot name, plot context (e.g. Tier 1)
  - Plot event: Plot Observation UUID, Plot UUID, event date.
- Review the data and fields that are published to the GBIF network with the aim of publishing more complete records.
- Consider splitting the data into multiple resources that are published to GBIF to support differences in data generalisations and acknowledgement of data owners. This is likely to assist publication and more complete records, and would support more granular metadata.

2.2.3 DOCMON

(Note: this section is based on the data samples extract from DOCMON, as a full analysis of DOCMON was out of scope for this report.)

Based on the DOCMON data sets the following fields are key for integrating data across these three sources:

- Place (Plot ID in CHR, Plot Name in NVS)
- StartDate
- Observer(s)
- Geospatial coordinates.

In addition, DOCMON has a numeric identifier for Place – MonitoringPlaceID – that could potentially be assessed for use as part of a unique identifier string for Place (given that the Place string is not unique when considering data from non-Tier 1 surveys).

Associated with the Tier 1 survey data is a comprehensive manual that provides a well-documented set of processes and good definitions of the survey variables to be measured by field teams. This manual would provide a rich source of information to form the basis of supporting the development of formal vocabularies.

Recommendation 16. In all data sets, DOC should publish as fully as possible the key fields that enable integration across information systems, especially place and event dates.

2.2.4 Integration

The following table outlines the methods that could be used to integrate data across the sources, and provides a brief statement of the strengths and weakness. The most robust methods are listed at the top of the table. Note that:

- **Plot ID** means the string assigned to a particular plot within a defined scope (e.g. Tier 1 AD172), and is usually intended for human references; this corresponds to Plot ID (CHR), Place (DOCMON), and PlaceName (NVS)
- **Plot UUID** means an identifier (frequently in the form of a 32-hexadecimal string) that is intended to uniquely and permanently (persistently) identify a record, and is particularly intended for machine/system use.

Table 1. The strength and weakness of using different field combinations to integrate DOCMON data held in three different repositories. Field combinations are listed from strongest to weakest.

Connectors	Strength	Weakness
Plot Obs UUID	Unambiguous connection of records from the same events. Enables connections even if slight variation in date (e.g. when survey work extends over more than 1 day at a particular site).	Needs a vocabulary to be made available and adopted by all 3 information systems.
Plot UUID + eventDate	Unambiguous connection of formal survey data at place and time.	Needs an accessible vocabulary. If a user needs to group data by survey events, they would need to consider acceptable data ranges for records to be considered part of the same event.
Plot ID + eventDate	Connection of formal survey data at place and time.	Plot ID is only unique within a specific context. Homonyms of these strings will occur, particularly in CHR and NVS data.
Plot UUID	Connection of formal survey data at a place.	

Connectors	Strength	Weakness
eventDate + geospatial coordinates	Connection of any biodata at a place and time.	Will not be limited to data resulting from Tier 1 surveys. Relies on consistent and correct recording and conversion of spatial coordinates.
eventDate + observer		Increasing mobility means observer able to visit multiple sites on same date, so is a weak connection. In most cases there will be different observers for different parts of the data (e.g. animal vs plant measurements). Relies on string matching of observer names.
Sample number	Enables direct links to be made between samples/plots and resulting specimens.	Limited to CHR and STMS.

Recommendation 17. DOC and MWLR should formalise, and make publicly accessible, vocabularies that support the integration and consistency of biodiversity data across systems. These vocabularies should include a persistent unique identifier and be governed using best practice.

Note: GBIF-NZ would be able to host the vocabularies (in simple format) on the GBIF-NZ portal, and sees the publication of these vocabularies as an important contribution DOC can make to the New Zealand biodiversity data community.

2.3 Other considerations

2.3.1 Sensitive data

Data sensitivity was not assessed for this report, but we were aware that the data provided included information that can be considered sensitive, and therefore feel the following commentary is pertinent.

Sensitivity of species occurrence records may result from, for example, the particular taxa being recorded (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 (private land or land with other restrictions).

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,[12] which has recently been revised.[13] Although focused on taxon-based sensitivities, many of the considerations – particularly the methods for generalising data – can be applied to other contexts.

[12] Guide to Best Practices for Generalising Sensitive Species-Occurrence Data 2008
[13] Current Best Practices for Generalizing Sensitive Species Occurrence Data 2023

The data standards used within the GBIF network *allow for omission or generalisation of data* and *provide ways of recording these actions* at both the data set and record level. However, 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'.

Recommendation 18. GBIF-NZ Node recommends DOC adopt an 'open-by-default, closed-by-necessity' stance to publishing biodiversity data.

Recommendation 19. GBIF-NZ node recommends DOC adopt and adapt the principles and practices outlined in the GBIF guides to sensitive data.

2.4 Indigenous data sovereignty and governance

Indigenous data sovereignty and governance are beyond the scope of this report, but some key points in the context of the GBIF network are provided below in brief.

- The GBIF network is a federated architecture when applied to data publishing. This architecture ensures that:
  - data holders have full local autonomy and flexibility as to what data they publish, which is key to enabling data holders, such as DOC, to respect agreements with iwi and other stakeholders
  - the primary (or master) data and the intellectual property are retained by the data holder, with only a transformed version of the data published to the GBIF network under a Creative Commons licence.
- GBIF has active programmes[14] on indigenous data governance to support the adoption of the CARE principles within the network,[15] as well as operational tools such as the traditional knowledge and biocultural labels and notices that have been developed by Local Contexts.[16]

Recommendation 20. DOC should maintain a watching brief on the indigenous data work being undertaken by the GBIF network so that it is aware, and can benefit from, guidelines and tools that may emerge from that work.

2.4.1 Data governance

Data governance processes for the data sets were not part of the assessment, but some governance-related aspects were noted during the work and are included in brief as recommendations and points below. Note that this is not intended to be a comprehensive analysis – only points encountered as part of the assessment are included.

- Ensure the necessary policies and processes are in place to enable publication. These should include:

[14] See for example, https://docs.gbif.org/2025-work-programme/en/#activity4-3

[15] Global Indigenous Data Alliance (GIDA) – CARE principles - https://www.gida-global.org/care

[16] Local Contexts - https://localcontexts.org/

    - publication of the names of people who are the observers and identifiers for the records
    - publication of data that relate to iwi whenua
    - publication of data that relate to private land
    - determining which licence to attach to published data (noting that the GBIF network utilises CC0 1.0, CC BY 4.0, or CC BY-NC 4.0 licences)
    - processes and approvals for generalising or withholding data
    - processes for prioritising data sets.
- Ensure the necessary data agreements are in place before data collection, and if not, retrospectively seek these agreements and/or amendments.
- Social licence within DOC: DOC can be expected to experience the same anxiety as other organisations embarking on data publication. This means that the social aspects of a project are likely to be more difficult than the technical aspects and require close attention. It is common to encounter responses such as 'my data', 'not good enough' and 'sensitive species'.

2.4.2 Capability and capacity

The management, delivery, and use of biodiversity data are streamlined by the presence of people with biodiversity informatics capability who apply data management, data science, biodiversity standards, and other domain skills to biodiversity data. GBIF does not directly address any lack of capability and capacity where these skills are missing within an organisation. However, utilising GBIF can connect organisations lacking this skill set 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 participation 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.

2.4.3 Benefits of GBIF participation

The benefits that can be obtained by the adoption of GBIF were identified for regional councils in a recent report.[17] These benefits would similarly accrue for DOC via active participation in the GBIF network. These are summarised here and include the following.

[17] Wilton AD, Jewell U, Goodsell B 2023. Potential for regional councils to use GBIF to access and share species occurrence data. Manaaki Whenua – Landcare Research contract report LC4381. Enivrolink Grant: 2329-ORC004.

- GBIF provides centralised services that enable discovery of species occurrence records and data sources supporting the need to underpin biodiversity and biosecurity policy, measurement, and management decisions.
- Improved data access is provided via the GBIF API and from the GBIF website, including as data downloads. Data can also be obtained directly via data holders' Integrated Publishing Toolkit publishing sites or via a hosted-portal or Living Atlas sites.
- Data are made accessible with consistent and well-supported data standards, which should not only reduce the handling difficulties experienced (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 data.
- Data downloads are available in Darwin Core Archives, ensuring metadata accompanies each download.
- Data downloads are issued with digital object identifiers (DOIs), providing the ability to declare the data that was used to support research, policy or management.
- Data are accessible in both raw and integrated form and are accompanied by data quality tests, which 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-mediated data for a specific community.
- Data publishing within GBIF uses a federated model, which ensures local autonomy and flexibility, enabling data holders, when publishing data, to meet the requirements of legislation, partners, and other stakeholders.
- GBIF provides free and open-source tools to help prepare and publish species occurrence data to a consistent and standards-based format.
- GBIF provides guides, manuals, example data sets, and training material to support data holders to become publishers.
- Data publishers can use the GBIF validators and data quality tests to identify potential data quality issues, enabling them to proactively address issues that may affect the long-term integrity and reuse of data.

3 Conclusions

DOC is well placed to adopt Darwin Core (and other standards used in the GBIF network/community) and GBIF as a primary means of preparing, sharing, and accessing biodiversity occurrence data. The key strengths of GBIF correspond to the pain points commonly experienced by staff across different organisations regarding biodiversity data, and which will also be common for DOC staff. These strengths are:

- discovering existing biodiversity (species occurrence) data
- accessing the data
- sharing and responding to requests for data
- integrating data of different provenance into a common standard and format to improve the usability of the data
- providing tools and information to help prepare and use data.

The DOC biodiversity data sets assessed were found to be compatible with the data standards used by GBIF and would be appropriate to be published to GBIF. The assessments clearly demonstrate the applicability of Darwin Core and related standards, and the ability to use these standards to bring data with different provenance together in a single standard and format at the time of re-/use, publishing and archiving.

This is particularly important because it allows the information systems and tools used by teams within DOC to be tailored to their specific needs rather than being forced to adopt Darwin Core-based structures to ensure compatibility. This does not preclude considering definitions from these standards, and the recommendations within this report, especially those regarding unique identifier and taxonomy, when a business unit tool or process is being modified.

The data mappings included here are preliminary, but nonetheless provide a good basis for transforming each of the data sets. In most cases we expect that these mappings will only require some minor effort to be finalised; in particular, with input from DOC's subject matter experts for each of the source data sets concerned.

Biodiversity data resulting from the Tier 1 monitoring programme are deposited in three different information systems: DOCMON (DOC), Collection Information system (Allan Herbarium, MWLR), and National Vegetation Survey Databank (MWLR). Data held in the latter two systems are already being published to the GBIF network. Publishing DOCMON using the Darwin Core standards and, preferably, to the GBIF network would be a major benefit for data users, both within and outside DOC, because it will enable discovery and access to these valuable data at a single place and in a consistent standard and format. The ability to accurately integrate these data would be enhanced with collaboration between the three data custodians on changes to their systems, and publication of supporting materials such as common vocabularies.

Our assessment indicates that there are no technical barriers to DOC's use of the standards and participation in the GBIF network. In short *DOC could start publishing data using these standards and publishing to the GBIF network immediately*. Below is a list of

recommendations, based on the authors' experience, to assist the adoption of these frameworks.

- Use a staged approach to adopting Darwin Core standards and GBIF.
    - Start as soon as possible with simple, quick mobilisations with known tangible benefits and frequently requested data, to build capability and enthusiasm for further mobilisations.
    - In parallel, undertake any necessary policy work, engagement with stakeholders, and training.
- Note that this is a social process as well as a technical one. As such, the work should include:
    - establishing data governance and management roles, so that it is clear who makes decisions about the data, including decisions about which data are accessible by whom; who is accountable for data quality; and who manages the data on a day-to-day basis (adding fields, changing frequency of publication, etc.)
    - change management, so that stakeholders are aware of the consequences for them that result from the changes
    - training, so that users understand why, when and how to use the data in their new format.
- Adopt policy settings, training, and technical support, and encourage staff to publish species occurrences to GBIF.
- Encourage staff to use GBIF to obtain species occurrence data.
- Once data are published to GBIF, encourage staff to use GBIF to fulfil internal and external 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.
- Collaborate with other New Zealand-based data publishers and GBIF-NZ to provide training and capacity building.
- Collaborate with other GBIF participants to develop common analytical and reporting tools based on GBIF services.
- Collaborate with appropriate GBIF participants, both within New Zealand and globally, to identify areas that may need to be expanded to support other species occurrence dimensions or sources.
- Collaborate with other New Zealand agencies to develop guidelines and, where necessary, vocabularies to support the publication and use of different types of data – particularly where common methodologies are used.

4 Acknowledgements

Some material for this report was derived from the following document:

Wilton AD, Jewell U, Goodsell B 2023. Potential for regional councils to use GBIF to access and share species occurrence data. Manaaki Whenua – Landcare Research contract report LC4381. Envirolink Grant: 2329-ORC004.

5 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[18] 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[19] 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.

5.1 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 these data from the publishers, integrates the

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

108 participants (including NZ)

2,367 publishing institutes

111,637 data sets

3,068,061,598 occurrence records

https://www.gbif.org/

New Zealand

Member since 2001

467 publishers of NZ occurrences

16 publishers within NZ

15,374,884 NZ occurrences

1,587 data sets that include NZ occurrences

https://www.gbif.org.nz, 12 Feb 2025

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

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

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[20] and Creative Commons[21] 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[22] 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.[23] 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 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) occurrence, identification, locality, and event data. Occurrence data sets are the most frequent data

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

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 use Darwin Core fields but differ in the arrangement, or structure, of the data. As a consequence, they have different required and recommended fields.

5.1.1 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

5.2 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 Metadata Language (EML), and the Darwin Core Archive.

5.2.1 Darwin Core

Darwin Core[24], sometimes abbreviated as DwC, is a data standard that has been developed by Biodiversity Information Standards (TDWG)[25], 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, identification of the organism, and more (Figure 1). Darwin Core is being actively maintained and extended by the TDWG community.

[24] https://www.tdwg.org/standards/dwc/
[25] https://www.tdwg.org/

GBIF uses Darwin Core as a 'stable, straightforward and flexible framework for compiling biodiversity data'[26]. 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 3. 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)

5.2.2 Ecological Metadata Language (EML)

Ecological Metadata Language (EML)[27] 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.[28] 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 analyse 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 includes an EML file as one of its components (see below).

5.2.3 Darwin Core Archive

Darwin Core Archive (sometimes abbreviated as DwC-A) is the preferred format for publishing data in the GBIF network. The Darwin Core Archive is a GBIF specification for a self-contained data set consisting of the metadata and data files, which are arranged using a star-schema approach (Figure 4). The four types of file in the archive are as follows.

[26] https://www.gbif.org/standards
[27] https://eml.ecoinformatics.org/
[28] https://knb.ecoinformatics.org/

- **Core data file**: the main or central data file, containing sampling-event, occurrence or checklist data. This file is formatted as a comma-separated value (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.
- **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[29]. The list of extensions available is maintained in the GBIF Extension Repository[30] (e.g., Humboldt Ecological Inventory, GBIF Relevé)
- **Metafile** (meta.xml in **Error! Reference source not found.**): an XML-formatted file that d escribes 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 4): an XML file that records a description of the data set using EML (see above).

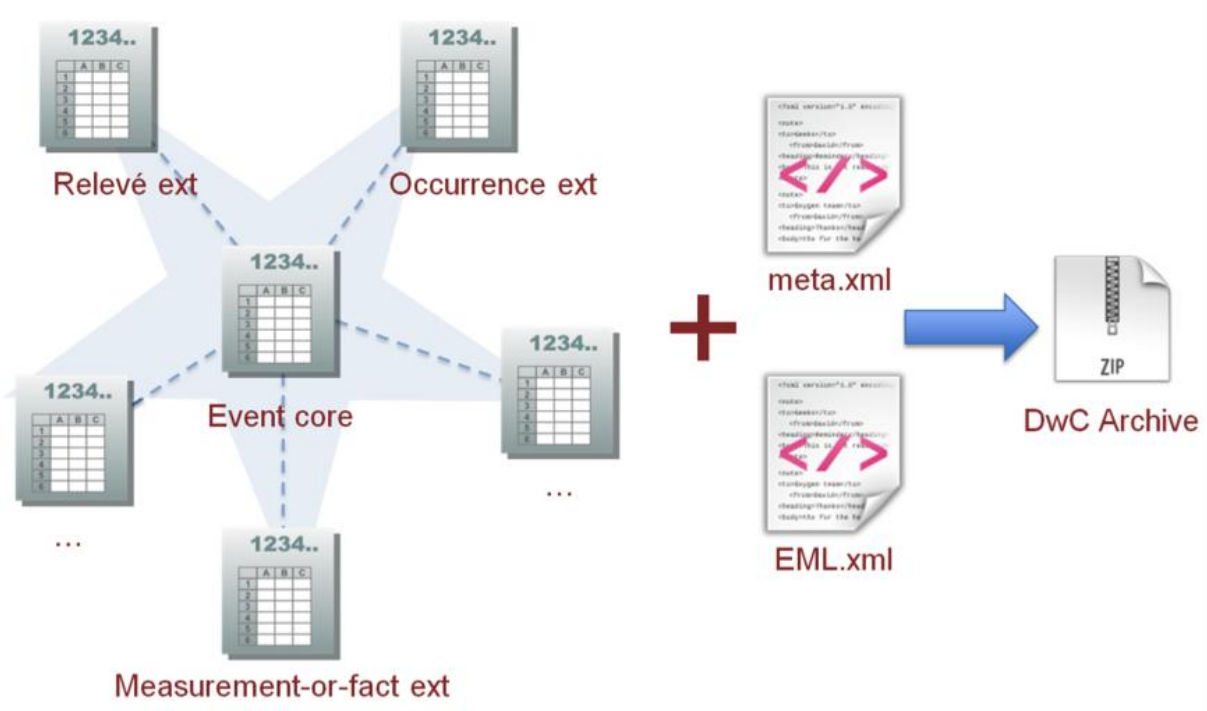

Figure 4. Structure and typical contents of a Darwin Core Archive.
(Source: GBIF IPT Manual, https://ipt.gbif.org/manual/en/ipt/latest/dwca-guide, CC-BY 4.0)

5.2.4 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/.

[29] https://www.tdwg.org/standards/ac/
[30] https://rs.gbif.org/extensions.html

5.3 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 5), or by creating Darwin Core Archives using other processes.

5.3.1 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.

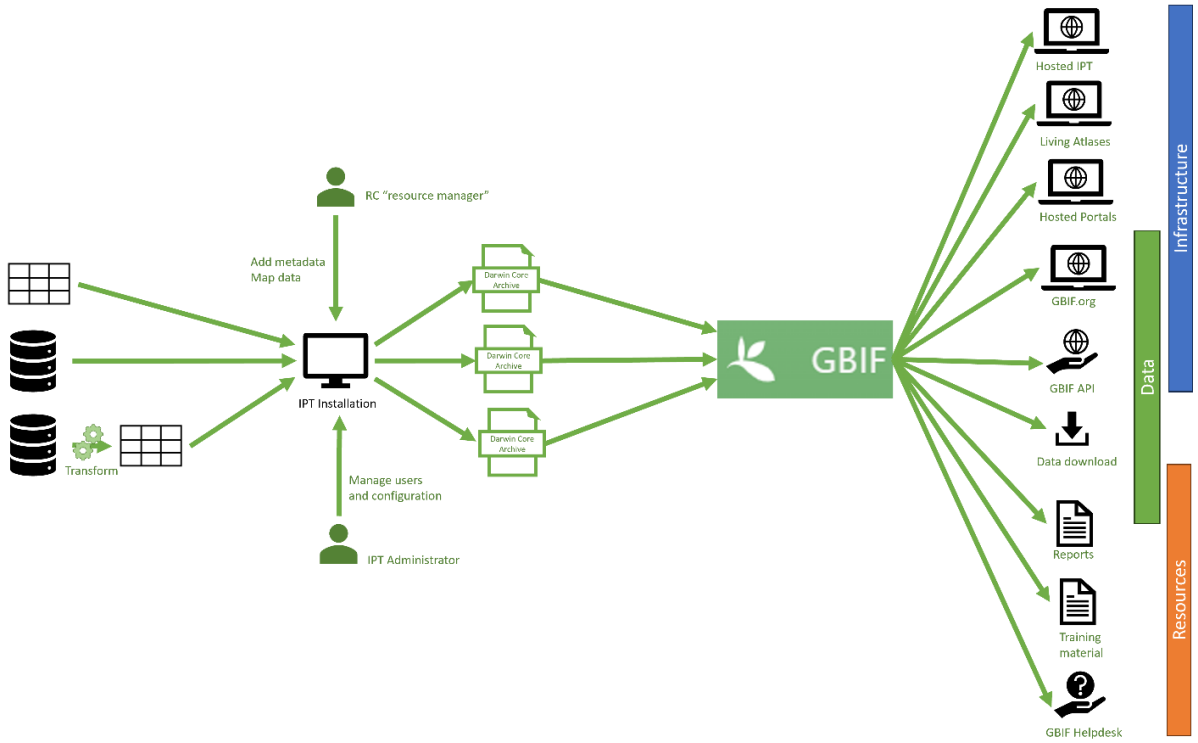

Figure 5. A conceptual overview of the GBIF network, showing publication using IPT, through to data, resources, and infrastructure provided by GBIF.

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

5.3.2 IPT deployment

IPT can be used 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 6). 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 use a hosted IPT installation. These installations can be hosted by another data publisher (*hosted installation* in Figure 6) or a participant node (*node-hosted* in Figure 6).

During 2023, GBIF-NZ worked with the Secretariat to establish a node-hosted instance of IPT for New Zealand.[31] 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. This installation is now available to New Zealand-based publishers.

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.

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

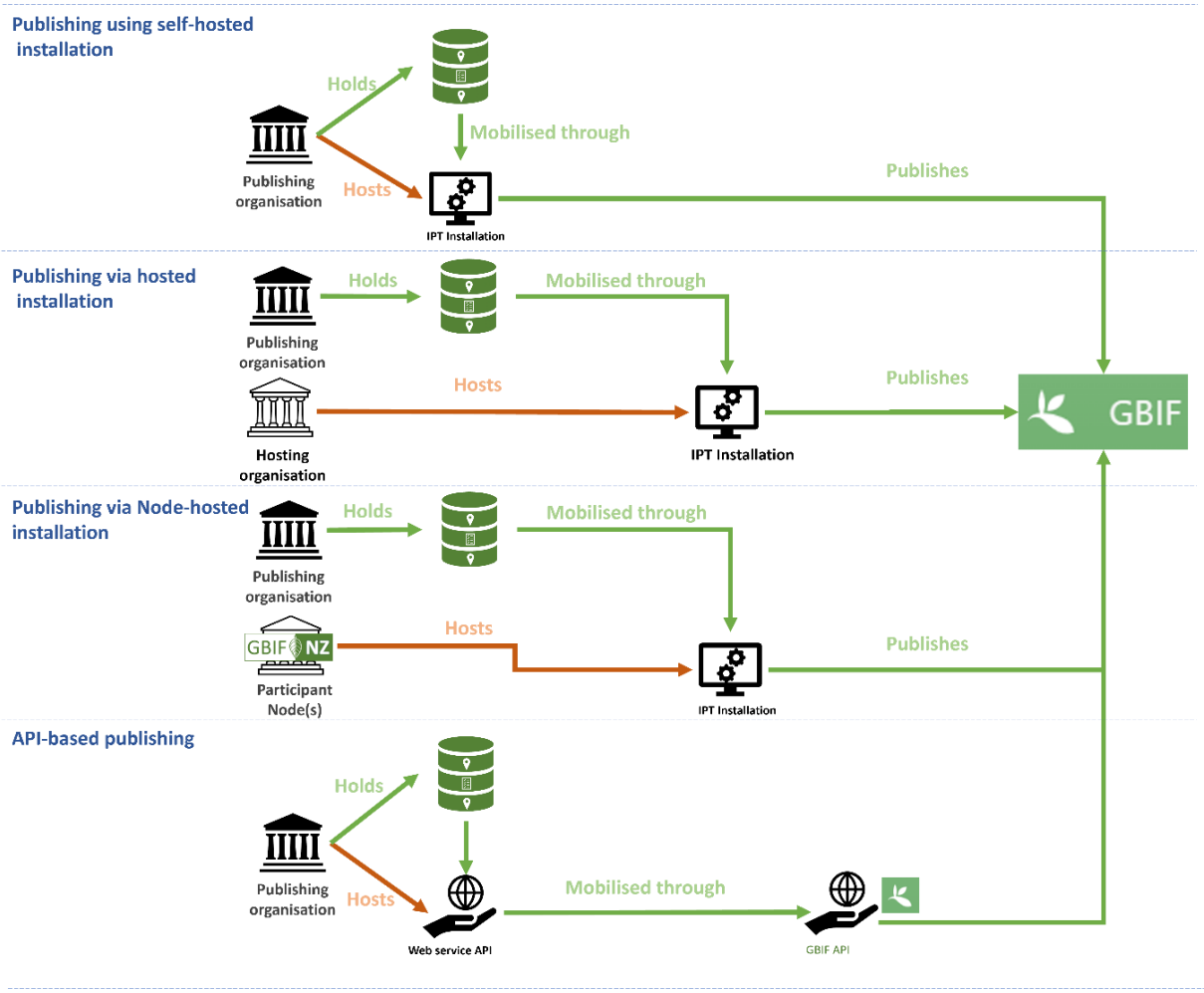

Figure 6. Summary of different approaches to publishing data to GBIF using IPT or the GBIF API.

5.3.3 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)[32] and agrees to the GBIF Data Publisher Agreement.[33] Application to become a publisher is made using a simple online process.

5.3.4 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 example 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

[32] https://www.gbif.org/become-a-publisher
[33] https://www.gbif.org/terms/data-publisher

    - 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 Darwin Core Archive 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

5.4 Infrastructure and services

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

- **Hosted portals**[34]**:** 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[35].
- **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[36] 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)[37]: This 'is a comprehensive and community-curated clearing house of information about scientific collections in the GBIF registry'[38].
- **Data access tools:** GBIF maintains a list of tools that facilitate data access and analysis[39]. These include, for example, an R library (rgbif)[40] and a python library (pygbif)[41] for accessing data from the GBIF API.

5.4.1 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/

---

[34] https://www.gbif.org/hosted-portals
[35] https://www.gbif.org.nz
[36] https://ipt.gbif.org.nz
[37] https://scientific-collections.gbif.org/
[38] https://scientific-collections.gbif.org/about
[39] https://www.gbif.org/resource/search?contentType=tool
[40] https://www.gbif.org/tool/81747/rgbif
[41] https://www.gbif.org/tool/OlyoYyRbKCSCkMKIi4oIT/pygbif-gbif-python-client

5.5 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[42] initiative, resulting in the development and publication of the GBIF-NZ portal[43]. 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[44] for New Zealand.

GBIF-NZ has worked with the GBIF Secretariat to establish a national hosted IPT installation[45]. 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.

5.5.1 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[46] (Figure 7). These providers are currently Crown Research Institutes, Museums, Regional Councils and community initiatives (Figure 8). 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 and five regional councils recently piloted publishing data to GBIF as part of a pilot investigating the potential to use of GBIF to publish and/or access their holdings of species occurrence data.

[42] https://www.gbif.org/composition/3kQFinjwHbCGZeLb5OhwN2/gbif-hosted-portals
[43] https://www.gbif.org.nz
[44] https://living-atlases.gbif.org/
[45] https://ipt.gbif.org.nz/
[46] 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 7. The number of New Zealand species occurrence records available via GBIF according to the publishing country. (Data accessed: 12 Feb 2025, https://api.gbif.org/v1/occurrence/counts/publishingCountries?country=NZ)

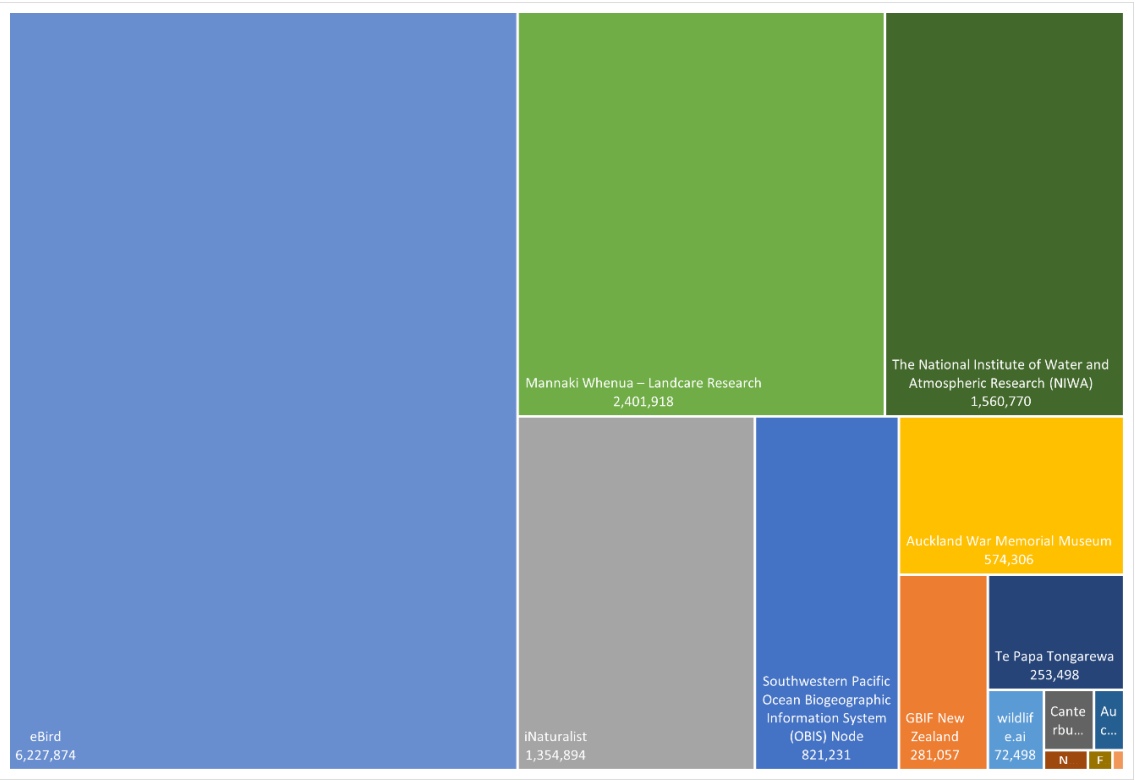

Figure 8. The number of records contributed by New Zealand-based data providers.[47] (Data accessed: 12 Feb 2025, https://api.gbif.org/v1/occurrence/search?publishingCountry=NZ&facet=publishingOrg&limit=0&facetLimit=50)

[47] 'New Zealand-based providers' is based on the country information included in the GBIF data set registration, even if the underpinning information infrastructure resides overseas (e.g. eBird).

Appendix 1 – Glossary of selected terms and abbreviations

API	Application programming interface: a software interface that allows information systems to communicate.
CSV	Comma-separated values: a text-based file in which records are separated by new lines, and fields are separated by commas.
Darwin Core	A data standard maintained by Biodiversity Information Standards (TDWG)
Darwin Core Archive	A self-contained data archive format defined by GBIF, which contains metadata describing the provenance and structure of the data as well as the biodiversity data.
eBird	A citizen science platform maintained by the Cornell Lab of Ornithology, represented in New Zealand by New Zealand eBird (https://ebird.org/newzealand/home)
EML	Ecological Metadata Language: 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.
Integrated Publishing Toolkit	Integrated Publishing Toolkit: a web-based application developed and maintained by GBIF. Usually abbreviated to IPT.
Living Atlas	The open-source platform that has been developed by the Atlas of Living Australia. This platform has now been adopted by other GBIF Nodes that 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 which 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 which 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	Biodiversity Information Standards: the abbreviation is based on the original name and scope of the organisation – Taxonomic Database Working Group.
TSV	Tab-separated value: a text-based file in which records are separated by new lines and fields are separated by tabs
UUID	Universal unique identifier: 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 considered to be unique.
vernacular name	An informal name, in any language, assigned to a taxon, or taxa, by a community. Also referred to as common name.
XML	Extensible Markup Language: a hardware- and software-independent specification for storing and transmitting data. It is maintained by the World Wide Web Consortium (W3C).

Appendix 2 – List of data sets assessed

Weeds	Weeds GIS application
Tier 1 5MBC	Tier 1 5-minute bird counts
Tier 1 Bats	Tier 1 acoustic recording for bats
Tier 1 BirdARD	Tier 1 acoustic recording for birds
Tier 1 BirdIncidentals	Tier 1 incidental bird detections
Tier 1 BirdDistance	Tier 1 5-minute distance sampling
Tier 1 DNA	Tier 1 Ungulate faecal pellet DNA sampling
Tier 1 Mammal sightings	Tier 1 Ground survey for introduced mammal pests
Tier 1 Mammal sign	Tier 1 Faecal pellet counts
Tier 1 Possum	Tier 1 Possum transect lines
Riverbird count summaries	
Kaki Master egg chick database	
Twizel Kaki Hide data	

Appendix 3 – Indicative data mappings

The following sections show more detailed *indicative data mappings* for each of the data sets. These mappings were developed to assist assessment of the data against the data standards. They were undertaken without consultation with the relevant domain experts, so it can be expected that some of these mappings will need to be refined with their collaboration. For most data sets these represent a partial mapping of DOC data fields using a limited number of key and/or problematic fields. The Weeds data provides the most complete mapping, and shows potential for internal and external versions of the mapped data.

Appendix 3.1 General

The following fields are, or should be considered, mandatory whenever available:

- occurrenceID
- eventID
- parentEventID
- basisOfRecord
- type (usually 'Event')
- occurrenceStatus
- organismQuantity and organismQuantityUnit (must be included in Sample Event)
- samplingProtocol
- eventDate
- institutionCode
- Country and CountryCode
- scientificName.

The following fields are not mandatory, but should be included as part of good practice (note that some of these fields are not included in all the following mappings for brevity):

- Rightsholder
- accessRights (when appropriate)
- informationWithheld (when appropriate)
- generalizations (when appropriate)
- modified
- institutionID
- stateProvince
- Kingdom
- Class
- Order
- Family
- Genus
- recordedBy and identifiedBy.*

* Ideally these would be the name of the person(s) who performed these actions, as can be useful for data validation, implying a degree of confidence, data quality and fitness. Where this must be withheld, consideration should be given to replacing it from a standardised vocabulary (e.g., DOC field staff, DOC science staff; DOC contractor).

Appendix 3.2 Weeds

Record filters

Further work is required to understand the meaning of retired infestation records and whether they represent a species occurrence. Some or all of them may need to be filtered out. Some of the patterns encountered in the data are illustrated below. This is by no means an exhaustive list.

OBJECTID	InfestationID	InfestationName	ScientificName	Adult Count	Adult Measure	Juvenile Count	Juvenile Measure	Seedling Count	Seedling Measure	Percentage Of Coverage	Flowering	Fruiting	RecOf Absence	INF_From_ Date	INF_To_ Date	GlobalID	Retired	SHAPE.area	SHAPE.len
801	WEEDINF1	ManiapotoDarwins	Berberis darwinii	546	Plants					Frequent (6-25%)	Yes	No	0	6/10/2007	3/10/2008	{40E5D360-CA7E-41BD-8B8E-092AE4EDE4CE}	0	7936096.185	15039.69394
802	WEEDINF1	ManiapotoDarwins	Berberis darwinii	667	Plants					Frequent (6-25%)	Yes	No	0	3/10/2008	16/10/2009	{0987CB0E-328B-4521-A264-33F1588842CC}	0	7936096.185	15039.69394
803	WEEDINF1	ManiapotoDarwins	Berberis darwinii	723	Plants					Frequent (6-25%)	Yes	No	0	16/10/2009	16/10/2010	{7F5D3273-EC3A-4591-944A-B5F1DF9ED40D}	0	7936096.185	15039.69394
805	WEEDINF1	ManiapotoDarwins	Berberis darwinii	703	Plants					Frequent (6-25%)	Yes	No	0	16/10/2010	14/10/2011	{71A71ED3-5B12-455F-9151-9BD9DA4470B0}	0	7936096.185	15039.69394
2402	WEEDINF1	ManiapotoDarwins	Berberis darwinii	888	Plants					Frequent (6-25%)	Yes	No	0	14/10/2011	19/10/2012	{DEBAAC47-DB98-4E9C-BE4C-7926366F3DEC}	0	7936096.185	15039.69394
2403	WEEDINF1	ManiapotoDarwins	Berberis darwinii	873	Plants					Frequent (6-25%)	Yes	No	0	19/10/2012	25/10/2013	{1E8B4D99-44E0-47DB-A6BA-0CB5D4291468}	0	7936096.185	15039.69394
2404	WEEDINF1	ManiapotoDarwins	Berberis darwinii	461	Plants					Frequent (6-25%)	Yes	No	0	25/10/2013	31/10/2014	{ADC9F0FE-B45C-4BE7-9BD4-C635EE2EACD8}	0	7936096.185	15039.69394
168967	WEEDINF1	ManiapotoDarwins	Berberis darwinii	268	Plants					Frequent (6-25%)	Yes	No	0	31/10/2014	31/12/9999	{D9E9A81B-6D6F-436A-A36A-DC02A3D4995F}	0	7936096.185	15039.69394
542910	WEEDINF10006	Pinus radiata Absence 20/10/2017	Pinus radiata										1	20/10/2017	31/12/9999	{2958F6AF-D8EC-43DE-92A1-FC4F865D6D6A}		24396.13148	890.4834253
542911	WEEDINF10007	Senecio elegans Absence 20/10/2017	Senecio elegans										1	20/10/2017	31/12/9999	{DB7A8074-A018-4E80-BDF7-C2DB06B23DB7}		24396.13148	890.4834253
553304	WEEDINF10127	Takapourewa cleavers	Galium aparine	20	Percentage Cover					Frequent (6-25%)	No	Yes	0	7/11/2017	8/11/2017	{76D2F320-1A5F-4442-877C-A37FA2FB444B}		113.6706961	69.42290574
811923	WEEDINF10127	Takapourewa cleavers	Galium aparine	20	Percentage Cover					Frequent (6-25%)	No	Yes	0	8/11/2017	31/12/9999	{FF80F02C-225E-4A95-AF4B-D6DCD42B7B0A}	1	12.50281466	12.5505002
987659	WEEDINF1031	LakeOhia_Pines_2015	Pinus radiata							Abundant (51-75%)			0	26/06/2020	26/06/2020	{6898BEF1-C0B0-47EF-8433-5E7827DB4131}		9787554.111	53154.50931
987660	WEEDINF1031	LakeOhia_Pines_2015	Pinus radiata							Abundant (51-75%)			0	26/06/2020	26/06/2020	{19BE1997-78DC-40E1-920A-73265D6097E5}		9838941.193	66127.60105
987662	WEEDINF1031	LakeOhia_Pines_2015	Pinus radiata							Abundant (51-75%)			0	26/06/2020	26/06/2020	{62560EEA-BF4E-4AC8-BD02-226964CF2C98}		9827655.349	65865.92029
987663	WEEDINF1031	LakeOhia_Pines_2015	Pinus radiata							Abundant (51-75%)			0	26/06/2020	31/12/9999	{7CC94B3B-AFCF-43CF-978F-DD52FCCF8982}	1	9796899.806	55749.33763
987664	WEEDINF1031	LakeOhia_Pines_2015	Pinus radiata							Abundant (51-75%)			0	26/06/2020	31/12/9999	{51E2EC9C-F3CC-4518-B435-8E97E5FAB89D}	1	9838784.909	66064.84931
683885	WEEDINF11186	Craigieburn Forest - P. contorta	Pinus contorta	1	Plants	1	Percentage Cover	1	Percentage Cover	Occasional (2-5%)	No	No	0	23/08/2018	15/06/2020	{8AADD447-1D77-4370-A98B-7AD88D63BF3E}		51743394.48	39481.1339
974852	WEEDINF11186	Craigieburn Forest - P. contorta	Pinus contorta	1	Plants	1	Percentage Cover	1	Percentage Cover	Occasional (2-5%)	No	No	0	15/06/2020	12/08/2020	{7FA9455D-591F-4ADF-8E16-F9921977157F}		51743394.48	39481.1339
998469	WEEDINF11186	Craigieburn Forest - P. contorta	Pinus contorta	1	Plants	1	Percentage Cover	1	Percentage Cover	Occasional (2-5%)	No	No	0	12/08/2020	31/12/9999	{EA6A1B1F-3481-4B1C-AC62-0D25DF4C4B7F}		51743394.48	39481.1339
133659	WEEDINF2521	KapitiBLA	Rubus fruticosus agg.	33	Plants					Frequent (6-25%)	No	No	0	16/07/2010	24/10/2018	{69B1C9E5-B8DA-4AE2-8609-B6912349D017}	0	4532.236711	660.4360381
740320	WEEDINF2521	KapitiBLA	Rubus fruticosus agg.							Scarce (1%)			0	24/10/2018	11/11/2020	{00ECD421-4002-4335-8F30-EEAC01B7ACA3}		78.14115398	31.37596185

OBJECTID	InfestationID	InfestationName	ScientificName	Adult Count	Adult Measure	Juvenile Count	Juvenile Measure	Seedling Count	Seedling Measure	Percentage Of Coverage	Flowering	Fruiting	RecOf Absence	INF_From_ Date	INF_To_ Date	GlobalID	Retired	SHAPE.area	SHAPE.len
1037281	WEEDINF2521	KapitiBLA	Rubus fruticosus agg.							Scarce (1%)			0	11/11/2020	27/01/2021	{BBD0B8C1-0CEF-43CD-AC5B-A7BF36762773}		390.7081698	94.12812637
1064904	WEEDINF2521	KapitiBLA	Rubus fruticosus agg.							Scarce (1%)			0	27/01/2021	18/01/2024	{2306BB35-C638-49A8-BCF3-E67BA322A37A}		312.5670158	62.75216452
1443763	WEEDINF2521	KapitiBLA	Rubus fruticosus agg.							Scarce (1%)			0	18/01/2024	31/12/9999	{BF2B119D-64BE-4875-A10C-5BCF152C1925}		625.1340316	125.504329
202982	WEEDINF4748	ATNP_Bay	Laurus nobilis							Frequent (6-25%)			0	19/10/2015	20/10/2015	{C51FFC5B-F221-4ED9-B132-82C3ECF3399B}		3750.811189	376.5133383
223425	WEEDINF4748	ATNP_Bay	Laurus nobilis							Frequent (6-25%)			0	20/10/2015	31/12/9999	{024BCCAB-B385-48DF-9F6B-B0F69AC70317}		5045.807297	445.9738495

Legend

Same infestation over multiple years. Number of plants varies by year but the area remains the same.

Records of absence.

Retired infestation, but no indication the weed has been eradicated

Same infestation over multiple years. Number of plants not provided. Area changes by year. The sequence ends with two retired records, both with the same date range but with different areas – reason unclear.

Different units of measure for different life stages

Preliminary field mapping and conversions

Weeds export		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
		Record	type	Event	Constant	The dc record type is 'Event', even though mapping to a Darwin Core occurrence.
Created_Date Last_Updated_Date	2014-08-26 9:37:52		modified	2014-08-26	IF Last_Updated_Date <> null then modified := Last_Updated_date ELSE modified := Created_Date	Use just the date component of the datetime, not the time.
			licence		Constant	Include this. Seek advice from Legal on its content (type of licence).
			institutionID	https://www.gbif.org/grscicoll/institution/[xxxx]	Constant	DOC's registration number from GRSciColl or similar once DOC is a registered publisher.
			datasetID			Optional. DOC to decide if it will have a standard of assigning a UUID to all its data sets.
			institutionCode	NZ Government Department of Conservation	Constant	There are other institutions with 'Department of Conservation' in their names (e.g. Missouri, Western Australia). Need to be specific.
			datasetName	NATIS Operational Weed Infestations	Constant	
			basisOfRecord	HumanObservation	Constant	

Weeds export		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
			informationWithheld	Some data withheld. See metadata for details.	Constant	These exclusions apply only to the external version. No exclusions from the internal version.
			dataGeneralizations	null		No need to reduce the precision for weeds. If there are data that should not be published, then they should be withheld rather than generalised – see previous point.
GlobalID	{40E5D360-CA7E-41BD-8B8E-092AE4EDE4CE}	Occurrence	OccurrenceID	{40E5D360-CA7E-41BD-8B8E-092AE4EDE4CE}		This is one of two IDs on the source row that are unique within the data set. Have used this one rather than OBJECTID. See section on identifiers and linkages.
			recordedBy	DOC Personnel	Constant	Individual details not available in source.
AdultCount JuvenileCount SeedlingCount AdultMeasure JuvenileMeasure SeedlingMeasure	5 27 562 Plants Stems Stems per hectare etc		individualCount organismQuantity organismQuantityType		See appendix for calculation. See also measurementOrFact	There is complexity around the combinations of different values in the input fields and that involves more detail than is appropriate for the body of this document. To avoid losing the thinking that has gone into this, a method of mapping inputs to outputs depending on combinations of input values is proposed in Appendix 4. That is subject to

Weeds export		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
						confirmation at time of mobilising the data. However, the conclusion is that the data can be mapped.
Flowering Fruiting			reproductiveCondition	Flowering Fruiting Flowering | Fruiting		This mapping is based on an assumption that this data element can take a list of values. This assumption is to be confirmed. If the assumption is incorrect and a list cannot be used then a Measurement or Fact could be used. Either way, it can be mapped.
			establishmentMeans	Introduced	Constant	By definition, weeds are introduced, directly or indirectly.
RecOfAbsence	0 1		occurrenceStatus	Present Absent	0 -> Present 1-> Absent	
Comment	ID'd by Peter de Lange on retaining wall behind red house. Exotic Allister Cameron 0274 330967 Road sides, Otira township / Horse Paddock, Rata Lodge backpackers		occurrenceRemarks			Include this in the internal Darwin Core Archive, but not in the external one because it is free text and may contain sensitive information.

Weeds export		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
			associatedOccurrences	null		See section on identifiers and linkages.
		Location	locationID			There is no ID in the source that uniquely identifies the location. See section on identifiers and linkages.
			country	New Zealand	Constant	
			countryCode	NZ	Constant	
			stateProvince	<region name>	Use GIS to determine this based on the polygon and on a reference source of regional and unitary council boundaries.	
			locality	<nearest named place>	Use GIS to determine this.	
			locationAccordingTo		Constant	The gazetteer(s) and version or other source(s) used for stateProvince, locality, etc
			decimalLatitude		The decimal lat./long. of the centroid of the footprintWKT representation of the polygon.	For consistency, it should be the centroid of the WKT representation, not of the original ArcGIS representation, if different. As a future enhancement, it may be desirable to modify this to ensure the centroid is within the polygon it represents (not necessarily the case by default for
			decimalLongitude			

Weeds export		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
						concave polygons) and/or in habitat that is representative of the polygon (e.g. not in a water body, when it represents terrestrial weeds, etc.). However, this would need to be considered in the context of a change of centroid possibly causing an unintended and undesired change of higher geography (region, etc).
			geodeticDatum		Constant	Obtain this from the team performing conversion of coordinates
			coordinateUncertaintyInMeters		Transform	GIS needed to calculate this. The horizontal distance between the centroid and the furthest point of the WKT representation of the polygon.
			pointRadiusSpatialFit		pointRadiustSpatialFit = $\pi r^2 / A$ Where r = coordinateUncertaintyInMeters A = sampleSizeValue (area of the polygon)	
			footprintWKT			GIS needed to calculate this.

Weeds export		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
						Not part of the existing source data, but an initial internet search indicates it can be created by ArcGIS.
			footprintSRS		Constant	Need information from GIS on this.
?		Event	eventID		TBD	See section on identifiers and linkages.
			eventType	site visit	Constant	
INF_From_Date	2012-10-19 0:00:00		verbatimEventDate	2012-10-19 0:00:00		
			eventDate	2012-10-19	Transform	Use just the date component of the date-time
SHAPE_area	7936096.185		sampleSizeValue	7936096.185		
			sampleSizeUnit	square metre	Constant	
CollectionCode	BERDAR LYCFER	Identification	verbatimIdentification			
		Taxon	taxonID		null	Have mapped the ID applicable to scientificName to scientificNameID, not to taxonID.
ScientificName	Berberis darwinii		scientificName		Mapped 1:1 for those scientific names that have an exact equivalent in the source of taxonomic reference data. Otherwise a human decision will be needed on which scientific name to use	

Weeds export		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
					that is valid in that reference source.	
			kingdom		Taken from the source of taxonomic reference data, based on the scientificName.	
			phylum			
			class			
			order			
			family			
			genus			
			specificEipthet			
CommonName	Darwin's barberry Chilean rhubarb		vernacularName			There is no equivalent of this in NZOR. Use the value from the source data.

There will be separate instances of Measurement or Fact for each of the populated life stages. For example:

Measurement or Fact				
		measurementType	**measurementValue**	**measurementUnit**
AdultCount = 100	AdultMeasure = Plants	Adult	100	Individuals
JuvenileCount = 30	JuvenileMeasure = Stems per Ha	Juvenile	30	Stems per Ha
SeedlingCount = 20	SeedlingMeasure = Percentage Cover	Seedling	20	Percentage Cover

Not mapped						
Weeds export		**Darwin Core**		**Output example**	**Processes required**	**Notes**
Field	**Class**	**Class**	**Field**			
OBJECTID	1130524 371556					This is 1 of 2 unique IDs on every row. Have used GlobalID for ocurrenceID instead.
SpeciesID	80 656					There will be a species ID in the Darwin Core, but not this one. It will be obtained by using the scientific name to reference the chosen source of taxonomic reference data and using the corresponding ID from that.

Appendix 3.3 Tier 1 5MBC

Recommended core: sampling event core (can also be mapped to occurrence core)

Worked example as sampling event

The following tables illustrate a worked example for 5MBC data expressed as a Sampling Event to illustrate the use of Extended Measurement or Fact. Use of the Extended version is necessary to maintain the linkage of data that relates to an occurrence, rather than all occurrences under an event.

In the example only a section of fields necessary to illustrate the example is included.

Place	Station	Season	DateStarted	Timestamp	SpeciesName	TemperatureDesc	Near	Far	VeryFar
BM7	D	2013-14	19/11/2013	09:16:00	Goldfinch	16 - 22 °C	0	0	3
BM7	A	2013-14	19/11/2013	09:37:00	Goldfinch	16 - 22 °C	0	0	4
BM7	A	2013-14	19/11/2013	09:37:00	Gull, Southern Black-backed	16 - 22 °C	0	0	2
BM7	A	2013-14	19/11/2013	11:38:00	Goldfinch	> 22 °C*	0	0	4
BM7	A	2018-19	3/11/2018	10:16:00	Shelduck, Paradise	11 - 15 °C	0	0	0

* Altered from original data for illustration

Sample Event core

eventID	Place	parentEventID	DateStarted	Timestamp
2013-14-BM7	BM7			
2	[BM7-]D	2013-14-BM7	19/11/2013	09:16:00
3	[BM7-]A	2013-14-BM7	19/11/2013	09:37:00
4	[BM7-]A	2013-14-BM7	19/11/2013	11:38:00
2018-19-BM7	BM7			
6	[BM7-]A	2018-19-BM7	3/11/2018	10:16:00

Occurrence extension

eventID	occurrenceID	vernacularName
2	1	Goldfinch
3	2	Goldfinch
3	3	Gull, Southern Black-backed
4	4	Goldfinch
6	5	Shelduck, Paradise

Measurement or Fact extension

eventID	measurementValue	measurementType
2	16 - 22 °C	Temperature category
3	16 - 22 °C	Temperature category
4	> 22 °C**	Temperature category
6	11 - 15 °C	Temperature category

Extended Measurement or Fact extension

eventID	occurrenceID	measurementValue	measurementType
2	1	0	Near individuals
2	1	0	Far individuals
2	1	3	VeryFar individuals
3	2	0	Near individuals
3	2	0	Far individuals
3	2	4	VeryFar individuals

eventID	occurrenceID	measurementValue	measurementType
3	3	0	Near individuals
3	3	0	Far individuals
3	3	2	VeryFar individuals
4	4	0	Near individuals
4	4	0	Far individuals
4	4	4	VeryFar individuals
6	5	0	Near individuals
6	5	0	Far individuals
6	5	0	VeryFar individuals

Record-level filters

Not all records are suitable for publication to GBIF as they do not represent species occurrences.

- Omit records with StationNotMeasured = Y.

Preliminary field mapping and conversions

5MBC export		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
		Record	type	Event	Constant	
			modified	[ISO date]		Recommended if available.
			institutionID	https://www.gbif.org/grscicoll/institution/[xxxx]	Constant	DOC registration number from GRSciColl [48]or similar.
			datasetID			
			institutionCode	New Zealand Government Department of Conservation	Constant	
			datasetName		Constant	
			basisOfRecord	HumanObservation	Constant	
			informationWithheld			
			dataGeneralizations			
ID	489393	Occurrence	OccurrenceID	489393		
			recordedBy			
			individualCount	3	Sum(Near, Far, VeryFar)	The total counts could be passed in the field individualCount or in the paired organismQuantity fields.
			organismQuantity	3	Sum(Near, Far, VeryFar)	
			organismQuantityType	individuals	Constant	

[48] https://scientific-collections.gbif.org/ - a world registry of scientific collection, also includes data holders publishing to GBIF.

5MBC export		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
			occurrenceStatus	present	Constant	Optional
			associatedOccurrences		Calculated	Optional
		Location	locationID	b278fcb5-3b17-4810-b813-602a612ae2c4		
Place	BM7		verbatimLocality	Plot BM7, station A		
Station	A					
			country	New Zealand	Constant	
			countryCode	NZ	Constant	
			stateProvince	<regional boundary>	GIS classification	
			locality	<nearest named place?>	GIS classification	
			locationAccordingTo	<name of spatial layer>		
BIRA_X	1274411.289	verbatimCoordinates		1274411.289 5092061.604		Fields dependent on station (BIR*_X + BIR_*_Y)
BIRA_Y	5092061.604					
		verbatimCoordinateSystem		EPSG:2193		
			decimalLatitude		Calculate from BIT*_X, BIR*_Y	
			decimalLongitude			
			geodeticDatum			
			footprintWKT			
			footprintSRS			
		Event	eventID			
			parentEventID	2013-14-BM7		

5MBC export		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
			eventType	Five-minute bird count	Constant	
DateStarted	19/11/2013		verbatimEventDate	19/11/2013		
			eventDate	2013-11-19	Transform	
TimeStarted	09:37:00		eventTime	09:37:00-12/09:42:00-12	Calculate end time, append time zone offset	Time zone indicator appended as -12, -12:00 or -12:00
			habitat			
			samplingProtocol	DOC Five-minute bird count		
			sampleSizeValue	5		
			sampleSizeUnit	minute		
Observer_1	Penelope Gillette		recordedBy	Penelope Gillette | Ashley Smith	Concatenate(Observer_1, Observer_2, Observer_3 using ' | ' separator).	
Observer_2	Ashley Smith					
Observer_3						
SpeciesName	Skylark	Identification	verbatimIdentification	Skylark		
			identifiedBy			
ScientificName	Alauda arvensis	Taxon	scientificName	Alauda arvensis Linnaeus, 1758		
			kingdom	Animalia		
			phylum	Chordata		
			class	Aves		
			order	Passeriformes		
			family	Alaudidae		
			genus	Alauda		

5MBC export		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
			specificEipthet	arvensis		
Measurement or Fact						
SunOverheadDesc	2 min		measurementValue			
TemperatureDesc	16 - 22 °C		measurementValue			
PrecipitationLevelDesc	None		measurementValue			
PrecipitationTypeDesc	Mist		measurementValue			
WindDesc	Leaves/branches in constant motion		measurementValue			
OtherNoiseDesc	Loud		measurementValue			
Near	0		measurementValue			
Far	0		measurementValue			
VeryFar	2		measurementValue			
Not mapped						
ResultMasterID	2346818					
Season	2013-14					
StationNotMeasured	N					
ReasonNotMeasured						
Remeasurement						
RemeasurementReason						
PrecipitationType	M					
StationNotes						
SunOverhead	2					
Temperature	5					

5MBC export		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
PrecipitationLevel	0					
Wind	2					
OtherNoise	2					
DistanceUnknown						
ResultNotes						
EntryOrder						
MonitoringPlaceID	870					
Seen	[No data in sample]					
Heard	[No data in sample]					
TotalCount						

Notes

- ResultMasterID: not unique within the data set; duplicate values across records.

Appendix 3.4 Tier 1 Bats

Recommended core: occurrence

Record-level filter(s)

- Exclude records where category = 'Non-bat'.

Preliminary field mapping and conversions

DOCMON		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
		Record	type	Event	Constant	
			basisOfRecord	MachineObservation		
			modified			
ID	55798	Occurrence	occurrenceID	55798		
AssignedSite	CM70	Location	locationID	8c791073-cdb7-4500-9fe6-e8420528c2c0		
			verbatimLocality	Plot CM70		
MonStationAttributes.Easting			verbatimCoordinates			
MonStationAttributes.Northing						
			verbatimCoordinateSystem	EPSG:2193		
			decimalLatitude		Calculated from MonStationAttributes.Easting and .Northing	
			decimalLongitude		Calculated from MonStationAttributes.Easting and .Northing	
			geodeticDatum	EPSG:4326		

DOCMON		Darwin Core		Output example	Processes required	Notes
Field	**Example**	**Class**	**Field**			
		Event	eventID			
			parentEventID	2022-23-CM70		
DateStarted	26/03/2023		verbatimEventDate	26/03/2023		
			eventDate	2023-03-26	Convert verbatim	
TimeStarted	0:13:28		eventTime	00:13:28-12	Add time zone	
Category	Long tail	Identification	verbatimIdentification	Long tail		
Observer	Moira Pryde		identifiedBy	Moira Pryde		
ScientificName	Chalinolobus tuberculatus		scientificName	Chalinolobus tuberculatus (Forster, 1844)		
Not mapped						
BatFolderName	.\CM70_BAT_2022					
BatFileName	.\20230326_001328.bmp					
MonitoringPlaceID	3130					
ResultmasterID	2407294					
Season	2022-23					

Appendix 3.5 Tier 1 BirdARD

Core: Occurrence or sampling event

Preliminary field mapping and conversions

DOCMON		Darwin Core		Output example	Processes required
Field	**Example**	**Class**	**Field**		
		Record	type	Event	Constant
			basisOfRecord	MachineObservation	
			Modified		
ID	255672	Occurrence	occurrenceID	255672	
Season	2022-23				
Place	T181	Location	verbatimLocality	T181, BIRP	Concatenate
Station	P				
PlotCornerP_Easting	1218415.739		verbatimCoordinates		
PlotCornerP_Northing	4786933.213				
			decimalLatitude	Calculated from the corresponding PlotCorner easting and northing fields.	
			decimalLongitude		
DateARD	30/03/2023		verbatimEventDate	30/03/2023	
			eventDate	2023-03-30	
TimeARD	3:00:07		eventTime	3:00:07-12	
			samplingProtocol	Acoustic recording	Constant
SpeciesName	Kiwi, spp		verbatimIdentification	Kiwi, spp	
ScientificName	Apteryx sp.		scientificName	Apteryx Shaw & Nodder	
Processor	Robin Long		identifiedBy	Robin Long	

DOCMON		Darwin Core		Output example	Processes required
Field	**Example**	**Class**	**Field**		
Not mapped					
Segment	f3				
FileNameARD	T181_BIRP_20230330_030007.wav.tier1.night.final.csv				
TypeARD	NOCTURNAL				
StationEasting	1218555.699				
StationNorthing	4787081.408				
StationLocationDate	29/03/2023				
PlotCornerP_LocationDate	29/03/2023				
ResultmasterID	2671358				
MonitoringPlaceID	4674				
MonitoringStationID	17654				

Notes

- ResultMasterID not unique across records.

Appendix 3.6 Tier 1 BirdIncidentals

Recommended core: occurrence

Record filter(s)

- Omit records where NoSpeciesRecords = 'Y'.

Preliminary field mapping and conversions

DOCMON		Darwin Core		Output example	Processes required
Field	**Example**	**Class**	**Field**		
		Record	type	Event	Constant
			basisOfRecord	HumanObservation	Constant
			modified		
ID	45563	Occurrence	occurrenceID	45563	
		Location	locationID		
Place	T181		verbatimLocality	T181	
Station	234m				
DateStarted	30/03/2023	Event	verbatimEventDate	30/03/2023	
			eventDate	2023-03-30	
			eventType	Site visit	Constant
			samplingProtocol	DOC incidental bird detections	Constant
StationNotes[49]	[Transferred from BIRP distance count record sheet: S-E [Silvereye] flew past.]		eventRemarks (internal only)	Station notes: [Transferred from BIRP distance count record sheet: S-E [Silvereye] flew past.]	

[49] Omit when Field contains values such as 'NA' or 'No species recorded'.

DOCMON		Darwin Core		Output example	Processes required
Field	**Example**	**Class**	**Field**		
ResultNotes[49]	Heard from plot			Result notes: heard from plot	Concatenate[50]
SpeciesName	Stewart Island brown kiwi | Rakiura Tokoeka	Identification	verbatimIdentification	Stewart Island Brown Kiwi | Rakiura Tokoeka	
		Taxon	vernacularName	Stewart Island brown kiwi | Rakiura Tokoeka	
ScientificName	Apteryx australis australis		scientificName	Apteryx australis australis	
NumberObserved	1	Occurrence	individualCount or organismQuantity + organism QuantityType	1	
Not mapped					
MonitoringPlaceID	4674				
EntryOrder	8168				
ResultMasterID	2377666				
NoSpeciesRecorded	N				
Season	2022-23				

[50] Concatenate StationNotes and ResultNotes for internal DOC use.

Appendix 3.7 Tier 1 BirdDistance

Core: Occurrence or sampling event

Preliminary field mapping

DOCMON		Darwin Core		Output example	Notes
Field	**Example**	**Class**	**Field**		
ID	20487959	Event	eventID	20487959	To be confirmed
			parentEventID	2023-24-CO94-A	
DateStarted	9/03/2024		verbatimEventDate	9/03/2024	
			eventDate	2024-03-09	
TimeStarted	9:38:00		eventTime	9:38:00	
Observer_1	Laura McIvor		recordedBy	Laura McIvor	
Observer_2	NA				
Observer_3	NA				
StationNotes	Some stream noise. Silvereyes fly in at end.		eventRemarks	Some stream noise. Silvereyes fly in at end.	
			samplingProtocol	Tier 1 Five-minute distance sampling	
			sampleSizeValue	5	
			sampleSizeUnit	minutes	
SpeciesName	Fantail, NZ / Black / Grey	Taxon	verbatimIdentification	Fantail, NZ / Black / Grey	
ScientificName	Rhipidura fuliginosa		scientificName	Rhipidura fuliginosa (Sparrman, 1787)	
ResultNotes	[Fantail recorded, assume New Zealand Fantail]	Identification	identificationRemarks	[Fantail recorded, assume New Zealand Fantail]	
Place	CO94	Location	verbatimLocality	Site CO94, station A, distance 25-45m	
Station	A				
DistanceDesc	26 - 45 metres				

DOCMON		Darwin Core		Output example	Notes
Field	**Example**	**Class**	**Field**		
WindDesc	Leaves still/move silently		MeasurementValue		
OtherNoiseDesc	Moderate		MeasurementValue		
SunOverheadDesc	0 min		MeasurementValue		
TemperatureDesc	11 - 15 °C		MeasurementValue		
PrecipitationLevelDesc	Dripping foliage		MeasurementValue		
PrecipitationTypeDesc	None		MeasurementValue		
ClusterPrecisionDesc	Accurate		MeasurementValue		
Not Mapped					
ClusterSize	1				
Distance	26-45m				
MonitoringPlaceID	3279				
ClusterPrecision	A				
StationNotMeasured	N				
ReasonNotMeasured	NA				
Remeasurement	Replaced				
RemeasurementReason	Old station not found				
ResultMasterID	2654003				
Season	2023-24				
OtherNoise	1				
PrecipitationLevel	1				
PrecipitationType	N				
SunOverhead	0				
Temperature	4				

DOCMON		Darwin Core		Output example	Notes
Field	**Example**	**Class**	**Field**		
Wind	0				
EntryOrder	108377				

Appendix 3.8 Tier 1 DNA

Core: occurrence.

This is a partial match only due to time constraints. Future work should investigate eDNA-specific extensions and workstreams.

Preliminary field mapping

DOCMON		Darwin Core		Output example	Notes
Field	**Example**	**Class**	**Field**		
ID	1042		occurrenceID		Tentative
Place	AD172	Location	verbatimLocation	Site AD172, Station 234m	
Station	234m				
Season	2022-23				
MeasurementDate	18/04/2023	Event	eventDate		
			parentEventID	2022-23-AD172-234	
			samplingProtocol	Tier 1 Faecal Pellet Monitoring	
			sampleSizeValue	3	
			sampleSizeUnit	Pellet swabs	
SampleLabel	ZZ01	MaterialSample	materialEntityID	ZZ01	Tentative
LabID	S1374_01		materialSampleID	S1374_01	Tentative
DNAResult	Cervus elaphus scoticus	Identification	verbatimIdentification	Cervus elaphus scoticus	
			scientificName	Cervus elaphus scoticus Lönnberg, 1906	
DNANotes	[RelevantGenBankMatch: MF872248.1]		identificationRemarks	[RelevantGenBankMatch: MF872248.1]	
Not Mapped					
HRMResult	NA				

DOCMON		Darwin Core		Output example	Notes
Field	**Example**	**Class**	**Field**		
Confidence	NA				
PCA	NA				
Sequencing	Cervus elaphus				
SampleNo	1				
ResultMasterID	2593305				
Match	100				
MonitoringPlaceID	870				
Seq	1171				

Appendix 3.9 Tier 1 Ungulate

Recommended core: Sampling Event

Worked example

ID	Place	Station	PlotNumber	Season	Observer	DateStarted	HabitatDesc	NumberOf RabbitPellets	IntactUngulate PelletsByGroup	Possum Pellets	Wallaby Pellets
1817731	CO94	AB	1	2023-24	Stephen Pilkington	9/03/2024	Forest	0	5*	N	Y*

* altered from original data

Sampling Event core

eventID	parentEventID	eventDate	Habitat	samplingProtocol	sampleSizeValue	sampleSizeUnit
2023-24-CO94		2023-07-01/2024-06-30				
2023-24-CO94-AB-UP-1	2023-24-CO94	2024-03-09	Forest	DOC Tier 1 Ungulate Pellet Count	1	m radius [tentative – exemplar only]

Occurrence extension

eventID	occurrenceID	verbatimIdentification	scientificName	occurrenceStatus	organismQuantity	organismQuantityType
2023-24-CO94-AB	?	Rabbit	Oryctolagus cuniculus (Linnaeus, 1758)	Absent	0	pellets
2023-24-CO94-AB	?	Possum	Trichosurus vulpecula (Kerr, 1792)	Absent	0	pellets
2023-24-CO94-AB	?	Wallaby	Notamacropus rufogriseus (Desmarest, 1817)	Present		
2023-24-CO94-AB		Ungullates	Euungulata	Present	5	Pellet groups

Preliminary field mapping and conversions

DOCMON		Darwin Core		Output example	Notes
Field	**Example**	**Class**	**Field**		
Place	CO94	Location	verbatimLocality	Site CO04, Station AB, Plot 1, Transect bearing 53	
Station	AB				
PlotNumber	1				
TransectBearing	53				
DateStarted	9/03/2024	Event	verbatimEventDate	9/03/2024	
			eventDate	2024-03-09	
Observer_1	Stephen Pilkington		recordedBy	Stephen Pilkington	
Observer_2	NA				
Observer_3	NA				
HabitatDesc	Forest		habitat	Forest	
			samplingProtocol	DOC Tier 1 Ungulate Pellet Counts	
			samplingSizeValue	1	Tentative
			samplingSizeUnit	metres radius	Tentative
Fields used to create occurrence record for each taxonomic group/species					
NumberOfRabbitPellets	0	Occurrence	Used to generate occurrenceStatus , organismQuantity, organismQuantityUnit, verbatimIdenfication, scientificName		
NumberOfHarePellets	0				
IntactUngulatePelletsByGroup	0				
Non-intactUngulatePellets	N				
PossumPellets	N				
RabbitPellets	N				
HarePellets	N				

DOCMON		Darwin Core		Output example	Notes
Field	**Example**	**Class**	**Field**		
PigDung?	N				
PigRooting?	N				
WallabyPellets	N				
Not mapped					
ID	1817731				
TotalPellets	0				
TotalGroups	0				
Season	2023-24				
EntryOrder	307927				
TurnPointID	NA				
Other	NA				
Habitat	F				
ResultMasterID	2654180				
ReasonNotMeasured	NA				
MeasuredReverse	NA				
MonitoringPlaceID	3279				
TransectNotMeasured	N				
TransectNotes	LJ-1, LJ-8, LJ-16, bluffs on either side of POSAA line so UNG count closer to POS line. [Data might be invalid because AB line was measured perhaps only 2 to 3m from POSAA line]				

Appendix 3.10 Tier 1 Mammal sightings

Core:Sampling event (or Occurrence)

Worked example

ID	Place	Season	DateStarted	NoSpeciesRecorded	Observer	SpeciesName	NumberObserved	AgeSexDesc
12149	AE137	2022-23	7/01/2023	Y	NA	No species recorded	0	NA
10620	AD139	2021-22	14/12/2021	N	Jess Randall	Chamois	4	Unidentified
10597	AB147	2021-22	15/01/2022	N	Megan Bogisch	Chamois	1	Unidentified
10598	AB147	2021-22	15/01/2022	N	Megan Bogisch	Chamois	5	Adult female
10599	AB147	2021-22	15/01/2022	N	Megan Bogisch	Chamois	5	Juvenile
10602	AB147	2021-22	15/01/2022	N	Katie Russ	Hare	1	Unidentified

Event core

eventID	parentEventID	eventDate
2022-23-AE137		2022-07-01/2023-06-30
2021-22-AD139		2021-07-01/2022-06-30
2021-22-AB147		2021-07-01/2022-06-30
A	2022-23-AE137	2023-01-07
B	2021-22-AD139	2021-12-14
C	2021-22-AB147	2022-01-15

Occurrence extension

eventID	recordedBy	scientificName	occurrenceStatus	organismQuantity	organismQuantityType	lifeStage	sex
A		Mammalia	absent	0	individuals		
B	Jess Randall	Rupicapra rupicapra (Linnaeus, 1758)	present	4	individuals	unidentified	unidentified
C	Megan Bogisch	Rupicapra rupicapra (Linnaeus, 1758)	present	1	individuals	unidentified	unidentified
C	Megan Bogisch	Rupicapra rupicapra (Linnaeus, 1758)	present	5	individuals	adult	female
C	Megan Bogisch	Rupicapra rupicapra (Linnaeus, 1758)	present	5	individuals	juvenile	unidentified
C	Katie Russ	Lepus timidus Linnaeus, 1758	Present	1	Individuals	unidentified	unidentified

Preliminary field mapping

DOCMON			Darwin Core		Output example 1	Output example 2	Notes
Field	**Example 1**	**Example 2**	**Class**	**Field**			
ID	12510	10620	Occurrence	occurrenceID	12510	10620	
Observer_1	NA	Jess Randall		recordedBy		Jess Randall	
Observer_2	NA	NA					
NumberObserved	0	4		organismQuantity	0	4	
				organismQuantityType	Individuals	individuals	
NoSpeciesRecorded	Y	N		occurrenceStatus	absent	present	
AgeSexDesc	NA	Unidentified		lifeStage		unidentified	
				sex		unidentified	
Place	CO94	AE139	Location	verbatimLocality	Site CO94, station 2x2km	Site AE139, station 2x2km	
Station	2x2km	2x2km					
Easting	NA	1306530		verbatimCoordinates		1306530 5123490	
Northing	NA	5123490					

DOCMON			Darwin Core		Output example 1	Output example 2	Notes
Field	Example 1	Example 2	Class	Field			
				verbatimCoordinateSystem		EPSG:2193	
				decimalLatitude	[add if possible]	Calculated from Easting/Northing	When absences are recorded the coordinate data are still important.
				decimalLongitude			
				geodeticDatum	epsg:4326	epsg:4326	
			Event	eventID	?	?	
				parentEventID	2023-24-CO94	2021-22-AE139	
DateStarted	8/03/2024	14/12/2021		verbatimEventDate	8/03/2024	14/12/2021	
				eventDate	2024-03-08	2021-12-14	
				samplingProtocol	DOC Mammal sightings	DOC Mammal sightings	
				samplingEffort			
StationNotes	No species recorded.	[The 7 Chamois observed by Gregory Whall GPS coordinate not recorded, entered as AE139 2016/2017 season corner P location]		eventRemarks	No species recorded.	[The 7 Chamois observed by Gregory Whall GPS coordinate not recorded, entered as AE139 2016/2017 season corner P location]	
SpeciesName	No species recorded	Chamois	Identification	verbatimIdentification	Exotic mammals	Chamois	

DOCMON			Darwin Core		Output example 1	Output example 2	Notes
Field	**Example 1**	**Example 2**	**Class**	**Field**			
				scientificName	Mammalia	Rupicapra rupicapra (Linnaeus, 1758)	
Not mapped							
Season	2023-24	2021-22					
NearestTransectID	NA	NA					
ResultMasterID	2654672	2148190					
EntryOrder	3736	3047					
MonitoringPlaceID	3279	879					
AgeSex	NA	U					Not required. Description used.

Notes

- Absence of sightings should be recorded as an occurrence with the scientificName = 'Mammalia' and occurrenceStatus = 'absent'.
- The taxonomic scope of mammal sighting data sets needs to be clearly stated in the metadata.

Appendix 3.11 Tier 1 Mammal sign

Recommended core:Sampling event

Preliminary field mapping

DOCMON		Darwin Core		Output example	Notes
Field	**Example**	**Class**	**Field**		
		Record	basisOfRecord	HumanObservation	
			type	Event	
ID	22220300	Occurrence	occurrenceID	22220300	Tentative
			occurrenceStatus	Present	
SignType	Pellets/dung		occurrenceRemarks	Pellets/dung	
Place	CP93	Location	verbatimLocality	Plot CP93, Station 234m	
Station	234m				
		Event	eventID		To confirm
			parentEventID	2023-24-CP93	
DateStarted	8/03/2024		verbatimEventDate	8/03/2024	
			eventDate	2024-03-08	
			samplingProtocol	DOC Mammal Sign	
			samplingEffort		To confirm
SpeciesName	Possum	Identification	verbatimIdentification	Possum	
			scientificName	Trichosurus vulpecula (Kerr, 1792)	
Not mapped					
MonitoringPlaceID	3339				
EntryOrder	20814				

DOCMON		Darwin Core		Output example	Notes
Field	**Example**	**Class**	**Field**		
Season	2023-24				
NoSpeciesRecorded	N				
StationNotes	NA				
ResultMasterID	2654680				

Appendix 3.12 Tier 1 Possum

Recommended core:Sampling event

Worked example

ID	Place	Station	DeviceNumber	Season	Night1Result	NonTargetSpecies
9998	CO94	AA	1	2023-24	NT	Rat
9999	CO94	AA	2	2023-24	P, NT	Rat

Sampling events

eventID	parentEventID
2023-24-CO94	
2023-24-CO94-AA	2023-24-CO94
2023-24-CO94-AA-1	2023-24-CO94-AA
2023-24-CO94-AA-2	2023-24-CO94-AA

Occurrence extension

occurrenceID	eventID	verbatimIdentification
9998-1	2023-24-CO94-AA-1	Rat
9999-1	2023-24-CO94-AA-2	Possum
9999-2	2023-24-CO94-AA-2	Rat

Record filter

Records with the following results would be excluded from publication to GBIF:

- BI – chewcard is beyond interpretation
- L – chewcard is lost
- NOT SET – chewcard was not set.

Preliminary field mapping and conversions

DOCMON		**Darwin Core**		**Output example**	**Notes**
Field	**Example**	**Class**	**Field**		
ID	386564	Occurrence	occurrenceID	386564	Tentative
Place	CO94	Location	verbatimLocality	Plot CO04, station AA, Device 1	
Station	AA				
DeviceNumber	1				
		Event	eventID		
			parentEventID	2023-24-CO94	
DateStarted	8/03/2024		verbatimEventDate	8/03/2024	
			eventDate	2024-03-08	
Night1ResultDesc	Non-target bite marks		eventRemarks	Non-target bite marks	
HabitatDesc	Forest		habitat		
DeviceType	Chewcard		samplingProtocol	Chewcard	Tentative. Requires more discussion.
TransectNotMeasured	N				
ReasonNotMeasured	NA				
RemeasurementReason	NA				

DOCMON		Darwin Core		Output example	Notes
Field	**Example**	**Class**	**Field**		
TransectNotes	Very steep in places				
Observer_1	Stephen Pilkington	Occurrence	recordedBy	Stephen Pilkington	
Observer_2	NA				
Observer_3	NA				
RainOvernight1Desc	None				
RainOvernight2Desc	NA				
TransectBearing	53				
TurnPointID	NA				
DeviceSet	RT				
DeviceSetDesc	Raised tree				
Night1Result	NT				
Night1Weight	NA				
Night2Result	NA				
Night2Weight	NA				
NonTargetSpecies	Rat	Identification	verbatimIdentification	Rat	
			scientificName	Rattus Fischer de Waldheim, 1803	
Remeasurement	Exactly Repeated				Tentative
Night2ResultDesc	NA				
Season	2023-24				
ResultMasterID	2654020				
ResultNotes	NA				
MonitoringPlaceID	3279				

DOCMON		Darwin Core		Output example	Notes
Field	**Example**	**Class**	**Field**		
EntryOrder	102556				
Habitat	F				
Night2Date	NULL				
KeaSafeTrapUsed?	NA				
Photos	NA				
TrapUsed	NA				
RainOvernight1	None				
RainOvernight2	NA				

Appendix 3.13 2013 River bird count summaries

Mapping

Mapping to either Event or Occurrence ore could be appropriate, but given the sparseness of the information occurrence core is the simpler option. Significantly no samplingProtocol data are provided (a required field for an Event Core data set), potentially precluding mapping to Sampling Event core.

Given the highly summarised nature of the data set a mapping table is not included here. If possible, it would be more appropriate to mobilise the original source data than using this highly summarised data.

The data could be mapped to Darwin Core and related standards.

Issues identified

- The data are highly summarised.
- The data lack precise dates, locality information, observers.
- There is a lack of unique identifiers.
- Observations are recorded against vernacular names.
- There is no habitat recorded.
- There is a lack of information on the sampling protocol.
- It appears to contain data from different sampling methodologies, time periods, and observers, which would make documenting (in the metadata) and interpreting more complex.

Appendix 3.14 2013 Kaki – Master egg chick database

Recommended core: Sampling Event

Worked example

For this data set two examples, Occurrence and Sampling event cores, are provided.

1. *Using Occurrence core*

Occurrence core

occurrenceID	organismID	eventDate	degreeOf Establishment	Pathway	sex	lifeStage	Location	Habitat	Country	recordedBy	Vitality
3253-01	BKBKW/GO	2010-10-19	wild		male	egg	MacKenzie Basin		New Zealand	DOC staff	alive
3253-02	BKBKW/GO	2010-11-10	captivity		male	hatched	DOC Bird Facility	Aviary	New Zealand	DOC staff	alive
3253-03	BKBKW/GO	2011-01-26	released	*Released / species management*	male	juvenile	MacKenzie Basin		New Zealand	DOC staff	alive
3254-01	10/48	2010-10-19	wild		male	egg	MacKenzie Basin		New Zealand	DOC staff	alive
3254-02	10/48	2010	captivity		male	egg	DOC Bird Facility	Aviary	New Zealand	DOC staff	Dead
3252-01	BKBKW/BKY	2010-10-19	Wild		male	egg	MacKenzie Basin		New Zealand	DOC staff	Alive
3252-02	BKBKW/BKY	2010-11-09	Captivity		male	hatched	DOC Bird Facility	Aviary	New Zealand	DOC staff	Alive
3252-03	BKBKW/BKY	2011-01-26	released	*Released / species management*	male	juvenile	MacKenzie Basin		New Zealand	DOC staff	alive

Resource relationship extension

resourceID	relatedResourceID	relationshipOfResource	relationshipEstablishedDate
BKBKW/GO	BKR/RO	parent (male)	2010-10-19
BKBKW/GO	RO/YW	parent (female)	2010-10-19
BKR/RO	RO/YW	paired with	
RO/YW	BKR/RO	paired with	
BKBKW/GO	BKBKW/BKY	nestling	2010-10-19
BKBKW/GO	10/48	nestling	2010-10-19
BKBKW/BKY	BKR/RO	parent (male)	2010-10-19
BKBKW/BKY	RO/YW	parent (female)	2010-10-19
BKBKW/BKY	BKBKW/GO	nestling	
BKBKW/BKY	10/48	nestling	

Measurement or fact extension

occurrenceID	measurementType	measurementValue	measurementUnit	measurementMethod
3253-03	Release weight	xxx	g	
3253-02	Management	hand raised		
3253-01	Estimated egg age	1	day	Candle
3252-02	Management	Hand raised		
3252-01	Estimated egg age	2	Days	Candle

2. Using Sampling Event core

Event

eventID*	eventDate	Location	Habitat	Country	recordedBy
E01	2010-10-19	MacKenzie Basin		New Zealand	DOC staff
E02	2010-11-10	DOC Bird Facility	Aviary	New Zealand	DOC staff
E03	2011-01-26	MacKenzie Basin		New Zealand	DOC staff
E04	2010	DOC Bird Facility	Aviary	New Zealand	DOC staff
E05	2010-11-09	DOC Bird Facility	Aviary	New Zealand	DOC Staff

* Arbitrary eventID assigned

Occurrence

occurrenceID	organismID	eventID	?degreeOfEstablishment	Pathway	sex	lifeStage	Vitality
3253-01	BKBKW/GO	E01	Wild		male	egg	alive
3253-02	BKBKW/GO	E02	Captivity		male	hatched	alive
3253-03	BKBKW/GO	EO3	Released	*Released / species management*	male	juvenile	alive
3254-01	10/48	E01	Wild		male	egg	alive
3254-02	10/48	E04	Captivity		male	egg	dead
3252-01	BKBKW/BKY	E01	Wild		male	egg	alive
3252-02	BKBKW/BKY	E05	Captivity		male	hatched	alive
3252-03	BKBKW/BKY	E03	Released	*Released / species management*	male	juvenile	alive

Resource relationship

resourceID	relatedResourceID	relationshipOfResource	relationshipEstablishedDate
BKBKW/GO	BKR/RO	parent (male)	2010-10-19
BKBKW/GO	RO/YW	parent (female)	2010-10-19
BKR/RO	RO/YW	paired with	
RO/YW	BKR/RO	paired with	
BKBKW/GO	BKBKW/BKY	nestling	2010-10-19
BKBKW/GO	10/48	nestling	2010-10-19

Extended Measurement or Fact

eventID	occurrenceID	measurementType	measurementValue	measurementUnit	measurementMethod
EO3	3253-03	Release weight	xxx	g	
E02	3253-02	Management	hand raised		
E01	3253-01	Estimated egg age	1	days	Candle
E02	3252-02	Management	Hand raised		
E01	3252-01	Estimated egg age	2	days	Candle

Appendix 3.15 Twizel Kaki Hide – DOCDM-707756

Recommended core: Sampling event

This data set could also be mapped to an occurrence core if the population statistics were not required, but a sampling event core is more suitable because it permits the species-level observations to be linked to an annual event.

This data set is highly summarised.

Preliminary mappings

This data set provided a challenge in terms of the best core for packaging the data set. Both mappings are included, but the Sampling Core is the recommended mapping.

A		Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
Year	/9	1995/9	1996/9	1997/9	1998/9	1999/200	2000/200	2001/200	2002/200	2003/200	2004/200	2005/200	2006/200	2007/200	2008/200	2009/201	2010/201
Adults		47	47	54	41	31	39	47	67	66	55	72	87	78	82	85	
Sub adults		10	12	5	7	17	22	50	42	46	68	67	48	54	69	71	
Juveniles		1	4	3	6	33	27	28	16	28	37	55	16	22	0	13	
Wild poulation		**58**	**63**	**62**	**54**	**81**	**88**	**125**	**125**	**140**	**160**	**194**	**151**	**154**	**151**	**169**	
Released adults																2	
Released sub-adults		18	10	14	13	16	37	39	45	60	61	53	77	80	88	95	
Released juveniles					5	33	29	31	18	28	35	56	12	16	0	13	
Captive poulation		27	27	27	27	27	27					21	19	15	15	13	
Captive breeding pairs							7				8	7	6	6	6	5	
Wild productive breeding pairs		6	5	9	5	4	7	8	9	13	11	14	17	20	10	16	
Non productive breeding pairs		2	4	0	1	2	1	0	4	2	1	3	4	1	11	2	

	/95	1995/96	1996/97	1997/98	1998/99	1999/2000	2000/2001	2001/2002	2002/2003	2003/2004	2004/2005	2005/2006	2006/2007	2007/2008	2008/2009	2009/2010	2010/2011
Wild poulation	72	58	63	62	54	81	88	125	125	140	160	194	151	154	151	169	0
Total releases	22	18	10	14	18	49	66	70	63	88	96	109	89	96	88	110	0
Adult population	52	47	47	54	41	31	39	47	67	66	55	72	87	78	82	85	0

1. Using Occurrence core

occurrenceID	eventID	eventDate	locality	occurrenceStatus	lifeStage	individuals
TMP1	XXX	1995/1996	Canterbury, Twizel Kaki Hide	present	adult	47
TMP2	XXX	1995/1996	Canterbury, Twizel Kaki Hide	present	sub-adult	10
TMP3	XXX	1995/1996	Canterbury, Twizel Kaki Hide	present	juvenile	1

- The individuals field is used in this example for brevity, but could also be provided in organismQuantity and organismQuantityType.

2. Using Sampling Event core

Sampling Event core

eventID	samplingProtocol	eventDate	locality
XXX	Bird survey	1995/1996	Canterbury, Twizel Kaki Hide

Occurrence extensions

eventID	occurrenceID	occurrenceStatus	lifestage	individuals
XXX	TMP1	present	adult	47
XXX	TMP2	Present	Sub-adult	10
XXX	TMP3	present	Juvenile	1

Measurement or Fact extension

eventID	measurementType	measurementValue	measurementUnit
XXX	Wild population	58	individuals
XXX	Total releases	18	individuals
XXX	Productive pairings in wild	6	Pairs
XXX	Non-productive pairings in wild	2	Pairs
XXX	Captive population	27	individuals