Summary Report

eDNABridge

Streamlining the transfer of New Zealand eDNA data into GBIF

PREPARED FOR
Ministry for the Environment

PREPARED BY
Dr Kiel Hards (Epi)
Dr Petra Muellner (Epi)

26 February 2026

Acknowledgements

This project would not have been possible without support from several organisations and individuals. The Ministry for the Environment funded the project. Organisations that provided information for the project were the Ministry for the Environment (MfE), Wilderlab, Hawke's Bay Regional Council (HBRC), Department of Conservation (DOC), Earth Sciences New Zealand (NIWA), GBIF New Zealand and the GBIF Secretariat.

The authors would specifically like to thank Aaron Wilton (GBIF-NZ) for his technical advice and guidance, Meredith Mckay (GBIF-NZ), Tobias Guldberg Frøslev and Matthew Blisset (both GBIF Secretariat) for advice and coordinating efforts with the wider GBIF network, Jay Barnard (HBRC), Vicki Lyon (HBRC), Cindy Baker (NIWA), Dave West (DOC) and James Muir (DOC) for supporting the pilot data upload, as well as Shaun Wilkinson and Susan Welsh (both Wilderlab) for supporting the extracting and mapping of eDNA samples for the pilot upload. Finally, the authors would like to thank Fiona Hodge and Carolyn Mander (both MfE) for coordinating the wider project and supporting discussions and connections between all parties.

Disclaimer

1 Executive summary

Environmental DNA (eDNA) offers a powerful, non-invasive way to monitor biodiversity at both local and global scales. While wide sharing of eDNA data is encouraged, several operational barriers prevent data sharing, including the lack of efficient pathways for bulk upload to global repositories.

Currently few options exist to submit large amounts of eDNA data to the Global Biodiversity Information Facility (GBIF). Enabling eDNA data to be automatically uploaded from providers of eDNA analysis to GBIF has the potential to improve the accessibility and interoperability of New Zealand's eDNA data. However, any such automation will require validation of data quality and integrity to be ultimately successful.

To support the reuse and sharing of New Zealand eDNA data, the Ministry for the Environment commissioned the development of an open-source software package, developed with the widely used programming language R. The *eDNABridge* R package enables data-producing laboratories and entities holding large amounts of eDNA data to perform bulk uploads to the GBIF.

As a proof-of-concept, pilot datasets containing public metabarcoding records from the Department of Conservation, Hawke's Bay Regional Council, and Earth Sciences New Zealand (ex NIWA), sequenced and analysed by Wilderlab, were uploaded to GBIF in December 2025. A total of 4,526 New Zealand eDNA samples were published, representing 764,133 individual organism detections.

This project demonstrates that large-scale, standardised publication of eDNA data to GBIF is both feasible and efficient, provided appropriate technical and organisational prerequisites are in place.

Contents

2 Background

Environmental DNA (eDNA) is a form of environmental data obtained by sampling the environment for trace DNA left behind by animals, plants, fungi, and microbes. The richness of eDNA data, covering large numbers of species, makes open access crucial to realising its full value, allowing these data to support a wide range of uses from ecosystem restoration to biosecurity surveillance.

Determining which organisms are present in an eDNA sample (typically referred to as typing/taxonomic identification) requires sequencing and analysis by specialist laboratories, using a method called tree-of-life (ToL) metabarcoding. In New Zealand, this service has predominately been provided by Wilderlab[1]. However, it is expected that an increasing number of laboratories will offer metabarcoding services in the future. For example, Hill Labs[2], a large New Zealand commercial laboratory, has recently added various eDNA tests to their offerings.

The New Zealand government (both central and local) funds the collection of biodiversity data in many different forms for a multitude of purposes. However, the full value of this data is currently not being realised as data is being stored in different places, which makes it difficult to link and / or re-use data. A first step to address this shortcoming is to include New Zealand eDNA into relevant data repositories, to provide easy access to the data generated by a wide range of stakeholders: such as researchers, regulators and government agencies. The Global Biodiversity Information Facility (GBIF) provides storage and access for a variety of biodiversity data, including eDNA. GBIF accepts eDNA-derived records that meet defined technical standards and provide storage and access for biodiversity data in general. In New Zealand, the GBIF Node (GBIF New Zealand) supports organisations wishing to publish biodiversity data by providing technical guidance, endorsement as a data publisher, and hosting infrastructure through a nationally managed Integrated Publishing Toolkit (IPT). This ensures that datasets meet international standards and remain persistently available for indexing via GBIF.org.

However, upload of data to GBIF requires a series of technical steps that are a barrier to data sharing, especially for one-off users or owners of smaller data sets. A lack of automation can also make the upload time consuming, even for experienced users. In brief, GBIF uses a decentralised publishing model, where submitting organisations or regional GBIF centres host the data themselves and make it available for discovery and indexing via the main GBIF website. The primary publishing mechanism is the GBIF Integrated Publishing Toolkit (IPT)[3], a server-based system that supports manual upload and management of datasets on a one-by-one basis. A newer option, the GBIF Metabarcoding Data Toolkit (MDT)[4], builds on this architecture by adding support for eDNA-specific processing and a programmatic

[1] https://wilderlab.co/
[2] https://www.hill-labs.co.nz/
[3] https://www.gbif.org/ipt
[4] https://www.gbif.org/metabarcoding

submission interface, but remains focused on individual dataset submissions and is currently in a pilot phase, and hence not yet ready for implementation at scale. Alternative publishing approaches are possible by exception, subject to approval by the GBIF Secretariat. Regardless of the upload method used, organisations submitting data must be approved by GBIF as publishers to ensure that published data are relevant, openly available, and hosted with persistent arrangements.

Currently few options exist to submit large amounts of eDNA data to GBIF, fewer still that aim to balance this with the need for the validation of data quality and integrity. To address this gap, the Ministry for the Environment commissioned the development of an R package that provides an end-to-end workflow from sequencing laboratories to GBIF, enabling scalable bulk uploads with configurable levels of automation. Such a technical solution (an "eDNA bridge") will enable New Zealand eDNA data to be automatically uploaded, from suppliers of eDNA analysis to GBIF. This provides an opportunity to leverage GBIF's existing publishing infrastructure to immediately improve accessibility and interoperability of New Zealand's eDNA data and support future stakeholder collaboration.

3 Project description

3.1 R package

The development of the *eDNABridge* formed part of a broader, stakeholder-centred project aimed at improving the accessibility and re-use of New Zealand eDNA data. The project was designed in consultation with participating sequencing laboratories, data-providing organisations, and representatives from the GBIF network, to ensure alignment with both operational needs and established GBIF publishing processes. Based on these consultations, the design of the package focussed on addressing the most labour-intensive technical part of the data sharing process: the preparation, validation, and bulk submission of data to GBIF. While anyone can use the package it is ideally suited to be used by laboratories, which can easily automate the bulk upload of the data they produce (assuming appropriate permissions by data owners are in place).

The role of the *eDNABridge* package in the overall GBIF submission workflow is explained in Appendix 1. In brief, it automates the process of creating a type of file called a Darwin Core Archive from sequencing laboratory data. This type of file follows a set of biodiversity standards outlined in the titular Darwin Core[5], to produce a single self-contained dataset, which contains all the data required for biodiversity reporting. The *eDNABridge* package finally handles the process of uploading and submitting this Darwin Core Archive to GBIF.

Overall, the *eDNABridge* package:

[5] https://dwc.tdwg.org/

- Provides an end-to-end, scalable workflow from sequencing laboratory outputs through to publication on GBIF, reducing manual effort and enabling bulk uploads of eDNA data.
- Is designed for flexible use, supporting both interactive, guided workflows and fully automated, scripted integration within laboratory or organisational pipelines.
- Produces GBIF-compliant Darwin Core Archives and supports programmatic upload to IPT servers, improving data quality, consistency, and interoperability.
- Is delivered as an open-source, extendable, R package with comprehensive user documentation, available both through function-level help and a documentation website (bundled with the package source code).

Technical details about *eDNABridge*, including its architecture, are described in Appendix 1. The package is freely available on GitHub at https://github.com/gbif/eDNAbridge.

3.2 Pilot upload

To test the *eDNABridge* package under real-world conditions and initiate large-scale publication of eDNA data, three New Zealand organisations participated in a pilot bulk upload of data produced by Wilderlab:

- Department of Conservation (DOC)
- Hawke's Bay Regional Council (HBRC)
- Earth Sciences New Zealand (ex NIWA)

With appropriate permissions in place, data were accessed via the Wilderlab API[6] and uploaded to the GBIF-NZ Integrated Publishing Toolkit (IPT). Where data sets overlapped, duplicates were excluded. Organisational publishing details were configured to associate datasets with the correct data owners.

Since some dataset-level metadata required for publication cannot be automatically derived from sequencing laboratory outputs, manual input from data-owning organisations was required for each data set. This includes descriptive information about the purpose of data collection, project context, involved personnel, data ownership, and applicable data-use conditions.

Data were mapped to the Darwin Core standard, validated, and structured into annual datasets (i.e. one Darwin Core Archive per year) following guidance provided by GBIF-NZ. Using the package, the end-to-end upload process for each data set can be performed very quickly (i.e. within minutes), including validation. This demonstrates that efficient bulk

[6] The Wilderlab API is an interface to allow clients to access their data programmatically and is described at https://wilderlab.co/api-instructions

publication of eDNA data to GBIF can be feasible at scale. The total sample numbers uploaded are described in Table 1.

Table 1: eDNA samples uploaded to GBIF in the pilot upload

Organisation	eDNA samples	Occurrences
DOC	2,710	430,886
HBRC	1,753	310,699
Earth Sciences	63	22,548
Total	**4,526**	**764,133**

4 Discussion

The pilot implementation of *eDNABridge* demonstrates that large-scale publication of eDNA occurrence data to GBIF is both feasible and efficient when supported by purpose-built software tooling. The package successfully automated the preparation, validation, packaging, and bulk upload of eDNA data, removing a major technical barrier that has previously made data uploads time-consuming and error-prone. Once data and permissions were in place, the end-to-end workflow operated reliably and completed within minutes, demonstrating that the approach is well suited to high-throughput use cases.

The package proved effective at identifying common data quality issues prior to publication, particularly those arising from human entry errors, and ensured that outputs conformed to GBIF publishing requirements through the generation of standards-aligned Darwin Core Archives. Structuring datasets into logical groupings, such as by sampling year, supported both performance and discoverability on the GBIF platform. Performance during the pilot was largely influenced by reliance on external online services, including sequencing laboratory APIs, GBIF validation services, and the IPT server itself. Where tighter integration is possible, particularly within sequencing laboratory systems, and where validation steps can be performed without repeated external service calls, the approach is expected to scale readily to much larger volumes of samples.

Differences between taxonomic systems used by sequencing laboratories and those recognised by GBIF were observed, and while these could be flagged during validation, full reconciliation remains outside the scope of an automated upload tool and may require further coordination or the use of shared identifiers.

While the *eDNABridge* is concerned with the technical process of uploading the data, agreements will need to be in place between stakeholders with roles and responsibilities clearly defined. While data may be uploaded on behalf of an organisation by another party

(for example, a sequencing laboratory), this can affect how easily staff within the data-owning organisation are able to access, update, or manage the published records. In addition, end-to-end automation using the *eDNABridge* requires users to hold the necessary permissions to access source data from the sequencing laboratory and to perform relevant actions within the selected GBIF IPT instance. The collaborative nature of many eDNA sampling efforts mean that sequencing laboratories associate eDNA sample records with multiple parties, which needs to be accounted for to prevent duplication of submitted information.

Overall, the *eDNABridge* addresses one of the most labour-intensive components of the GBIF publishing process by automating the technical steps required to transform sequencing outputs into publishable datasets. However, the process is not completely hands-off and users are still required to complete standard GBIF publishing steps, including securing appropriate permissions, defining dataset scope, and providing project-level metadata. When used alongside established GBIF processes and clear organisational coordination, *eDNABridge* provides a practical and scalable pathway for publishing large volumes of eDNA data.

5 Future directions

To enhance the usefulness of *eDNABridge* for the sharing of New Zealand eDNA data, the package could be updated so that it can fully accommodate the inclusion of details about New Zealand's local context, to support initiatives to address data sovereignty concerns. Enhancements to validation and archive generation to support additional GBIF extensions will be important to achieve this, including support for planned GBIF extensions for the Local Contexts initiative[7].

Future development could also explore features that can further streamline integration of the tool into the workflow of data producing laboratories in New Zealand. For example, by providing an additional pathway to reconcile taxonomic assignments between sequencing laboratories and GBIF, via shared taxonomic and barcode reference systems such as the Barcode of Life Data System (BOLD). From a global perspective, future development of the *eDNABridge* could focus on expanding its applicability and deepening the integration of the package with the GBIF workflow, such as with support for updating GBIF datasets in bulk. The addition of new ingestion modules to support other sequencing laboratories and data providers will allow the package to be used across a broader range of eDNA production pipelines without custom preprocessing.

Support for alternative GBIF publishing endpoints, including the Metabarcoding Data Toolkit (MDT) as it matures beyond the pilot phase, would further improve flexibility for users with different hosting arrangements. Notably, the MDT system is becoming increasingly integrated

[7] https://techdocs.gbif.org/en/data-publishing/local-contexts-integration

with FAIR (Findable, Accessible, Interoperable, and Reuseable) eDNA data (called FAIRe)[8], a data standard tailored for the complexities of eDNA records[9]. As mandatory FAIRe fields were not available in the pilot dataset (namely, the DNA sequence), support for FAIRe data structures was not developed in this version of the package. However, the modular nature of this package would readily allow the ingestion module to be extended to support FAIRe datasets. This could be developed in concert with support for MDT servers, while also providing the benefit of supporting IPT servers when the former is not available.

Further optimisation of validation workflows to reduce reliance on external services, such as the GBIF species validation API that validates individual taxonomic names but is one of the most time-consuming steps in the validation process. Improvements like this, along with closer integration with laboratory systems, would improve performance and support larger-scale automation. In combination, such additions would strengthen the role of *eDNABridge* as a general-purpose, standards-aligned bridge between eDNA data production and global biodiversity infrastructure.

[8] https://onlinelibrary.wiley.com/doi/10.1002/edn3.70100
[9] https://fair-edna.github.io/index.html

Appendix 1: Technical details of the *eDNABridge* R package

Where the *eDNABridge* fits in the eDNA-GBIF workflow

eDNA samples are handled and processed by various parties taking on specific roles, before the species (taxa) identified in each sample can be made available to GBIF (Figure 1). These include, for example, the parties collecting the sample, the sequencing laboratory and staff linking the sequencing results with the associated metadata. Full definitions of roles are provided in Appendix 2.

The *eDNABridge* addresses the process from ingestion of data from the 'Identifier' (commonly a testing laboratory), through to the 'Metadata provider' and 'Data uploader' and finally the 'Hosting institition'. This involves standardisation/validation, preparation of data in the correct format for GBIF and finally uploading and publishing data to GBIF via an Integrated Publishing Toolkit (IPT) server (Figure 1). For ease of use by different users, the package includes detailed user guidance in the command line interface, high-level functions to automate uploads with sensible default values and low-level functions that can be incorporated into custom R scripting workflows. The package facilitates the ingestion of eDNA data, validation against the Darwin Core data scheme, generation of a Darwin Core Archive and submission/publication with a selected GBIF IPT server.

Submission of data to GBIF follows established publishing processes, including organisational approval and appropriate permissions for data upload and management. Users are expected to have these approvals and access arrangements in place prior to submission. Submission of data to GBIF is a public release, that cannot be easily reversed and should be treated with the same considerations as any other public release of documents. Editing and deleting data is possible through direct interaction with GBIF.

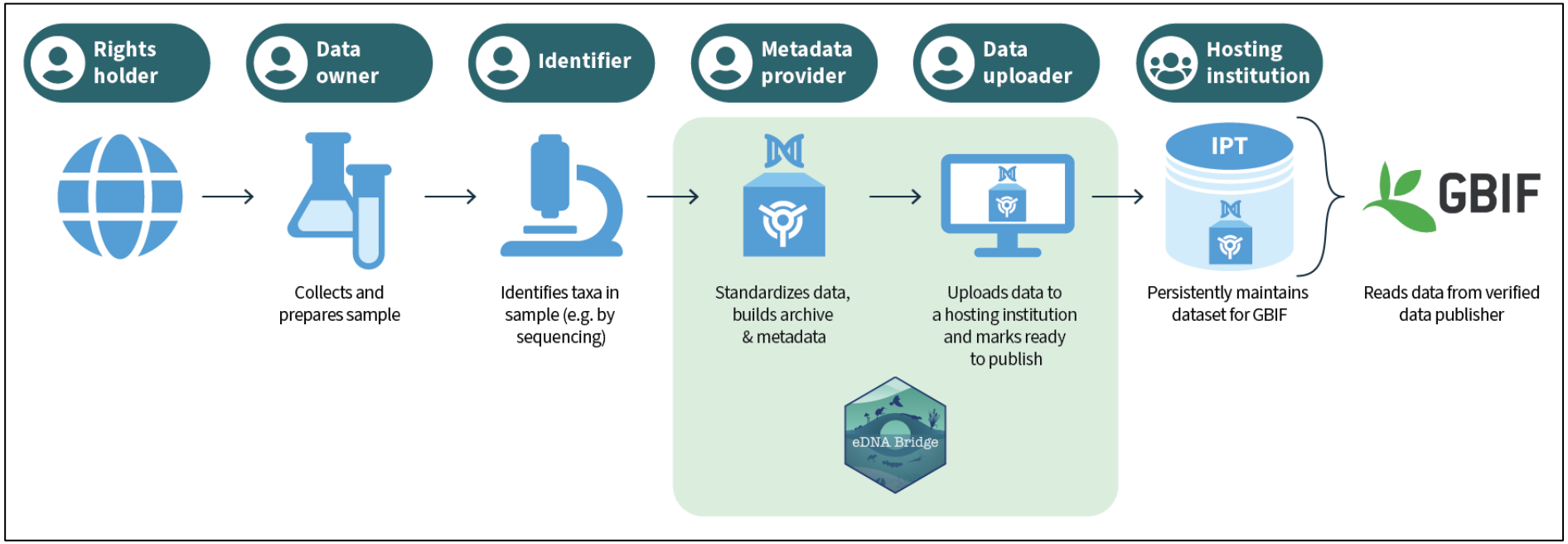

Figure 1: Simplified high-level schematic of the workflow for collecting and publishing eDNA occurrence data to GBIF. Notable roles in the process are indicated, with full definitions for each role provided in Appendix 2.

eDNABridge architecture

The *eDNABridge* is an R package that can be executed by any user that has a version of R > 4.1 installed on their system. The workflows provided by the R package can be run interactively via a terminal session with prompts or programmatically by calling functions with parameters in custom R scripts. Data can be exported and re-imported at various stages. The primary data output is a Darwin Core Archive, but intermediate tables of data can also be exported/imported for advanced users.

The R package contains six modules (Figure 2), numbered according to the step in the eDNA workflow that the module addresses. The code is designed to be easily extendable with additional functionality, should this be desired. For example, the ingestion module could be extended to support additional sequencing laboratories, the validation module could be extended to support other types of data, and the upload module could be extended to support an MDT server.

The R package contains minimal dependencies, with the only specialist R package requirement being the *insect* R package, which supports the conversion of taxonomy information from the Wilderlab database into the format expected by GBIF.

During package development, we explored R packages with overlapping functionality, such as the *EML* package for preparing the Ecological Markup Language metadata file. Several packages were evaluated but were found to be unsuitable to link with the *eDNAbridge* e.g. because they were not considered stable enough, were outdated or could not fully capture the tasks required. Hence required capabilities were implemented from scratch in in *eDNABridge*, adding new capability and allowing the code to be tailored to the specific requirements for GBIF datasets.

Modules 01-04 represent the core workflow and contain most of the functions that advanced R users will interact with:

- **Module 01 – Ingestion:** Ability to source data directly from Wilderlab, using the Wilderlab API, with authentication details supplied by the user. Ability to source data directly from CSV files or an Excel file that matches Wilderlab's multi-file format & column names. Ability to source from a CSV or Excel file as a single file, with column specifications supplied by the user.
- **Module 02 – Validation:** Rename columns to match Darwin Core terms, either from user-supplied mappings or default mappings for Wilderlab data. Check for required fields and correct data types, including taxonomic names. Show a summary of the data and any issues found via a common table aggregating issues.
- **Module 03 – Archive Generation:** Create the necessary files and structure for a Darwin Core Archive (.zip), including Core occurrence file (csv), optional Extended Facts or Measurement file (csv), optional DNA extension file (csv), archive metadata file (xml), Ecological Metadata Language file (xml).

- **Module 04 – Archive Upload:** Using the user's IPT credentials and URL, upload the Darwin Core Archive file to directly into their selected IPT instance. Depending on the IPT configuration, the dataset may be published immediately or require further review and publication by an administrator.

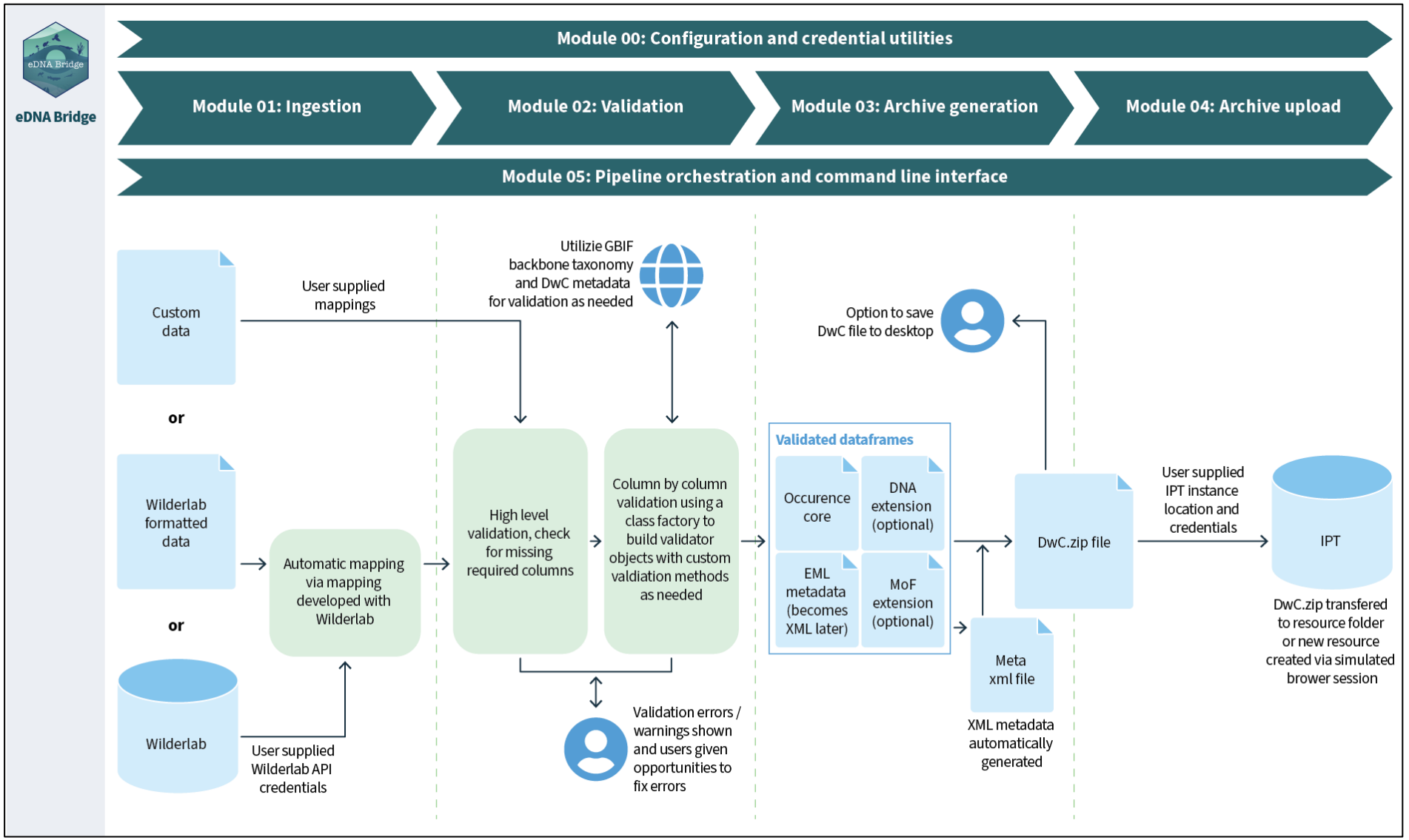

Figure 2: Overall architecture and information flows in the eDNABridge *R package.*

Appendix 2: Descriptions of individual roles in the GBIF publishing workflow

Rights holder: Has rights to the sample collected; may not be involved in sample collection and may not hold all rights. Typically, whoever has rights to the sampling site.

Data owner: Entity that collected or generated the data; primary contact for the dataset. Typically, the NZ entity (e.g. council) uploading the data.

- Must be the Data Uploader and/or part of the Hosting Institution to have permissions to modify data on an IPT instance.

Identifier: Entity that identified taxon occurrences within samples; may differ from the Data Owner if identification is a commercial service. Typically, the sequencing laboratory.

Metadata provider: Entity that produces dataset metadata for GBIF submission; may also be the Data Owner or Identifier. Typically, the entity performing the data upload.

Data uploader: Person/organisation who prepares and uploads datasets to an IPT instance or GBIF node. This could be either the NZ entity (e.g. council) or the sequencing laboratory.

- Requires credentials and approval to upload data on behalf of a Data Publisher.
- Typically has permission to edit or delete data on the IPT server (subject to Hosting Institution rules).

Data publisher: Organisation formally approved by GBIF to publish data to GBIF.org; cannot be an individual. This is typically the NZ entity (e.g. council) uploading the data but could also be the sequencing laboratory if an automated upload is performed.

- Requires endorsement from a GBIF Node and stable hosting arrangements.
- In New Zealand, endorsement of Data Publishers is provided by GBIF New Zealand in its role as the national GBIF Node.
- Without a valid Data Publisher, data will not appear on GBIF.org.

Hosting institution: Institution that hosts data for GBIF indexing, usually via an always-accessible IPT server. This could be the NZ entity (e.g. council) itself, or a regional GBIF node (e.g. GBIF-NZ).

- Manages IPT hosting and configuration (e.g. validation extensions).
- Must ensure continuous public availability for GBIF re-indexing.

GBIF node: National or thematic GBIF body supporting publishers and biodiversity initiatives.

GBIF Secretariat: Central GBIF governing body maintaining GBIF.org and coordinating global indexing. Typically, not involved unless escalated support or hosting exceptions are required.