Data Release Policy

POLICY TITLE: DATA RELEASE
FIRST APPROVAL: MARCH 14, 2023
LAST REVISION APPROVAL: N/A
NEXT SCHEDULED REVIEW: MARCH 14, 2026

Purpose

The Centre for Biodiversity Genomics (CBG) serves members of the international research community who are employing DNA-based identification systems. It is resolutely committed to the rapid release and sharing of sequence data and collateral specimen information. This policy document outlines the CBG's approach to ensuring its data are available to the research community in a timely way.

Guidelines

As the CBG is a recipient of funds from Genome Canada, this policy aligns with Genome Canada Data Sharing Policies. Public data release means that data generated at the CBG is made accessible in an appropriate online platform or data repository:

a) DNA barcoding data is deposited in the CBG's Barcode of Life Data Systems (BOLD) where it is made publicly available after an embargo period (see below). Any individual can create a BOLD user account which provides access to both the BOLD Public Data Portal and its APIs. Sequence data, as well as specimen provenance and taxonomic data, are released and are used by BOLD's data aggregation, analytical, and reporting tools.

b) Metabarcoding data is deposited in the CBG's Metabarcoding Research and Visualization Environment (mBRAVE) platform and is made publicly accessible by uploading to the Sequence Read Archive after the embargo period.

Data definitions

DNA barcode data is stored in BOLD as specimen records, each consisting of two components: (i) specimen data – information and images for the physical specimen, each designated by a unique Sample ID, and (ii) sequence data – information resulting from and related to the molecular analysis of a specimen, each designated by a unique Process ID. Specimen records for animals usually have a third component: (iii) Barcode Index Number (BIN) – BOLD's assignment of each qualifying record to a persistent operational taxonomic unit known as a BIN, each designated by a unique identifier.

More information on these data elements and definitions is provided in the BOLD Handbook.

Method of data release

Some components of a nascent DNA barcode record are immediately available on BOLD at the time of submission through various tools and pages. For example, upon submission of a specimen to BOLD, its provenance information and image(s) become available to the public through the BOLD Taxonomy Browser. This information is used to generate summary statistics and illustrative distribution maps, but does not disclose the content of individual research projects and specimen data until expiry of the embargo period.

Upon successful sequencing of a specimen, its specimen and sequence data immediately enter the BOLD Identification Engine. Reports generated by this tool include similarity scores and tree-based identification with branch labels containing detailed taxonomy, geographic localization (e.g., province), and corresponding BOLD Process IDs and Sample IDs. Information on individual specimens (e.g., museum catalogue number and voucher repository) and their detailed geographic origin is not disclosed through the BOLD Identification Engine. Finally, once a specimen record is assigned to a BIN, this data is added to the associated BIN page that aggregates information on its members (see example at https://portal.boldsystems.org/bin/BOLD:AAA9566).

Data is batch released to the public on BOLD on a twice-per-year schedule, December 15th and July 15th, pending data preparation. These data releases are accessible in various formats on BOLD's Data Packages page. Data may then also be uploaded to GenBank and/or GBIF if initiated by the project manager.

Embargo period

The embargo period is the amount of time following data upload to BOLD that the data remains private to the primary user. This period enables the primary user to interpret and publish their results before they become fully available to the entire user community.

Data generated by the CBG is released to the public through the BOLD Data Portal, BOLD Data Packages, and their associated tools twice per year. The official embargo period is 12 months. Data is not released publicly until the embargo period has expired.

Option for data to stay private

The CBG enters into some fee-for-service agreements where it commits to hold client data private. The CBG reviews these arrangements on a case-by-case basis for justification. Such services are only provided by the CBG on a full-cost recovery basis.

Data preparation period

The data preparation period is the interval required for the preparation and validation of data prior to public release. Barcode data requires management and validation prior to its release to ensure its accuracy so it can be used with confidence by the research community. In some instances, data preparation may take longer than the official embargo period.

Data which do not pass key standards are not released publicly (i.e., failure to amplify; misidentified records).

Data revision

Updates to taxonomic assignments and level of resolution occur on an ongoing basis as a result of iterative data validation processes.

Notification

Users do not receive notification that their data has been uploaded or released. User must check BOLD to ascertain the status of their records.

Acknowledgement

The CBG expects users of its data and resources to acknowledge such usage in publications and to abide by any terms and conditions of use.

These terms and conditions are made explicit on BOLD's Data Packages page.