Data sharing and DNA barcodes

February 2021

Convention on
Biological Diversity

Data sharing and DNA barcodes

Data sharing in support of CBD goals

The preamble to the Convention on Biological Diversity (CBD) in 1992 highlighted "the general lack of information and knowledge regarding biological diversity and the urgent need to develop scientific, technical and institutional capacities to provide the basic understanding upon which to plan and implement appropriate measures"[1].

In subsequent years, the importance of such information has been repeatedly recognised as an essential foundation to meet the goals of the CBD. Target 19 of the Aichi Biodiversity Targets[2] proposed that "by 2020, knowledge, the science base and technologies relating to biodiversity, its values, functioning, status and trends, and the consequences of its loss, are improved, widely shared and transferred, and applied". Paragraph 204 of the UN *Future We Want*[3] notes the need "to provide the best available policy-relevant information on biodiversity to assist decision makers". The Global Assessment Report on Biodiversity and Ecosystem Services[4] from the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) highlights the need for richer information to understand threats to biodiversity and ecosystem services. Appendix IV of its Summary for Policymakers[5] lists many significant knowledge gaps, including: "Data on changing interactions among organisms and taxa", "Monitoring of many listed species in the Convention on International Trade in Endangered Species of Wild Fauna and Flora", "Basic data on many taxa (86 per cent of existing species on Earth and 91 per cent of species in the ocean still await description)" and "Extinction risks and population trends for the following taxonomic groups: insects, fungal species, microbial species (microorganisms) and parasites". *Achieving the SDGs with Biodiversity*[6] from the Swiss Academy of Sciences calls for better reporting on environmental dimensions in country reports to the UN (e.g., Environmental-Economic Accounts Experimental Ecosystem Accounting) as an essential first step towards formulating evidence- and data-based biodiversity-centred pathways towards sustainability. Collectively, these documents make clear the need for more detailed biodiversity data to direct actions that will enhance environmental sustainability.

In parallel, there has been a growing international trend towards increased public access to research results and datasets that can inform planning and decision-making, generated by remote-sensing, the Internet of Things (IoT), and citizen science supported by novel sensor technologies, high-throughput sequencing, and machine learning. In the last few years, there has been a further recognition that data should not only be openly available but also FAIR[7] - Findable, Accessible, Interoperable and Reusable.

---

[1] https://www.cbd.int/doc/legal/cbd-en.pdf
[2] https://www.cbd.int/sp/targets/
[3] https://sustainabledevelopment.un.org/futurewewant.html
[4] https://ipbes.net/global-assessment
[5] https://www.ipbes.net/sites/default/files/ipbes_7_10_add.1_en_1.pdf
[6] https://doi.org/10.5281/zenodo.4457298
[7] https://doi.org/10.1038/sdata.2016.18

Since ratification of the CBD, there have been important advances in access to digital information on biodiversity. Most significantly, the Global Biodiversity Information Facility[8] (GBIF, established in 2001) has assembled more than 1.6 billion data records from some 56,000 datasets, each recording the occurrence of species in space and time. These data are openly accessible to researchers, policymakers, and the public, allowing them to analyse and interpret biodiversity patterns and change. A recent study by Heberling et al.[9] shows how these data not only contribute to biodiversity science but also to "diverse knowledge domains, including environmental sciences and policy, evolutionary biology, conservation, and human health".

Despite these advances, existing datasets are heavily biased towards regions with well-documented biotas and strong community engagement. Even in these situations, data collection is focused on vertebrates (especially birds) and vascular plants. Hence, understanding of biodiversity patterns and change at regional and global scales is largely limited to signals detectable in these groups. Nevertheless, as noted in the IPBES Global Assessment Report, our ability to manage and conserve biodiversity and ecosystem services depends on understanding the composition of whole communities and their complex responses to environmental change. Recognition has grown in recent years that we lack the long time-series necessary to determine, for example, the extent and significance of major insect declines noted in several regions[10].

The challenge faced with all but the most readily recognised organisms is two-fold. First, there are too few experts with the knowledge required to describe the millions of unnamed species. The CBD, through the Global Taxonomy Initiative (GTI), has consistently recognised this taxonomic impediment as a limiting factor in its efforts to implement the goals of the convention. Secondly, and closely linked to the first challenge, there are too few people who can identify even described species. In fact, there is no list of all species, the fundamental biological units that comprise the planet's life system and that are essential for human survival. Hence, we cannot measure how species are distributed across regions and ecosystems and how their distributions are changing over time in response to human activities and other pressures. As a result, there is an urgent need for transformational approaches that can detect, explore, and monitor biodiversity across all groups of organisms at scales that allow rapid data collection even in the absence of taxonomic expertise.

DNA barcodes as a global tool

As sequencing technologies have become more accessible, rapid, and inexpensive, DNA barcoding has become increasingly important in assisting taxonomic work. It has overcome the challenge of identifying specimens in groups where taxonomic expertise is limited or in situations where only damaged specimens, cryptic life stages, or trace DNA are available. Clustering specimens based on their DNA barcodes allows unnamed species to be recognised

[8] https://gbif.org

[9] Heberling et al. 2021, *Data integration enables global biodiversity synthesis*, PNAS February 9, 2021 118 (6) e2018093118; https://doi.org/10.1073/pnas.2018093118

[10] Wagner et al. 2021, *Insect decline in the Anthropocene: Death by a thousand cuts*, PNAS January 12, 2021 118 (2) e2023989118; https://doi.org/10.1073/pnas.2023989118

and tracked through space and time. Advances in the scale and power of sequencing technology have enabled the development of metabarcoding as an efficient method for analysing whole communities of species from mixed samples or environmental DNA (eDNA).

The International Barcode of Life (iBOL)[11] consortium has developed the Barcode of Life Data Systems (BOLD)[12], an informatics platform that includes barcode sequences from specimens held in natural history collections around the world. Whenever possible, specimens have been identified to a named species, but many are only assigned to a higher taxonomic group. Clustering these sequences using the Barcode Index Number (BIN)[13] system allows related sequences to be associated, even in the absence of a formal scientific name. By organising sequences and specimens into BINs, DNA barcoding accelerates the description of new species. Regardless of the order of events, as BOLD expands its content and as identified voucher sequences are added or new species are described, this reference library will gain power until it allows users anywhere to identify any species from a short sequence of DNA. This will radically transform how society uses taxonomic knowledge by enabling detailed monitoring of biodiversity for conservation, biosecurity, product certification, and all other processes that depend on accurate species identification. BOLD is a key component of an increasingly interconnected ecosystem of molecular data repositories, including the International Nucleotide Sequence Database Collaboration (INSDC[14], comprising DDBJ, EMBL-EBI, NCBI) for all DNA and RNA sequences, UNITE[15] for fungal ITS sequences, and SILVA[16] for ribosomal RNA sequences across all domains of life.

Advances in DNA-based identification require the assembly of reference sequences for species across complete branches of the tree of life. In the case of animals, this is a 648 base pair segment of the cytochrome *c* oxidase 1 (COI) gene from the mitochondrial genome. Other markers (ITS, rbcL, matK) are used in a similar way for plants, while ITS is employed for fungi. While these sequences allow species identification, they do not code for biological compounds with commercial value.

Complete records in BOLD include these sequences (and metadata demonstrating the quality of the sequence), the best available taxonomic identification, locality coordinates and collection date, and an image of the specimen (which serves as an important quality check for those reviewing BOLD). A complete record of this kind contains all the information necessary to serve as a digital voucher for the DNA barcode. BOLD also contains records that lack images and those with only a higher taxonomic identification. The BIN system adds value to these records since each BIN demonstrates the geographic range, seasonal occurrence, ecological associations, and abundance of the associated species. Metabarcoding uses the reference sequence library on BOLD to deliver both species identifications and BIN assignments.

[11] https://ibol.org/
[12] http://www.boldsystems.org/
[13] https://doi.org/10.1371/journal.pone.0066213
[14] http://www.insdc.org/
[15] https://unite.ut.ee/
[16] https://www.arb-silva.de/

Completing the barcode reference library for all organisms will make it possible to survey and monitor biodiversity everywhere, providing essential information for implementing the goals of the CBD internationally and nationally, as well as supporting the Global Taxonomy Initiative.

iBOL's current seven-year research program, BIOSCAN[17], directly supports these goals through massive expansion of the geographic and taxonomic coverage on BOLD and by laying the foundation for a global network of DNA-based biodiversity monitoring stations[18]. BIOSCAN is supported by iBOL member institutions and networks on every continent[19] and is backed by major project investments in multiple counties. Many of these projects are focused on developing comprehensive national reference barcode datasets using BOLD as a shared data repository. In Costa Rica, the *BioAlfa*[20] initiative aims to barcode all species to enable this nation to become the world's first "bioliterate" country. Its President and cabinet signed a decree in 2019 recognizing *BioAlfa* as a national priority and that the DNA barcode library will be in the public domain for the good of all. As more countries make progress in barcoding their biota, each additional record expands coverage, providing sequences of vital importance to all countries where each species is found.

DNA barcodes and data sharing

iBOL offers an online Handbook[21] documenting the processes for data submission and sharing through BOLD. The Consortium also adopted formal Data and Resource Sharing Policies for DNA barcode records in August 2011[22].

The data covered by these policies and managed by BOLD include:

1. Specimen data
    - Unique BOLD Process Identification Number and Sample Identification Number
    - Country/Ocean where each specimen was collected, ideally with GPS coordinates of its collection site
    - Date of collection
    - Digital image of the specimen (whenever possible)
2. Taxon Name/Identifier
    - Provisional taxonomic assignment (e.g., family level or even purely descriptive – e.g., “environmental sample”) – these identifications are refined over time.
3. Genetic Data
    - Gene region sequenced
    - Primer sequences and PCR conditions
    - Electropherogram (trace) files
    - Sequence “contig” assembly (DNA barcode)
    - Barcode Index Number (BIN)

[17] https://ibol.org/programs/bioscan/
[18] https://doi.org/10.1139/gen-2020-0009
[19] https://ibol.org/about/ibol-consortium/
[20] https://documentcloud.adobe.com/link/track?uri=urn%3Aaaid%3Ascds%3AUS%3A584dcfa3-66a4-4fc6-83d0-514741858faf
[21] https://www.boldsystems.org/index.php/resources/handbook
[22] https://ibol.org/wp-content/uploads/2011/11/110822_-iBOL-Data-and-Resource-Sharing-Policies1.pdf

As noted above, these data elements serve as reliable evidence that a particular species occurred at a location, allowing related occurrences to be mapped and interpreted. Those records with expert identifications are treated as reference sequences for establishing the BIN clusters and for identifying other sequences when they are added to BOLD.

The barcode library is of high value for taxonomy, conservation, biosecurity, and regulatory applications, but the individual records do not have commercial value for biotechnology, medicines, novel food crops, or other applications that merit Access and Benefit Sharing (ABS) arrangements in accordance with the Nagoya Protocol[23].

Recognizing the high value of barcode records for researchers, governments, and communities from all nations, the iBOL Data and Resource Sharing Policies strongly promote early open access to data for public use, following a two-phase process:

- **Phase I** – automated early release (within one week of sequence generation) to liberate enough information to aid other researchers and monitor progress:
  - Location information: all available information
  - Temporal information: date of sample collection
  - Taxonomic information: order-level assignment with BIN
  - Sequence information: sequence, trace files, primers used, and the centre that carried out sequencing
  - Database identifiers: BOLD process ID and specimen ID (voucher number, depository, and collection code)
- **Phase II** – release of additional data elements that require manual curation or detailed taxonomic studies (as the data become available or following publication of a manuscript):
  - Location information: GPS coordinates, elevation/depth, province/state, exact site of collection, and individual(s) who collected the specimen
  - Taxonomic information: species-level assignment (and subspecies, if appropriate) and individual who made the identification
  - Sequence information: manually assembled and curated barcode sequence

Data can be submitted directly to the BOLD workbench and pipelines are in place to support data exchange with NCBI (GenBank) and other platforms. Information on species distribution from BOLD also contributes to other international open data initiatives, including GBIF and the Group on Earth Observations Biodiversity Observation Network (GEO BON)[24].

Benefits from DNA barcode sharing

The primary applications of DNA barcoding and metabarcoding have involved supporting taxonomic research, species discovery, and inventory as well as in conservation, biosecurity, and product certification. The acquisition of DNA barcode data is now a standard component in taxonomic studies, and these sequences and their associated BINs are referenced in hundreds of new publications each year. DNA barcoding also supports major regional surveys

[23] https://www.cbd.int/abs/
[24] https://geobon.org/

of the composition of the fauna and flora (e.g., the review of terrestrial arthropod diversity in Canada[25] and Germany's *GBOL III: Dark Taxa*[26] initiative).

The remainder of this section describes three of many possible use cases to show how DNA barcoding is helping to address important societal and environmental concerns.

Target Malaria

Target Malaria[27] is working to reduce the devastating health impacts of malaria by employing gene drive to suppress populations of mosquitoes in the *Anopheles gambiae* complex as they are important vectors of this disease. DNA barcodes are being employed by teams from the Universities of Ghana and Oxford to understand and monitor how the removal of *A. gambiae* impacts the wider ecosystem, a critical issue to ensure the conservation of biodiversity and to detect possible changes, including increases in other harmful species[28].

STREAM

STREAM (Sequencing the Rivers for Environmental Monitoring and Assessment)[29] is a collaboration between WWF-Canada, Environment and Climate Change Canada, the University of Guelph, and Living Lakes Canada to enable local communities to monitor changes in water quality. DNA barcodes from aquatic invertebrates are compared with reference sequences in BOLD to generate reports that assess water quality by evaluating the sensitivity of the species found at each site to physical and biochemical disturbances. Because DNA barcoding allows accurate species identification in a fraction of the time required for morphological study, this program can readily scale to a national level.

Shark Conservation

Some 73 million sharks are harvested each year to meet the market demand for their fins. This activity is threatening the survival of many species and disrupting local ecosystems. The fact that shark fins are difficult to identify to a species based solely on their appearance has limited efforts to control this illegal harvest. DNA barcoding of shark fins in multiple countries has upended previous assumptions on the source of this commodity[30]. While it was believed that most sharks were harvested from international waters, DNA barcoding revealed that many are inshore species. This significantly alters priorities for conservation and regulatory action since much of the fishing activity is taking place in waters under national jurisdiction. This means that national enforcement and legal processes have an important role to play in reducing this illicit trade.

---

[25] https://zookeys.pensoft.net/issue/1251/
[26] https://bolgermany.de/home/gbol3/
[27] https://targetmalaria.org/
[28] https://ibol.org/barcodebulletin/features/2021-27-01-the-important-interactions-behind-the-itch/
[29] https://wwf.ca/habitat/freshwater/stream/
[30] https://news.mongabay.com/2020/10/efforts-to-tackle-shark-fin-trade-need-to-focus-closer-to-shore-study-says/