Data Standards for BARCODE Records in INSDC (BRIs)

By Robert Hanner, Chair
Database Working Group, Consortium for the Barcode of Life

Proposed 6 November; 2005; Revised 26 March 2009

Background. The Consortium for the Barcode of Life (CBOL) formed a Database Working Group (DBWG) at its inaugural meeting, held at the Smithsonian Institution in May 2004. The DBWG was created to pursue one of CBOL's principal goals: a global reference library of DNA barcode sequences that is integrated with other systems of biodiversity information (e.g., databases of specimens, species, biogeographic information). At this inaugural meeting, the DBWG participants and Chair agreed that DNA Barcode data should be archived in the public domain, preferably by the International Nucleotide Sequence Database Collaboration (INSDC)[1]. At this initial meeting, the DBWG also endorsed the need to link barcode records to voucher specimens and valid species names. In September of 2004, the DBWG convened a meeting on the campus of the US National Institutes of Health, hosted by GenBank at the National Center for Biotechnology Information (NCBI). This meeting outlined a proposal for new data standards that would apply to DNA barcode records submitted to INSDC members in the future. In April 2005, the DBWG consulted with representatives of leading taxonomic initiatives[2] and refined its data standards proposal based on their input. In May 2005, GenBank presented the proposal at the INSDC annual meeting where it was greeted with strong support and swift approval. The DBWG subsequently met with representatives of major museum database initiatives to discuss implementation of the proposed data standards[3]. Participants at this meeting endorsed the proposed standards without reservation.

If the following proposal is approved, the DBWG would work with NCBI to develop a more detailed set of user guidelines, to be posted on the CBOL and INSDC websites.

Proposal. The proposed standards include five major components:

1) **Creation of a reserved keyword ("BARCODE")**. NCBI and its collaborators will add the BARCODE 'flag' to new submissions that meet the standards established in consultation with CBOL. Data records that meet these criteria will be known as BARCODE records in INSDC (BRIs);

[1] GenBank, the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ)

[2] DBWG meeting at the Smithsonian Institution's Center for Research and Conservation, Front Royal, Virginia, 27-29 April 2005. Participants represented: The University of Guelph Barcode of Life Database (BoLD); Species2000; Integrated Taxonomic Information System (ITIS); the Global Biodiversity Information Facility (GBIF); the Duke University National Evolutionary Synthesis Center (NESCENT); NCBI; the Ocean Biogeographic Information System, the Census of Marine Life; ZooRecord of Thomson Publishing; International Plant Names Index (IPNI); iPlants; the International Commission on Zoological Nomenclature (ICZN); uBio of the Marine Biological Lab, Woods Hole; the National Biological Information System of the US Geological Survey; the US Department of Agriculture's GRIN database; the Natural History Museum, London; the Royal Botanic Gardens, Kew; the Smithsonian Institution; and CBOL.

[3] DBWG meeting at NCBI on 3 October 2005. Participants represented: the Global Biodiversity Information Facility (GBIF); the Zoological Information Management System (ZIMS); BoLD; the Taxonomic Database Working Group (TDWG); NESCENT; CBOL; and database initiatives at the University of California, Berkeley Museum of Vertebrate Zoology; the University of Kansas Biodiversity Research Center; University of Alaska Museum.

2) **Required documentation data elements.** DBWG proposes that the following data elements be required of all BRIs. In requiring these data, DBWG seeks to provide the user community with unambiguous links to essential information related to the BARCODE sequence and the voucher specimen from which it was derived. DBWG proposes that each BRI must:

   A. Include a unique identifier for the voucher specimen using a structured field specified by CBOL and NCBI[4]. This unique identifier can to used to create a linkage with the record for the voucher specimen in a biorepository and its associated metadata in a public online database.
   B. Include the name of a formally described species or a provisional label for an unpublished species. These elements will permit linkages to records found in one of the sources specified by CBOL and NCBI[5] and databases of provisional taxon concepts;
   C. Include Country-Code using the controlled vocabulary used by GenBank;
   D. Come from a gene region accepted by CBOL as an effective barcode (see process for approving candidate barcode regions, 3b, below).
      i) The 'Folmer region' of the cytochrome c oxidase 1 ('COI') is the default barcode region for eukaryotes until and unless a non-COI region is approved by CBOL. COI is defined relative to the mouse mitochondrial genome as the 648 bp region that starts at position 58 and stops at position 705;
      ii) In November 2009, CBOL approved *rbcL* and *matK* as the barcode regions for vascular plants. They are defined relative to the *Arabidopsis thaliana* chloroplast NC_000932 sequence annotation as follows: the *rbcL* barcode region is at the 5' end of the *rbcL* gene between bp1-599 (27-579 excluding primer sequences); the *matK* barcode region is between bp205-1046 (227-1019 excluding primer sequences).
   E. Include at least 75% contiguous, high quality bases from within the approved barcode region being amplified (see recommended guidelines below, 4A-D). However, if requested, GenBank (or another INSDC member) could assign the BARCODE flag to records with shorter sequences following guidelines defined by CBOL (see 5a, below);
   F. Include the name of the region used;
   G. Be associated with trace files for the forward and reverse sequencing runs submitted to the NCBI Trace Archive or the Ensembl Trace Server; and
   H. Include the sequences of all forward and reverse primers used. For records in which the contiguous sequence was assembled from more than one amplicon or when a cocktail of multiple primers was used for amplification, multiple sets of primer pairs must be provided. In addition, submission of the names of the forward and reverse primers with the primer sequences is strongly recommended.

[4] The voucher specimen identifier uses a triplet structure based on elements of the Darwin Core separated by a colon (:) as specified by INSDC (institutionCode:collectionCode:catalogNumber). This triplet field is parallel to the Life Science Identifier (LSID) that is an Object Management Group (OMG) standard.

[5] DBWG proposes a hierarchy of sources of species names, including vetted checklists such at Catalog of Life, nomenclators such as IPNI and the Zoological Record; lists of all published names such as uBio and the proposed NameBank; recent publications that have not yet been incorporated into compilations; and pre-publication data resources.

3) **Strongly recommended data elements**. The following data elements have been added to the INSDC at CBOL's request for validation of the voucher specimen, and are strongly recommended but are not required:

   A. Latitude and longitude;
   B. Name of the identifier;
   C. Name of the collector; and
   D. Date of collection.

4) **Sequence quality and coverage elements.** In contrast with the documentation elements described above (2A-I), quality and coverage elements are not required for BARCODE designation. They are presented as recommended 'good practices' that support the consistency and repeatability of the data records.

   Discrimination of closely related species using DNA barcodes often relies on differences in one or a very few base-pair sites. The sequences reported in BARCODE records are interpretations of the raw data produced by sequencing systems, so the post-processing of these data are critical elements in determining the accuracy and reliability of the sequences themselves. In requiring the submission of forward and reverse trace files with all BARCODE records, DBWG is ensuring that researchers and users will have access to these original data. In addition, DBWG seeks to ensure that appropriate procedures were used in the creation of BARCODE sequences from electropherogram trace files. The simple submission of bidirectional trace files does not ensure the reliability, accuracy and precision that researchers and users will seek in BARCODE records. The quality scores of, and degree of the overlap between, the bidirectional sequencing runs are therefore important metadata in the interpretation of analyses of barcode data.

   The DBWG recommends that in preparing BARCODE records, submitters should:

   A. Prior to creating the contig sequence, trim the sequence in each trace file using the following procedural standards:

      i. Ends should be trimmed to minimize low quality base calls on each end of each read. PHRED scores > 20 are generally considered to be high quality base calls, and scores > 30 are very high quality; and

      ii. Primer sequences should also be trimmed.

   B. In editing the single reads, base calls with quality score of less than 20 should be recorded as N. In creating the contig sequence, base calls of high or very high quality in one directional read should be maintained over those with lower quality in the other read.

   C. In deciding whether a record will be repeatable and reliable for species identification, submitters should select as potential BARCODE records only those for which the contig was based on bi-directional coverage with non-N base calls at no less than 40% of the reported sequence. As described below (5D), CBOL can direct GenBank (or another INSDC member) to remove the BARCODE designation from records which have all required elements (1A-I) but have been shown to be unreliable for species identification due to low sequence quality and coverage.

5) **Governance rules**. The INSDC provides an archive of records that can only be changed by the submitter. In the case of BRIs, the following modifications to the rules governing changes to data records are proposed to assure and maintain data quality and consistency:

   A. CBOL will define the circumstances under which records shorter than the recommended length could be BRIs. These might include sequences from type specimens or specimens of extinct or extremely rare species;
   B. CBOL has developed and implemented a process[6] whereby research groups could propose and justify a non-COI gene region to which the BARCODE flag can be given;
   C. BRIs that are assembled on and submitted to GenBank from the University of Guelph's Barcode of Life Systems (BOLD) will be considered by GenBank to be submitted jointly by the individual researcher and BOLD;
   D. BRIs that are submitted to GenBank may only be modified by the submitter. However, GenBank will remove the BARCODE flag from these records at CBOL's recommendation if a record is found to be unreliable for species identification. These records would remain in GenBank as non-BARCODE records; and
   E. DBWG and NCBI will develop a proposal to CBOL for attaching third-party comments, criticisms, and suggested corrections to BRIs, thereby providing the research community with additional quality indicators. These third-party comments would also support CBOL's review of BRIs from which the BARCODE flag might be removed.

[6] See protocols for approval of non-COI BARCODE regions at http://www.barcoding.si.edu/PDF/Guidelines%20for%20non-CO1%20selection%20FINAL.pdf