Proposed Standards for BARCODE Records in INSDC (BRIs)

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

November 6, 2005

Background. The Consortium for the Barcode of Life (CBOL) formed a Database Working Group (DWG) at its inaugural meeting, held at the Smithsonian Institution in May 2004. The DWG 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 DWG 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 DWG also endorsed the need to link barcode records to voucher specimens and valid species names. In September of 2004, the DWG 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 DWG 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 DWG 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 DWG 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 three 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] DWG 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] DWG 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 data elements.** DWG proposes that the following data elements be required of all BRIs. In requiring these data, DWG seeks to provide the user community with reliable, retrievable and verifiable information concerning the barcode sequence itself, the specimen from which it was obtained, and the species name that was applied by the submitter.

DWG proposes that each BRI must:

a) Include a link to a voucher specimen using a structured field specified by CBOL and NCBI[4], and to the metadata associated with that specimen and contained in the public database of the voucher specimen's repository.
b) Include a link to a documented species name found in one of the sources specified by CBOL and NCBI[5];
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). Initially, only cytochrome c oxidase 1 is approved as a barcode region, defined relative to the mouse mitochondrial genome as the 648 bp region that starts at position 58 and stops at position 705.
e) Include at least 500 contiguous unambiguous base-pairs from bidirectional sequencing within the approved barcode region. However, if requested, GenBank could assign the BARCODE flag to records with shorter sequences following guidelines defined by CBOL (see 3a, below);
f) Include no more than 1% ambiguous sites for the entire submitted sequence;
g) Include the name of the gene region used;
h) Be associated with trace file submitted to the NCBI Trace Archive or the Ensembl Trace Server; and
i) 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.

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 will be strongly recommended but not required:

j) Latitude and longitude;
k) Name of the identifier;
l) Name of the collector; and
m) Date of collection.

[4] The voucher specimen identifier uses a triplet structure (institution|collection|item) as used in the DarwinCore, advocated by GBIF. This triplet field is parallel to the Life Science Identifier (LSID) that is an Object Management Group (OMG) standard.

[5] DWG 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) **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 500 base-pair minimum could be BRIs. These might include sequences from type specimens or specimens of extinct or extremely rare species;
   b) CBOL will be responsible for establishing, implementing, and offering a process 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 Database (BoLD) will be considered by GenBank to be submitted jointly by the individual researcher and BoLD. These records can be modified by either party;
   d) BRIs that are submitted to GenBank directly by individual researchers may only be modified by the submitter. However, GenBank will remove the BARCODE flag from these records at CBOL's recommendation. These records would remain in GenBank as non-BARCODE records; and
   e) DWG 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.