

**Standards for BARCODE Records in GenBank (BRGs):
Voucher Specimen Identifiers and Species Names
Report of Meetings Convened by the Database Working Group
Consortium for the Barcode of Life**

Smithsonian Center for Research and Conservation, Front Royal, VA
Wednesday and Thursday, 27-28 April 2005

Background. The Consortium for the Barcode of Life (CBOL) formed a Database Working Group (DWG) at its inaugural meeting in May 2004. Within a short time, DWG approached the National Center for Biotechnology Information (NCBI) at the National Institutes of Health's National Library of Medicine with an inquiry concerning collaborative activity. NCBI responded positively and a DWG meeting was held there in September 2004. At that meeting, officials in GenBank, the US national repository for nucleotide sequence data, offered to act as an archival repository for DNA barcode records and to use "BARCODE" as a reserved keyword to identify these records. This keyword would be the flag on all BARCODE records in GenBank (BRGs).

Participants at the September 2004 NCBI meeting argued that DNA barcode data records should adhere to a different standard than typical GenBank records that are used in exploratory biomedical research. Barcode records, according to the meeting participants, should be usable as authority references that connect DNA sequences to the names of species. To serve this function, each barcode record in GenBank should be linked to a voucher specimen, preserved and available for further study in a museum, herbarium, zoo, frozen tissue collection, or other repository of biological reference material. GenBank records can include references to such voucher specimens, but the data field is not structured and is therefore not easily searchable. Associating voucher specimens with GenBank records is not routine practice across all taxonomic groups.

The First International Barcode Conference was held at the Natural History Museum in London in early February 2005. Preliminary plans for data standards for barcode records in GenBank were presented at that meeting. NCBI proposed to use the keyword BARCODE to indicate GenBank records that adhere to the protocols and higher data standards established in collaboration with CBOL. Participants in the London conference were enthusiastic at the prospect of a new and stronger tradition of linking GenBank records to voucher specimens. They challenged CBOL's leadership and the DWG in particular to create and maintain barcode records in GenBank as a high quality reference data resource. In particular, they noted two pressing needs:

- unambiguous linkages between GenBank records and voucher specimens, and
- improvement in the sources of species names linked to GenBank records.

In response to the challenges voiced at the London meeting, CBOL and DWG convened two meetings at the Smithsonian Institution's Center for Research and Conservation at Front Royal, Virginia. These meetings were devoted to the problems of voucher specimen identifiers and sources of species names submitted with barcode records in GenBank. The reports of those meetings follow.

Meeting Report:
Voucher Specimen Identifiers in BARCODE Records in GenBank
(BRGs)

Wednesday and Thursday, 27-28 April 2005
Smithsonian Institution Center for Research and Conservation, Front Royal, VA

Robert Hanner, Chair of the CBOL Database Working Group, chaired a small meeting devoted to developing a proposal for linking barcode records in GenBank to corresponding voucher specimens. (See Appendix 1, Participant list and agenda A). The meeting organizers had decided that for the present purposes, there were three important systems of identifying voucher specimens:

- GenBank uses a free-text field but is considering the addition of a structured field that concatenates institution acronyms, collection codes within institution, and specimen catalog IDs within collections. This is the approach used in the [DarwinCore](#) standard and is parallel to the Life Science Identifier (LSID) [proposed by the Objects Management Group](#) (presentation by Scott Federhen);
- BoLD (Barcode of Life Database, University of Guelph) used four data fields (institution, collection, catalog number and collector) but now uses the specimen ID provided by the collector. BoLD has installed DiGIR software that will allow retrieval of specimen data from DiGIR data providers using the specimen ID (presentation by Sujeevan Ratnasingham);
- GBIF is providing web access to specimen data held in museums around the world. They have been discussing the need for a Globally Unique ID (GUID), perhaps one that extends the LSID approach (presentation by Donald Hobern); and
- ISIS is a database of zoo specimens from more than 1,000 zoos. It tries to deal with the problem that a single animal can provide multiple tissue and blood samples during its life. Zoo animals are commonly transferred among zoos with the result that samples from a single animal can be collected at different zoos, giving the appearance that the samples came from different animals. The Zoological Information Management System (ZIMS), a zoo database system under development, will implement a system of GUIDs (presentation by Crispin Wilson).

In the ensuing discussion, participants noted that the triplet [institution-ID collection-ID specimen-ID] could, in principle, be a workable solution but there were a number of important obstacles. GBIF has found more than 650,000 combinations of institution-IDs and collection-IDs because there is no universally accepted controlled vocabulary for either institutions or collections within them. A number of authority files for institutional acronyms have been published (e.g., Index Herbariorum), but these were created separately within disciplines and have never been merged. As a result, a museum may have one acronym used by mammalogists and another used by botanists. There may also be overlapping use of an acronym by different institutions in different countries.

NCBI is willing to compile these authority files and resolve inconsistencies involving institutional acronyms and collection IDs. GBIF and the Taxonomic Data Working Group (TDWG) have received a grant from the Moore Foundation for implementation of data standards in biodiversity databases, including work on the GUID question. NCBI stated its eagerness to work with GBIF and TDWG in assembling and resolving the

authority files. Participants agreed that at this point in time, a “gold standard” system of identifying specimens would be a triplet that concatenates the institution, collection and specimen identifiers. The provisional syntax, subject to the approval of GenBank, EMBL and DDBJ would be delimited by brackets and internally separated [institution-ID collection-ID specimen-ID].

In many cases, newly collected specimens will be barcoded and submitted to GenBank long before they are accessioned into a museum or herbarium. As a result, they will not have received their catalog identification numbers by the time the barcode record is submitted to GenBank. Participants agreed that interim specimen IDs could be used and later replaced with the final ID. The interim ID would use the same triplet structure, composed of [‘Collector name’ collection specimen-ID]. Participants termed this the “silver standard” for identifying voucher specimens.

Some repositories of biological specimens use a single registration system that is not segregated by collection. In this case, the specimen ID would be [institution-code registration-number].

Recommendations. The participants agreed to the following recommendations to the CBOL Database Working Group:

- BARCODE records in GenBank (BRGs) would be designated by the keyword “BARCODE”, if and only if they adhere to the data standards agreed to by CBOL and NCBI. In designating “BARCODE records” with the “barcode flag”, CBOL aims to provide the research community with a validated set of reference records that connect DNA barcode sequences to voucher specimens (and by inference to their species name);
- GenBank should not go back through existing records to identify those that merit the barcode flag. Only new submissions should be considered as potential BRGs;
- Initially, barcode records must come from the 5’ end of the mitochondrial cytochrome c oxidase gene region. Barcoding projects that elect to use a different gene region must justify their proposal to submit non-COI barcode records to CBOL;
- BARCODE records must be produced with bi-directional sequencing with less than 1% ambiguous bases and neither frame shifts nor stop codons;
- BARCODE records must include a minimum sequence length of 500 homologous bases at 5’ end of COI (specified relative to the mouse genome); records with shorter sequences can only be a BARCODE record if it comes from a type specimen or a representative of an extinct species. In these cases, a comment that justifies the shorter sequence must be included in the source feature note. Requests for barcode status in other cases must be approved in advance by CBOL;
- Primers should be identified in all BARCODE records. If multiple primer pairs were used on shorter fragments, these should be noted;
- Entry of following required locality data:
 - GenBank’s country code list, or possible alternative system (blank entries will be recorded as unknown); and/or
 - Geospatial reference such as GPS coordinates (unknown or withheld are acceptable entries)

- Trace files (sequencer electropherograms) are the raw data from which sequences are inferred. For this reason, trace files should be uploaded to NCBI's or the Sanger Centre's Trace Archives and linked to the BRG's accession number;
- Quality scores should be submitted with all BRGs;
- Links to web-based voucher specimen records that adhere to the gold or silver standard as described above;
- Optional links to web-based records of associated preserved tissue and/or DNA samples from the same voucher specimen;
- Links to the following on-line metadata elements are desirable but not required in barcode records:
 - Collector name
 - Collection date
 - Identified by
 - Original field collection ID
 - Tissue type
 - Quality score for sequence (if the sequence was assembled from shorter fragments, a note to this effect should be provided)

Participants also recommended that the species name provided in the barcode record should be drawn from an authority file. The second meeting focused on this issue.

Implementation. Several large scale, longer-term barcoding "campaigns" have been launched recently. These are the All Birds Barcoding Initiative (ABBI) and the All Fishes Barcoding Initiative (FishBOL). ABBI hopes to obtain barcode data for 10,000 bird species by 2010, and FishBOL has targeted all species of marine fishes within five years. CBOL agreed to work with the leaders of these and other barcoding campaigns to implement the proposals approved by the Database Working Group. In this way, high quality BARCODE data records would flow into GenBank and would establish best practices that could be adopted by the research community.

Participants agreed to the following near-term action items (i.e., April-September 2005):

- CBOL will:
 - Inform ABBI/FishBOL and other barcoding campaigns of proposed standards for BARCODE records;
 - Provide GenBank with the proposed requirements for BARCODE flag,
 - Work with GenBank on documentation of the data standards;
 - Report to the Database Working Group and hold an on-line comment/discussion period;
 - Establish a process for reviewing proposals to give the BARCODE flag to records with non-COI regions, and for conferring BARCODE status to records with sequences shorter than 500bp;
 - Consider holding a Database Working Group in fall 2005, possibly in conjunction with GBIF's October meeting in Sweden in October, or a November NIST meeting in Charleston; NESCENT will be issuing a call for workshop proposals and this may be an appropriate mechanism; GBIF has agreed to provide \$10K toward this workshop; and
 - With the Database Working Group's approval, disseminate the specifications for BARCODE records in GenBank.

- GenBank will:
 - Propose the structured voucher ID for adoption;
 - Propose additional barcode qualifiers to EMBL, DDBJ;
 - Write specifications for BARCODE flag;
 - In collaboration with GBIF and TDWG, develop the authority list of gold and silver standard codes for institutions and collections; and
 - Explore entry of multiple primer pairs;
- BoLD and Coriell Institute of Medical Research will revise the GenBank records they submitted recently in order to merit the BARCODE flag. This will require:
 - Submitting trace files; and
 - Providing online access to the required metadata.
- GBIF will:
 - Incorporate the CBOL proposal in GUID discussions;
 - consider incorporation of new metadata elements proposed by CBOL into TDWG/GBIF data models

Participants agreed to the following mid-term action items (September 2005 – April 2006):

- CBOL will discuss with BoLD the potential to develop distributable software for barcoding labs to use instead of web-based BoLD. This might be possible as part of BoLD's conversion to DB2;
- GenBank will add the capability to submit trace files at the same time they submit BRGs;
- BoLD will complete and begin testing its LIMS (Laboratory Information Management System); and
- GBIF will continue to work on developing specimen IDs and "soft" identifiers, and will solicit input from CBOL's Database Working Group as part of that effort.

Meeting Report: Species Names in BARCODE Records in GenBank (BRGs)

Thursday and Friday, 28-29 April 2005

Smithsonian Institution Center for Research and Conservation, Front Royal, VA

Immediately following adjournment of the meeting on voucher specimen IDs, Robert Hanner, Chair of the CBOL Database Working Group, convened a larger meeting devoted to developing a proposal for linking BARCODE records in GenBank to valid species names. (See Appendix 2, Participant list and agenda B). Dr. Hanner recounted the desire expressed at the London barcode conference for more thorough screening of the taxonomic names associated with barcode records in GenBank.

Current Status of Species Names in Barcode Records. Scott Federhen (National Center for Biotechnology Information, NCBI) and Sujeevan Ratnasingham (University of Guelph) began the meeting by describing, how GenBank and BoLD currently manage the species names submitted with barcode records. GenBank now has 132,000 bionomials in the NCBI Taxonomy database, along with a classification system above the species level. Submitters of GenBank records are not required to demonstrate the validity of the species name they submit. At the time of submission, a search is routinely run against various name lists using the text-string of the species name. If a match is found, GenBank's link-out capability is used to create a hyperlink to that name. There are many such link-outs from GenBank to checklists such as the Integrated Taxonomic Identification System (ITIS) and Species2000. Nothing is done if the submitted name does not match a valid name in a checklist. Misidentifications and misspellings can only be corrected by the record's original submitter or by the submitting lab, if the record was submitted by a graduate student who has subsequently left the lab.

Third parties contact NCBI on a regular basis to point out incorrect species names or other errors in GenBank records. The original submitter is contacted with a request to review the record and consider correcting it. They are under no obligation to do so and GenBank does not adjudicate between third parties and submitters.

Participants pointed out that in addition to the species name submitted to GenBank with the BARCODE record, there is a species name associated with the voucher specimen on its catalog entry. A specialist could examine the specimen in its museum or herbarium, correct or update the specimen's identification, record the new species name on the specimen label and the museum catalog, thereby producing a discrepancy between the GenBank record and the voucher specimen record in its repository. In principle, the link between a BARCODE record and the corresponding catalog entry for its voucher specimen would allow changes in the species determination on the voucher to be incorporated automatically into the GenBank record. However, this would run counter to GenBank's policy to only allow submitters or submitting labs to make data corrections.

BoLD's main function is to serve as a workbench for assembling collections of specimens and their associated barcode sequences and specimen data from barcoding projects. As such, it is a tool for barcoding researchers and is not meant to be a general purpose archival database. BoLD has imported the NCBI taxonomy database as an initial source of species names. Participating taxonomists in a project in the BoLD database can

make corrections directly to a taxonomic identifications. In these cases, both the old and new names are maintained as part of the data record's audit trail history.

Sources of Species Names. Significant effort has been devoted to compiling published taxonomic names in nomenclators, and reviewing the validity of these names and assembling them into vetted taxonomic indices. Representatives of the following major taxonomic names initiatives gave presentations: Global Biodiversity Information Facility (GBIF), Catalog of Life, Species2000, Integrated Taxonomic Information System (ITIS), Ocean Biogeographic Information System (OBIS), The Zoological Record, and the International Plant Names Index (IPNI). In addition, participants received a presentation on a proposed compilation of all taxonomic names called NameBank. These presentations are summarized in Appendix 3.

In the ensuing discussion, participants agreed that there are two reasons for requiring that BRGs use species names from the highest quality sources. First, stricter quality standards are being developed for BRGs so they can act as authoritative references for identifying unknown specimens. It should therefore contain the most current valid taxonomic names that are available. Second, when a GenBank user finds a BARCODE record and wants more information on that specimen, the species name will provide a link to information on exactly the species indicated by the data submitter. Without authoritative lists of species' names, the submitter's intent may not be clear and the path to retrieving relevant data will be obscure.

Participants agreed that the GenBank data submission system should present submitters with not just one, but the full range of authoritative lists of species' names, for two reasons:

1. There are mature and well-maintained vetted indices of species names for some taxonomic groups, but other taxa have only unreviewed lists of species name. Some taxa have no compiled list of species names at all, so a general source such as NameBank would be the only reference; and
2. A certain portion of BARCODE records submitted to GenBank will be associated with the publication of new species names. As a result these names will not be listed in any authority file. The publication in which the new species name is proposed will be the only available source of information on that species at the time the BRG is submitted.

This led the participants to define the following hierarchy of sources for taxonomic names:

- **Gold Standard.** Sources of taxonomic names that have been reviewed for their adherence to taxonomic standards, objective and subjective synonymy, and reflect expert opinions. Gold standard sources have the added value of being well-maintained; that is, as a name is revised in subsequent releases of an index, new information on the status of that name will be retrievable through the GenBank record. The species name in the BRG will not have to be updated because the authority file will be kept current, and the history of the species name of a voucher specimen will be available to GenBank users through that record;
- **Silver Standard.** Sources of taxonomic names compiled in published nomenclators that may be reviewed for adherence to taxonomic standards of the nomenclatural

Code and monotypic synonymy. These lists provide links back to the publication in which the name was proposed.

- **Bronze Standard.** Sources of all published names (such as the proposed NameBank). This would include new names that have been recently published for taxonomic groups covered by gold and silver standard sources, but have not yet been incorporated through the normal compilation and review processes. In some cases, a submitter may want to attach a provisional name (e.g., *species A*) to a voucher specimen because it is a new species that is awaiting formal description. The Bronze Standard would include published provisional names. Linkage to the publication will ensure that the provisional name is unique and retrievable.
- **Tin Standard.** In many cases, submitters will want to put a provisional species name on a BRG but will not publish the data. In these cases, the provisional names may not be retrievable and may not be globally unique within genus. In these cases, GenBank will add a unique string to the provisional species name at the time of submission.

When a BRG includes a provisional name (either Bronze or Tin Standard), the name may be replaced subsequently with a formal name by the submitter.

Enforcing data standards and quality. Participants agreed that there were several ways that BRGs could attain and maintain high data quality standards:

- *During data assembly.* BoLD is a publicly available database that provides barcode projects with the software tools needed to record, edit, and test barcode data records. Paul Hebert's lab has received funding from the Canadian government for a national barcoding network, and they are making BoLD available to the Canadian network as a workbench for in-progress projects. Their plan is to allow projects to assemble their data records and submit them *en masse* to GenBank. Several hundred records have already been submitted to GenBank in this way. By incorporating data quality checks (e.g., preliminary clustering of records and checks for stop codons to weed out contaminated sequences), use of controlled vocabularies, and links to authority files, many common data errors can be avoided. GenBank would also accept direct submissions from individuals and labs that chose not to use BoLD as a workbench;
- *Upon arrival at GenBank.* GenBank regularly checks for some errors in the sequence and is willing to put in place some additional data checks. For example, they could run institution-ID, collection-ID and species names in submitted records against authority files; and
- *After release on GenBank.* GenBank is an archival database and its policy is to only make corrections to data records if and when the submitter requests a change. GenBank is approached with suggestions for corrections or complaints about inaccurate data by third parties. In these cases, GenBank attempts to contact the submitter of the original record but in some cases they can't be found or disagree with the proposed changes. There are two ways that GenBank's policies can be honored while maintaining strong data integrity for BRGs:
 - When an investigator or lab decides to prepare and submit their records through BoLD, they could be asked to agree that BoLD would share the right to correct data errors. As the co-submitter, either BoLD or the originating lab would have the ability to make corrections. For example, BoLD is capable of importing data from DiGIR data providers and this would allow BoLD to update the species

- names of voucher specimens from their museum records. For those BRGs submitted to GenBank through BoLD, GenBank would allow either the submitter or BoLD to make corrections to the species name; and
- When an investigator submits records directly to GenBank and does not agree to correct an error noted by a third party, the third party may submit their proposed change to CBOL. GenBank is willing to remove the BARCODE flag at CBOL's request. The data record would remain in GenBank but would no longer be a BRG.

Recommendations. Participants put forward the following recommendations:

1. In addition to the changes made to taxonomic names through the updating of gold standard indices, GenBank will need protocols for updating species names proposed for other reasons and from other sources. For example, a voucher specimen may be re-identified and the original misidentification will need to be corrected. Many of these re-identifications may be proposed by people other than the original submitter. Taxonomic revisions and changes in classification schemes that are not associated with a gold standard names index may also create the need to change a name record. The system of maintaining data quality, described above, should provide adequate means for updating taxonomic names.
2. Earlier versions of a species name should not be deleted; the full history of names applied to a BARCODE record should be maintained. The most recent version of the name should be displayed when a record is retrieved, and previous versions should be available at that time. BoLD currently maintains an audit trail of changes to species names during assembly of data records, but only the final name is submitted to GenBank. A decision is needed on whether or not to retain the name history that was associated with a specimen prior to submission to GenBank, including the name that was first attached to a specimen.
3. When a species name is changed, the source and reason for that change should be recorded and retained with the name;
4. The Database Working Group and GenBank should explore possible systems for third-party correction, comment or annotation of BRGs. This could improve the overall data quality of BRGs and could make data curation a distributed responsibility;
5. The full name citation (binomen, author, literature citation) is needed to link a data record to gold, silver and bronze standard records. Citations that are not at this level of specificity will not allow the long-term tracking of changes to the name/concept;
6. GenBank's instructions to submitters should provide non-taxonomists with a Guide to Best Practices that would help them include the most accurate taxonomic information in their data submissions;
7. There is a need to create incentives to publish new names that are attached to BARCODE records in compliance with Nomenclatural Codes. This adherence will be needed to create clear linkages to the complex of a species name/taxon concept/barcode sequence;
8. The GenBank submission process should include an initial triage, assignment of an accession number, checking of the species name by running the species name against the relevant source files (gold or silver standards), and confirming that records with newly proposed species names (bronze standard) include links to PubMed journals or

- citations to non-PubMed publications;
9. The normal GenBank option of “hold until published” on a specified date should be available to barcode records; and
 10. The release of BARCODE records associated with new species names could be tied to a registration system of taxonomic names, such as the ITIS TSN or the system under consideration by the International Code of Zoological Nomenclature.

Implementation. The participants agreed to the following action items:

1. GenBank will build working relationship with sources of gold, silver and bronze standard taxonomic names (Species2000, ITIS, Zoological Records, BioAbstracts, GBIF, IPNI, GRIN, other nomenclators, etc. (6-12 months);
2. CBOL, NCBI and BoLD representatives will meet to explore use of RefSeq and to review the status of potential BARCODE records already submitted GenBank (1 month);
3. CBOL will develop a process through which third parties may propose removal of the barcode flag from records submitted directly to GenBank;
4. The CBOL Database Working Group Chair will:
 - Report back to the meeting participants;
 - Work with GenBank on documentation;
 - Report to the WG through NBII Portal;
 - Hold an online comment/discussion period; and
 - Establish contact with TDWG;
5. GBIF/TDWG will develop software for the exchange of species concepts (6-9 months).

Appendix 1. Participant List and Agenda A

Voucher Specimen Identifiers in BARCODE Records in GenBank

Bob Hanner, Chair, Database Working Group
University of Guelph
rhanner@earthlink.net

Per Bjorn
Global Biodiversity Information Facility (GBIF)
pdpbjorn@gbif.org

Scott Federhen, GenBank
National Center for Biotechnology Information (NCBI), NIH
federhen@ncbi.nlm.nih.gov

Donald Hobern (GBIF)
dhobern@gbif.org

Scott Miller, Chair, Consortium for the Barcode of Life (CBOL)
Smithsonian Institution
millers@si.edu

Sujeevan Ratnasingham
University of Guelph
sratnasi@uoguelph.ca

David Schindel, Executive Secretary
Consortium for the Barcode of Life (CBOL)
SchindelD@si.edu

Crispen Wilson
National Biological Information Infrastructure (NBII), USGS
cwilson@usgs.gov

AGENDA

Wednesday, 27 April:

12:00 - Lunch and introductions

1:00 - Statement of the problem (Bob Hanner)

1:30 - Informal presentations on Specimen ID systems used by:

- GenBank (Scott Federhen)
- GBIF (Donald Hobern)
- BOLD/Guelph (Sujeevan Ratnasingham)

3:00 - Coffee break

3:30 - Discussion of requirements for a specimen ID system in Barcode Section of Genbank

6:00 – Adjourn

7:30 - Dinner at nearby restaurant

Thursday, 28 April:

7:30 – Breakfast

8:30 - Group design effort on specimen ID data schema

10:30 - Group drafting of [proposed](#) data schema

11:30 – Adjourn

Appendix 2. Participant List and Agenda B Species Names in BARCODE Records in GenBank (BRGs)

Bob Hanner, Chair, Database Working
Group
University of Guelph
rhanner@earthlink.net

Frank Bisby, Species2000
University of Reading
F.A.Bisby@sp2000.org

Per Bjorn
Global Biodiversity Information Facility
(GBIF)
pdpbjorn@gbif.org

Cliff Cunningham, Director
National Evolutionary Synthesis Center
(NESCENT), Duke University
cliff@duke.edu

Scott Federhen, GenBank
National Center for Biotechnology
Information (NCBI), NIH
federhen@ncbi.nlm.nih.gov

Daphne Fautin
University of Kansas
fautin@ku.edu

Fred Grassle, Ocean Biogeography
Information System (OBIS)
Rutgers University
grassle@imcs.rutgers.edu

Joel Hammond
Thomson Publishing
joel.hammond@thomson.com

Sally Hinchcliffe, International Plant
Names Index (IPNI)
Royal Botanic Gardens, Kew
s.Hinchcliffe@kew.org

Donald Hobern (GBIF)
dhobern@gbif.org

Chris Lyal
Natural History Museum, London
C.lyal@nhm.ac.uk

Scott Miller, Chair, Consortium for the
Barcode of Life (CBOL)
Smithsonian Institution
millers@si.edu

Alan Paton
Royal Botanic Gardens, Kew
a.paton@rbgkew.org.uk

Andrew Polaszek
International Code of Zoological
Nomenclature (ICZN)
iczn@nhm.ac.uk

Sujeevan Ratnasingham
University of Guelph
sratnasi@uoguelph.ca

David Remsen
Marine Biological Lab
dremsen@mbl.edu

Michael Ruggiero
Integrated Taxonomic Information
System (ITIS)
ruggierm@si.edu

David Schindel, Executive Secretary
Consortium for the Barcode of Life
(CBOL)
SchindelD@si.edu

Anna Weitzman
Smithsonian Institution
Weitzman.Anna@nmnh.si.edu

Crispen Wilson
National Biological Information
Infrastructure (NBII), USGS
cwilson@usgs.gov

John Wiersema
USDA Agricultural Research Service
jwiersema@ars-grin.gov

Phoebe Zhang
OBIS, Rutgers University
phoebe@marine.rutgers.edu

AGENDA

Thursday, 28 April:

- 12:00 – Lunch and introductions
- 1:00 – Statement of the Problem (Bob Hanner)
- 1:15 – Goals of meeting, framework for decision-making: (David Schindel)
- 1:30 – Current status - species names attached to barcode records:
 - The NCBI Taxonomy database: (Scott Federhen)
 - The Barcode of Life Database, University of Guelph (Sujeewan Ratnasingham)
- 2:00 – Sources of lists of species names for Barcode Section of GenBank:
 - Global Biodiversity Information Facility (GBIF; Per Bjorn and Donald Hobern)
 - Species2000 (Frank Bisby and Daphne Fautin)
 - Integrated Taxonomic Information System (ITIS; Michael Ruggiero)
 - Ocean Biogeographic Information System (OBIS; Fred Grassle)
 - The Zoological Record (Joel Hammond)
 - International Plant Names Index (IPNI; Sally Hinchcliffe and Alan Paton)
- 3:00 – Coffee break
- 3:30 – MBL Names Bank proposal (David Remsen, MBL)
- 4:15 – Discussion: How do major initiatives fit the needs of the Barcode Initiative?
- 6:00 – Adjourn
- 6:30 – 7:30 - Guided driving tour of CRC
- 7:30 – Dinner

Friday, 29 April:

- 7:30 – Breakfast
- 8:30 – Group drafting of proposal and data schema and implementation plan for species names in Barcode Section of GenBank and/or BoLD
- 10:00 – Coffee break
- 10:30 – Group drafting continues
- 12:30 – Lunch
- 1:00 – Formal drafting of data schema and implementation plan
- 2:00 – Walking tour of CRC
- 3:00 – Finalize data schema and implementation plan
- 5:00 – Adjourn

Appendix 3. Summaries of presentations on taxonomic names initiatives

Global Biodiversity Information Facility (GBIF; presentation by Donald Hobern) is an international organization devoted to providing web access to species and specimen data. One of GBIF's four main activities is ECAT, an electronic catalog of all taxonomic names and their sources. The goal is to concatenate and unify these sources, provide a consistent view across them, and eliminate redundancy among them. ECAT will not attempt to create synonymies or consensus lists of names, but rather consistent access to all sources. ECAT will be offering seed grants for projects that assemble names in untreated taxonomic groups. GBIF is working with TDWG to develop a taxonomic concept schema that will facilitate exchange of information across sources of names. For example, legal lists of names (e.g., endangered taxa, invasives) may not reflect the most current taxonomy and they need to be mapped onto the most current names and classifications.

Catalog of Life and Species2000 (presentation by Frank Bisby). The Catalog of Life is a partnership between Species2000 (an independent global NGO) and ITIS (a US government activity). Its goal is a list of valid species names for all species (1.75 million?) by 2011. It has just reached the 500,000 name milestone. CoL publishes an annual checklist in April and maintains a database that undergoes constant updates. The CoL list presents objective and subjective synonymies provided by domain experts. It also manages a process for identifying the best taxonomy available for a taxonomic group. Species 2000 uses taxonomic experts who contribute their synonymized species checklists. For those taxa with multiple expert sources and competing classifications, Species2000 will conduct a peer review that selects preferred solutions. Competing taxonomies can be presented in Species2000.

Integrated Taxonomic Information System (ITIS; presentation by Michael Ruggiero) began with the goal of creating a validated list of species names for use by US government agencies but it has expanded to the global level. ITIS uses taxonomic contributors who review taxonomic names, compile taxonomies, and resolve conflicting taxonomies. Based on the results of this review process, valid species names are given taxonomic serial numbers (TSNs).

Ocean Biogeographic Information System (OBIS; presentation by Fred Grassle) was born out of the Census of Marine Life and has grown to include a million records of species occurrence. These records come from published studies that can include ecological data (e.g., abundance, environmental parameters, migratory pathways, habitat classification) and museum data. Oceanographic field projects can be highly variable with respect to sources of taxonomic determinations. Some (e.g., NGISA, a nearshore project in Japan, and coral reef projects) maintain voucher specimens and have their own local taxonomic experts. OBIS uses a system of ten regional data nodes. It makes no attempt to synonymize names or to adjudicate conflicting opinions.

The Zoological Record (presentation by Joel Hammond) is a commercial bibliographic compendium that is now owned and marketed by Thomson. It is a taxonomic index derived by scanning journals, meeting publications, monographs, and books. ZR now contains 1.4 million names.

International Plant Names Index (IPNI; presentations by Sally Hinchcliffe and Alan Paton) is the concatenation of three lists of plant names: Index Kewensis, Gray Cards, and the Australian Plant Names Index. IPNI is a nomenclator compiled by scanning publications, i.e., a list of published names filtered for nomenclatural synonymies but not subjective synonymies. It covers all vascular and includes 1.4 million names. Corrections can be made by the editors based on third-party suggestions. Kew is considering the creation of a vetted checklist of names based on IPNI. The checklist would compile names and citations and would create a list of accepted names with standardized spelling, adherence to the International Code of Botanical Nomenclature, and synonymies based on priority of literature and opinions of peer reviewers.

NameBank (presentation by David Remsen) is a proposal by the Marine Biological Laboratory to compile all names, valid and invalid, formal and informal, peer reviewed or not, as the base-level source of all names and their published source.

International Code of Zoological Nomenclature (presented by Andrew Polaszek) is currently being revised (preparation of a 5th edition by 2008). In connection with the preparation of the new edition of the Code, the International Commission on Zoological Nomenclature (ICZN) is in the process of discussing a future registry for all scientific names of animals. The development of such a facility would be in close collaboration with both GBIF and CBoL.