Workshop Report: CBOL Database Working Group

Interoperability of Museum, Taxonomic, and DNA Databases

Field Museum of Natural History, Chicago, IL; 7-9 January 2009

The Database Working Group (DBWG) of the Consortium for the Barcode of Life (CBOL; www.barcoding.si.edu) convened a meeting at the Field Museum of Natural History in Chicago to explore ways to improve the connectivity among DNA barcode data and databases of voucher specimens and taxonomic names. Twenty-two representatives of database initiatives and four members of CBOL’s Secretariat Office participated in the workshop (see participant list, Appendix 1). Discussions produced a set of action items, many of which use the 1-3 June 2009 e-Biosphere 09 conference as a deadline.

The workshop had three components (see agenda, Appendix 2):

1. **Presentations.** The first component included presentations by representatives of three categories of database initiatives: DNA barcodes, taxonomic names, and museum/herbarium specimens. The following database initiatives presented information on the degree to which they were dynamically interoperable with each other:

   A. Databases of DNA barcode records:
      - the Barcode of Life Data Systems (BOLD);
      - GenBank/NCBI;
      - DNA Data Bank of Japan; and
      - The Barcode of Life Initiative’s (BOLI) Data Portal;

   B. Databases of taxonomic names:
      - The Catalogue of Life (Species2000 and ITIS);
      - Encyclopedia of Life;
      - ZooBank;
      - GBIF/Global Names Index and Global Names Architecture (GNI/GNA); and
      - World Registry of Marine Species (WoRMS);

   C. Databases of museum specimens:
      - ARCTOS;
      - Specify Software; and
      - KE-EMu.

2. **Discussions and Action Items.** The workshop’s second component was devoted to discussions of how interoperability between barcode data in BOLD, GenBank, DDBJ, and EMBL and databases of voucher specimens and taxonomic names can be improved in the near-term and mid-term. CBOL is eager to demonstrate improved interoperability at the e-Biosphere 09 Conference on Biodiversity Informatics in June 2009. The specific functions that CBOL is promoting are:

   - Adherence to the BARCODE data standard, especially use of a structured triplet form of voucher specimen identifiers;
Active hyperlinks between BARCODE records in GenBank and voucher specimen records in museum or herbarium databases;
Easy import/export of records between BOLD and museum/herbarium databases;
More automated systems of checking BOLD and GenBank records for compliance with CBOL’s BARCODE data standards; and
Easy checks on the species identifications in museum/herbarium databases using records in BOLD and GenBank

Participants strongly endorsed the use of the Darwin Core triplet (a structured combination of institutional acronym, collection code, and specimen ID number) as a ‘natural key’ that readers could interpret. However, the underlying link will in most cases use static URLs, or LSIDs and GUIDs as they are implemented.

The barcoding community relies on provisional taxonomic labels to a great extent during the process of reconciling patterns of barcode variation with formal taxonomic names. CBOL has been taking an active interest in regularizing the structure and use of these non-Linnaean labels for clusters of variation that may develop into formal taxon concepts. Participants considered this a very relevant topic for consideration by TDWG (perhaps under its Taxon Names and Concepts Working Group and encouraged CBOL to work with them.

Short presentations were given on databases of controlled vocabularies of biorepository acronyms and collection codes:

- NCBI’s compilation of institutional acronyms from diverse sources;
- Biorepositories.org, an online registry based on the NCBI list, created by CBOL in consultation with NCBI and GBIF; and
- Biodiversity Collections Index (BCI), hosted by Royal Botanic Gardens Edinburgh with GBIF support.

The following action items emerged from that discussion:

A. BOLD will explore the web services now provided by Catalogue of Life. These services may allow BOLD to expand the ways it conducts quality checks of taxonomic names.

B. BOLD has plans for a series of new features but it is unclear which can be completed before the June 2009 e-Biosphere conference. These features include new web services which could provide the interoperability functions described above.

C. Participants agreed on the potential of GNI/GNA as a potential access system to taxonomic names and their status. CBOL, BOLD and GenBank will closely follow progress on the development of the Global Names Index/Global Names Architecture. The barcoding community can help define requirements for web services that will improve the quality of taxonomic names.

D. WoRMS will create linkages to BOLD, GenBank, and the Ocean Genome Legacy.

E. BOLD and GenBank will consider building connections to WoRMS as a system for checking the quality of taxonomic names of marine species.
F. ARCTOS is the first and only museum specimen database system that can easily link BARCODE records to online specimen databases. CBOL will promote large-scale barcoding initiatives in ARCTOS collections to take advantage of this functionality. ARCTOS and BOLD will also discuss web services that can create several of the other forms of interoperability listed above.

G. The Specify Software Project committed to developing a plug-in for web service integration between Specify and BOLD in the spring of 2009. Specify 6 which is designed as a pluggable, modular, Java platform will be easily extended for full integration with BOLD web services. A demonstration BOLD plug-in was shown at the meeting. The collaboration with BOLD developers, scheduled in early March will address several specific use cases involving barcode data of value to the biological collections community. CBOL expressed its willingness to consider financial support for some of this feature development. Specify and CBOL will co-author some marketing information conveying the value of BOLD integration to Specify clients.

H. CBOL will propose action plan to KE Software that includes a list of new features related to barcode data that could be added to KE-EMu. CBOL will also offer to visit some EMu users to give presentations on barcoding and to participate in the October 2009 meeting of the natural history EMu User Group. In the near-term, a new field may be added to the Darwin Core data standard for the URL of the specimen record. If it is approved, KE Software will consider adding this field which would enable easy linkage with BARCODE records.

I. CBOL will continue to provide updates from biorepositories.org to NCBI and BCI. CBOL will also approach BCI to determine whether BCKI has issued LSIDs for biorepositories, and if these IDs can be imported into biorepositories.org.

J. CBOL will expand its interactions with TDWG, especially in the area of provisional taxon labels.

3. Planning for e-Biosphere 09 conference. The final component of the workshop focused on plans for the e-Biosphere conference in early June 2009. The conference represents an opportunity to showcase the most advanced capabilities of barcode-related databases and to plan new initiatives. The conference also includes two days of high-level planning among biodiversity informatics initiatives and the outcomes of the Chicago workshop can feed directly into those discussions.

D. Schindel provided an overview of the purpose, agenda and content of the e-Biosphere 09 conference and invited the workshop’s participants to submit poster abstracts and rent display booths and exhibit areas for software and database demonstrations. He also invited participants to join the electronic discussion for on the Online Conference Community. The OCC will enable people to communicate in preparation for break-out discussion groups of users that is scheduled for the second day of the conference.

C. Parr and D. Patterson led a discussion of real-time demonstrations that could be in progress during the conference’s three days and other ways to engage participant interest
in the conference. A full description of these ideas will be transmitted to the e-Biosphere 09 Conference Steering Committee. They include:

- Challenging the taxonomy community to use cybertaxonomy to document and publish a new species during the three-day conference;
- Issuing broader challenges to users of biodiversity informatics to design the workflow of the future;
- Creating a tracking and display system that can be used during the conference and in projects like bio-blitzes to show progress along a workplan, like the lights on displays in the London Underground;
- Linking demonstrations to the information and software providers on display in the e-Biosphere exhibit area, so that booth inhabitants can demonstrate and articulate the value-add they bring to community workflows;
- Promote real-time projects with tangible connections to the conference, such as using DNA barcodes during the conference to:
  - test species identification of food being served in local restaurants;
  - survey the microbes in the Thames River or
  - survey the fungi in the air and carpets of the conference center.
- Promote discussions that focus on the use of biodiversity patterns in hypothesis-testing and hypothesis-driven research.
- Have flat-panel display monitors distributed around the conference venue on which the dynamic growth and use of biodiversity informatics could be displayed throughout the conference. Possible examples are:
  - The ongoing acquisition of Flickr images into the Atlas of Living Australia database;
  - The dynamic process of building EOL species pages;
  - Automated and manual data cleaning in EOL and other projects;
  - Automated checking of taxonomic names against the Global Names Index or other web services; and
  - Progress of the workflows of cybertaxonomy demonstrations described above.
# Appendix 1: Participant List

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>E-mail Address</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appeltans, Ward</td>
<td>WORMS and EUROBIS</td>
<td><a href="mailto:ward.appeltans@vliz.be">ward.appeltans@vliz.be</a></td>
<td>Belgium</td>
</tr>
<tr>
<td>Beach, Jim</td>
<td>Specify Software Project/Univ. of Kansas</td>
<td><a href="mailto:beach@ku.edu">beach@ku.edu</a></td>
<td>USA</td>
</tr>
<tr>
<td>Bisby, Frank</td>
<td>Catalogue of Life</td>
<td><a href="mailto:f.a.bisby@reading.ac.uk">f.a.bisby@reading.ac.uk</a></td>
<td>UK</td>
</tr>
<tr>
<td>Blum, Stan</td>
<td>TDWG/Calif. Academy of Science</td>
<td><a href="mailto:sblum@CalAcademy.org">sblum@CalAcademy.org</a></td>
<td>USA</td>
</tr>
<tr>
<td>Federhen, Scott</td>
<td>NCBI</td>
<td><a href="mailto:federhen@ncbi.nlm.nih.gov">federhen@ncbi.nlm.nih.gov</a></td>
<td>USA</td>
</tr>
<tr>
<td>Fusaro, Abigail</td>
<td>Ocean Genome Legacy</td>
<td><a href="mailto:fusaro@oglf.org">fusaro@oglf.org</a></td>
<td>USA</td>
</tr>
<tr>
<td>Hanner, Robert</td>
<td>iBOL/Biodiversity Institute of Ontario</td>
<td><a href="mailto:rhanner@uoguelph.ca">rhanner@uoguelph.ca</a></td>
<td>Canada</td>
</tr>
<tr>
<td>Hernandez, Francisco</td>
<td>WORMS and EUROBIS</td>
<td><a href="mailto:francisco.hernandez@vliz.be">francisco.hernandez@vliz.be</a></td>
<td>Belgium</td>
</tr>
<tr>
<td>Ivanenko, Viatcheslav N.</td>
<td>Moorea Biocode Project/ WoRMS</td>
<td><a href="mailto:ivanenko.slava@gmail.com">ivanenko.slava@gmail.com</a></td>
<td>Russia</td>
</tr>
<tr>
<td>Jarrell, Gordon</td>
<td>ARCTOS/Museum of Southwest Biology</td>
<td><a href="mailto:gordon.jarrell@gmail.com">gordon.jarrell@gmail.com</a></td>
<td>USA</td>
</tr>
<tr>
<td>Kwiatkowski, Andy</td>
<td>BOLD</td>
<td><a href="mailto:andrewk@uoguelph.ca">andrewk@uoguelph.ca</a></td>
<td>Canada</td>
</tr>
<tr>
<td>Lickman, Brad</td>
<td>KE Software</td>
<td><a href="mailto:brad.lickman@kesoftware.com">brad.lickman@kesoftware.com</a></td>
<td>USA</td>
</tr>
<tr>
<td>Ma, Juncai</td>
<td>BOLD Mirror, Chinese Academy of Sciences</td>
<td><a href="mailto:ma@sun.im.ac.cn">ma@sun.im.ac.cn</a></td>
<td>China</td>
</tr>
<tr>
<td>Orrell, Tom</td>
<td>ITIS and KE-EMu/Smithsonian</td>
<td><a href="mailto:Orrellt@si.edu">Orrellt@si.edu</a></td>
<td>USA</td>
</tr>
<tr>
<td>Parr, Cyndy</td>
<td>EOL/Smithsonian</td>
<td><a href="mailto:parrc@si.edu">parrc@si.edu</a></td>
<td>USA</td>
</tr>
<tr>
<td>Patterson, David</td>
<td>EOL/Marine Biological Lab</td>
<td><a href="mailto:paddy@mbl.edu">paddy@mbl.edu</a></td>
<td>USA</td>
</tr>
<tr>
<td>Pyle, Rich</td>
<td>ZooBank/Bishop Museum</td>
<td><a href="mailto:deepreef@bishopmuseum.org">deepreef@bishopmuseum.org</a></td>
<td>USA</td>
</tr>
<tr>
<td>Ratnasingham, Sujeevan</td>
<td>BOLD</td>
<td><a href="mailto:sratnasi@uoguelph.ca">sratnasi@uoguelph.ca</a></td>
<td>Canada</td>
</tr>
<tr>
<td>Remsen, David</td>
<td>GBIF</td>
<td><a href="mailto:dremsen@gbif.org">dremsen@gbif.org</a></td>
<td>Denmark</td>
</tr>
<tr>
<td>Roberts, Vince</td>
<td>MycoBank/CBS</td>
<td><a href="mailto:robert@cbs.knaw.nl">robert@cbs.knaw.nl</a></td>
<td>Netherlands</td>
</tr>
<tr>
<td>Spears, Rod</td>
<td>Specify Software Project/Univ. Kansas</td>
<td><a href="mailto:rods@ku.edu">rods@ku.edu</a></td>
<td>USA</td>
</tr>
<tr>
<td>Sugawara, Hideaki</td>
<td>National Institute of Genetics</td>
<td><a href="mailto:hsugawari@genes.nig.ac.jp">hsugawari@genes.nig.ac.jp</a></td>
<td>Japan</td>
</tr>
</tbody>
</table>

**CBOL Secretariat Office staff**

- Schindel, David   schindeld@si.edu
- Fritzsche, Meg    fritzschem@si.edu
- Jett, Kristin     jettk@si.edu
- Trizna, Mike      triznam@si.edu
Appendix 2: Workshop Agenda
Interoperability of Museum, Taxonomic, and DNA Databases
Database Working Group, Consortium for the Barcode of Life
Field Museum of Natural History, Chicago, Illinois
Wednesday to Friday, 7-9 January 2009

Participants are encouraged to ask for time in any session if they wish to make a presentation

**Wednesday, 7 January 2009:**
2:00pm: Welcome, overview, goals of workshop

*This session will introduce the history and current status of barcode data and their interconnectivity, and wider issues of interoperability in taxonomic and museum databases*

2:00: R. Hanner, Chair, Database Working Group (DBWG), Univ. Guelph
2:20: D. Schindel, Executive Secretary, Consortium for the Barcode of Life (CBOL)

2:30pm: Session 1. Status of barcode data and connections to other databases

*Three 15-20 minute presentations will introduce the main repositories of barcode data and their connections to each other and databases of taxon names and voucher specimens*

2:30: Barcode of Life Data Systems (BOLD): S. Ratnasingham, Univ. Guelph
2:50: NCBI/Barcode Submission Tool (BarSTool): S. Federhen, NCBI/NIH
3:30: Discussion

4:00pm: Coffee Break

4:30pm: Session 2: Status and plans for BOLD platform: S. Ratnasingham, Univ. Guelph

*This will be a more detailed, technical presentation of BOLD and its plans for expansion and improved interoperability.*

6:00pm: Adjourn for dinner at a nearby restaurant

**Thursday, 8 January 2009:**
9:00am: Session 3. Panel discussion of representatives of databases with potential connections to barcode data
Four 15-minute presentations will provide overviews of how major databases of taxonomic names are currently providing connectivity to other databases. The emphasis will be on web services, not queries by web visitors.

9:00: D. Patterson or C. Parr, Encyclopedia of Life
9:15: T. Orrell and F. Bisby, Catalogue of Life
9:30: D. Remsen, GBIF
9:45: R. Pyle, ZooBank
10:00: Discussion

10:30am: Coffee break

11:00am: Session 4. Status of museum databases and their potential connectivity to barcode data

Four 15-minute presentations will provide overviews of how some leading museum collection databases are linking to databases of taxonomic names and gene sequence databases. The emphasis will be on web services rather than queries by collection managers or web visitors.

11:00: KE-EMu: B. Lickman and T. Orrell
11:15: Specify: J. Beach and R. Spears
11:30: ARCTOS: G. Jerrell
11:45: Mycobank/BioloMICS: V. Roberts
12:00: Discussion

12:30pm: Lunch

2:00pm: Session 5. Enhancing connectivity with barcode data: Conceptual issues and challenges

Five brief presentations (five minutes, five slides, maximum) will introduce specific conceptual issues that will need consideration by the group. These are meant to propose new directions and stimulate discussion.

- Controlled vocabulary for repositories: D. Schindel
- Biodiversity Collection Index: R. Hyam (presented in absentia)
- Interim/provisional taxon labels: D. Schindel
- Barcode Identification Numbers: S. Ratnasingham
- Linking to updated versions of taxonomic names: S. Blum, F. Bisby, T. Orrell

2:30pm: Discussion

3:30pm: Coffee break

4:00pm: Session 6. Enhancing connectivity with barcode data: Technical issues and challenges
**Brief presentations of technical issues will set the stage for discussion and group planning.**

- DNA Data Bank of Japan (DDBJ): H. Sugawara

6:00pm: Adjourn for dinner on your own

**Friday, 9 January 2009:**

9:00am: Session 7. Interoperability and data curation: How can data QA/QC be improved while we build interoperability?

*This session will be a moderated discussion with emphasis on brainstorming and collective planning.*

10:30am: Coffee break

11:00am: Session 8. Enhancing connectivity with barcode data: Planning (division of labor, funding, timeline)

*This session will be a more structured moderated discussion with emphasis on producing action plans and timelines for deliverables.*

12:30pm: Lunch

2:00pm: Session 9. e-Biosphere: Background, online communities, conference and workshop agendas, plans for interoperability demonstration

*Two 10-15 minute presentations will provide background for open discussion. The emphasis will be on generating ideas for events that will take place during the e-Biosphere conference.*

- Program, Exhibits, Discussion Groups, Side-events: D. Schindel, CBOL
- Cybertaxonomy demonstration: D. Patterson and C. Parr, EOL; R. Pyle, ZooBank

3:30pm: Coffee break

4:00pm: Session 10. Planning and preparing for e-Biosphere 09

*This session will be a more structured moderated discussion with emphasis on producing action plans and timelines for the following deliverables:*

- Including barcode data in interoperability demonstration
- Database and software demonstrations
- Exhibits
- Poster presentations
- Breakout discussion group
- Side-event

5:00pm: Adjourn for departures