SSCE Position Paper: Data release for DNA barcoding

The SSCE strongly endorses the view that in promoting rapid data release, ***iBOL has the opportunity to drive taxonomy towards the collaborative open-access model that has proved so successful in other branches of science like genomics and physics, and to facilitate scientific enquiry in other areas of biology.***

The view among SSCE members is that iBOL should work towards immediate data release as the normal default practice, with rare exceptions based on specific good reasons. This position paper weighs alternative approaches and recommends specific options for consideration by the full SSC and iBOL's Board of Directors.

In the presence of a functioning DNA barcode, the endeavor of DNA barcoding rests on four cornerstones:

- Data quality standards for the DNA barcode sequences that makes them accurate, repeatable and reliable; standardized protocols that renders results comparable across laboratories;
- Traceability of data back to voucher specimens (for the taxonomic identification), electropherogram trace files (for sequence quality) and other metadata (for field data and lab procedures);
- A well identified reference set of samples to form the library, to provide a biological context for the sequences;
- Scalability of the approach to millions of specimens and hundreds of thousands of species. This requires more than large investments and economies of scale. It requires new ways of working that engage teams of researchers contributing their expertise and effort through networks and 'crowd-sourcing'.

The enormous scale achieved by the genomics research community required more than high-throughput sequencing facilities. It required a new way of working, starting with early release of raw data followed by annotation by anyone in the research community. The Wellcome Trust's 'Fort Lauderdale Principles' on data sharing for the International Human Genome Project described this new way of working and a code of conduct to guide members of the community:

- immediate release of data following basic quality control;
- crowd-sourcing of added value annotation; and
- implicit understanding of proprietary use by data producers.

The iBOL data release policy developed by Genome Canada was based on the Fort Lauderdale Principles but its application to barcode data has met with resistance. After six years of high-throughput data production:

- The BOLD (Barcode of Life Data Systems) public workbench contains more than 1.3 million records, but
- There are 514,000 public barcode records in GenBank, of which
- Approximately 100,000 bear traditional Linnean taxonomic names.[1]

This pattern suggests that while barcode researchers are willing to release data on voucher specimens and their barcode sequences and localities, they resist the release of their taxonomic identifications. This reflects a basic difference between genome and barcode records. Genomic data are just that -

[1] See blog by Rod Page on the recent emergence on "Dark taxa: GenBank in a post-taxonomic world"

sequence data from a few very well known model organisms produced by standardized protocols that either pass or don't pass standard tests of data quality. The taxonomic identity of these model organisms aren't in doubt.

In contrast, barcoding is done on the full range of species from the well-known to the rare and poorly described to those that are newly discovered through barcoding. The taxonomic identifications in barcode records are value-added judgments. These judgments result from the process of comparison between the barcode voucher specimen and taxon concepts described in the literature and based on reference collections. Immediate judgments may have low probabilities of being correct. The longer the process of comparison, the greater the added value. Under normal circumstances, rapid data release would be incompatible with highly reliable taxonomic identifications.

DNA samples tend to mainly come from either museums with generally reliable taxonomic identifications but degraded DNA, *or* from new field collections with intact DNA but only preliminary taxonomic identifications. That is, barcoding projects that seek high numbers of specimens can easily have either quality sequence data or reliable taxonomic identifications, but having both is more difficult. Many of the DNA barcode records in BOLD have not received adequate taxonomic verification. As a result, submitters of these records are hesitant to release them with taxonomic identifications that may not be accurate.

The challenge, then, is to create a community dynamic in which data submitters are willing to release their data rapidly with very preliminary identifications. At that point the community can assist in the process of adding value by making critical comparisons and providing feedback to the submitter who would improve the identification over time.

There are two additional factors which may also contribute to the resistance among barcode researchers to rapid data release:

1) *Academic protection.* The standard working model for taxonomy is to accumulate a substantial body of data which is then released at the point of publication. There is a cultural barrier to real-time release of data based on the fear of being 'scooped' by other researchers.
2) *Inertia.* A downside to the 'free sequencing' offer from iBOL's major core facility is that it can encourage researchers to request sequencing without giving much thought to their own subsequent downstream input. Data can be generated which then simply sit there because adding the value of a valid taxonomic identification is a low priority for the researcher concerned.

Together, these issues are preventing rapid release of DNA barcode data. This is a critical issue for iBOL. The success of the project depends on effective transfer of barcode-quality sequences into the public domain. iBOL is based on harnessing distributed efforts to build a shared resource. It will only work, if the data are shared. In addition, the project needs to demonstrate added-value, by creating a resource than has re-use value beyond the project's primary goals. Barcode data are useful for much more than species identification and delimitation, e.g., for phylogenetics, ecological forensics, conservation, and macro-analysis of biodiversity patterns.

The SSCE considered a range of models for possible iBOL data release policies and the benefits and costs associated with each (see Table 1). Two preferred options emerged from this process:

- ***Preferred Option 1*** *(a combination of options 3 and 5 on Table 1): All data are embargoed as private records in BOLD or other workbench databases for a pre-specified period (majority view was one year). Taxonomic identifications should be accompanied with a new metadata field that specifies confidence of the identification.* The embargo period would allow researchers to conduct critical comparisons with the literature, reference collections and other data resources to reduce the possibility of misidentification. The embargo period would also permit preparation of manuscripts for publication, thereby reducing the likelihood of unauthorized use by others before the submitter publishes. Full data release would be automatic at the end of the embargo period. This option would also involve creation of a metadata field as a confidence indicator on the identification. This would work most efficiently in the context of a feedback system, in which community members could use the data and provide feedback to the submitter on the taxonomic identification. Based on this crowd-sourced input, the submitter could correct errors in the identification and increase the confidence level in the metadata field. SSCE members who favored this option, felt that the time-limited embargo could be implemented for a fixed number of years during which the community would become acclimated to the new culture of rapid data release. The data release policy could then transition to the following preferred option.
- ***Preferred Option 2*** *(option 3 on Table 1): All data are released immediately following automated quality control, with a new metadata field* (described above) *indicating a confidence level ascribed to the taxonomic identification by the submitter.* This is very similar to the option described above, except with a more aggressive data release strategy which would deliver DNA barcode data into public domain immediately, enabling rapid community access to the data. It implements the data quality identification field outlined above, and the associated benefits associated with this.

In both cases, the SSCE stress the importance of:

(a) Future investment in bioinformatic frameworks to enable community annotation and feedback on biological identifications

(b) A clear statement in the spirit of the Fort Lauderdale Principles, encouraging researchers to publish Project Description papers that, in essence, stake a claim to an area of academic enquiry that will use their barcoding data, giving them academic protection after rapid public release of their data.

Additional points/notes

Assigning confidence levels to identifications. There are several issues that will determine how well a specimen is identified (Table 2). This makes it difficult to assign confidence to identifications in a standardized and straightforward fashion. One approach would be to create a controlled vocabulary that transmits the basis of the identification (e.g., expert taxonomist using reference collections; trained taxonomist using published revisions; non-specialist using field guide; non-specialist using identification key. Alternatively, the quality of identifications could be represented using summary categories (e.g., very confident, probably correct, tentative). Either approach to taxonomic identifications is parallel to the use of metadata for other biodiversity information. Latitude/longitude readings usually come with an error estimate. Geographic locations are qualified by the size of the error polygon. Gene sequences have trace files with quality scores for each base position. By putting a confidence qualifier on taxonomic identifications, researchers may be more willing to release their preliminary judgments and enable a community-based approach to data curation and collaborative research.

Recognition that there is no one-size-fits-all data release policy that will be acceptable to all iBOL partners. As iBOL expands, new core facilities and BOLD mirror databases will be launched and each will have data release policies that reflect the goals and concerns of their host institutions and funders. The SSCE encourages them to adhere to the principles presented here even though local conditions/concerns may vary. Thus while local adjustments may be necessary, the SSCE stress that the basic principle of public data release is a core component of the iBOL project, and that this philosophy should be reflected in any iBOL core-facilities' data-release strategy.
The identification of facilities and databases with significantly different data release policies as being 'iBOL participants' will need careful consideration.

Data release policy for barcode sequences in BOLD that are user submitted rather than being generated through iBOL support to CCDB. Sequencing done outside core facilities represents a potentially important contribution to the iBOL project. It is clearly desirable for such data to go into BOLD or other iBOL mirror sites. However, these submitters will not be bound by the data release policies adopted by core facilities since they did not derive the benefit of free sequencing from them. There is thus a risk of the 'free rider problem' (use of BOLD for data management and analysis without sharing data). One option is to allow a generous embargo period for private records in BOLD (e.g., five years), after which all data are automatically made public. This is long enough to not dissuade users from data submission and avoids creating a legacy of private data persisting in BOLD.

Potential for immediate data release to cause societal problems. There are some situations in which release of badly identified data could lead to major problems. One example is the establishment of trade-barriers due to mistaken concern about risks from (misidentified) pests/pathogens. The SSCE view is that this issue should be addressed via effective communication to regulatory agencies, and effective annotation of data (identification confidence), and in itself, should not be used as an argument against rapid data release.

TABLE 1: IBOL OPTIONS FOR DATA RELEASE			
Option	**Pros**	**Cons**	**Timing**
1. Immediate release of all data following automated quality control	Immediate access to data by the research community	Potential damage to DNA barcoding and iBOL 'brands' and criticisms leveled at data submitters for flooding GenBank with poor quality ID. Not dissimilar to Genbank at present, but scale of iBOL will rapidly increase number and diversity of mis-identified samples in Genbank.	Immediate full data
2. Immediate release of 'DRAFT' data following QC. This is the same as Option 1 but all records would be labeled 'DRAFT' in GenBank for a specified period~~s~~, after which the DRAFT annotation would be removed.	Immediate access to data, useful for search/blast type studies. Makes clear to community that ID checks have not been completed; provides some protection to IP of researchers, as harder for people to publish on these data at draft stage.	Requires second round updating on live records, would effectively restrict use of data until verification as it would be difficult to publish the unverified data (also viewed as plus from the point of view of the data owner). Would need to establish when the Barcode flag would be given (post verification seems most appropriate).	Immediate full data
3. Immediate release of all data following QC but with a new 'confidence estimate' metadata field associated with taxonomic ID. This is the same as Option 1 but a new field would be added to allow submitters to qualify the basis of/confidence in the taxonomic name, preferably using a standardized controlled vocabulary. The metadata would change as the submitter improves the ID.	Makes absolutely clear where researchers are worried about ID quality, potentially removes this barrier to data release. Could be used in combination with '2 draft release', or '5 Time-limited embargo'	Doesn't get at the issue of people wanting to have short term protection of their data, difficult to apply completely objectively and comparably (see Table 2 for key factors), would be most effective if linked to a community update option in which broader community efforts could be harnessed to improve ID quality	Immediate full data
4. Two-phase release (A): immediate partial data release and specified timing of full data release: Immediate release of most data with a time limited embargo on taxonomic identifications. High-level (ordinal?) name plus BINs would be used as the unique name-string for a determined period (one year?), after which submitters must release the best available taxonomic name.	Time limited hold on data, guarantees all data will be released in relatively short time period, gives reasonable time for checking of identifications, gets sequence and BIN data etc to public domain immediately	Lack of immediate release causes lag phase in development of shared resource. Most SSCE members questioned the value of Phase 1 data in GenBank, especially since taxonomic IDs for those records could be obtained from BOLD (in an inefficient way). Requires construction and maintenance of a more complex data management system that ensures second phase data release on time. Partial data release creates perception of information being 'deliberately withheld'. If embargo is short (e.g. 1 year), then benefits to the partial release of data are minimal, compared to the costs of administering a partial initial release, followed up by full release.	Short delay for full data
5. Time-limited full embargo: All data would remain private in BOLD for a determined period (one year?) following immediate automatic QC, after which all data must be made public on BOLD and GenBank.	Time limited hold on data, guarantees data will be released in relatively short time period, gives reasonable time for checking of identifications, reduces risk of being scooped.	Lack of immediate release causes lag phase in development of shared resource.	Short initial delay for full data but guaranteed release

6. Two-phase release (B): immediate partial data release and unspecified timing of full data release: immediate data release following QC of data except taxonomic IDs and other specified elements. Release of embargoed data occurs on publication or at discretion of submitter. (Current iBOL policy*)*	Gets at least some data to public domain rapidly. BINS are part of phase 1, and might do away with the need for names in some groups for some samples, given sufficient density of named records in database	With no time limit on phase 2, nothing is there to stop the data languishing in BOLD for years. In practice, not really different to Traditional Model.	Medium/ Long delay for full data
7. Full embargo until publication. This is the traditional model used by non-iBOL barcoding projects.	Ensures that all data are verified and have been through peer review, minimising likelihood of errors	May take years/decades before data are released. Reduces ability of community to benefit from shared resource. Misses the point of the barcoding initiative; iBOL would effectively fail if this option is adopted.	Long delay for full data

TABLE 2: FACTORS TO CONSIDER WHEN ASSESSING IDENTIFICATION QUALITY (Green = good, amber = medium, red = poor)						
Experience of identifier	**Access to reference material**	**Taxonomic maturity and complexity**	**Specimen quality**	**Identification method**	**Laboratory competence (lab mix ups)**	**Alignment of results from barcodes and different data types**
Expert taxonomist, deep knowledge of study group	Recent taxonomic revisions or monograph accessible; holotype and other reference collections have been examined	Straightforward well-studied group, general lack of ambiguity about species limits	High quality sample, all necessary parts present for identification, voucher available	Detailed character based analysis in comparison with well established reference set, involving formalized analytical routine	Close adherence to established best-laboratory practice;	Identifications based on barcodes and other data types (morphology, geography, ecology) are consistent and convincing based on adequate reference barcode records
Intermediate knowledge level (informed biologist working on known study group, or expert on regional flora/fauna)	Literature resources beyond field guides or keys not accessible; revisions or monographs exist but not recent; some reference collections examined but not type material	General agreement about species limits, but identifications potentially prone to errors due to hybridization, or some of the character differences being subtle	Imperfect sample, but still sufficient for determination with confidence, voucher available	Used clear cut diagnostic characters but not in an analytical framework	Generally well run laboratory facility, even if not adhering to formalized laboratory best-practice standards	Identifications based on barcodes and other data types are not inconsistent but lack convincing resolution due to lack of adequate reference barcode records or lack of clear species separation
Limited knowledge, non-specialist using third-party identification resources	Minimal reference material used	Taxonomically complex group, major uncertainty as to species limits/or dearth of taxonomic characters makes all determinations very difficult, even in expert hands	Poor quality sample, lacking important features needed for identification, and/or no voucher available	Visual inspection and opinion-based identity	Informal laboratory environment with poor documentation of sample tracking,	Identifications based on barcodes and other data types are inconsistent