Data Management Plan: Origins of Hawaiian Reef Fishes

In accordance with NSF philosophy and policy for the Dissemination and Sharing of Research Results, we are committed to open and persistent access to all data generated as part of this project. The primary product from this project will be DNA sequence data. In consultation with BCO-DMO, we will make this data available through GenBank and Indo-Pac Research Coordination Network, a NSF-funded NESCent project.

1. **Data Capture.** As part of the NSF-funded BiSciCol project (DBI-0956415), collaborator Richard Pyle and colleagues have developed software tools to allow highly efficient and accurate data capture in the field, including the assignment of persistent globally unique identifiers to all specimens and subsamples (i.e., individual organisms and material samples), occurrence instances, sampling events and associated locations, evidence records (e.g., videos, images, literature reports, observation records, etc.), taxonomic determination assertions, associated people in various rolls (collectors, taxonomic determiners, etc.), and other relevant metadata. All specimens are processed through a standard workflow beginning with collecting event metadata recording (including georeference coordinates) and individual specimen processing (including tissue subsampling).

2. **Data Products.** We will share programming code, data manipulation algorithms, data and metadata in a standards-compliant fashion. Data will be stored in relational database systems (such as MySQL and MS SQL Server) and will be made available through formats compatible with semantic web/Linked Open Data initiatives (RDF, RDFa, etc.). The data model includes robust support for cross-linking objects to external identifiers, mechanisms for managing “Meta-Authority”-asserted subjective taxonomies and classifications, annotations, and a variety of other related or supporting data. Extensive use of persistent, actionable Globally Unique Identifiers (GUIDs) will ensure access to metadata for core data objects.

3. **Data Pathways, Storage, and Availability.** All algorithms and programming code produced during this project will be publicly available. All DNA sequence data will be stored in our bioinformatics facilities at HIMB (see Facilities statement), archived in GenBank, with accession numbers forwarded to BCO-DMO. In GenBank we will batch submit the original haplotypes by specimens, so that haplotype frequency data and the full data set can be recovered. These data will also be placed in the archives of the Indo-Pac Research Coordination Network, a NSF-funded NESCent project spearheaded by Eric Crandall and Cynthia Riginos. All past projects have the same data accessibility.

4. **Data Standards Compliance.** All data management components of this proposal will conform to standards established by the Taxonomic Database Working Group (TDWG) , including the latest extensions and iterations of the DarwinCore and DarwinCore Archive standards.

5. **Persistence and Long-Term Preservation.** Data persistence is assured by incorporating data and services in replicates and mirrors on several different servers. Data will be curated for data-rot, standards compliance, and open access.

6. Documentation and Metadata. Metadata standards will comply with TDWG standards, such as Darwin Core Archive guidelines (http://rs.tdwg.org/dwc/terms/guides/text/index.htm), by using validators developed by GBIF (http://tools.gbif.org/dwca-validator). Metadata will be mostly generated automatically, using human manual input and corrections if necessary. Parsed metadata will be available through web user interfaces, shared files, and API requests. Ontologies and controlled vocabularies will be used for data curation, exchange, and publishing data. Thorough documentation is integral to all GitHub deposits and will be made available through the project website.

7. Data Product Dissemination. Data products and services will be disseminated to the scientific community through presentations in scientific meetings, publication of scientific results, and through various forms of social networking, as appropriate. Through direct affiliations and secondary propagation, data from this project will be disseminated to the following organizations and initiatives (among others): the Global Names Architecture (DBI-1062441), the BiSciCol project (DBI-0956415); the Catalog of Fishes (Eschmeyer 1998); Consortium for the Barcode of Life (CBOL); FishBase (Froese 2001); FishNet II; the Global Biodiversity Information Facility (GBIF); the Integrated Taxonomic Information System (ITIS); and various other biodiversity data initiatives, as appropriate.

8. Data Security. Security will be ensured by proper System Administration support and managed through best software coding practices, password protection, encrypted data transfer protocols, firewalls and intrusion detection systems.

9. Roles and Responsibilities. The PI (Bowen) will oversee data management implementation for all technical aspects of this project, with input from appropriate consultants and the project developers. System administration support will be responsible for backups and archival of the data.

10. Ownership, Copyright, IP, Licensing, Privacy and Confidentiality. Data content from this project contains only factual data, which are not subject to copyright laws. We implement the Principle of Open Access (http://mitpress.mit.edu/catalog/item/default.asp?tid=10611&ttype=2), with our default license being the Creative Commons Public Domain Dedication (CC-0; (http://creativecommons.org/publicdomain/zero/1.0/). Nevertheless, we will maintain an attribution trail for all data coming into the project environment. User account data that contain private information (e-mail address, user name) will be replicated but not revealed to third parties. Copyright of media generated through this project will be retained by the creators of such media.