Specimen Identifiers: Linking tissues, DNA samples, and sequence data to voucher specimens in publicly accessible databases

Daniel G. Mulcahy ‡

‡ Museum für Naturkunde, Berlin, Germany

Corresponding author: Daniel G. Mulcahy (mulcahyd@si.edu)

Abstract

Nearly all disciplines of biology now have some form of molecular genetic analyses incorporated into areas of their research, from systematics, ecology, and behavior, to physiology and conservation. In order for science to be transparent, the source and provenance of the genetic material used must be easily identifiable and traceable, following the FAIR principles of being Findable Accessible, Interoperable, and Reusable (Wilkinson et al. 2016). Natural history collections are ever-increasingly facilitating the use of genetic components from collection objects, and in some cases increasing the number and types of collection objects under their care (i.e., tissue/DNA-only, e.g., blood, feather, skin- fin-clips, environmental samples). Most natural history collections are now making their holdings available online, either on their own platforms or via aggregate search engines like the Global Biodiversity Information Facility (GBIF) and the Global Genome Biodiversity Network (GGBN).

Many natural history collections are also now using digital management systems, where digital identifiers such as Digital Object Identifiers (DOIs) and Uniform Resource Identifiers (URIs) are assigned to objects in collections (Güntsch et al. 2017), including multiple objects derived from the same individual organism (e.g., voucher specimen, images, and genomic samples). Associated objects may receive different digital identifiers in order to be uniquely identifiable in the digital management system, but sharing this information on third-party platforms (e.g., GBIF, GGBN) is challenging, especially in avoiding duplicate entries. Furthermore, genetic materials are often sought out by researchers external to the holding collection institution, and molecular sequence data are then generated and deposited in third-party public repositories, such as GenBank and the Barcode of Life Database (BOLD). Making genomic material digitally discoverable to researchers, and linking them and data generated from these samples back to the associated voucher specimens is another challenge. Fortunately, there is a current international initiative to collate the different forms of data surrounding a voucher specimen, as a Digital Extended Specimen (DES), across multiple institutions worldwide (Hardisty et al. 2022).

The National Center for Biotechnology Information (NCBI), which hosts GenBank, has created a BioCollections Database to curate metadata for natural history collections and linking sequence data to voucher specimens (Sharma et al. 2018). Institution codes, collection codes, and catalog numbers are linked to create a "structured voucher" annotation (following the Darwin Core Triplet) to standardize usage across interconnected databases (e.g., GenBank, European Nucleotide Archive, and the DNA Databank of Japan). However, duplicate institution codes can make this complicated, and collection institutions that use digital identifiers instead of traditional catalog numbers make this even more challenging.

The NCBI BioCollections Database curators have resolved the duplicate institution codes problem, by adding the three-letter country code (or state code, within the same country). However, this database is used only for sequence data, in GenBank, and related databases (e.g., ENA, DDBJ), which raises the question, is there a need for a more universal biocollection codes database? Additionally, as museums move towards using digital identifiers, in the place of catalog numbers, confusion can arise when multiple digital identifiers are assigned to parts of the same "specimen" (e.g., specimen voucher, tissue, DNA, images, etc.). For instance, if a given specimen has unique URIs for the voucher specimen, the DNA, and an image, a researcher borrowing the DNA, might use the DNA URI as an identifier for the genetic database. A different researcher, at a later date, might see that specimen (or image) in the museum's collection, and think it is a different specimen of that species, when in fact it is the same specimen. This could result in a second researcher borrowing the sample and publishing it as a "new" sequence. Researchers already have difficulties in submitting sequences to GenBank, as several have confused field numbers for catalog numbers from the National Museum of Natural History, Smithsonian Institution (Mulcahy et al. 2022).

Some museums, using modifiable codes, can append a primary code (from the specimen voucher) for additional "parts" of that specimen. For example, if the primary code for an insect specimen ends in "…6d15ce", a leg taken for DNA extraction could be modified as "…6d15ce_leg" and "…6d15ce_dna" for the extract. This minimizes the chances for mistaking these as being from different specimens. However, if completely different codes are assigned to different parts of the same specimen, the chance increases for mistaking two objects from the same specimen as being from different specimens.

Collections staff must carefully consider the hierarchical relationships of objects in their collections, and how they are assigned URIs, especially when considering long-term operability in current and future aggregate database structures (e.g., GBIF, GGBN, NCBI, and the DES).

In this presentation, these issues are raised and the difficulties in linking specimens, genomic resources, and associated data in aggregate databases and data repositories are discussed.

Keywords

institution code, collection code, catalog number, digital identifiers

Presenting author

Daniel G. Mulcahy

Presented at

TDWG 2022

Conflicts of interest

References

- Güntsch A, Hyam R, Hagedorn G, Chagnoux S, Röpert D, Casino A, Droege G, Glöckler F, Gödderz K, Groom Q, Hoffmann J, Holleman A, Kempa M, Koivula H, Marhold K, Nicolson N, Smith V, Triebel D (2017) Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. Database 2017 https://doi.org/10.1093/database/bax003
- Hardisty AR, Ellwood ER, Nelson G, Zimkus B, Buschbom J, Addink W, Rabeler RK, Bates J, Bentley A, Fortes JAB, Hansen S, Macklin JA, Mast AR, Miller JT, Monfils AK, Paul DL, Wallis E, Webster M (2022) Digital Extended Specimens: Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet. BioScience https://doi.org/10.1093/biosci/biac060
- Mulcahy D, Ibáñez R, Jaramillo C, Crawford A, Ray J, Gotte S, Jacobs J, Wynn A, Gonzalez-Porter G, McDiarmid R, Crombie R, Zug G, de Queiroz K (2022) DNA barcoding of the National Museum of Natural History reptile tissue holdings raises concerns about the use of natural history collections and the responsibilities of scientists in the molecular age. PLOS ONE 17 (3). https://doi.org/10.1371/journal.pone.0264930
- Sharma S, Ciufo S, Starchenko E, Darji D, Chlumsky L, Karsch-Mizrachi I, Schoch C (2018) The NCBI BioCollections Database. Database 2018: 1-8. https://doi.org/10.1093/database/bay006
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18