BIODIVERSITY COLLECTIONS NETWORK

EXTENDING U.S. BIODIVERSITY COLLECTIONS

TO PROMOTE RESEARCH AND EDUCATION

ON THE FRONT COVER. With more than 10 million specimens, the McGuire Center houses one of the world's largest Lepidoptera collections, representing most of the world's 20,000 butterfly species and many of the estimated 245,000 moth species. *Credit: Eric Zamora, Florida Museum of Natural History*

Contents

Verity Mathis, collection manager for mammalogy, holds a skull of an impala, *Aepyceros melampus*, in the mammals collection at the Florida Museum. *Credit: Kristen Grace, Florida Museum*

Summary

Our national heritage of approximately one billion biodiversity specimens includes such diverse objects as dinosaur bones, pressed plants, fish preserved in alcohol, pinned insects, eggshells, and microscopic pollen grains. Collectively these specimens form a powerful resource for exploring Earth's biota across taxonomic, temporal, and spatial scales.

There is an urgent need to **build a network of extended specimen data** that represents the depth and breadth of the more than one billion biodiversity specimens held in U.S. collections institutions. The Extended Specimen Network (ESN) would include the physical specimen and its associated physical and digital genetic, phenotypic, and environmental data. The network would rely on new data integration mechanisms necessary to link all of the dynamic components together. The ESN will help researchers understand the rules that govern how organisms grow, diversify, and interact with one another, and how environmental change and human activities may affect those rules. As a resource for formal and informal education (including citizen science), the ESN will provide scalable learning opportunities for K–12 and lifelong learning in data literacy as well as the life sciences and humanities.

The ESN will require long-term funding for a central organizing unit with responsibility for community coordination, education, mobilization, and maintenance of the central data repository and the network infrastructure. Also critical to the success of the ESN will be new infrastructure to provide the linkages between the specimen and derivative data, continued specimen digitization and data standardization, new collection protocols, new approaches to data sharing and collaboration, and enhanced partnerships with data providers, computer and data scientists, and educators.

ABOVE. A variety of specimens in an invertebrate zoology collection. C*redit: Kristen Grace, Florida Museum*

RIGHT. Field notes from Rachel Hackett's graduate research on prairie fen biodiversity. Specimens were deposited in the Central Michigan University Herbarium. *Credit: Peggy Brisbane, Central Michigan University*

Background

Investment by the U.S. federal government through the National Science Foundation's Advancing Digitization of Biodiversity Collections program (ADBC) has facilitated the digitization of approximately 62 million U.S. biodiversity specimens since 2011. The program has provided support to the University of Florida and Florida State University for hosting the Integrated Digitized Biocollections (iDigBio). IDigBio is the central coordinating unit for the digitization effort and hosts the national data repository. The final ADBC grants will be awarded in 2021. During the past year, the Biodiversity Collections Network (BCoN) has led an effort to gather input from primary stakeholder communities regarding future directions for collections and their use in research and education. The effort culminated in a workshop held 30 October – 1 November 2018 at Oak Spring Garden in Upperville, Virginia, during which a strategy was developed to maximize the value of collections for future research and education by building on and leveraging the accomplishments of ADBC.

THE EXTENDED SPECIMEN

Science and industry rely on physical specimens housed in U.S. biodiversity collections. Rapid advances in data generation and analysis have transformed understanding of biodiversity collections from singular physical specimens, to dynamic suites of interconnected resources enriched through study over time. The concept of the "extended specimen" conveys the current perspective of the biodiversity specimen as extending beyond the singular physical object, to potentially limitless additional physical preparations and digital resources.

Abigail Hollingsworth, Central Michigan University (CMU) undergraduate, digitizing plant specimens in the CMU Herbarium. *Credit: Peggy Brisbane, Central Michigan University*

Keith Curry-Pochy discusses snake patterns with a child during a snake-themed Tot Trot event in affiliation with the Discovery Zone at the Florida Museum. *Credit: Kristen Grace, Florida Museum*

The Extended Specimen Network

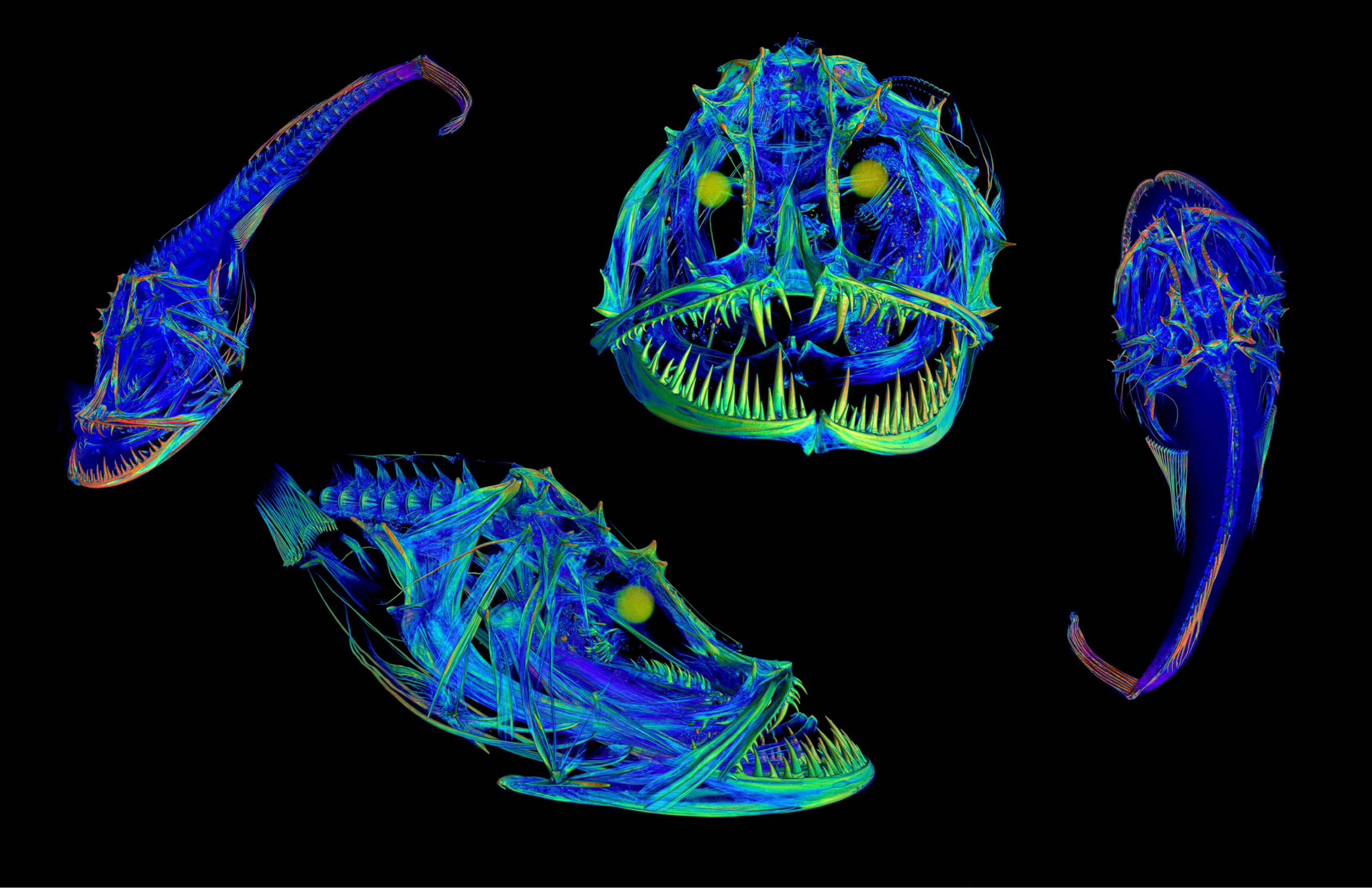
Building on the accomplishments of the past decade, the next step in advancing and enhancing collections infrastructure should focus on the production of extended specimens that represent the depth and breadth of biodiversity held in U.S. collections. Existing specimens are extended through digitization and linkages with associated data, including genetic, phenotypic, behavioral, and environmental. New specimens will be collected with these extended attributes in mind. Combined with and even driving data integration technologies and relevant data layers, extended specimens will form the core of a powerful Extended Specimen Network, or ESN.

ENABLING SCIENTIFIC DISCOVERY

The ESN will stimulate new avenues of investigation, expedite existing ones, and provide a more usable resource for making science-based policy decisions. By linking physical specimens to the data derived from them, such as gene sequences, images, behavior and species interactions, we can better define and understand the traits that comprise an organism. Such information has direct benefits to society and quality of human life, including how zoonotic diseases are transmitted and controlled, how crops can be more effectively and efficiently grown in changing climates, and how we can sustain and use biological resources in our ocean. The envisioned result is an integrated, global enterprise of specimen-based data that will also benefit non-biodiversity-related research and commercial enterprises today and into the future.

The ESN will facilitate the as yet incomplete work of documenting and naming the organisms that make up Earth's biodiversity. Machine learning and other innovative data science and engineering techniques can use the full range of data available about known species to find hidden novelties among existing and newly collected specimens.

Current data portals for digitized specimen data have user interfaces designed primarily for use by taxonomists and collections professionals. To take full advantage of the richer data content and broader relevance of the ESN, we must develop interfaces for a more diverse user base. With existing interfaces, a user can retrieve a set of relevant occurrence records from a search based on taxonomic name or a geographic unit. However, the ESN must provide an interface that will allow users to ask such questions as: Do the available data comprise a representative set, or are critical data lacking because key specimens have not been collected and/or digitized? How many different species, as opposed to different species names, occur in a given region? Do the organisms of the region occur in populations that are genetically distinct from other populations of the same species? Have unique interactions among organisms been documented in the region of interest?



High-resolution computed tomography reconstruction of an Angler, *Lophius piscatorius* (UF 118531). *Credit: Zachary Randall, FLMNH*

ENABLING SEAMLESS DATA INTEGRATION, ATTRIBUTION, AND USE TRACKING

The system of identifiers and specimen tracking protocols needed to implement the ESN will enable the linking of extended specimen objects, elucidation of relationships between items in disparate collections (e.g., tissue:voucher, plant:pollinator, host:parasite, and facilitation of interoperability with data sources) outside of our immediate realm. It will also allow collections institutions to identify the use of their specimens and develop metrics for measuring their impact. Such metrics permit collections to demonstrate their value and better acquire and manage resources. Currently, collections lack access to the data that demonstrate their full contributions to specimen-based research through citation in publication, vouchering of Genbank sequences, or products created from direct specimen use (e.g., images, CT scans). A reliable mechanism of tracking such use must be built.

Improved specimen tracking will create the potential for cost recovery when specimens are used in commercial enterprise. The ability to work with novel research (e.g., pharmacology, human health, food security) and commercial communities (e.g., pharmaceuticals, agriculture) will demonstrate the value of collections outside the immediate stakeholder community and contribute to the increased sustainability of biodiversity collections.

A comprehensive data tracking system will also allow biodiversity collections to meet new requirements for documenting the use of specimens and data derived from them. Such a requirement is the Nagoya Protocol, which is a supplementary agreement to the Convention of Biological Diversity (CBD) that establishes an international legal framework for access and benefit sharing of genetic resources. It requires that countries providing specimens define their access procedure and requires users to report on and share benefits upon their use.

Foundation for the Extended Specimen Network

COMPLETE AND IMPROVE EXISTING DIGITIZED DATA

A significant proportion of the existing digitized specimen records are incomplete. Some records, having been entered as "skeletal" records, lack important data fields (e.g., locality, date, collector) while geocoordinates are lacking for most. These data gaps limit the utility of these data, and thus data records must be completed and standardized to maximize their value in the ESN. Development of computational tools that can help fill in, or at least infer, missing values must be a high priority, perhaps taking advantage of a combination of image analysis (including optical character recognition) and data pattern matching.

FILL GAPS IN BIODIVERSITY DATA

Biological collections comprise the most comprehensive record of life on Earth, yet their potential cannot be fully realized until the data contained within them are revealed and made computationally accessible. Many of the approximately 1,469 U.S. collections are not yet digitized, and some do not have accurate estimates of the size or taxonomic content of their holdings. We must take stock of the holdings in U.S. collections by characterizing their holdings in terms of taxonomic, temporal, and geographic emphases, and compile these data into a national collections index. The existing *Index Herbariorum*, an index to plant collections, is a model. Such a reference enables prioritization of collections for digitization to create an effective national resource.

Biodiversity specimens are a resource for documenting environmental change, and researchers must continue to collect them into the future. New collections are especially important in areas that are undergoing rapid change, such as the Arctic. A more holistic, next-generation approach to the collection of new biodiversity specimens is needed. As Schindel and Cook (2018) note, a next generation approach could focus on nested sampling that extends beyond the single organism (e.g., a single plant) to its biotic associates (e.g., soil microbes, epiphytes, endophytes, and parasites spanning from viruses to insects and fungi) and its environment (e.g., community composition, microclimate, macroclimate, habitat quality). The downstream integrative linkages between these nested samples will open up new and dynamic research opportunities using these collections.

Distinguished scientist, Pamela Soltis, works on the Tree of Life – a collaborative effort of biologists and nature enthusiasts from around the world to understand biological diversity. *Credit: Kristen Grace, Florida Museum*

BUILD AND STRENGTHEN STRATEGIC PARTNERSHIPS

The ESN will require expanding existing partnerships and creating new strategic engagements.

Computer and data science research communities. The ESN requires strong interdisciplinary development between biology, biological collections, and computer and data sciences communities in order to build and maintain next generation collections infrastructure and accessibility.

Aggregators of related data. The data needed to extend biodiversity collections are held in databases such as the Catalogue of Life checklist, the Biodiversity Heritage Library (BHL), the Barcode of Life Data System (BOLD), National Center for Biotechnology Information (NCBI), and the Encyclopedia of Life Traitbank. Creating an effective two-way means of data exchange with these resources is required. Collaboration with programs such as the NSF-funded National Ecological Observatory Network (NEON), Long Term Ecological Research Network (LTER), and Critical Zone Observatories (CZO) will ensure that standards and protocols enable interoperability between collections data and historical and current occurrence records. Further, standards for the extended specimen data that are driven by the ESN will make the data collected by future researchers using these centers more computable.

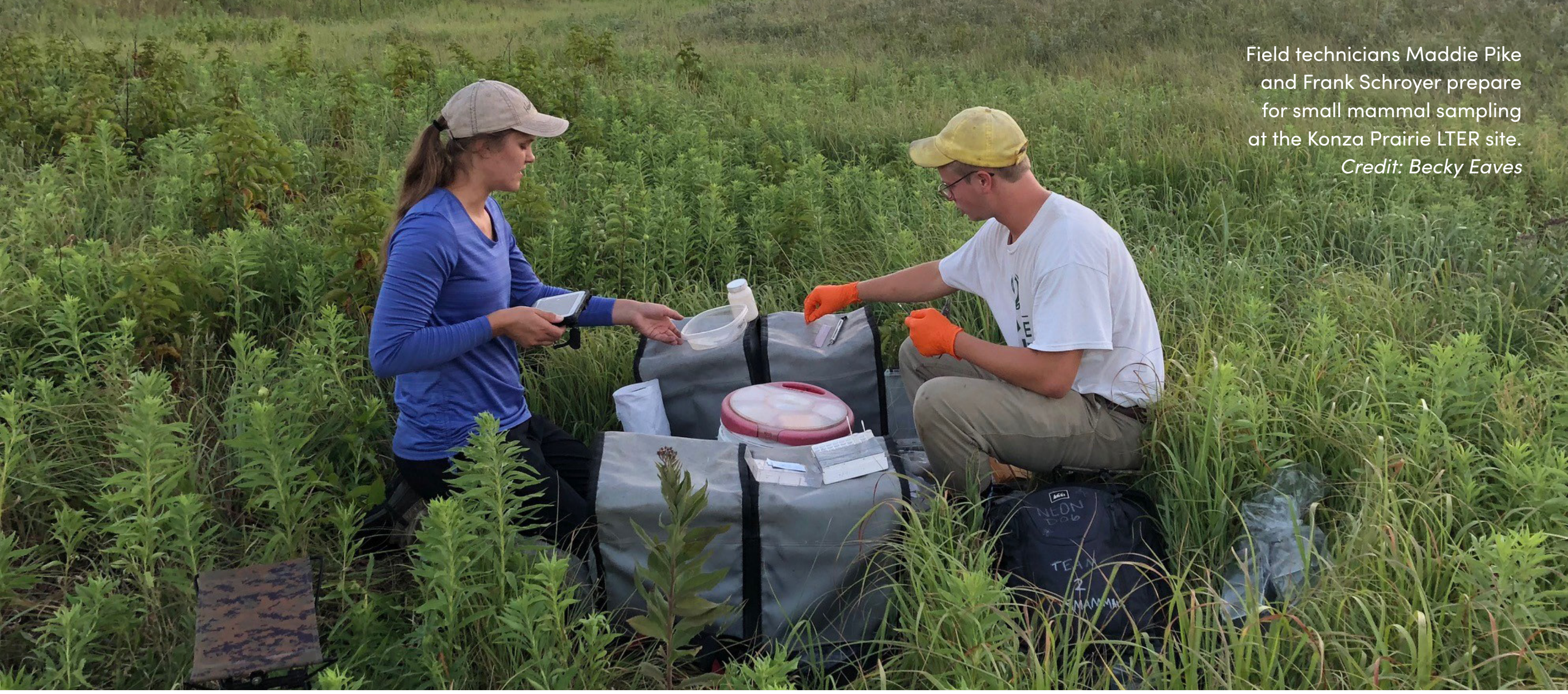
International Biodiversity Organizations. Participation to the fullest extent possible in the Global Biodiversity Information Facility (GBIF) proposed alliance for biodiversity knowledge will facilitate local work and help align efforts to observe and measure U.S. biodiversity in relation to other global efforts, including the Atlas of Living Australia (ALA) and the Distributed System of Scientific Collections (DiSSCo), a new European Union program. At a local scale, we must pursue collaboration with data aggregators in Mexico (CONABIO) and Canada (Canadensys) to permit the seamless transfer of data needed for continental-scale understanding of the breadth of biodiversity and its distribution and change over time.

Education and Broadening Participation. Working with national educators organizations as well as groups, such as the Society for Advancing Chicanos and Native Americans in Science (SACNAS), Minority Serving Institutions (MSIs), Historically Black Colleges and Universities (HCBUs), and other higher learning institutions, will promote the development of a more diverse scientific workforce. Partnerships among robust and fledgling citizen science initiatives will promote projects that can be widely implemented with compatible data standards and protocols.

Other Partnerships. An openness to collaboration with non-traditional partners in academia and industry could lead to wider application of pertinent technology. As the breadth of users and the value and diversity of products generated from digitized collections data grows, so too will support for investments in biodiversity collections.

ABOVE. Lichen students at Highlands Biological Station, North Carolina. *Credit: James Lendemer, NY Botanical Garden*

RIGHT. Closeup photo of *Tetracha virginica* collected from Kansas field site by NEON staff. *Credit: Eddie Minnick*



Field technicians Maddie Pike and Frank Schroyer prepare for small mammal sampling at the Konza Prairie LTER site. *Credit: Becky Eaves*

Empowering 21st Century Learners through Biodiversity Knowledge

FORMAL EDUCATION

Through education, the ESN and collections community has the largest potential to engage, educate, and empower the next generation of biodiversity data stewards, biodiversity researchers, and ESN users. Biodiversity data and ESN usage aligns naturally with core content common to all K–12 science curricula and the undergraduate biology core, and can be included seamlessly in evolution, biodiversity, systematics, taxonomy, and ecology subject matter. Specimen-based data make science accessible through the specimen itself, which is tangible, place-based, and engaging, as well as through aggregated specimen data that are verifiable, relevant, and a logical gateway to data literacy (Monfils et al., 2017). The place-based capacity of collections data combined with the social and societal relevance of biodiversity science can play a role in creating inclusive, culturally relevant, and socially conscious educational materials that engage a broad and diverse audience in biodiversity science. By defining and infusing biodiversity data literacy skills, creating a learning progression that incorporates ESN data and biodiversity data literacy into formal education, and providing accessible materials, teacher training, and educator interfaces that facilitate incorporating ESN data in the classroom, we support a biodiversity literate society and the training of the next generation of data literate scientists.

INFORMAL EDUCATION

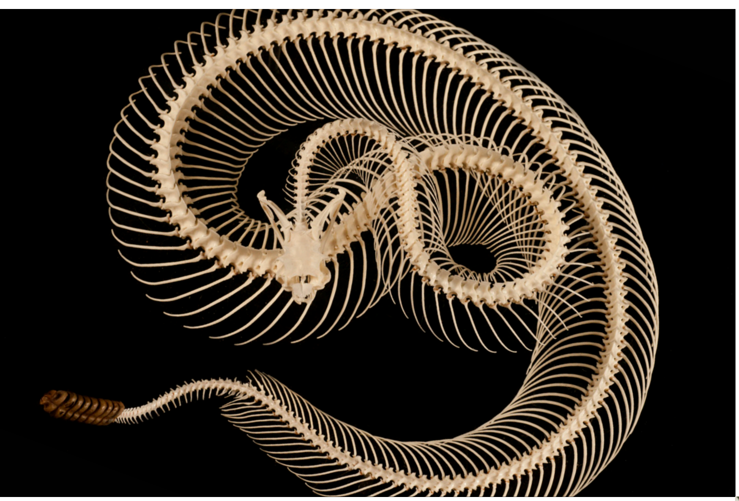
As our digital resources continue to expand, so too will informal education opportunities. Indeed, we probably cannot realize the full expression of the ESN without the strong involvement of the citizen science community.

A number of citizen science projects are based on monitoring biodiversity. Ebird, ebutterfly, iNaturalist, and the U.S. Phenology Network, for example, provide platforms for minimally trained individuals to contribute sightings or recordings of organisms or a particular attribute of an organism. Internet-based projects can involve the public directly in contributing to collections-based science and databases. Projects such as Notes from Nature, Smithsonian Transcription Center, and CitSciScribe are platforms that invite the online public to add digital data to images of specimens. Tasks range from transcription, to morphological measurements, to phenological annotation, and the majority relate directly to an active research project. Such programs engage participants from a wide range of ages, abilities, and interests, and with minimal startup costs. This approach creates an inclusive and diverse group of individuals working to advance biodiversity science.

Implementing and Sustaining the New Agenda

Key to the success of the ESN is a securely funded, long-term coordinating unit that builds upon the current model of the digitization iDigBio hub, but that has expanded functionality. The ESN hub would manage the national specimen data repository and network infrastructure and, in partnership with the primary stakeholder community, carry out the activities needed to implement the ESN described in this document:

- Develop a robust, comprehensive specimen identifier system in collaboration with other international data aggregators and providers.
- Work with data creators and providers to extend biodiversity specimens to create stable data linkages.
- Develop new user interfaces that take advantage of the power of the ESN and serve a broader range of potential users of these data for education, research, and conservation.
- Create an authoritative, comprehensive, and self-updateable index of U.S. collections institutions with structured metadata to describe their holdings to expedite discovery of critical undigitized collections.
- Continue digitization of existing material focused on underrepresented taxa (e.g., insects, fossils), including specimens held in small regional, personal, and individual researcher collections.
- Improve previously digitized specimen data by imaging specimens, completing skeletal records, and augmenting data with georeferencing.
- Develop new protocols for the collection and accession of data-rich samples that provide greater context for understanding the biotic and abiotic interactions of organisms and create comprehensive data sets for research and education.
- Champion broad-scale adoption of core biodiversity data literacy skills and competencies in K–12 and undergraduate curricula.
- Support enhanced training of emerging and established professionals for interdisciplinary work in biodiversity, data science, and informatics.



ABOVE. An articulated eastern diamondback rattlesnake skeleton (*Crotalus adamanteus*). *Credit: Kristen Grace, Florida Museum*

RIGHT. Aldo Rincon and Alex Whiting move fossil jackets into a prep lab. *Credit: Kristen Grace, Florida Museum*

Larry Page, curator of ichthyology, demonstrates the use of a photo box. Using the Photo Box Plus and a copy stand, specimens are photographed and cataloged digitally and added to an online database for scientists and researchers around the world to search, allowing better access to collections and reducing the need to loan and ship delicate items. *Credit: Kristen Grace, Florida Museum*

References

Monfils, A.K., K.E. Powers, C.J. Marshall, C.T. Martine, J.F. Smith, and L.A. Prather. 2017. Natural history collections: Teaching about biodiversity across time, space, and digital platforms. *Southeastern Naturalist* 16(sp10), https://doi.org/10.1656/058.016.0sp1008.

Schindel, D.E., and J.A. Cook. 2018. The next generation of natural history collections. *PLoS Biology* 16(7):e2006125, https://doi.org/10.1371/journal.pbio.2006125.

Webster, M.S. 2017. The extended specimen. Pp. 1–9 in *The Extended Specimen: Emerging Frontiers in Collections-based Ornithological Research*. Studies in Avian Biology (no. 50), M.S. Webster, ed., CRC Press, Boca Raton, FL.

Preferred Citation

Biodiversity Collections Network. 2019. *Extending U.S. Biodiversity Collections to Promote Research and Education*. American Institute of Biological Sciences, Washington, DC, 8 pp.

This material is based upon work supported by the National Science Foundation under Grant No. DBI-1441785. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

BIODIVERSITY
COLLECTIONS NETWORK

https://bcon.aibs.org

Biodiversity Collections Network
c/o American Institute of Biological Sciences
1201 New York Avenue, NW, Suite 420
Washington, DC 20005

Graduate student Jordan Hoffman sampling reindeer lichen (*Cladonia*) mats on Whiterock Mountain in Nantahala National Forest, North Carolina, as part of population genomics research. *Credit: James Lendemer, NY Botanical Garden*

March 2019