Extending U.S. Biodiversity Collections to Promote Research and Education

Revised Draft, 15 February 2019

□

I. Executive Summary

Our national heritage of approximately one billion biodiversity specimens, once digitized, can be linked to emerging digital data sources to form an information-rich network for exploring earth's biota across taxonomic, temporal and spatial scales. A workshop held 30 October - 1 November 2018 at Oak Spring Garden in Upperville, VA

under the leadership of the Biodiversity Collections Network (BCoN) developed a strategy for the next decade to maximize the value of our collections resource for research and education. In their deliberations, participants drew heavily on recent literature as well as surveys, and meetings and workshops held over the past year with the primary stakeholder community of collections professionals, researchers, and educators.

Arising from these deliberations is a vision to focus future biodiversity infrastructure and digital resources on **building a network of extended specimen data** that encompasses the depth and breadth of biodiversity specimens and data held in U.S. collections institutions. The extended specimen network (ESN) includes the physical voucher specimen curated and housed in a collection and its associated genetic, phenotypic and environmental data (both physical and digital). These core data types, selected because they are key to answering driving research questions, include physical preparations such as tissue samples and their derivative products such as gene sequences or metagenomes, digitized media and annotations, and taxon- or locality-specific data such as occurrence observations, phylogenies and species distributions. Existing voucher specimens will be extended both manually and through new automated methods, and data will be linked through unique identifiers, taxon name and location across collections, across disciplines and to outside sources of data. As we continue our documentation of earth's biota, new collections will be enhanced from the outset, i.e., accessioned with a full suite of data. We envision the ESN proposed here will be the gold standard for the structured cloud of integrated data associated with all vouchered specimens. These permanent specimen vouchers, in which genotypes and phenotypes link to a particular environment in time and space, comprise an irreplaceable resource for the millenia.

Collectively, data linked through the ESN will enhance the capacity to explore research questions across taxonomic, temporal and spatial scales. The ESN will allow researchers to explore the rules that govern how organisms, grow, diversify and interact, and enable scientists to ask more nuanced research questions specific to how environmental change and human activities may affect those rules. The engaging vouchered specimen, coupled with the open access ESN, and immediate and relevant science resulting from the ESN, can play a unique role in promoting STEM education, engaging citizen scientists, and empowering a scientifically literate society. The specimen and the associated data provide a relatable and engaging entry point to participate in iterative data driven science, learn core data literacy skills, and build open, transdisciplinary collaboration.

Creating the ESN requires new infrastructure to provide the linkages between the specimen and data derived from it. On the established foundation of existing digital data from collections it will require the development of new standards, connections, and resources such as ontologies to facilitate discovery, and implementation of a robust specimen identifier tracking system. Finally, continued digitization of established, as well as new collections, is necessary to ensure the grounding of extended specimen data in the framework of when and where it was collected. The ESN will also require new approaches to data sharing and collaboration, partnerships with national and international data providers, computer and data scientists, and educators. These will enable new relationships with investors who may exploit these data for commercial interests, thereby advancing the interests of national health, prosperity and welfare.

The ESN will benefit from research-driven episodic funding for the collection of new specimens, which in turn will require digitization and curation. For the ESN to function as envisaged above, it may require long-term support for a central organizing unit with responsibility for community coordination, education and outreach, data mobilization, and maintenance of the central data repository and the network infrastructure.

II. Background

In 2010, the U.S. National Science Foundation (NSF) convened several meetings of researchers and collections professionals to consider a national plan for the digitization of biodiversity collections. The result was the Network Integrated Biocollections Alliance (NIBA) strategic plan, which laid out the need for the digitization of collections and proposed a structure for the effort. The plan called for the creation of thematic collection networks that would digitize specimens to create a dataset for addressing defined research questions, as well as for a central hub that would organize the effort and provide training and support for the networks. The NSF responded to this NIBA plan by creating the Advancing Digitization of Biodiversity Collections (ADBC) program, a ten year, $100 million commitment that has made annual awards since 2011. Several years later, representatives of the biodiversity collections community crafted the NIBA Implementation plan. The Biodiversity Collections Network (BCoN), a five year Research Coordination Network (RCN) NSF award (DBI-1441785), was funded in 2014 to bring members of the community together to address collections needs that fell outside the scope of ADBC and its strict digitization mandate.

Since the release of the NIBA plan, the collections community has made transformational progress in the digitization of specimens and sharing of specimen data. The ADBC program has funded 23 Thematic Collection Networks (TCNs) and 29 Partners to Existing Networks (PENs) to add additional collections to the network. To date, the TCNs and associated PENs have collectively digitized 62 million specimens from 915 collections held in 317 institutions and have provided training and work experience for thousands of students and emerging professionals. The ADBC program has provided continual support to University of Florida and Florida State University for hosting Integrated Digitized BioCollections (iDigBio; DBI-1115210; DBI-1547229), the central coordinating unit for the digitization effort. iDigBio provides training in digitization and data mobilization to participating institutions, and shares digitized data through its iDigBio Portal, which provides access to more than 115 million specimen records and 27 million associated media records from 1,576 datasets. iDigBio has also carried out a vigorous outreach program, with symposia, webinars and workshops on digitization methods, data manipulation and sharing, and use of digitized data in research. It has also developed and disseminated a vast array of best practices and standards associated with digitization through its website. There is a growing body of research utilizing digitized data served through iDigBio and other ADBC-sponsored portals, numbering more than 500 citations in 2018 (see https://www.idigbio.org/research).

The end of the ADBC funding program is on the horizon. Grants funded in 2021, the last year of competition, will conclude in 2023 or 2024, and iDigBio will continue to function through the end of the last ADBC grant. The processes, tools, and protocols developed through this initiative provide a robust foundation from which innovative new research fields may be cultivated and solutions to national problems identified.

This report summarizes the results of a workshop held 30 October - 1 November 2018 at Oak Spring Garden in Upperville, VA. BCoN initiated the workshop, which was co-hosted by the Oak Spring Garden Foundation. The goal of the workshop was to identify a national strategy or agenda for the next phase of the effort to deploy the data held in U.S. natural history collections for research, policy and education. A full list of workshop participants, a summary of progress toward NIBA goals, plus feedback from the collections community and researchers and educators with their work, and list of recent lare included as Appendices A--D.

III. The Extended Specimen Network (ESN)

Building on the accomplishments of the past decade, we propose to transform our rich heritage of collections and associated data into a powerful new source of knowledge to address national priorities. This resource will have transformative potential to address long-standing questions about the breadth and complexity of biodiversity and provide for monitoring and research far into the future. Central to this agenda is the creation of extended specimens that capitalize on the depth and breadth of biodiversity held and digitally accessible in U.S. collections.

The concept of the extended specimen, introduced by Webster (2017), elevates and expands the physical specimen with an augmented digitized specimen record by associating genotypic, phenotypic and environmental data types. These may include media collected in the field along with the specimens (images, sound and video recordings, field notes, etc.), computable and semantically rich descriptive content, and internet-scale connected data resources that support discoverability of this this networked content in novel ways at multiple scales.

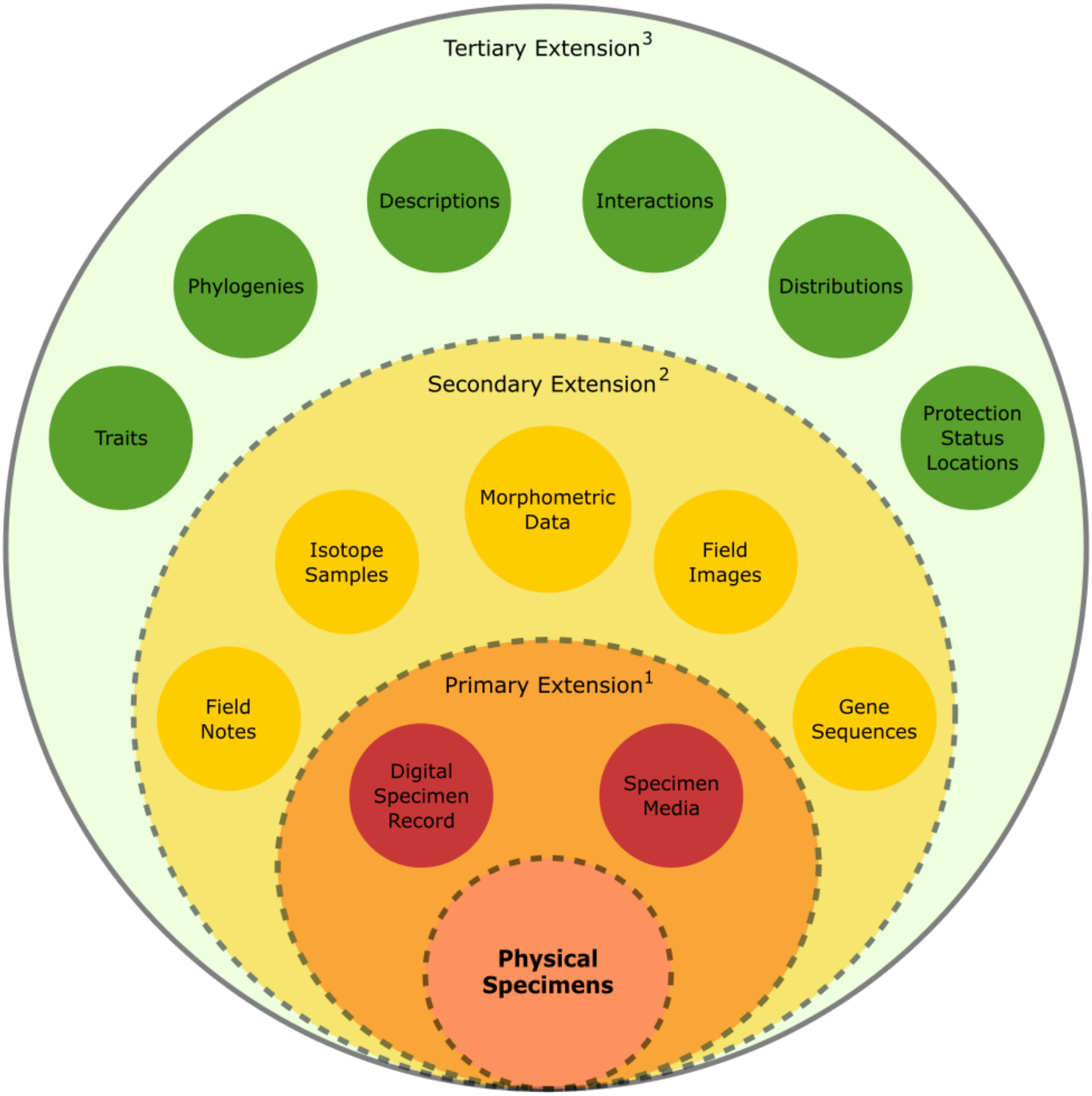

Figure 1. Components of the extended specimen. Physical Specimens. Approximately 1 billion physical specimens reside in U.S. museums, including derivative collections such as tissue samples. [1]Primary Extension. The digitized version of the specimen metadata in the form of a digital specimen record, as well as data and images, audio and video files are held by collections that hold physical specimens and also contribute to the national specimen database. [2]Secondary Extension.Data that reside in private collections, institutional repositories, and national/international repositories, often disconnected physically and digitally from the original specimen, which may link to individual specimens. [3]Tertiary Extension. Data from other repositories that link to taxonomic names applied to specimens or collection event (including collector, collections location)

Enabling Scientific Discovery

Specimen digitization provides critical information on species occurrence and co-occurrence over space and time, enhances species discovery, and provides a means to scale up and reduce uncertainty when examining overall patterns of diversity. Recognizing the value of digitized specimen data, members of the stakeholder community identified **increasing its use in research** as the top priority for the next phase of data mobilization from collections.

A significant body of new research based on data from collections has emerged in the past decade including, most prominently, research that involves genetic data. One of the most frequent requests to collections is for tissue from particular species for the purpose of extracting genetic data. Tissue holdings, however, are not digitized and publicly available outside of a handful of institutions, and the derivative genetic data, though usually deposited in the National Center for Biotechnological information (NCBI, or Genbank) are not consistently linked back to the digitized specimen collection data. Funding for the ESN would lift these constraints, enabling repurposing and linking back of new (and legacy) genomic data. The value of these genomic data relate to climatology, genomics, human health and phenology and trait characterization.

Similarly, biodiversity efforts involving the collection and standardization of phenotypic/trait data (e.g., Encyclopedia of Life, oVert, Phenoscape), particularly when linked back to the specimen through the ESN, will provide a rich resource for fundamental and applied sciences. Finally, constructing the vast tree of life, only partially constructed through efforts to date (e.g., OtoL, ATol, AVaTol, GoLife, etc.) will be enabled through these deep and critical linked specimen data. Using the FAIR guiding principles to make these specimen-centered collections data Findable, Accessible, Interoperable and Reusable (FAIR), the ESN will facilitate knowledge discovery (Wilkinson et al, 2016).

Combined with and even driving data integration technologies and relevant data layers, extended specimens will become part of a powerful data network. By anchoring this network in the physical specimen, which places an organism at a particular place and time, we can link data spanning the ecosystem to the genome and microbiome to investigate systems that structure biodiversity. Specimen and habitat images, video, audio recordings, and other media connected to physical specimens can be studied in concert and reveal undiscovered traits, undetected mutualistic or parasitic associations, and previously undetected phenotypic variation. When applied to museum specimens, techniques such as CT scanning allow researchers to view internal structures in 3D, revealing previously hidden suites of characteristics, which can enable novel comparisons across diverse groups of organisms.

Perhaps the most exciting new research area to which extended specimens can contribute is the characterization of the traits that compose an organism, i.e., its structures, physiological processes, behavior, and interactions with other organisms (e.g., parasites and symbionts). A better understanding of traits has direct benefits to society and quality of human life, informing 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 oceans. We need to develop knowledge graphs (ontologies) to make trait data gathered from specimens interoperable, which requires new infrastructure and computer science expertise (Kissling et al. 2018). The envisioned result will be an integrated global network of specimen-based data that can be accessed and applied to non-biodiversity related research and commercial enterprises. The ESN will foster new avenues of investigation, expedite existing research efforts, and provide an enriched resource for making science-based policy decisions.

A fully developed ESN requires scientists to complete the work of documenting and naming the organisms that make up earth's biodiversity. We will achieve this documentation by new exploration and further documentation of biodiversity in tandem with mobilization of specimen data in existing collections. Estimates suggest that 75% of undescribed species in the world are already represented in collections, but are misidentified or languish in an unidentified backlog (Bebber, 2010, Kemp, 2015). Incorporation of machine learning and other data science and engineering techniques, including pattern-matching techniques that can use the full range of data available about specimens of known species, can help to find hidden novelties from both images and specimen data records, including geospatial and genomics-linked resources (McAllister et al., 2018).

Current portals for digitized specimen data have user interfaces designed primarily for taxonomists and collections professionals. To effectively deliver the enriched ESN content to a broader audience, 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: Does 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?

Enable Seamless Data Integration, Attribution and Use Tracking

The key to assembling extended specimens and linking them into a robust data network is the standardization of existing digital data and implementation of a universal specimen identifier tracking system. Data integration is required at multiple levels - at the specimen level to link individual preparations or derived products (e.g., vouchers, tissues, gut contents, media, etc.), at the database level to integrate interactions (e.g., predator:prey, host:parasite, plant:pollinator, etc.) and beyond our own collections to link other sources of relevant data (e.g., observation data, environmental data, geographic data, ecological data, satellite data).

New technologies, such as blockchain (https://en.wikipedia.org/wiki/Blockchain, Halamka et al., 2017, Hilsberg et al., 2018), may provide viable solutions for how data and products can be stored, linked and shared with collaborators and stakeholders, facilitating transparency and traceability while maintaining data integrity, standardized auditing, and formalized contracts for data access. This decentralized approach to data exchange has already gained significant traction in the corporate, academic, and federal arenas. Initial discussions on how museums might use blockchain include preserving data in a low-cost manner while enhancing the reachability of those data to students, researchers, teachers, etc. and as a tool for enhancing collections management (Campbell, 2017, Lemle, 2018).

Attribution measures are an important tool for collections to track usage, demonstrate value, and advocate for collections management resources (e.g., staff, funding,). Collections are generally unable to showcase their full contributions to specimen-based research through citation in publications, vouchering of Genbank sequences, products created from direct specimen use (images, CT scans etc.) or use of data downloads from data aggregators due to limited connectivity between the collections and research, aggregator and publication communities. The overarching reason for this is the lack of reliable tracking mechanisms of such use. There is also a need for attribution metrics for the individuals involved in data curation and collections management activities (collectors, catalogers, determiners, georeferencers, data users and augmenters) to allow for individual acknowledgement and assessment of the work done in providing or augmenting digitized records.

The proliferation of data aggregators makes it increasingly difficult for collections to track use of their data effectively by the increasingly varied end user communities now using biodiversity data. This duplication of mechanisms for publishing records and data manipulation tasks has created an environment of uncertainty and incompatible versions of individual records for the data providers while providing ease for users. The necessary data integration and attribution infrastructure requires both technological and social solutions -- technological solutions to facilitate the linking of disparate data elements, and social advocacy to ensure that all those involved in creating and using the data pipeline are aware of the need and benefits of repatriating products of use back to collections. The Global Biodiversity Information Facility (GBIF) is attempting to address issues of attribution and integration by proposing that the major aggregators work towards unifying their disparate data caches and data tasks operating on that data stream (taxonomic authority, geographic authority, Darwin Core standards, etc.) to simplify the data aggregator landscape and bring about data consistency (Hobern et al., 2019).

A consequence of the greater discoverability of biodiversity collections through digitization is new requirements for the acquisition and use of these data. The Nagoya Protocol is a supplementary agreement to the Convention of Biological Diversity (CBD) that establishes an international legal framework for access and benefit sharing (ABS) of genetic resources. It requires that countries providing specimens define their access procedures, calls for users to share benefits upon their use, and establishes an international framework to ensure compliance with domestic access and benefit-sharing laws. By necessity, this new regulatory landscape will affect how biodiversity collections are managed and used. Because many of these issues are beyond the traditional scope and expertise of the current biological collections community, it will require an interdisciplinary approach to the development of technological solutions and training.

Improved specimen tracking will not only facilitate global compliance with the emerging policy and legal issues such as the Nagoya Protocol, but it will also support better metrics for data use and create the potential for commercial use cost-recovery where necessary. The ability to engage 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 sustainability of biodiversity collections. A mechanism for how this data integration and attribution architecture could be achieved is presented in the report from a BCoN workshop held on this topic in February 2018 in Lawrence, KS (http://bcon.aibs.org/wp-content/uploads/2018/05/BCoN-Needs-Assessment-workshop-report-1.pdf).

IV. Foundation for the Extended Specimen Network

Complete and Improve Existing Digitized Data

The digitization effort during the past decade on the part of U.S. collection holding institutions has been truly remarkable, and the value of this resource has wide-ranging implications for research, education, data- and collections management, and commercial interests. However, a large proportion of these newly digitized specimen records can be improved and augmented. Some lack locality, date and collector information, while geocoordinates have not been provided for most. The incomplete transcription and georeferencing of specimens inhibit the comprehensive use of these data, and we must complete this work in order to maximize their value in the ESN. Development of computational tools, perhaps taking advantage of a combination of image analysis (including optical character recognition) and data pattern matching methods that can help complete, or at least infer, missing values, must be a high priority.

The stakeholder community considers the inability to query databases reliably using key metadata criteria to be one of the greatest obstacles to greater use of specimen data in research and education (Appendix C). Because the use of specimens and associated data by the research community is predicated on the accessibility and accuracy of those data, more work is required to improve data standards, increase access to taxonomic authority sources, and ensure adoption of standard vocabularies and data models for published data. The core structure of specimen occurrence metadata across biodiversity collections is standardized through publication by use of the Darwin Core schema. However, some data values, gathered across vast temporal, geographic and taxonomic ranges, are usually not standard, even within a collection. Momentum is gaining in this area with the implementation of standardized data quality test and assertion tools by the major aggregators and driven by work done by the TWDG Data Quality committee. Further improvement to these foundational data may necessitate the development of additional tools that build upon the technical groundwork and data pipeline construction completed during the first round of ADBC funding.

Fill Gaps in Biodiversity Data

Biological collections hold the most comprehensive record of life on Earth; their full potential can only be fully realized when the data contained within these collections are revealed and made computationally accessible. Most of the approximately 1,469 U.S. collections are not yet fully 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, perhaps using *Index Herbariorum*, an index to plant collections, as a model. Such a reference enables prioritization of collections for digitization to create an effective national resource.

Digitization progress across different organismal groups reveals significant disparities in numbers of digitized specimens relative to known diversity patterns, taxonomic challenges, and overall specimen numbers. Through ADBC funding, essentially all mycological collections have digitized some of their specimens. An estimated 43% of all fish collections (60 out of 143) share some digitized data through an aggregator. Rough estimation reveals that at least 437 of the 659 herbaria in the U.S. (66%) have digitized some of their holdings, and at least partial digital records exist for approximately 34% of the estimated 76 million herbarium specimens. In contrast, estimates indicate that only 6% of all insect collections have been digitized (Cobb et al., in preparation). A recent article on the EPICC TCN (Eastern Pacific Invertebrate Communities of the Cenozoic), reports, “Our group...quantified just how much "dark data" are present in our joint collections. We found that our 10 museums contain fossils from 23 times the number of collection sites in California, Oregon and Washington than are currently documented in a leading online electronic database of the paleontological scientific literature, the Paleobiology Database.” (Marshall 2018 and Marshall et al., 2018)

We must explore new strategies to reach a critical mass of digitized specimen data for all groups of organisms. For collections of some of the larger taxonomic groups, rapid digitization using skeletal records with minimal fields of information or lot-based digitization, as well as technological solutions (e.g., robotic imaging stations for specimens in drawers, lot imaging) have been, and will continue to be developed. These records can act as useful proxies for the undigitized components of biological collections, with specific emphasis placed on the data types that augment primary occurrence data (i.e., observation data, images, field notes, movement data, slide collections, vocalizations, video, DNA sequences). This call to action to expand collections and their access via digitization will also require mobilizing new researchers and personnel to enact project-focused collections that are digitized upon acquisition (e.g., Contreras 2018); this will require an increase in person-power applied to digitizing existing collections.

Biodiversity specimens are a resource for documenting environmental change, and researchers must continue to collect new specimens in perpetuity. New collections are critical for documenting areas of rapid change, such as the Arctic and marine systems. To be optimally useful, a holistic, next-generation approach to the collection of new biodiversity specimens is needed. As noted by Schindel and Cook (2018) 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.

Incorporating biodiversity data effectively into research and other initiatives will depend on data accessibility through a permanent central portal, as well as continued integration of new and existing data across the community. In the realm of collections, this includes the seamless connection of specimens and derived data at all levels. In the ESN, specimen preparations (e.g., tissues, skeletal preparations, gut contents) will be linked back to the original voucher specimen to associate all products of collections research. Data derived from these preparations, e.g., genomic data from tissues and phenotypic data from skeletal preparations, should be linked back to the specimen. The integration between different collection types (e.g., predator-prey, host-parasite, tissue-voucher) and datasets outside the collections-based research realm ensures the future utility of collections to a rapidly expanding population of end-users. Over the past ten years, massive molecular, trait, and environmental datasets have been collected that can now be coupled with digitized specimen data and used in ways not previously imagined. When combined, these resources will enable new discoveries and facilitate computational connections between these different data types, unleashing inquiry-based science in novel, and perhaps even unexpected, ways.

The National Ecological Observatory Network (NEON) collects and archives approximately 80,000 samples annually through its continental-scale environmental monitoring platform. Each sample receives appropriate physical and digital curation to enhance sample discoverability and enable transformative research. NEON facilitates the extended specimen paradigm by ensuring that all archived specimens are associated with precise spatial and temporal information from their collection along with any available genetic or taxonomic metadata. Depending on the data product and sample type, NEON provides specimen images, morphometric data, genetic sequences and taxonomic data. The NEON Biorepository uses a Symbiota platform for its specimen portal. Researchers will have access to NEON samples to conduct additional analyses; results of these analyses will be linked to the samples themselves on the portal.

Inspire an Open, Integrated Collections Alliance

To realize the full potential of digitized biodiversity collection data, a more community-minded approach to data gathering and sharing must evolve, and the training and reward structures for collection community professionals must be adjusted to place higher value on collaborative activities. Such changes will foster an agile, connected community capable of supporting curation, research, and the mobilization and use of data by a much broader array of scientific and commercial interests. We must train collectors of new specimens to approach their work in a way that maximizes efficiency and accuracy of downstream digitization and analysis to prevent continued accumulation of undigitized and unprocessed material accruing within collections.

Collections must insist that data from new collections be structured according to community standards and that data flow seamlessly into the institutional, national, and international data streams. The open sharing of collections data should be the rule, with embargos on data accessibility for well justified and documented reasons only an limited exception.

Research communities and other groups that are involved in the collection of specimens and data must stay abreast of best practices and standards governing the collecting of rich, augmented data sets through consultation with collections professionals at the initial phases of planning such collecting. For example: field notes, data spreadsheets, images, additional specimens outside of targeted species ("by-catch"), and possible ancillary environmental/ecological datasets should not only be gathered intentionally, but be compiled and stored in such a way that they are easily retrievable and link seamlessly to the core specimen metadata. Managing collections in light of these new data streams will require new best practices as well as tools, frameworks, and pipelines to accommodate and link these aspects of the extended specimen.

In a future scenario of greater collaboration and data exchange, collections professionals will divide their attention between addressing needs of their own collections and participating in activities of the greater collections network. Consequently, they will be well positioned to adjust administrative practices in their own collections to national norms. By developing the ability to place their own collection in a national context, collections professionals can help senior institutional leaders (e.g., provost, research administrators, trustees) understand the central roles that biodiversity collections play in supporting research. Realizing the value of collections will motivate buy-in on funding campaigns to support education and research that is only possible because of an institution's scientific collections.

Build and Strengthen Strategic Partnerships

The NIBA Strategic Plan called for the development of a web of partnerships among the stakeholders for digitized collections data to ensure the success of a digital collections network, and indeed such a web has developed in the past ten years (see Appendix B). To create the ESN envisioned here, requires that we build on existing partnerships and foster new collaborations.

Computer and data science research communities. Building the ESN will require collaboration with the computer and data sciences community to build and maintain next generation collections infrastructure. This might include, for example, implementing blockchain technology for specimen tracking and developing semi-automated approaches to data completion and standardization.

International Biodiversity Organizations. Surveys of the national stakeholder community identified greater international collaboration as a key need for the next phase of collections data mobilization in the U.S. The Global Biodiversity Information Facility (GBIF) now in its 20th year of operation, recently surpassed one billion species occurrence records available for searching through their data portal. GBIF has served as a model and advisor for the implementation of the NIBA plan and the iDigBio data portal. Given their scope and vision, GBIF is arguably our most important global partner in the implementation of our new national agenda. Participation to the fullest extent possible in GBIF's proposed alliance for biodiversity knowledge (Hobern et al. 2019) will facilitate local work and help align our efforts to observe, measure and mU.S. biodiversity in relation to other global efforts.

Among other country or regional data aggregators, the Atlas of Living Australia (ALA) is the most mature national biodiversity resource and provides an excellent model for user interfaces that meet the needs of the broader community. The Distributed System of Scientific Collections (DiSSCo) is a new European Union initiative that aims to "position European natural science collections at the centre of data-driven scientific excellence and innovation in environmental research, climate change, food security, health and the bioeconomy." Inspired in part by the ADBC program, DiSSCo is poised to explore new uses for collections data from which we can learn and with which we will collaborate.

On a hemispheric scale, we must pursue collaboration with data aggregators in Mexico (CONABIO) and Canada (Canadensys) to permit the seamless transfer of the data needed for continental-scale understanding of the breadth of shared biodiversity, its distribution and change over time. Common approaches among all three North American countries would help to fill gaps through new collections and digitization of existing ones. Opportunities for collaboration in the development of research tools and training programs would strengthen research capabilities of U.S. collections strengthen the capabilities in all three North American countries.

Aggregators of related data. Integrating historic occurrences and interactions of organisms with current observations is crucial for understanding environmental change. To do so will require new tools for data integration and analysis. Collaboration with programs such as the NSF-funded National Ecological Observatory Network (NEON), Long Term Ecological Research Centers (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 this new initiative, will inform and make more computable, the data collected by many future researchers using these centers.

Examples of other databases that enrich the context of biodiversity specimens include taxonomic resources such as the Catalogue of Life, Integrated Taxonomic Information System (ITIS) and the Encyclopedia of Life (EOL), published literature such as the Biodiversity Heritage Library (BHL), publishers of genetic information such as the

Barcode of Life Data System (BOLD) and NCBI (Genbank), and the Encyclopedia of Life Traitbank.

Professional Societies. Strengthening the relationship with taxon-based societies, as well as collections and data oriented organizations (e.g., Natural Science Collections Alliance, Society for the Preservation of Natural History Collections and the Taxonomic Databases Working Group) will facilitate the greater use of collections data in research and training. The ESN campaign will require broadening these alliances to include organizations that have developed international standards for genome and phenotype curation and annotation (e.g., International Society for Biocuration (ISB), International Society for Biological and Environmental Repositories (ISBER)) and others that provide the best practices for developing ontologies (Open Biological and Biomedical Ontology (OBO) Foundry).

In the case of taxon-oriented societies, greater involvement with the digitization process will inculcate a greater sense of responsibility for the protection and development of collections in their domain. Developing an efficient means of tracking specimens through their use in scientific publications will require partnership with publishers of primary scientific literature such as society journals as well as their editorial boards. The American Institute of Biological Sciences (AIBS), a federation of scientific societies spanning the biological sciences, works to unify the community around common interests. Working with organizations like AIBS and iDigBio, the new agenda can further expand linkages to a wider diversity of scientific fields, journal editors, funders and policymakers.

Education and Broadening Participation. Working with national educators organizations such as the National Association of Biology Teachers (NABT), the Association for Biology Laboratory Education (ABLE), and Quantitative Undergraduate Biology Education Synthesis (Qubeshub.org) will expand the awareness of ESN resources, and foster training of next generation of scientists in biodiversity literacy. Groups such as the Society for Advancing Chicanos and Native Americans in Science (SACNAS) as well as Minority Serving Institutions (MSIs), Historically Black Colleges and Universities (HBCUs) and other higher learning institutions that primarily serve groups underrepresented in science, will be critical partners as we continue efforts to recruit and retain a diverse 21st century workforce. Partnering with programs to help people gain the skills necessary for self-sufficiency should be a priority for all collections institutions. Partnerships among robust and fledgling citizen science initiatives will promote science literacy and foster a culture of compatible data standards and protocols.

Other Partnerships. The development of new partnerships beyond the familiar primary user groups is key to developing the new capabilities required for research and for sustaining the U.S. biodiversity data store and the collections themselves. The stakeholder community identified an extensive list of the scholarly and business-oriented groups with which potential partnerships can be formed for knowledge sharing and financial support, summarized in Appendix C. It must be noted that many of these groups were also identified as potential partners in the NIBA plan. Although we have some direct evidence of use of collections data by these groups, the extent of such use is unknown, and formal partnerships have not developed in most cases. 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 value and diversity of products generated from digitized collections data grows and becomes more visible, so too will support for investments in biodiversity collections. As our ESN becomes more representative of U.S. national biodiversity holdings, and the data are standardized and easily searched and combined, we will be able to offer the most relevant and complete data sets to users, while owners of the data can be credited and compensated for their use. The first partnership that needs to develop in this context is with business development advisors who can help us with marketing as well as cost- and profit-sharing agreements.

V. Empowering Biodiversity Science through 21st Century Learners

Through education, the collections community has the largest potential to engage, educate, and empower the next generation of biodiversity data stewards, biodiversity researchers, and ESN users. Towards this end, the collections community has made notable strides to integrate physical collections and collections-based data into formal and informal education (Appendix B). As a result of these educational efforts, we are well positioned to implement focused and inclusive education strategies and materials that engage a rich array of learners and result in improved biodiversity literacy content and skills. To train the biologists of the future, who will use the ESN to address large scale research questions, we must make it a priority to infuse biodiversity data literacy skills into all levels of formal and informal education.

We define education broadly in this report. For our purposes, *Formal Education* encompasses K-12, undergraduate, graduate, and post-doctoral education and training. *Advanced Professional Training* includes existing scientists, curators and museum professionals with skill gaps. We use the term *Informal Education* as a broad category that spans everything from traditional museum experiences to rapidly developing citizen science initiatives with incredible potential to engage a broad audience in gathering and transcribing biodiversity data.

Formal Education. Natural history specimen-based data can be integrated into the core biology curriculum and enable training and education of data savvy scientists and engaged biodiversity enthusiasts. Specimen-based data make science accessible through the specimen itself, which is tangible, place-based, and interesting, as well as through aggregated specimen data that are verifiable, relevant, and a logical gateway to data literacy (Cook et al., 2014, Powers et al., 2014. Monfils et al. 2017). Biodiversity data from natural history collections is well suited to learning core content in the K-12 and biology undergraduate curriculum including evolution, biodiversity, systematics, taxonomy, and ecology. Biodiversity data, skills and competencies can be integrated into the curriculum without a dramatic increase in the content to be covered.

Biodiversity Literacy in Undergraduate Education (BLUE) is a growing network that fosters partnerships among biodiversity and education researchers. The goal is to identify strategies, centralize resources, and develop, assess, and support implementation of educational materials that promote biodiversity data literacy. BLUE has initiated a series of Open Education Resources, that are available through the QUBEShub.org, that use digitized natural history collections-based data to teach content in topics such as island biogeography, pollination biology, co-evolution, climate change, form and function, etc. and data skills relative to the biodiversity data pathway, digitization, data discoverability and utility, data standards, etc. Materials developed or endorsed by BLUE, directly support the ESN efforts by implementing modules that facilitate training of diverse, competent, and engaged young biologists who are well prepared for a broad set of career paths generating and utilizing biodiversity data to address scientific issues of critical national and global importance.

The biological collections community can engage a larger set of stakeholders and extend the reach of the biodiversity data beyond the biology classroom by providing the datasets, experiences, and materials that engage data-centric non-biological sciences and bridge the liberal arts disciplines. As within the biology classroom, we need to be conscious of the different resource needs of prospective students and educators as we create inclusive materials that meet those requirements and provide individualized support for successful implementation.

As we look to novel uses of specimens in research, we need to avail ourselves of the highly diverse millennial workforce. Diversity in research teams provides new perspectives and ideas leading to new solutions and increased productivity (Al Shebli et al., 2018). Increasing the presence of historically underrepresented groups within the biodiversity sciences requires a concerted community-wide effort of those actively involved in these fields. The place-based capacity of collections data combined with the social and societal relevance of biodiversity science can serve a role in creating inclusive, culturally relevant, and socially conscious educational materials that engage a broad and diverse audience in biodiversity science. When creating educational materials or speaking on the value of biodiversity data, we must document and share the history of the contributions of diverse scientists to natural history and engage a diversity of natural history researchers and collections professionals as spokespeople for the field of biodiversity science. We need to engage indigenous communities to capture and preserve knowledge and perspectives, and we need to be collaborating internationally to create programs with a universal design. Providing effective rationales for pursuing biodiversity related professions, clear models of career paths within these professions, outlines of the available educational opportunities leading to these professions, and sustained support as students move through the educational system are critical for diversifying and empowering the emerging workforce.

More than any time in the past, graduate students in research-driven Master's and Ph.D. programs can gain diverse and practical skill sets via the analysis of biodiversity data, including those related to taxonomy and natural history, (paleo)genomics, statistics, machine learning, and computer science. Large databases that the broader natural history collections community have been developing, maintaining, and extending in recent decades perfectly suit graduate training that incorporates data analysis, including statistical analysis, machine learning, and artificial intelligence. Further, trained students will go on to mentor the next-generation of natural history collections-based scientists, or assume careers outside of our core biodiversity science community (e.g., as data analysts in biotech, agriculture, or other industry), where they will bring awareness to the importance and promise of natural history collections.

With economically relevant issues of climate change, invasive species, zoonotic diseases and food availability and security facing the U.S. (Wuebbles et al. 2018), we need to expose the next generation to the best data and encourage them to think creatively, be open to new ideas and new technologies, and work with others to generate ideas and problem solve. Biodiversity scientists will require advanced training in a diversity of fields (e.g., integrative and comparative biology, systems biology, evolution, ecology, developmental biology, computational biology). Soft skills, including flexibility, communication, management, responsiveness, and the ability to cooperate with others, are fundamental to success. Students will need to work with multiple mentors with diverse expertise in order to train for addressing large-scale multidisciplinary science questions, and thus we need transformative programs to affect this paradigm shift in graduate education.

The NSF Postdoctoral Research Fellowship in Biology Track 2: Research Using Biological Collections (2015–present) is an excellent example of the type of advanced training programs needed to advance this agenda. This program funds postdoctoral researchers who perform collections-based research, often in novel or cutting-edge ways. Importantly, by explicitly funding novel uses of collections, this program has resulted in a variety of uses of collections that a "typical" museum researcher (i.e., phylogeneticist, taxonomist) would not necessarily envision, including evaluation of plant responses to rising CO_2 availability using herbarium collections, understanding the evolution and development of tetrapod olfaction through histology and imaging of museum specimens, and testing hypotheses of invasive species establishment using trait data from digitized, curated ants.

Advanced Professional Training. Collections community survey respondents cited the lack of programming or other data science expertise as one of the greatest obstacles to the digitization of their specimens (Appendix C). These deficits further limit the use of the digitized data in research. Therefore, companion training programs are needed to prepare the current and next generation of collections managers and curators to digitize, annotate, augment and maintain the enormous digital data sets that will radiate as part of the extended specimen.

Specifically, the collections community will need programs that integrate fundamental data and data management skills as well as the basics of coding. With the rapidly changing requirements of collections professionals and the considerable skill set associated with digitization, aggregation, and management of large biodiversity data sets, we need to consider training that can meet the needs of professionals at different stages in their career. Since 2011, iDigBio has been highly effective at creating inclusive and widely applicable training workshops and webinars. These opportunities to learn new skills have helped to create a broadly inclusive collections community as well as improved understanding between data managers and data users (Seltmann et al. 2018). The Carpentries have been useful in training the research community in coding and foundational data science. The future needs for the extended specimen research program, however, far exceeds current curricula. Integrating continuing education with ongoing formal education, drawing on the skill sets of science education researchers, creating standard materials and gateway workshops to teach about different aspects of collections management and biodiversity science are all critical in leveraging our past successes to create a sustainable knowledgebase.

Informal Education. An increasing number of citizen science projects and focused programs at, or hosted by, museums and natural history collections are engaging the public in biodiversity science. As our digital resources continue to expand, so too will informal education opportunities. The voluntary contributions of citizen scientists can likewise bring a 'multiplier effect' to substantial collections funding via secondary uses of specimen information (such as for educational purposes), increased personal investment in collections, and the opportunity to directly contribute to scientific research. Indeed, we cannot obtain the full expression of the ESN envisioned in this agenda without strong involvement of the citizen science community.

A number of effective citizen science projects are based on monitoring biodiversity. eBird, eButterfly, iNaturalist, and U.S. National Phenology Network, to name just a few, provide platforms for minimally-trained individuals to contribute sightings or recordings of organisms or a particular attribute of an organism, e.g., its phenological state. For individuals interested in a deeper understanding of biodiversity, offerings such as Master Naturalist programs can provide the opportunity for members of the public to become experts in their local flora and fauna. Biodiversity monitoring and naturalist programs led by museum and collections staff can provide opportunities for amateurs to closely investigate organisms, thereby teaching skills critical to biodiversity science, namely species identification, taxonomy and systematics.

The Urban Nature Research Center at the Natural History Museum of Los Angeles County is a strong example of what can be done when museum professionals collaborate with citizen scientists to investigate urban biodiversity. Notably, UNRC researchers and participants in the citizen science projects they have developed have discovered numerous species new to science (e.g., Hartop et al., 2016), documented range expansions of introduced species (e.g., Vendetti et al., 2018), and have demonstrated natural history museum-based citizen science contributions to species conservation (Ballard et al., 2016). Specimens remain the gold standard for verifiable data, yet as UNRC researchers and collaborators find (Spear et al., 2017), citizen science observations in concert with specimen data are perhaps the most effective and efficient way to scale up data collection to address many of today's challenges problems. In addition to the direct research implications of these citizen science projects, participants also often help provide specimens to collections. In doing so, citizen scientists learn collections skills while improving the representation of recent records within our collections.

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. While tasks are online, researchers and scientists regularly communicate with participants to answer questions and ensure that volunteers complete tasks as designed. WeDigBio, an annual global digitization event, has united these efforts to work collectively towards digitization goals over the course of four days each year. Such programs engage participants from a wide range of ages, abilities, and interests, and with minimal start-up costs. This approach creates an inclusive and diverse group of individuals working to advance biodiversity science.

Additionally, the success of small scale efforts by institutions to engage recent immigrants, developmentally challenged, and previously incarcerated citizens in efforts to improve language, keyboarding and workplace skills indicates work in collections can improve quality of life in addition to serving as an entry point for new scientists (*Moving the Needle: Broadening Participation in the Biodiversity Sciences* webinar series). Many students engage in citizen science platforms as part of formal education activities, creating a point of intersection between formal and informal education that can be leveraged for addressing the needs in both communities.

By design, citizen science projects turn to established best practices and procedures for quality assurance and quality control. Such practices include training documents and videos, within-project quality checks, and post hoc assessments. Citizen science data standards have been created and vetted by members of international organizations such as the Data and Metadata Working Group with the U.S. Citizen Science Association and the Citizen Science Interest Group with the Biodiversity Information Standards (TDWG) organization. These measures ensure that the work completed by citizen scientists is research-ready. However, we must continue to develop and improve training materials, best practices and ongoing support from the scientific community to maximize the impact of citizens.

VI. Implementing and Sustaining the New Agenda

Building a comprehensive network of extended specimen data that integrates the wealth of biodiversity held in U.S. collections and associated data repositories will require a monumental effort, comparable to building a new telescope for planetary exploration. However, the lack of physical infrastructure on the scale of a telescope makes the magnitude of the effort required to build and sustain such a resource harder to comprehend. A major challenge to the implementation of the new agenda and maintenance of the resulting information network will be to convey the unity of the resource and the scope of the effort required to build and maintain such a resource.

Initially building the core infrastructure needed for the ESN might be accomplished through established grant programs. Episodic funding for particular digitization or data augmentation projects would also be effective for filling research needs and knowledge gaps, for improving data and for development of new educational initiatives. However, sustaining the ESN requires a central organizing unit funded over a much longer time horizon than currently supported by any existing grant program. Building upon the current model of the digitization hub (iDigBio) established in the ADBC program, a securely funded coordinating center could maintain the data network and partner with collections institutions and professional societies to support the stakeholder community to share techniques, resources, strategies for outreach and demonstrating the value of collections. We suggest that the creation of a distributed platform as ubiquitous and indispensable as the NCBI-managed GenBank database, with similar open-ended funding, is required for the ESN to reach its full potential. Indeed, through more extensive data integration mechanisms, Genbank data would be integrated within the larger framework of the proposed ESN , thus enhancing the utility of these genetic data by virtue of the links to phenotype and environment. Similarly, data from other new and ambitious efforts such as the Earth BioGenome Project that aims to sequence, catalog, and characterize the genomes of all of earth's eukaryotic biodiversity over a period of 10 years (Lewin, 2018) and isotopic data from IsoBank would be accessible through the ESN, as would specimen data from new, comprehensive collecting initiatives such as the NEON Biorepository.

Although securing the long-term funding base for the ESN will take some time to develop, there are steps we can take now to set the process in motion, including the following:

- Determine the primary specimen-based needs of researchers to develop the primary requirements for the ESN, including enhanced curation practice.
- Develop a robust, comprehensive specimen identifier system in collaboration with other international data aggregators and providers to enable transparent and uncomplicated integration of biodiversity data with other data sources while facilitating the attribution of collections' role in discovery and policy and promoting transparency for broader issues involving multiple stakeholders (e.g., access and benefit-sharing).
- Create an authoritative, comprehensive, and self-updateable index of U.S. collections institutions (similar to Index Herbariorum for global herbaria) with structured metadata to describe their holdings as a first step toward expediting the discovery of undigitized collections and revealing these to the research community.
- Continue digitization of existing material focused on underrepresented taxa (e.g., those in entomology and paleontology) and including incorporation of specimens held in small regional, personal, and individual researcher based collections. Additional efforts need to include improvement of 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 datasets for research and education.
- Develop data access tools designed to maximize the educational potential of collections and collections-based data to inspire interest in the natural world and enable a diverse, digitally fluent workforce, and allow citizen scientists to contribute meaningfully to documenting Earth's biodiversity.
- Champion broad scale adoption of core biodiversity data literacy skills and competencies in k-12 and undergraduate curricula to foster a biodiversity savay future workforce, engage new end-users in novel uses of ESN data, and sustain and promote careers advancing collections science, biodiversity research, and data literacy.
- Support enhanced training of emerging and established professionals for interdisciplinary work in biodiversity, data science, and informatics. Emphasize skills needed to work collaboratively in interdisciplinary teams and other soft skills related to communication, creativity, and critical thinking as these skills are paramount to conducting transformative science, communicating research, and cultivating and leveraging relationships with new research and user communities.

VII. The Extended Specimen Agenda and the National Science Foundation's 10 Big Ideas

If implemented, the new agenda for biodiversity collections proposed here will provide a scientific tool for biological sciences as well as a resource that enables progress toward cross-cutting and frontier challenges reflected by the National Science Foundation's 10 Big Ideas. For example:

Understanding the Rules of Life. To understand the rules by which biological and environmental factors influence the wide range of organisms upon which humans depend on will require diverse data from specimens across the tree of life. Specimens of rare or possible extinct organisms are a source of genetic, phenotypic, and environmental data, the three basic data types required for identifying causal and predictive relationships across these scales. Collections provide these fundamental data, though as described previously, they have not yet been exposed through comprehensive digitization and data linkage. When the vision of the ESN is fully realized, data patterns may be discovered using e.g., new AI methods, and specific research questions such as how genetic regulators of similar traits (phenotypes) evolve in relation to environment, can be addressed computationally.

Harnessing data for 21st Century Science and Engineering. The digitization of natural history collections has already contributed to the "deluge of data" from the nation's scientific facilities, and as we broaden the knowledge bank of specimens with additional genotypic, phenotypic, and environmental data, the amount of data will increase exponentially. The extended specimen network will be a prime example of a "cohesive, national scale approach to research data infrastructure," and will inform efforts in other domains. In addition, specimen based data provides a unique opportunity to educate students in data science and train a data enabled workforce. Collections data alongside the archived specimens are an engaging and accessible data source that can provide students an opportunity to experience the entirety of the data pathway, while practicing verifiable and testable science.

Midscale infrastructure. The level of support required for the ESN coordinating center falls in the gap between what the NSF currently funds as either small and large infrastructure. The computer and human infrastructure required for data acquisition, deployment and training to support the ESN represent a midscale infrastructure need, both in terms of funding level and the scientific research it will support.

Navigating the New Arctic. The Arctic biota has been minimally sampled over the past two centuries. Arctic collections are stored in relatively few national collections and those specimens provide our best baselines for understanding the implications of rapid environmental perturbation on life (e.g.. Bond et al. 2015) in a geopolitical region that is becoming critical to global economy and security. Due to the rapid rate of annual warming in the Arctic, we should immediately build international coalitions that will invest in building spatially extensive, site intensive natural history collections that will provide the biodiversity infrastructure necessary to critically assess the current changes. Over time, specimens and their associated digitized data would provide the samples required by emerging technologies (e.g., genomics, isotope ecology) to provide a robust historical context for the proposed network of observational platforms, and a reference for the identification, distribution, behavior, and response of species and their pathogens in the Arctic (Cook et al., 2013, Hoberg et al., 2013).

Growing Convergent research at NSF. Enhancement of the roles that collections-derived data can play in understanding and protecting human health and in education are key objectives of the new agenda, achieved through the central efforts to integrate and share data more effectively and widely. The ESN, combining diverse data sources linked to a physical museum voucher specimen, is a physical manifestation of convergent research and will amply demonstrate the power of this approach.

The Future of Work at the Human Technology Frontier. The use of biodiversity collections in projects to both refine current techniques in machine learning and to document variation among organisms as exhibited by specimens holds great promise (e.g., Carranza-Rojas et al., 2017, McAllister et al., 2018). Use of these tools will lead to new avenues of data analysis, requiring new skills for researchers and data managers. Collections professionals will find themselves at this frontier and supporting their adoption of new technology and documenting the experience through training and best practices is a key aim of the agenda presented here. The work done in this community may prove transferable to other communities.

Enhancing Science and Engineering through Diversity. The place-based capacity of collections specimens and associated data combined with the social and societal relevance of biodiversity science can serve a role in creating inclusive, culturally relevant, and socially conscious educational materials that engage a broad and diverse audience in biodiversity science. Collections institutions have great potential to engage a broad range of young people to create a scientifically literature work force and build the ranks of our scientific and engineering communities.

VIII. References

AlShebli, B. K., T. Rahwan, W. L. Woon. 2018. The preeminence of ethnic diversity in scientific collaboration. Nature Communications. 9: 5163.

Ballard, H. L., Robinson, L. D., Young, A. N., Pauly, G. B., Higgins, L. M., Johnson, R. F., and Tweddle, J. C. 2017. Contributions to conservation outcomes by natural history museum-led citizen science: Examining evidence and next steps. Biological Conservation. Volume 208: 87-97. Doi: 10.1016/j.biocon.2016.08.040 https://www.sciencedirect.com/science/article/pii/S0006320716303512?via%3Dihub

Bebber, D. P. et al. 2010. Herbaria are a major frontier for species discovery. Proceedings of the National Academy of Sciences 107: 22169-22171.

Bond, A.L., K.A. Hobson, B.A. Branfireun, 2015. Rapidly increasing methyl mercury in endangered ivory gull (*Pagophila eburnea*) feathers over a 130 year record. Proceedings of the Royal Society B: Biological Sciences. 282(1805), p.20150032.

Bond, et al 2015
Campbell, P. 2017. Archaeology and blockchain: a social science data revolution? https://www.theguardian.com/science/2017/oct/02/archaeology-and-blockchain-a-social-science-data-revolution Mon 2 Oct 2017 07.20 EDT

Carranza-Rojas, Jose, H. Goeau, P. Bonet, E. Mata-Montero, Joly, A. 2017. Going deeper in the automated identification of herbarium specimens. BMC Evolutionary Biology 17: 181.

Contreras, D.L. 2018. A workflow and protocol describing the field to digitization process for new project-based fossil leaf collections. Appl Plant Sci 6:e1025.

Cook, J. A., C. Brochmann, S. L. Talbot, V. Fedorov, E. B. Taylor, R. Väinölä, E.P. Hoberg, M. Kholodova, K. P. Magnusson. 2013. Genetic Perspectives on Arctic Biodiversity. Pp. 459-483 in *Arctic Biodiversity Assessment. Conservation of Arctic Fauna and Flora Committee*, Copenhagen.

Cook, J. et al. 2014. Natural history collections as emerging resources for innovative education. *BioScience* 64: 725-234.

Halamka, J. D., A. Lippman, Ekblaw, A. 2017. The Potential for Blockchain to Transform Electronic Health Records. Harvard Business Review. https://hbr.org/2017/03/the-potential-for-blockchain-to-transform-electronic-health-records (accessed 28 November 2018).

Hartop, E. A., Brown, B. V., and Disney, R. H. L. 2016. Flies from L.A., The Sequel: A further twelve new species of Megaselia (Diptera: Phoridae) from the BioSCAN Project in Los Angeles (California, USA). Biodiversity Data Journal 4.e7756

Hilsberg, T., Robinson, A., Robinson, M., Haynes, D. 2018. Biocoin.Life Foundation. White Paper V0.85. https://static1.squarespace.com/static/5a338858d7bdcea6d17ccc0f/t/5a7f8d45652deaf340ccdd98/1518308678928/White+Paper+-+Developing+a+Global+Biocoin.life+20180211.pdf

Hoberg, E. P., S. J. Kutz, J. A. Cook. K. Galaktionov, V. Haukisalmi, H. Henttonen, and S. Laaksonen. 2013. Parasites in Terrestrial, Freshwater, and Marine Environments. Pp. 420-449 in *Arctic Biodiversity Assessment. Conservation of Arctic Fauna and Flora Committee,* Copenhagen.[TB1]

Hobern, D., B. Baptiste, K. Copas, R. Guralnick, A. Hahn, E. van Huis, E.-S. Kim, M., McGeoch, I. Naicker, L. Navarro, D. Noesgaard, M. Price, A. Rodrigues, D. Schigel, C. Sheffield, J. Wieczorek.2019, Connecting data and expertise: a new alliance for biodiversity knowledge. *Biodiversity Data Journal.* 7: e33679

Kemp, C. 2015. The endangered dead. Nature. 158:292-294.

Kissling, W.D., Walls, R., Bowser, A., Jones, M.O., Kattge, J., Agosti, D., Amengual, J., Basset, A., Van Bodegom, P.M., Cornelissen, J.H. and Denny, E.G., 2018. Towards global data products of Essential Biodiversity Variables on species traits. *Nature ecology & evolution*, p.1. https://doi.org/10.1038/s41559-018-0667-3

Lewin,H.A., G. E. Robinson, W. J. Kress, William J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. Thomas P. Gilbert, M. M. Goldstein, I. V. Grigoriev, K.J. Hackett, D. Haussler, E. D. Jarvis, W. E. Johnson, A. Patrinos, S. Richards, J. C.Castilla-Rubio, M.-A. Sluys, P. S. Soltis, X. Xu, H. Yang, G. Zhang 2018. Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences 115 (17) 4325-4333; DOI:10.1073/pnas.1720115115

Marshall, C.R. 2018. Digitizing the vast 'dark data' in museum fossil collections https://phys.org/news/2018-09-digitizing-vast-dark-museum-fossil.html#jCp)

Marshall C.R,, S. Finnegan, E.C .Clites, P.A. Holroyd, N. Bonuso, C. Cortez, E. Davis, G.P. Dietl, P.S. Druckenmiller, R.C. Eng, et al. 2018 Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. Biol Lett 14:2–5.
https://royalsocietypublishing.org/doi/full/10.1098/rsbl.2018.0431

McAllister, C.A, M. R. McKain, M. Li , B. Bookout & E. A., Kellogg. 2018 Specimen- based analysis of morphology and the environment in ecologically dominant grasses: the power of the herbarium. Phil. Trans. R. Soc. B 374: 20170403.

Monfils, A., Powers, K., et al. 2017. Natural History Collections: Teaching about Biodiversity Across Time, Space , and Digital Platforms. Southeastern Naturalist 16: 47-57. https://doi.org/10.1656/058.016.0sp1008.

Powers, K.E., et al., 2014. Revolutionizing the Use of Natural History Collections in Education. Science Education Review, 13 (2): 24-33 2014

Schindel, D.E. & J.A. Cook. 2018. The next generation of natural history collections. PLoS Biol 16(7): e2006125.

Seltmann K., Sara Lafia, Deborah L. Paul, Shelley A. James, David Bloom, Nelson Rios, Shari Ellis, Una Farrell, Jessica Utrup, Michael Yost, Edward Davis, Rob Emery, Gary Motz, Julien Kimmig, Vaughn Shirey, Emily Sandall, Daniel Park, Christopher Tyrrell, R. Sean Thackurdeen, Matthew Collins, Vincent O'Leary, Heather Prestridge, Christopher Evelyn, Ben Nyberg. 2018. *Georeferencing for Research Use* (GRU): An integrated geospatial training paradigm for biocollections researchers and data providers. Research Ideas and Outcomes 4: e32449. https://doi.org/10.3897/rio.4.e32449

Singer R.A., Love K.J., Page L.M. 2018. A survey of digitized data from U.S. fish collections in
the iDigBio data aggregator. PLoS ONE 13(12): e0207636.
https://doi.org/10.1371/journal.pone.0207636.

Spear, D. M., Pauly, G. B., and Kaiser, K. 2017. CItizen Science as a Tool for Augmenting Museum Collection Data from Urban Areas. Frontiers in Ecology and Evolution 5:86.

Vendetti, J.E., Burnett, E., Carlton, L., Curran, A.T., Lee, C., Matsumoto, R., Mc Donnell, R., Reich, I., and O. Willadsen. 2018. The introduced terrestrial slugs Ambigolimax nyctelius (Bourguignat, 1861) and Ambigolimax valentianus (Férussac, 1821) (Gastropoda: Limacidae) in California, with a discussion of taxonomy, systematics, and discovery by citizen science. Journal of Natural History. https://doi.org/10.1080/00222933.2018.1536230.

Webster, M.S. 2017. The Extended Specimen: Emerging Frontiers in Collections-based Ornithological Research. Boca Raton, FL: CRC Press/Taylor & Francis Group.

Wilkinson, M.D. *et al*. 2016.   The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3: 160018.

Wuebbles, D.J., D.R. Easterling, K. Hayhoe, T. Knutson, R.E. Kopp, J.P. Kossin, K.E. Kunkel, A.N. LeGrande, C. Mears, W.V. Sweet, P.C. Taylor, R.S. Vose, and M.F. Wehner. 2017. Our globally changing climate. In Climate Science Special Report: Fourth National

Climate Assessment, Volume I. D.J. Wuebbles, D.W. Fahey, K.A. Hibbard, D.J. Dokken, B.C. Stewart, and T.K. Maycock, Eds. U.S. Global Change Research Program, pp. 35-72, doi:10.7930/J08S4N35.

Authorship of this document

Thiers, Barbara, Anna Monfils, Jennifer Zaspel, Elizabeth Ellwood, Andrew Bentley, Katherine Levan, John Bates, David Jennings, Dori Contreras, Laura Lagomarsino, Paula Mabee, Linda Ford, Robert Guralnick, Robert Gropp, Marcy Revelez, Neil Cobb, James Lendemer, Katja Seltmann and Mary Catherine Aime.

Appendices (Separate Documents)

Appendix A: Workshop Participants

Appendix B: NIBA Digitization Progress

Appendix C: Stakeholder Outreach Summaries

Appendix D: Selected References on Recent Uses of Biodiversity Collections