A Strategic Plan for Establishing a Network Integrated Collections Alliance

Executive Summary

This report is a strategic plan for a 10-year effort to digitize and mobilize the scientific information associated with biological specimens held in U.S. research collections. The primary objective of the initiative is to create a national collections resource that will contribute critical information to U.S. scientific research and technology interests, and will aid in understanding the biodiversity dimensions and societal consequences of climate change, species invasions, natural disasters, the spread of disease vectors and agricultural pests and pollinators, and other environmental issues. Network Integrated Collections Alliance (NICA) resources such as databases, network portals, and analytical tools will synthesize information contained in the nation's collections and place them into national service for stakeholders in government, academia, business, K-12 education, informal science education, and the public.

Biological collections across the U.S. are united by over two centuries of common purpose in research vision, curatorial methods, and field protocols. Digitizing the nation's collections represents a grand challenge that will require development of technical and human resources, such as automated workflows, a robust data publishing and error-checking infrastructure and professionals networked to support the creation of an enduring digital alliance of collections institutions. These challenges can be addressed, in partnership with federal agency and other stakeholders, in order to create an organizational structure and processes that reflect the long-standing biological collection community values of inclusiveness, scientific empowerment and open data access, while allocating credit to data owners and editors.

Digitization of biological specimens will take place within the nation's collections facilities, which will be organized into networks having shared interests in geographic scope, taxonomic research domain, or specimen preservation type. These collaborations will be supported by a national digitization hub, whose responsibility will be to assure the successful implementation of the collaborative and inclusive digitization vision. The digitization hub will: establish collaboration protocols for consensus-based decision making among collections; proactively form working relationships and synergies with U.S. and international partners; grow scientific and collection institutional engagement; oversee new digitization technology development; establish protocols for ensuring data quality and proper crediting of data owners and editors; prioritize digitization efforts based on advice from stakeholders, oversight committees, and collections professionals; and define metrics for measuring progress against explicit goals while also reporting progress to all stakeholders.

This strategic plan is the outcome of a deliberative community process that has included surveys of 291 federal and ~600 federally supported collections (see reports referenced in Appendix II), along with multiple workshops. These have recently included community engagement meetings on "Future Directions in Biodiversity and Systematics Research", and an NSF-funded Research Coordination Network meeting entitled "Collections Data Integration." Directed planning activities began with a workshop held at the National Evolutionary Synthesis Center (NESCent) on February 5-7, 2010. The product of that first meeting was an outline for the digitization plan.

Input from the community further shaped the vision, and a second workshop at NESCent on 28-30 April provided the input and guidance for this strategic plan.

VISION STATEMENT: *The Network Integrated Collections Alliance will develop an inclusive, vibrant, partnership of U.S. biological collections that collectively will document the nation's biodiversity resources and create a dynamic electronic resource that will serve the country's needs in answering critical questions about the environment, human health, biosecurity, commerce, and the biological sciences.*

Call to Action

Collections of biological specimens gathered over two centuries of field exploration document the nation's biological diversity and represent a monumental societal investment for research and applied environmental science. Identification of new species and documentation of the properties and distribution of life forms is possible only through research involving curated biological specimens. The knowledge derived from specimens contribute vitally to studies of invasive species, biological conservation, land management, pollination, biotic responses to climate change, spread of pathogenic organisms, and research and management activities of many kinds.

NICA Use Scenario 1: A massive oil spill occurs off the coast of Louisiana. Critical information is needed on the potential impact of the oil on living systems. With the national biological database completed, EPA, Coast Guard, Louisiana environmental responders, oil executives, marine fisheries personnel, fishing boat companies, estuary researchers, public health personnel, and others have instantaneous (millisecond) access to all life forms that have been recorded from the region actually or potentially threatened by the spill. Included are data on natural history, geographic distributions, protected and endangered status, position in food chains, and physiological limits of the species. Such data are vital to act quickly to mitigate damage. Without the digitized national database in place, it will take months or even years to gather the same data.

Large-scale digitizing of the nation's biological collections and mobilizing their images and data through the Internet has never been more urgent or achievable. Technological advances with scanning and research information management systems, decades of experience managing collections data in electronic form, and recent collection data standards have positioned the collections community to address the challenge in a coherent and efficient way. Further, the community has committed to the mission of open access to those data in networked environments. A national focus on collection digitization will transform the practice of collections-based biological research, and international research collaboration by implementing electronic protocols and information channels for real-time communications of all collection data.

The community of natural history research collections has already developed the social and technological infrastructures to provide open access to species occurrence data through broad community participation. Projects such as VertNet (http://vertnet.org; Constable 2010), which already provides access to over 50 million species occurrence records from more than 70 institutions, accessed at a rate of nearly 2.5 million records per week, make clear that such national endeavors can succeed.

The Scope of Collections Digitization

U.S. biological collections are an incomparable national treasure and source of knowledge. The collections contain a cross-section of the world's biodiversity including fossils, invertebrates, vertebrates, protists, fungi, plants, and human cultures. This immense knowledge base is underutilized due to the difficulty of obtaining and analyzing data within and across collections. Digitization and mobilization of specimen and associated data (e.g., field notes, illustrations, gene sequences) removes this impediment, but presents technical and organizational challenges. The largest of these is how to capture specimen data fast enough to achieve digitization of entire collections while maintaining sufficient data quality (*Overcoming the Digitization Bottleneck in Natural History Collections Workshop*, September 2006).

Collections digitization is defined broadly to include transcription into electronic format of various types of data associated with specimens, the capture of digital images of specimens, and the georeferencing of specimen-collection localities. To assess the scope of undertaking to digitize the nation's collections, the collections community continues to conduct surveys to document the number and diversity of specimens contained in U.S. collections. Additionally, the community has held three workshops on "Future Directions in Biodiversity and Systematics Research". These, in addition to two recent reports (see Appendix II for references), highlight the scale of the challenge, the need to address the integration of digitized biological data, the need to coordinate the capture of specimen data and images, and the necessity of providing broad accessibility to specimen data by scientists worldwide. Estimates of collection size range as high as three billion specimens globally, with as many as one billion preserved and cared for by U.S. institutions, most of which (~ 90%) are not accessible online.

NICA Use Scenario 2: The Yosemite National Park mammal survey (Moritz et al., 2008) is a valuable example of how biological collections data can be mobilized to evaluate the threat of climate change on living animal populations. Samples collected on the Joseph Grinnell expeditions of 1914-1920 and deposited in the Museum of Vertebrate Zoology, UC Berkeley, served as the baseline data to compare with recent surveys. The results showed that over the intervening 90+ years temperatures for the park had risen by as much as 7^0 F and several key mammal species ranges had changed dramatically. Some were in danger of extirpation from the park due to a narrowing of their habitat. These comparable data gathered across a century of time throughout the western US show how global climate change is affecting U.S. national and state parks. The collections will enable effective new conservation strategies to be developed to provide additional management of threatened species and other organisms found in the national parks.

Prior to this initiative, there has been no nationwide coordination of the effort to digitize and electronically mobilize data from biological research collections. Episodic and incremental funding has yielded limited success with digitization, addressing mainly specific, localized projects. Such past efforts were not designed to have an impact across collection types or institutions in any efficient or supportable way.

This strategic plan emerges at a critical period of accelerated environmental change. Understanding the impact of this change creates new research challenges that must draw upon the massive store of knowledge of life on earth, past and present, that is held in our nation's

biological collections. This digitization initiative will be a unified campaign involving a coordinated funding program and well defined strategy for execution. In addition to improving the physical care of collections and supporting collections-based research (see references, Appendix II), it is vital to increase online accessibility of U.S. biological collections through an integrative and focused digitization effort.

The Challenge of Collection Digitization

The nation's biological collections have developed over more than two and a half centuries, and standards for collection acquisition, preservation, and documentation are well formalized and consistent. However, different types of organisms require different methods of preservation to ensure that key biological traits will be preserved. Below, some of the similarities and differences across collections types are highlighted.

Plants and fungi are usually prepared as dried, flattened specimens attached to archival quality paper or are stored in archival envelopes or boxes. All collecting information about such specimens is printed on a label attached to the specimen. These preservation methods pre-adapt plants and fungi specimens for rapid digitization, since they can be handled and digitized individually with all data attached. Herbaria have organized themselves into regional networks, such as the California Consortium of Herbaria (http://ucjeps.berkeley.edu/consortium/), the Consortium of Pacific Northwest Herbaria (http://www.pnwherbaria.org/) and the Southeast Regional Network of Expertise and Collections (SERNEC, http://www.sernec.org/). These regional networks will be united under the nascent U.S. Virtual Herbarium initiative which will serve as an information and technology conduit for the regional networks.

NICA Use Scenario 3: An airliner encounters a bird strike while taking off from Reagan National Airport. Both engines are damaged but the plane makes a safe emergency landing. How can the serious problem of bird strikes be solved? Upon examination, the motors are found to contain bits of feathers and tissues. The feathers are compared with bird specimens in a museum and identified, while DNA from tissues is compared with the museum's collection of genomic resources. Within a short time it is determined that the birds were not from a nearby resident flock of Canada geese, which might require removal or mitigation, or even delay of air travel, but were in fact a flock of large migratory birds that had strayed from their normal migratory route. This fictional account has actual antecedents in bird strike studies that have utilized collections to help redesign jet engines for commercial and military use, and to plan safer air travel (e.g., Dove 1999).

Mammals and birds are prepared differently. Most are skinned, preserving the exterior appearance of the animal. Some are stored in alcohol as whole specimens. Mammals, both modern and fossil, often have their skulls and skeletons prepared and stored separately. The size of birds and mammals varies from hummingbirds and insect-sized bats to large specimens of ostriches and whales. This physical scaling requires heterogeneous storage and preparation techniques, which can make specimen imaging challenging and time consuming. Historically, data from mammals and birds were individually recorded in leather-bound museum master catalogues. Today, those catalogues can be used as primary sources and starting points for digitizing data from the collection. Ancillary data such as vocalizations, habitats, food habits, field notes,

climatic data associated with specimens, age and reproductive data, and other observations of each specimen are often available and will be digitized. Specimen data from at least 30 U.S. collections has been published through both the Mammal Network Information System (MaNIS, http://manisnet.org) and Ornithological Information System (ORNIS, http://ornisnet.org).

Reptiles and amphibians are typically catalogued as individual fluid preserved specimens, but historically large series of specimens of a single species were preserved as a "lot" and stored in jars of alcohol without reference to individual animals. Other collections, such as fish and many invertebrates, are also preserved and cataloged as lots. Although plants, birds, reptiles, amphibians, and mammals are often identified to species, fish—the largest vertebrate group when numbers of species is considered—are often identified only to family or genus. In contrast, fossils may have multiple identifications associated with a single sample of rock. Data sharing networks for herpetological, fish and fossil data have been developed (HerpNET, http://herpnet.org; FishNet 2, http://fishnet2.net; PaleoPortal, http://paleoportal.org).

NICA Use Scenario 4: A deadly virus suddenly appears in the southwestern United States, infecting hundreds of people, at least half of them fatally. It is quickly identified as a new Hantavirus. Field biologists quickly determine that it is carried by several species of rodents. By utilizing the digitized database, public health officials and zoologists are able to delineate the geographic area in which the virus is likely to occur, the habitats of the rodent species, associated species, and other parameters of their life cycle that will influence the spread of the virus. The disease is contained. This scenario occurred with the Sin Nombre Hantavirus in New Mexico in 1993 and the disease was understood rather quickly because of extensive mammal collections from the region, as well as ongoing field research with the species involved (Yates et al., 2002). Today, the extensive collections of genomic resources, mammals, and viruses would provide a ready resource for identifying viruses associated with any of thousands of species.

Insects (the most numerous organisms in collections) are curated primarily by pinning individuals and printing the basic information on tiny tags beneath the specimens. Until recently, individual catalog numbers were typically not assigned to insect specimens. Identification is often only to the level of order or family, although the better curated collections are known to genus or even species. Cataloguing the nation's insect collections will likely require the development of hardware and software to speed digitization while assuring specimen safety. As a group, the insects (and other small invertebrates) involve hundreds of millions of specimens and present an enormous digitization challenge.

Special specimen preparations, such as marine species (e.g., jellyfish, giant squid), or thin sections generated from modern and fossil specimens, also require special digitization techniques, as do microbes and extremely small invertebrates and protists (e.g., unicellular organisms such as diatoms, yeasts, prokaryotes, microinvertebrates).

Across the major taxonomic groups, innovations in digitizing hardware, software, process engineering, workflows and networked data interactions will be needed. Such innovations will lead to collaborative community approaches sensitive to local needs but able to leverage efficiencies and synergies of industrial-scale collection data processing and publishing.

Strategic Plan For Collections Digitization: Objectives

The key objectives of this strategic plan are:

• ***Digitize data from all U.S. biological collections, large and small, and integrate these in a web accessible interface using shared standards and formats.***

Estimates suggest that there are on the order of 1 billion specimens held in U.S. biological collections, residing in thousands of institutions. A significant number of these institutions have embarked on the digitization of some specimens, but very few of the smaller and none of the larger collections have a complete digital accounting of their specimen holdings and associated ancillary information (e.g. field notes). Although some institutions share data through a common web portal, such collaborative projects are limited to particular regions (e.g., SEINET, http://swbiodiversity.org/seinet/index.php) or a particular taxonomic or thematic group (e.g., VertNet, http://vertnet.org; PaleoPortal, http://paleoportal.org).

• ***Develop new web interfaces, visualization and analysis tools, data mining, georeferencing processes and make all available for using and improving NICA resources.***

In order to hasten completion of the digitization of U.S. collections in a timely and efficient manner, the development of new techniques will be essential to develop new web interfaces, visualization and analysis tools, data mining tools, and georeferencing processes and make all available for using and improving the NICA resources. As the body of digitized specimen data grows, so will the types of questions that this data pool can be used to answer. New tools for analyses of these data will permit a deeper knowledge of species distribution, biological interactions, and response to environmental change and crises management. The Network Integrated Collections Alliance will be a platform on which to build innovative applications that support data improvement such as collaborative georeferencing tools, species identification, data visualization, and data and image analysis. For example, a simple application on a mobile cellular device could access NICA resources and provide information about local biodiversity. Auto-updating analysis and modeling tools that link NICA resources to physical and chemical environmental layers derived from *in situ* sensor networks or remote sensing products provide an unparalleled opportunity to generate a process-oriented view on biodiversity change over space and time.

• ***Create real-time upgrades of biological data and prevent the future occurrence of non-accessible collection data through the use of tools, training, and infrastructure.***

Specimen collection for scientific study continues because many species remain undocumented and large gaps remain in our knowledge of the earth's biodiversity, especially in marine and tropical regions. Reference to the digitized collections repository will help to target future collecting efforts by providing an excellent baseline of inventory completeness. The tools developed for efficient data and image collection will be designed to accept data from field sites, to ensure that data from field collections, including specimen images and genomics data, stream directly into digital archives.

Elements of the Plan

An estimated one billion specimens are held in more than 1600 collection institutions in the United States. As discussed above, the physical preparations of these modern and fossil specimens include skins and hides, wet and dry skeletons, pinned insects, taxidermied mounts, fluid-preserved organisms in vials, bottles or tanks, dried in packets or boxes, pressed on sheets or mounted on microscope slides. The core elements of the plan to accomplish this monumental task are discussed below.

Organization, Leadership, Governance and Collaboration

The organizational and leadership structure will have as its highest priorities:

Accessibility. Enable new science and provide more effective monitoring and regulatory activities by networking collections and mobilizing specimen data to the Internet in order to interconnect and integrate specimen information across laboratory, institutional and governmental boundaries.

Inclusiveness. Maximize the number of collaborating biological collections by creating value for institutional participation and engaging broader stakeholders by addressing their needs for specimen data and their integration.

Efficiency. Recognize and address issues of technology, scale, error-checking, and staging of data computerization activities to maximize quantity and quality of collection data produced and published.

Accuracy. The participation of expert biologists and collections professionals in combination with new technology for image acquisition, data capture, and error-checking routines is critical in order to make the digital resources scientifically accurate and reliable.

National Digitization Hub (NDH): The national hub will serve as the administrative home for the digitization effort, fostering partnerships and innovations, facilitating best practice standards and workflows, serving as a repository for data and techniques, and establishing cohesion and interconnectivity among digitization projects. Through collaborative and inclusive processes, the NDH will: determine the scope and staging of technology development priorities; promote the development of collections-level metadata at all institutions; ascertain priorities for collections institution engagement and collaboration; identify key external partners; measure progress against explicit goals; establish and promote credit for data publishers and editors and; develop a plan for long-term sustainability and report on progress to all stakeholders. It is likely that the NDH will consist of a small number of full-time staff, including an Administrative Director, Cyberinfrastructure Director and Director of Engagement and Outreach. A governing advisory board will have authority for decision making, oversight, and guidance.

Regional and Thematic Collections Networks: Translating the vision of the *National Digitization Hub* and its advisors and collaborators will be Regional and Thematic Collections Networks

which will directly organize and support the digitization effort at a group of institutions. Regional collaborations may consist of institutions housing both large and small collections that are united to focus on digitization and mobilization of collections data from the same geographic area. They may also address scientific or environmental questions pertinent to that region and may share resources and expertise among collections located in proximity to one another. Thematic collaborations may be driven by the specific needs of collections of a particular clade or preservation type, or motivated by a particular scientific question to be addressed by the use of collections images and data. These thematic networks will: define and delineate subprojects for technology development or content generation; identify deliverable goals, metrics for assessment, and specific needs for community support; provide technical support; and strengthen communications and outreach to other collections.

Collections Institutions: Within biological collections themselves, where the information associated with specimens is sequestered, collections researchers and curatorial staff will be incentivized and supported with state-of-the-art technologies and workflows, and with benefits for participation, to value and undertake the kinds of baseline-level digitization activities of the initiative. The heart of this effort will reside in the galleries and among the cabinets and drawers of the thousands of collections that comprise this national enterprise. Collections personnel will prioritize their specimen holdings for digitization, select the technological solutions that are most effective for their collections, and share their data and digitization experiences. They will define the functions and capabilities that add value to the collections and present them to the collection networks and Hub administrators as requirements for the initiative, provide feedback on workflows, and suggest best practices for the project.

Technology Development

In order to achieve the goal of digitizing U.S. biological collections, this initiative must foster the creation of technological innovations to increase the rate of data capture (while maintaining data quality). Such efforts can dramatically lower both the time and cost of specimen digitization. The community must seize the moment to develop specific best practices, standard methodologies for data capture and workflow technologies that can likely achieve orders of magnitude increases in the rate of digitization processes.

Three broad challenges with technology development can be foreseen. The first is the need to develop hardware and software tools for automated data entry, quality control and publication workflows that are generalizable and extensible across different types and sizes of collections. This effort will require expertise from process engineers and from biological and curatorial domain experts to determine data entry process bottlenecks and efficiencies as recommendations for the technology solutions. Goals will be to minimize complexity, cost, and damage to specimens, while maximizing the quantity and quality of specimen data produced and mobilized.

To initiate and further develop technology and workflow optimization, the NDH will create a working group with collections experts and workflow process engineers. The working group will be charged with the rapid creation of a plan that describes where the greatest efficiences in specimen digitization and mobilization can be achieved for the least amount of overhead. The

output of this working group report will be used to prioritize technology development that will occur within the first two years of the project.

A second technology challenge will be developing new systems or evaluating and deploying existing prototype systems that can move to production stage in a short time-frame. In order to digitize as many specimens as possible during the 10-year time frame of this initiative, it is essential that such technologies are ready for deployment within 2 years of the project start. Accommodating institutions that already have successful digitization programs will require careful coordination.

A third technology challenge will be the logistics surrounding deployment, training and helpdesk support of the data entry and publication workflow technologies. These components must have high levels of usability, low learning curves, and be robustly engineered to reflect physical collection and traditional local procedures and protocols for specimen handling.

With limited grant funding, a few biological collections have already designed data entry workflows with innovative technology. For example, the botanical community has developed prototype systems such as Herbis (http://www.herbis.org/) and Apiary (http://www.apiaryproject.org/content/apiary-home) that perform 'one-button' specimen imaging and data capture. Such 'one-button' systems automate all steps in the workflow so that positioning specimens and clicking a button to capture an image are the main human operations. The images and label data are automatically sent to a structured database. These systems show the promise of workflow automation that remove onerous and repetitive tasks often performed by humans, and better done by computers

Workforce development and training

The human resources required for this initiative include faculty and curators, collection managers, information technologists, and a diversity of staff, students, and volunteers who are involved with collections. The goal of the workforce development and training element of the plan is to have a corps of people who are enabled to perform efficient and accurate collections digitization and manage that process according to established standards so that data may be readily shared and integrated with that generated by other institutions. Some of these people may be based in regional centers.

In order to realize the vision of the Network Integrated Collections Alliance, a significant investment in training of collections personnel, ranging from digitizers to collections administrators will be needed. Training will also be required in the use of software and hardware tools that are currently available for collections digitization and mobilization projects, and for those that will be developed in the course of the project. Collections personnel must receive instruction in best practices for data capture, editing for conformance to established data standards, and how to store and share data. Collections administrators will need training in how to develop a workable plan to accomplish their digitization objectives, and how to plan for the long-term maintenance and updating of the digital resource created by their institution.

Scientists and students who generate new collections need training in how to create digital documentation of new specimens as they are collected so that these data immediately become part of the available collections data stream and not part of the backlog of collections to be digitized. Because the national NICA resources created by this plan will be universally accessible, students at the high school and university level should be introduced to it during formal coursework; furthermore, students may represent a significant part of the digitization workforce in university collections, and thus should be trained to participate in this work and in how to use digital collections data to answer research questions. Finally, citizen scientists may contribute significantly to the digitization workforce and they too must have access to training in digitization and best practices.

Training will take place at designated facilities, such as the National Digitization Hub or in lead institutions comprising the Regional and Thematic Collections Networks, and may also be available as a distance-learning activity or offered in conjunction with national society meetings and workshops. Training must be followed up with on-going support, in the form of on-line instruction, document repositories, discussion or help resources using social media, and deployment of problem-solvers who can travel to collections as necessary.

Products

Digitized specimen data: The primary product of coordinated biological collections digitization initiative will be an openly accessible digital archive of the diversity and distribution of life on earth. Stewarded by the individual institutions that curate the specimens and their digital representation, these data will be available over the Internet through multiple interfaces—through institutional, regional, thematic, and national portals, and as content within international biodiversity caches, such as those run by the Global Biodiversity Information Facility (GBIF). Public web interfaces will serve as gateways for education, applied management, and research. The digital specimen data collections resulting from the initiative will be more than just repositories of archived information. Web services and peer-to-peer network communications through standardized application programming interfaces will facilitate cooperative data entry, duplicate specimen discovery, shared authority files, and a new level of interactivity to the desktops of collections researchers and data consumers. Machine-to-machine data transfer will port relevant data from collections to a broad range of applications from scientific to commercial to educational and recreational.

New Applications to Support Rapid Specimen Digitization and Data Mobilization: Hardware and software products to accomplish high throughput specimen digitization, as well as optimized workflows for all types of specimens, will be key products of this initiative. Equally as important will be tools for data editing and standardization, to ensure that data captured at different institutions will be compatible. In order to assure that a backlog of non-digitized specimens does not recur in the future, these products must be designed to be sustainable and scalable into the future. The community can leverage current innovations such as cloud-based data publishing networks. VertNet for example, has begun migrating to the cloud—a virtualized data center—in order to solve current impediments with fully distributed network architectures (see Constable, 2010 for details).

Network Integrated Collections Alliance (NICA) Virtual Communities: An essential product of this plan is a vibrant, virtualized community that can effectively share data and digitization experiences, while collectively developing requirements for new tools. By using the Internet as a research platform, existing and new social media and user/content management approaches will be utilized to allow for persistent collaboratoriums where members of the community can self organize to accomplish the goals of this product. An early example of such an approach is GEOLocate: Community Edition (http://www.museum.tulane.edu/coge/), a collaborative georeferencing application.

The Network Integrated Collections Alliance will build on a durable legacy of protocols and formalisms by which collections have always collaborated with one another in physical space, but NICA will transform those historical methods of cooperation into a highly-interactive, network-based research enterprise.

A Workforce Enabled for 21st Century Life and Work: Digitizing the nation's biological collections will involve a large and diverse workforce ranging from volunteers to students to museum professionals to research faculty. Training in the use of technology for digitizing existing biological collections and for acquisition of new data that can seamlessly enter the digitized collections dataset will empower the workforce for life in a world with continual changes in technology. Through the training and practical experience gained through this project, the participants will be well-equipped for future challenges and will use the experience gained to train the next generation of biodiversity specialists and enthusiasts, as well as make the best possible uses of technology in their personal lives.

Partnerships

Digitizing the nation's natural history collections and managing the resulting digital collections network will be a monumental task that does not end when the specimens are digitized. New collections will continue to be added, new techniques for specimen and data analysis will surely lead to re-digitization of some specimens, and new questions will require new means of delivering collections information. Ensuring the long term success of this venture will require the development of a web of partnerships of the stakeholders for these data. NICA will have a significant association with numerous partner organizations and agencies, discussed in detail below. The first focus is on federal agencies given the value of such data for national priority environmental, scientific, security, and related issues. Then, other stakeholders (e.g., NGOs, state and local governments, educational institutions) who will also dramatically benefit from this initiative are addressed.

The Department of Homeland Security (DHS) is responsible for dealing immediately with acts of terrorism, protection of the borders against the introduction of damaging or deadly pests or pathogens, interdicting shipments of forbidden items, and measuring the impact of security activities on the natural environment. Each of these operations, and many more, require accurate and rapid access to the nation’s NICA resources. DHS will utilize the data continuously to identify protected or endangered species and other contraband involving plants and animals. DHS scientists will be using the database to discern species distributions and

associations of possible pathogenic species and their hosts. Access to data will be instantaneous so that quieries can be made in the field by officers on the line.

The Department of Defense already requires good and rapid data on species associated with operations that occur on military lands and may affect personnel. With the database in operation, military activities throughout the world would have instant and complete information of all specimens that occur in the area and that may impact operations.

The Department of Energy (DOE) requires accurate data on species identifications and distributions to assess the impact of any large energy generating operation, such as wind power, on organisms. DOE may have immediate need for information on species that migrate during particular times of year. Cross-taxa data are difficult to access at present. Should radiation, petroleum, or other accidents occur, models can be developed and predictions made about the effects, spread, and potential damage of such occurrences in the natural environment and in human populations. Should new sources of energy be considered, such as wind energy, possible conflicts with wildlife can be assessed readily.

The Department of Agriculture already maintains many biological collections, but like all such collections at present, the data are not easily obtained and cannot easily be accessed across taxonomic associations. The proposed NICA resources will be used to locate potential biological control agents for pests, predict the spread and influence of damaging species, assist with pest and crop management, control damaging vertebrates, invertebrates, plants, and micro-organisms, and document the status of managed and unmanaged pollinators. The data will also be a primary resource for developing new foods, biocides, and biological control agents. As managers of Forest Service lands, knowledge of verifiable species occurrence is vital to informed decision making in our nation's forests and wilderness areas. Similarly, extension agents will have instant access to species identification for pests, competitors, and other organisms that impact agricultural lands.

The Department of Commerce will be a major user of these data, especially through *NOAA* activities related to global climate change research and marine sciences. The data to be made available through the Network Integrated Collections Alliance are the finest record of organisms across time in numerous climatic situations. The complete database will be an important record of species distributions associated with climate changes. The largest collections of marine organisms reside in the collections, so any attempts to describe oceanic biodiversity and identify trends will require access to the totality of the nation's collections, something not possible at this point in time.

The Department of Interior includes the Bureau of Indian Affairs, Bureau of Land Management, Bureau of Reclamation, Minerals Management Service, National Park Service, Office of Surface Mining, U.S. Fish and Wildlife Service, and the U.S. Geological Survey which hosts the Integrated Taxonomic Information System and the National Biological Information Infrastructure, which in turn is home to the U.S. node of GBIF. The DOI is a natural partner in terms of the large number of collections that it maintains, the integral part that collections data plays in its management of land and natural resources, and its mandate to provide infrastructure and leadership in the management of national biological information.

The Department of State is constantly monitoring threats to US citizens overseas or travelers returning to or entering the United States. Biological data provide a rapid assessment of the threat of the spread of disease or other contaminants in foreign countries and the likelihood that such organisms pose a danger to the nation.

The Department of Health and Human Services, which includes the Center for Disease Control and the National Institute of Health, will utilize natural history collections data in multiple ways. Specimen records can provide essential baseline information on zoonotic disease transmission, for example. Digitized specimen records will serve as essential vouchers for much of the organismal data available via NIH-supported Genbank. The many collections of genomic resources will be readily linked to provide access to frozen living tissues specimens from thousands of species worldwide, as well as any viral or pathogenic associates of those species.

The Department of Justice will utilize the data continuously to assist with forensic investigations (identifying insects, bacteria, fungi, and many other organisms associated with a crime scene), as well as the vast collections of genomic resources linked to NICA resources which will provide genetic sequence data for research related to a crime.

The Department Transportation is often involved in the unexpected transport of pests across the nation. Efforts to control such possibly disastrous introductions or dispersals will require regular access to the collections database.

The Environmental Protection Agency is charged with protecting the environment which involves large efforts to monitor the effects of environmental change on all biota. From invasive species tracking to the characterization of ecosystem and biotic diversity change in response to pollution and climate change, there is a significant need for the Network Integrated Collections Alliance.

The Smithsonian Institution, the nation's national museum, is the largest collection repository in the United States. However, its collections, too, are not completely digitized and available. As the largest collection in the Network Integrated Collections Alliance, the Smithsonian would be expected to be a primary provider and user of collections data, directly benefitting the many federal and non-federal agencies, organizations, and individuals that require access to such data.

The possibility for partnerships that could support the digitization of national biological collections extends beyond the Federal government. For example, data from NICA resources are required each day by almost all non-governmental organizations (NGOs) that deal with environmental issues. Whether one is planning or selecting new reserves to protect species or habitats, measuring the effects of habitat destruction on ecosystems and individual species, seeking indicator species for such vital issues as extinction, determining changes in distributional patterns, identifying keystone species for special protection efforts, assessing bellwethers of global climate change, or devising long-term conservation strategies, georeferenced and associated data of species past and present are required. Rapid access to such data allows timely responses to threats and greater precision when developing plans to conserve nature.

As with governmental departments and NGOs, a digitized database of biological organisms is vital to a host of state and private organizations that require species distribution data, species associations, and other data related to plants, animals, and microbes. Private businesses—from large petroleum, logging, fishing, and real estate development corporations to private businesses and individual entrepreneurs concerned with such activities as tourism, sports fishing, farming, pollination services, and hunting—need rapid access to reliable information on nature's species and habitats.

State and local government (e.g., state and county wildlife, forestry, and health departments) constantly require extensive information on species distributions and habitat and landscape associations, whether dealing with issues of fishing and hunting, agricultural extension programs, water quality, state and county health department needs, and quality of life matters of importance to such groups as local chambers of commerce.

Providers of educational services have a direct interest in the data that will be made available through this project, as do landscapers, artists, television and film producers and historians. Nature lovers at all levels, from professional biologists to bird watchers, journalists, wildlife artists, and photographers would be regular users of and contributors to a digitized database of nature. Automated comparisons of images and observational records placed online by the public and validated through comparison with verified specimen records to be captured by this project will allow citizen science data to be mobilized to an unprecedented extent while maintaining professional standards of data quality. The NICA resources envisioned here can be a powerful tool for increasing science literacy, which empowers Americans to make informed decisions in their personal lives and at the polls.

This new and powerful association of museums and collections that will develop in the United States will link seamlessly with efforts that are further advanced in Europe and Australia. Such organizations as the National Biological Information Infrastructure (NBII) and Encyclopedia of Life (EOL) in the United States, and the Global Biodiversity Information Facility (GBIF), will be continuous users of these data. GBIF in particular has paved the way for this initiative through the creation of its Data Portal (http://data.gbif.org), which demonstrated that large data sets from a range of collections institutions could be mobilized effectively. GBIF has also devoted a great deal of effort to developing and promulgating standards and best practices, and data editing techniques. GBIF's newly published "Global Strategies and Action Plan for the Digitization of Natural History Collections" will be very influential in the development of the Network Integrated Collections Alliance. Collaborations for these and other international organizations (e.g., the Atlas of Living of Australia; http://www.ala.org.au/) can be foreseen to support continued development of digitization and mobilization best practices.

Most profoundly affected by NICA will be science itself, particularly science that is specimen based or that discerns patterns in nature to understand the past and predict future trends. Ecologists, marine biologists, botanists, crop scientists, pharmaceutical researchers, and wildlife biologists, among others, will benefit enormously as new patterns of species and habitat associations develop when investigators are able to formulate questions across taxa and across geographic boundaries to quickly seek new and unexpected patterns of species coexistence, interactions, evolutionary trends, distributional changes, or species associations. Evolutionary

biology will be a great beneficiary of these data as vast amounts of associated morphological, ecological, and genetic data become available, in addition to information on present and past geographic distributions. Medical researchers will examine distributions over time, as well as assess the possibility of the spread of a disease into new geographic areas and into new hosts. In summary, the Network Integrated Collections Alliance will have significant, positive impacts on biological science, since the data will be immediately applicable to everything from basic taxonomy to astrobiology, from contemporary understanding of nature to the history of the development of life's diversity.

Long-Term Sustainability

It is essential that the U.S. Network Integrated Collections Alliance be maintained for society in perpetuity. Doing so requires a sustainability plan past the proposed ten-year time frame. One of the first priorities for the National Digitization Hub will be to develop a sustainability plan that will be comprehensive and adaptive, given a changing technological and social landscape.

Multiple avenues for sustainability can be considered. The National Science Foundation has made significant investments in DataNets to achieve long term digital presevation. One of the first funded DataNets, called DataOne (http://dataone.org), is explicitly focused on environmental data. A partnership with DataOne would be one obvious avenue for sustaining the technological infrastructure and data generated by this project. Forming strong partnerships with mission-oriented federal agencies, who already manage collections or serve as national nodes for sharing species data (e.g., National Biological Information Infrastructure), would be another natural collaboration.

A key component of this plan is developing technologies that increase the rate of digitization, and that provide simplified, scalable, and sustainable data publishing networks. Providing scalable and sustainable data publication methods may rely on commercial data centers (e.g., the cloud). Other commercial partnerships are possible that can leverage resources and expertise that may be limiting within our community or within the sciences more broadly.

Sustainability is greatly enhanced through continued development of data curation as an emerging discipline that links traditional library and information science with management of an explosion of new types of biological data (e.g., genomics and proteonomics data). Training and workforce development are initial and essential steps towards formal career opportunities that become integrated into the fabric of academic institutions.

Community Involvement in this Planning Effort

Creating NICA resources will require intellectual input from collections curators and researchers and an understanding of diverse perspectives and requirements from all collection data stakeholders. The initial draft of the digitization plan was shared with the community via wide distribution to individuals, institutions, agencies, and professional societies. A summary of the full community input process is below.

Feedback from the biological collections community was brisk and constructive, generating many discussions in 60+ responses posted to the project's blog (http://digbiocol.wordpress.com/). Detailed responses were received from a diversity of collections professionals and biologists, from both large and small institutions, and from representatives of societies, departments, and administrators across the nation. Collections community enthusiasm is high, with replies such as "this is a fantastic proposal that would have a tremendous impact on natural history collections in the U.S." and "a national collection resource like this will make new kinds of biodiversity science possible". Others saw the value to both scientific and public audiences: "this project would be of great value to multiple communities and many different kinds of users with disparate goals—from schoolchildren seeking to learn about species found in their geographic area to scholars investigating biodiversity". The Society for the Preservation of Natural History Collections (SPNHC), an international membership of over 600 professionals including collection curators, collection managers, conservators, and registrars, submitted a letter strongly endorsing the concept. They wrote "Our Society enthusiastically supports an initiative that increases access to collections and that promotes novel uses of specimens and specimen data". Strong letters of support for the final strategic draft plan were also received from the American Institute of Biological Sciences (AIBS) and the Natural Sciences Collections Alliance (NSCA).

Stakeholders have also provided perspective and caution. The primary concern voiced was that the logistics and electronic products of this digital initiative do no harm to the physical specimens. Other voices argued for identification and collections curation activities be supported within the context of the initiative, so as to maximize the quality of the resulting digital resource and to help with preservation of the actual specimens upon which the digital data are derived. An articulate blogger wrote that the plan "seems to reflect a vision of one-way data mobilization: from the dusty shelves of museums to the eager hands of scientists. I would argue that this initiative should explicitly allow and encourage informatics research for the purpose of collections improvement, with the goal of positive feedback between specimen curation and taxonomy." This interaction between the digital resource and the physical specimens is clearly an important component of an accurate and vital collections data resource; it should be recognized as an important principle of the initiative.

The public comment period remains open; additional perspectives are cordially solicited and highly valued. They will contribute to the further development of this plan. Continued community feedback on the initiative outlined here is critical. Professional opinion can still be transmitted by adding a comment to the blog page (http://digbiocol.wordpress.com/). Perspectives on institutional priorities or taxon-based needs are welcomed. Specific feedback is needed to critique or amplify the proposed organizational model, to offer suggestions for revision, priorities for collection digitization, and to suggest ways to maximize collaboration across institutions and federal agencies, and at the international level. Ongoing discussion will continue to be aggregated as the effort to establish NICA resources continue. The approach for broad involvement, across the biological collections community and among stakeholders that has been integral to the shaping of this strategic plan will continue to be a hallmark of the effort.

A draft final plan was distributed to all participants of both NESCent workshops for comment and feedback was received from multiple contributors. This feedback led to the development of

this final plan, which is being delivered to federal agencies for further action. Once delivered, the community will continue to discuss and further develop efforts related to this plan in the form of additional workshops and meetings.

This outstanding transformational opportunity to digitize the specimens held by the nation's biological collections is matched in significance only by the massive societal, govermental, and research investment that has already been made with over two centuries of U.S. exploration and curation. The nation's biological collections institutions are prepared to contribute their knowledge to the most pressing science and environmental issues of our day.

Leveraging this national and monumental investment in the information contained in biological collections by digitizing and mobilizing specimen data to the Internet will renew the long-standing shared vision and purpose of biological collections institutions and take them to a new era of research communication, collaborative research, and societal engagement, while impacting science in all disciplines and at all levels.

Appendices

Appendix I: Terms

Specimen: an item of biological origin that is stored in a collection. A specimen is documented with information about the name of the species, where, when and by whom it was collected.

Collection: a set of specimens held by an institution. Institutions holding collections include museums, herbaria, universities, government agencies.

Collections Digitization: broadly defined to include transcription into electronic format various types of data associated with specimens, the capture of digital images of specimens, and the georeferencing of specimen collection localities,and other associated data quality enhancement activities.

Appendix II. Relevant Reports and Scientific Publications (in order cited in main text)

Report from the National Science Foundation based on a survey of collections which had received federal support for projects over the past twenty years http://www.nsf.gov/pubs/2009/nsf09044/nsf09044.pdf

Report from OSTP and the Interagency Working Group on Scientific Collections based on the survey of federally-held collections: http://www.nescent.org/wg/digitization/images/d/d1/Collections2.pdf

Stevenson, J. W. and D. W. Stevenson. 2003. Development of a national systematics infrastructure: a virtual instrument for the 21st century. Report to the National Science Foundation, Biodiversity Surveys and Inventories Program. New York, December, 2003.

Constable, H., R. P. Guralnick, J. Wieczorek, C. Spencer, A. T. Peterson and the VertNet

Steering Committee. 2010. VertNet: A New Model for Biodiversity Data Sharing. *PLoS Biology* 8(2): e1000309

Moritz, C. et al. 2008. Impact of a Century of Climate Change on Small-Mammal Communities in Yosemite National Park, USA. Science 322: 261-264.

Dove, C. 1999. Feather identification and a new electronic system for reporting US Air Force bird strikes. Paper posted at DigitalCommons@University of Nebraska - Lincoln. http://digitalcommons.unl.edu/birdstrike1999/13.

Yates, T. E. et al., 2002. The ecology and evolutionary history of an emergent disease: Hantavirus pulmonary syndrome. BioScience 52: 989-998.

Page, L., Funk, V., Jeffords, M., Lipscomb, D., Mares, M., and A. Prather. 2004. Workshop to produce a decadal vision for taxonomy and natural history collections, Gainesville, November 2003. Report to the National Science Foundation, Biodiversity Survey and Inventories Program, Gainesville, November, 2003.

Page, L., et al. 2005. LINNE: Legacy Infrastructure Network for Natural Environments. Illinois Natural History Survey Publication, Pp. 1-16.

Appendix III. Workshop Participants and Funding Support

NESCent digitization workshop I (Feb 5-7th 2010) participants: Hank Bart, James Beach, Stan Blum+, Andy Deans, Michael Donoghue*+, Linda Ford+, Gerald Guala, Rob Guralnick+, Pat Holroyd, Jennifer Leopold, Michael Mares*+, Chuck Miller, Bob Morris, William Piel, Babara Thiers+, Todd Vision, Mark Westneat+, Quentin Wheeler+, Tim White, Brian Wiegmann.

* - *Workshop leaders,* + *Initial strategic plan draft writing*

NESCent digitization workshop II (April 28-30th 2010) participants: John Ascher, Mark Barkworth, Hank Bart, James Beach+, Joel Cracraft, Andy Deans, Linda Ford+, Gerald Guala, Rob Guralnick*+, Mari Kimura, Michelle Koo, John Long, Michael Mares+, Bob Morris, Paul Morris, Christopher Norris, William Piel, Alan Prather, Kate Rachwal, Randall Schuh, Barbara Thiers*+, Paul Tinerella, Mark Westneat+, Tim White, Brian Wiegmann*+, Jean Woods.

* - *Workshop leaders,* + *Final strategic plan writing*

Workshop funding support to: Allen Rodrigo (workshop I), and Andy Deans and Brian Wiegmann (workshop II).