Strategic Plan for Integrated Digitized Biocollections (iDigBio)

Prepared By:

David Jennings, iDigBio Project Manager

Dr. Larry Page, iDigBio Project Director

Dr. Bruce MacFadden, iDigBio Principal Investigator – Education & Outreach

Dr. Pam Soltis, iDigBio Principal Investigator – Serving the Research Community

Dr. Greg Riccardi, iDigBio Principal Investigator – Digitization

Dr. Jose Fortes, iDigBio Principal Investigator – Cyberinfrastructure

Revision Date:

6/1/2015

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

1. Project Overview

Integrated Digitized Biocollections (iDigBio) is the national coordinating center for the Advancing Digitization of Biodiversity Collections (ADBC) program funded by the U.S. National Science Foundation (NSF). iDigBio is enabling digitization of data and media for millions of biodiversity specimens from U.S. collections and is integrating those data to make them available online for the research community, government agencies, students, educators, citizen scientists, and the general public to promote understanding of biodiversity and societal consequences of environmental issues.

1.1 Vision

The vision for ADBC is a permanent repository of digitized information from all U.S. biodiversity collections that leads to new discoveries through research and a better understanding and appreciation of biodiversity through improved outreach, which then leads to improved environmental and economic policies. Creation of the permanent digitized repository is occurring in four stages:

1. An initial stage where the effort to digitize U.S. biodiversity collections is catalyzed by funding from NSF and enabled by iDigBio activities that foster collaborations, identify priorities, demonstrate the value of biodiversity and collections, and generate information on best practices related to standards, workflows, and data management.

2. An intermediate stage where digitization at Thematic Collections Networks (TCNs), Partners to Existing Networks (PENs), and other participating institutions/networks improves methods and strategies and demonstrates the scientific and societal benefits of validated and readily accessible data.

3. A third stage in which the vision for ADBC is realized through participation by all U.S. institutions with biodiversity collections.

4. A fourth stage in which digitization is a routine and sustained practice in all institutions with biodiversity collections, and the national database is easily accessible as an up-to-date source of information on biodiversity.

1.2 Mission

The mission of iDigBio is to develop a national infrastructure that supports the vision of ADBC by overseeing implementation of standards and best practices for digitization; building and deploying a customized cloud computing environment for collections; recruiting and training personnel, including underserved groups; engaging the research community, collections community, citizen scientists, and the general public through outreach activities; and planning for long-term sustainability of the national digitization effort.

1.3 Horizon

iDigBio is enabling digitization of data from all U.S. biodiversity collections and is integrating those data to make them broadly available and useful with shared standards and formats. Ultimately, ADBC is

furthering the discovery and understanding of biodiversity, and iDigBio is engaging the research and collections communities in a spirit of collaboration in an effort to open biodiversity research collections to new downstream user communities.

iDigBio involves the development of a permanent and powerful cloud computing infrastructure to link biodiversity data from collections across the U.S. into a single unified web interface, overcoming the "data silos" that exist across the country. Search and analytical tools enable users to mine diverse data including taxonomy, geographic location, 2- and 3-dimensional images, vocalizations, and molecular resources tied to specimens in collections. These data promote integrative biodiversity research on living and fossil species and provide an immense resource for agricultural science and land use management, human health, and assessing the impacts of climate change, invasive species, and other natural resource management issues.

Key partners in this effort are the Thematic Collections Networks (TCNs), which form a national grid of institutions that are digitizing specimens and associated resources. Integration with the greater community of biocollections resources, tools and organizations is critical to accomplishing the grand challenge of digitizing and integrating data from all U.S. collections, large and small. For more details regarding the larger community that encapsulates ADBC, please refer to the Network Integrated Biocollections Alliance (NIBA) strategic plan and implementation plan.

2. Project Organization

The iDigBio project is organized into five domains to achieve its goals: (1) Project Administration & Management, (2) Education & Outreach, (3) Serving the Research Community, (4) Digitization, and (5) Cyberinfrastructure. The Principal Investigators (PIs) provide leadership within each domain, and the Project Director and Project Manager provide overall leadership for the project.

2.1 Project Oversight

The project incorporates several oversight committees to provide governance, advice, and leadership concerning goals, strategies, implementation, activities, and progress. The Project Director, in consultation with the Project Manager, Executive Committee, and Steering Committee, makes project decisions that affect scope, budget, and/or risk. The Project Manager effects day-to-day planning, execution, and decision-making.

2.1.1 Executive Committee

The Executive Committee (EC) is composed of the iDigBio Project Director, Principal Investigators, Project Manager, and key Senior Personnel. The Executive Committee is responsible for overall project management, general oversight of iDigBio activities, managing conflicts of interest, and implementation of the strategic plan, including assurance that the digitization, research, training, and outreach missions of iDigBio are integrated and accomplished.

2.1.2 Steering Committee

The Steering Committee (SC) is composed of the iDigBio Project Director, Principal Investigators, Project Manager, and Senior Personnel. The Steering Committee is responsible for reviewing progress and coordinating activities in digitization, research, training, and outreach at iDigBio, and for advising others in iDigBio on overall resource allocation, strategic directions, and management policies.

2.1.3 Internal Advisory Committee

The Internal Advisory Committee (IAC) is composed of the Project Manager, Biodiversity Informatics Manager, and representatives from the Thematic Collections Networks (TCNs) and Partners to Existing Networks (PENs). The IAC meets regularly to report on progress in digitization efforts, share and identify best practices and standards, identify gaps in digitization areas and technology, and enhance training efforts.

2.1.4 External Advisory Board

The External Advisory Board (EAB) is composed of members selected by iDigBio and approved by the NSF program officer. The EAB meets annually and is responsible for advising iDigBio on its strategic directions, management policies, and activities, including progress and integration of digitization projects, research, training, and outreach activities among all funded institutions.

3. Strategies and Objectives

3.1 Project Administration & Management

iDigBio Project Administration & Management decisions and activities are executed under the leadership of Dr. Larry Page, iDigBio Project Director, who is based at the University of Florida's Florida Museum of Natural History (FLMNH). The Project Administration & Management domain provides administrative and logistical support to the other project domains and ensures that activities are coordinated within and across iDigBio domains and with collaborative organizations as appropriate. The key objectives of the Project Administration & Management domain are:

a. Coordinate project oversight committees and other key meetings;
b. Plan for long-term sustainability, including participation with NIBA;
c. Comply with NSF cooperative agreement and reporting requirements;
d. Build and maintain strategic partnerships;
e. Plan and monitor project budget and expenditures;
f. Implement a structure to measure progress against goals;
g. Manage internal and external communications, including broad dissemination of outcomes;
h. Develop and maintain collaboration and communication capabilities; and
i. Serve as the central resource for ADBC and the collections community, including promotion of cohesion and interconnectivity.

3.2 Education and Outreach

iDigBio Education and Outreach (E&O) decisions and activities are executed under the leadership of Dr. Bruce MacFadden, iDigBio co-PI, based at who is based at the University of Florida's Florida Museum of Natural History (FLMNH). iDigBio E&O activities are focused on digitization curricula development, stakeholder identification, E&O materials and protocols, and public speaking engagements, which are recorded and published for national impact. The key objectives of the iDigBio E&O domain are:

a. Foster project awareness within the scientific and collections communities;
b. Engage the public and collections community through resources and opportunities that highlight the importance of biodiversity collections and digitization;
c. Identify and assess the needs of target audiences, downstream user groups, and other stakeholders, and execute outreach activities to meet those needs;
d. Identify and assess the needs of downstream partners and stakeholders, and execute outreach activities to meet those needs;
e. Develop outreach resources related to digitization and biodiversity; and
f. Measure and track the impacts and outcomes of outreach efforts.

3.3 Serving the Research Community

iDigBio decisions and activities related to serving the research community are executed under the leadership of Dr. Pamela Soltis, iDigBio co-PI, based at who is based at the University of Florida's Florida Museum of Natural History (FLMNH). The Research domain leverages existing relationships to deliver

branding and messaging related to iDigBio to foster collections/research community adoption of iDigBio's services, infrastructure, tools, resources, and data including: specimen portal, website, Wiki, Listservs, Adobe Connect teleconferencing, Workshop coordination/funding, Working Group coordination/funding, Appliance development, and Hosting of services. In addition, the Research domain highlights developments and opportunities for research using specimen data alone or in conjunction with other types of data. The key objectives of the Research domain are:

a. Engage the research community to promote community adoption of iDigBio services, infrastructure, tools, resources, and data;
b. Promote and facilitate both traditional and novel uses of specimen data;
c. Produce use cases of research applications of specimen data, provide them to the Cyberinfrastructure team, and help validate the effectiveness of implementation;
d. Seek opportunities for integration of iDigBio specimen data and API services with key data and research services from other projects and organizations;
e. Identify strategic partners in the research and collections community and develop synergistic relationships with those partners; and
f. Measure and track the outcomes of research use efforts.

3.4 Digitization

iDigBio decisions and activities related to development and optimization of digitization workflows/processes, digitization documentation, and efforts to share/improve digitization tools within the collections community are executed under the leadership of Dr. Greg Riccardi, iDigBio co-PI, based at Florida State University's Institute for Digital Information and Scientific Communication (iDigInfo). iDigBio digitization experts work with the community to understand gaps in workflows, processes, practices, and tools that prevent effective and efficient specimen digitization. These digitization-related activities are catalyzed by site visits, virtual and on-site workshops, working groups, contact with tool developers, and user contact to foster a high-degree of community involvement. The observations and subsequent analyses from these activities yield documentation, papers, training materials, and presentations that promote effective digitization practices and workflows, including qualitative and quantitative measures of success. The key objectives of the Digitization domain are:

a. Engage the collections community to market and build interest in adopting iDigBio services, infrastructure, tools, resources, and data;
b. Establish minimum information standards and data fitness for use parameters;
c. Develop, optimize, and disseminate digitization workflows;
d. Conduct digitization training and produce online training materials;
e. Increase awareness and utility of digitization tools and resources that can improve efficiency and scalability of digitization efforts;
f. Evaluate, document, and publish analyses related to digitization hardware and software tools; and
g. Identify gaps, bottlenecks, and challenges in digitization efforts, standards, and best practices.

3.5 Cyberinfrastructure

iDigBio Cyberinfrastructure decisions and activities are executed under the leadership of Dr. José Fortes, iDigBio co-PI, based at the University of Florida's Advanced Computing and Information Systems (ACIS) laboratory. Cyberinfrastructure decisions are informed by the experience and expertise of ACIS personnel, key stakeholders in iDigBio, biodiversity community input, and information technology community input. The Cyberinfrastructure implementation process follows a formula that balances strategic planning with the agility to meet new challenges, short-term project needs, and enhanced/clarified specifications. The cyberinfrastructure team releases new features and upgrades to the specimen portal on a semi-annual basis. The key objectives of the Cyberinfrastructure domain are:

a. Research, implement, and maintain a scalable cloud infrastructure for text (data/metadata) and object (media) storage;
b. Implement infrastructure to enable hosting of web services and/or websites for strategic partners, such as TCNs;
c. Deploy iDigBio appliances and services via multiple channels (e.g., web services, locally-run virtual machines, infrastructure-as-a-service cloud implementations) to enhance, simplify, and/or improve activities of data providers and data consumers;
d. Implement a comprehensive authentication and access control system to enable data tracking and a cohesive user experience among iDigBio systems;
e. Develop, implement, and maintain iDigBio APIs to access text and media data stores;
f. Develop, implement, and maintain a Graphical User Interface (GUI) to provide end-users, including data contributors and data consumers, access to search, visualize, and download text and media data from the cloud infrastructure (i.e., the Portal);
g. Integrate iDigBio services and user portals with key strategic partners and other collaborators;
h. Secure infrastructure resources to maintain adequate performance and capacity;
i. Serve as a central site for aggregation of digitized collections data; and
j. Plan for the management and long-term preservation of iDigBio's digital data.

4. Assessment and Evaluation

iDigBio's Project Evaluator, a professional with training in social science and expertise in evaluation, oversees assessment and evaluation of iDigBio-supported activities. The evaluator employs a multi-method approach including participant observation, surveys, interviews, analysis of project records, and tracking of data use in publications. The evaluator participates as an observer in annual Summits and Retreats, meetings of the Steering Committee, Internal Advisory Committee, External Advisory Board, periodically in IT meetings, and as an advisor in Core Team meetings and Working Groups (e.g., E & O and Georeferencing). The evaluator collaborates with iDigBio personnel and partners (e.g., Data Carpentry, Small Collections Network, and WeDigBio) on the development of instruments for needs assessment and evaluation and provides summary reports. The evaluator also conducts evaluations of project management, leadership, and communication.

The purposes of iDigBio's assessment and evaluation efforts are to:

a. Document progress against goals in the five domains: (a) project management, (b) education and outreach, (c) serving the research community, (d) digitization, and (e) cyberinfrastructure;
b. Perform assessments to determine the needs of iDigBio's various constituencies, stakeholders, and audiences in order to develop effective tools, services, training, outreach, and communication;
c. Conduct formative evaluation of iDigBio efforts spanning the five domains listed above; and
d. Measure the impact of iDigBio on the national effort to digitize natural history collections.

Key metrics used to evaluate the impact of iDigBio include (but are not limited to):

a. Effectiveness in building cohesion among the digitization community;
b. Awareness of the collections digitization effort among the collections and scientific community;
c. Engagement of the public and collections community in digitization efforts related to biodiversity;
d. Engagement with resources related to digitization and biodiversity;
e. Understanding the value of digitization of biodiversity collections;
f. Use of iDigBio services, infrastructure, tools, resources, and data by the research community;
g. Rates of digitization;
h. Efficiency of digitization;
i. Use of digitization best practices; and
j. Use of the specimen portal by down-stream users.

5. Environmental Scan

The following is a summary of the major strengths, weaknesses, opportunities, and threats that have been highlighted by the External Advisory Board, NSF Site Review team, and surveys of the collections community:

	POSITIVE	NEGATIVE
INTERNAL	**Strengths – iDigBio has…** • Successfully engaged with the collections community, particularly through workshops and working groups; • Developed a vast repository of resources and information; • Developed and optimized digitization protocols and workflows; • Actively reduced barriers to digitization; • Demonstrated strong leadership and management; and • Fostered partnerships, networks, and collaborations within the community and has developed significant synergy with TCNs.	**Weaknesses – iDigBio has…** • Non-intuitive website navigation and inconsistent arrangement of resources; • Minimal data in specimen portal; • Inconsistent internal communication; • Insufficient involvement of smaller collections and institutions; • Insufficient effort to promote diversity and include under-represented groups; • Lack of explicitly defined of project leadership evaluation plans; and • Poor coordination of outreach activities across ADBC.
EXTERNAL	**Opportunities – iDigBio should…** • Highlight research uses of data; • Develop digitization standards and encourage compliance; • Advocate for greater digitization funding and formation of cooperative groups across institutions and projects; • Coordinate efforts across institutions, disciplines, and agencies to bridge initiatives and facilitate networking; • Integrate directly with collection management software and tools; • Capture “dark data” from both large and small institutions; and • Collaborate with key existing projects, such as Global Biodiversity Information Facility (GBIF) and Biodiversity Information Serving Our Nation (BISON).	**Threats – iDigBio should…** • Develop a strategy for financial sustainability to reduce dependence on federal funding; • Develop a strategy for long-term storage and maintenance of its cyberinfrastructure, metadata and images; • Be mindful of its extremely broad project scope, which includes a vast array of collaborations and partnerships; and • Be mindful that iDigBio depends on a number of external projects, tools, and initiatives whose long-term sustainability could be questionable.

6. Sustainability

To ensure long-term sustainability of the national digitization effort, iDigBio is implementing strategies to maintain its strengths, address its weaknesses, capitalize on available opportunities, and mitigate potential threats.

6.1 Maintaining Strengths

The Integrated Digitized Biocollections (iDigBio) project has made significant progress since the initiation of funding in 2011. iDigBio's innovations include both sociological and technological accomplishments with wide-ranging benefits to the collections community.

First, iDigBio has established successful ***communication*** between the Information Technology (IT) and biodiversity collections communities. Having bridged this "cultural" barrier, iDigBio personnel are working together to identify challenges and to design appropriate solutions. This communication extends beyond the personnel specifically working on iDigBio to other partners, such as the TCNs, which allows for collaboration, synergy, and effective training throughout the community.

Second, perhaps the most successful innovation of iDigBio is the series of ***training workshops*** that have been organized and sponsored by iDigBio personnel. These workshops have delivered effective training on digitization-related methods and practices, as well as on other topics contributed by the workshop participants. These workshops and training materials, publically available on the iDigBio Wiki, have provided a wealth of new resources and have secured iDigBio's leadership role in ***workforce development*** within the collections community. During its first four years, iDigBio has sponsored 58 workshops, summits, symposia, and other events and that have been attended by 2,094 participants (1,160 of which are unique) from 485 unique institutions to its. In addition, iDigBio has held 31 webinars that have been attended by 1,103 participants.

Finally, the forthcoming ***availability of massive amounts of specimen data*** has energized the collections community about the use of specimen data for a variety of big research questions that have been intractable to this point. This renewed energy within the community has fostered iDigBio to produce significant innovations in IT design and implementation, including:

- Creating the practice of introducing identifiers in the data stream to enable data linking
- Development of emerging data models for ingestion and integration of data sets from diverse collections
- Cloud architecture for data storage, retrieval, and management
- Open schema infrastructure to offer flexibility and agility in handling an evolving data model
- Appliance framework to respond to the needs of biocollections informatics

6.2 Addressing Weaknesses

iDigBio has recognized that its website and specimen portal are the primary interfaces between the project and its audiences. Completely revised versions of the iDigBio website and specimen portal were released to the community in December 2013, and minor changes have continually been made since

then. The website was redesigned to focus on making it easier to understand and to use and to be more approachable to a lay visitor. The specimen portal and APIs were redesigned to create the foundation for a system that will serve the community for years to come, including user interface improvements and improved stability and flexibility.

Availability of specimen data in the portal is a top priority, and iDigBio has worked diligently to improve the efficiency of the data mobilization and ingestion process. iDigBio has the goal of ingesting data sets within 2 weeks of data mobilization completion. Data mobilization is dependent on the provider, although iDigBio staff continually track the process from initiation to closure using Redmine. From December 2013 thru May 2015, the number of collections in the portal has increased from 121 to 438 (↑262%), the number of specimen records has increased from 4,410,237 to 28,036,830 (↑536%), and the number of media records has increased from 1,005,679 to 4,511,503 (↑349%).

iDigBio senior personnel have acknowledged that communication, both within iDigBio and with our clients, could be improved. iDigBio believes that less-than-ideal communication is the result of having a small staff charged with multiple tasks who are unable to find the time to report on activities in a meaningful way and in a reasonable timeframe. In response, iDigBio hired a Communications Coordinator who was charged with tracking iDigBio activities and broadly disseminating information about those activities. The Communications Coordinator improved communications by defining categories of participating audiences and then focusing high-impact messages tailored for those audiences using a variety of media outlets. iDigBio personnel are making a concerted effort to write and post concise reports on our website in a timely fashion, and the Communication Coordinator announces their release using Twitter, Facebook, and other more traditional means.

iDigBio has initiated efforts to provide national leadership in developing methods to enhance the cultural and ethnic diversity of participation in the biodiversity sciences. iDigBio is hosting a series of workshops targeted at undergraduate students in underrepresented populations focusing on opportunities for careers and graduate study in environmental biology, biodiversity, ecology, and evolution.

iDigBio has initiated efforts to annually evaluate the project and project leadership. The project is evaluated by the collections community to gauge interest and involvement with iDigBio and to solicit input on iDigBio's progress. The project leadership is evaluated by iDigBio staff and students, giving them an opportunity to share their views regarding their own roles and responsibilities and to solicit input on project leadership and management.

iDigBio is actively recruiting an Education & Outreach Coordinator who will be responsible for coordinating and implementing the Education & Outreach (E&O) activities of iDigBio. In addition, the Education & Outreach Coordinator will be responsible for communicating, coordinating, and networking across ADBC (i.e., iDigBio, the TCNs, and other stakeholders) to promote, encourage, develop, and implement E&O activities that achieve broad reach and high impact.

6.3 Capitalizing on Opportunities

iDigBio is working to highlight research uses of the data in the specimen portal. iDigBio has begun implementing features in the specimen portal that are based directly upon input from biologists and how they anticipate using the data. For example, iDigBio is working to implement advanced mapping features in the specimen portal and is working with GenBank to create links between specimen data and genetic data. In addition, iDigBio is actively recruiting a Data Management Coordinator to facilitate the use of natural history collections data in addressing big-science questions by integrating tools and services into a computational environment for data integration, analysis, and visualization. The Data Management Coordinator will act as a liaison between museum collections staff/researchers and IT/cyberinfrastructure staff/developers; develop and document use cases for the use of natural history collections data, including translating these into cyberinfrastructure requirements; prioritize requirements for iDigBio's cyberinfrastructure and data products; provide advice on cyberinfrastructure recommendations and direction; provide advice on cyberinfrastructure data standards and linked data; and serve on the iDigBio Steering Committee to provide strategic advice and direction.

iDigBio is continuing to advocate for, coordinate with, and facilitate networking among the collections community through its involvement in professional meetings and societies. iDigBio is also working to continue direct representation in key organizations, such as The Society for the Preservation of Natural History Collections (SPNHC), Natural Science Collections Alliance (NSCA), and the American Institute of Biological Sciences (AIBS).

iDigBio is continuing to support a variety of Working Groups and Interest Groups. iDigBio is allowing these groups to form organically to dynamically address the needs of the collections community. Some of iDigBio's groups are focused on delivery of short-term objectives, while others are tasked with ongoing research, development, and improvement activities.

iDigBio is continuing to develop its partnerships with Specify, Symbiota, and KE EMu to introduce globally unique identifiers and to create export features specifically tailored to iDigBio. The goal of these partnerships is to streamline the path to data ingestion with iDigBio.

iDigBio has started to mobilize data sets beyond the TCNs by working with the Global Registry of Biorepositories (GRBio) to make contacts with U.S. institutions outside the TCNs. In addition, iDigBio is advancing its existing partnerships with the Global Biodiversity Information Facility (GBIF) and Biodiversity Information Serving Our Nation (BISON).

6.4 Mitigating Threats

iDigBio, in concert with the Network Integrated Biological Alliance (NIBA), is examining and developing business models that create an environment for long-term economic sustainability. Examples of such business models include institutional commitments, cost-recovery, grants, endowments through benefactors, and engaging in public/private partnerships. The goal is to establish a multifaceted funding base for long-term sustainability through community engagement and partnerships with government agencies, private foundations, and international stakeholders.

Building a community of use around iDigBio is paramount to sustaining the infrastructure. As the success of ADBC continues, collections-based data are becoming much more generally available, used, and appreciated, not only for their scientific value, but also for understanding the importance of biodiversity to the economic stability and health of the planet. Coupled with tools from informatics and computer science, collections data are being made available for a host of applications related to biodiversity research, natural resource management, and public policy development. iDigBio has developed a data mobilization and ingestion process to ensure the vast number of collections being harvested are tracked from initiation to closure using Redmine.

iDigBio remains vigilant in managing its scope. To reduce uncertainty and to prevent scope creep, iDigBio has published the Project Scope along with a series of agreed upon in-scope and out-of-scope activities on the iDigBio website. The iDigBio Steering and Executive Committees regularly review tasks, objectives, and progress to insure they are within scope.

iDigBio is establishing strategic partnerships, including international, with external projects, tools, and initiatives. These strategic partnerships not only promote integration of tools and services into iDigBio but also promote long-term sustainability of both iDigBio. Ultimately, these strategic partnerships will contribute towards a robust comprehensive infrastructure.