Summary Document: Integrated Digitized Biocollections (iDigBio) Summit

Meeting Held: November 29th – December 2nd, 2011

Presented by the National Resource (Home Uniting Biocollections, or HUB, at the University of Florida and Florida State University) for Advancing Digitization of Biological Collections (ADBC)

Table of Contents

Overview

The goals of the iDigBio Summit were to foster communication, cohesion, clarification of iDigBio HUB objectives, clarification of Thematic Collections Network (TCN) objectives, identification of challenges and needs, and preliminary documentation of best/common practices related to the digitization and integration of existing, vouchered biological/paleontological collections data within the context of the Advancing Digitization of Biological Collections (ADBC) program. The Summit focused on three broad themes: "Communication and Collaboration", "Enabling the TCNs and Collaborators", and "Enabling the National Resource".

Summit Materials

Summit activities consisted of presentations from the iDigBio Principal Investigators, TCN Principal Investigators and senior personnel, representatives from associated organizations and tool providers, and the National Science Foundation. Attendees also participated in break-out sessions that focused on specific topics.

Presentations

Visit www.idigbio.org, navigate to the Forums, and review the "iDigBio 2011 Summit" forum. There you will find topics that were created for each presentation; each topic links to both the recorded presentation and the PowerPoint used in the presentation. Community comments, feedback, discussion, and the creation of new topics are welcomed and encouraged within the Forum. The iDigBio Forum is a community resource, and the growth and trajectory of the Forum will comply with the needs of the community.

Breakout Materials

Raw breakout session notes, captured real-time during the sessions, are also available within the "iDigBio 2011 Summit" forum. Community feedback via Forum comments are welcomed, and will be taken into account during iDigBio project planning activities and during the ongoing development of the project's functional requirements. Direct links to the four documents that contain breakout notes are provided for your convenience:

Geo-Referencing
Specimen Imaging & Post-Processing
Label Capture & Post-Processing
Data Management & Archival
Enabling the National Resource

Resources

The following resources are currently available from iDigBio.

iDigBio Website

Resource: https://www.idigbio.org
Purpose: Overview of iDigBio and ADBC; advertisement of upcoming events; publication of codified documentation and standards; blog articles related to digitization, project progress, and biodiversity. Blog contributions are welcomed from all sources. Formal editorial guidelines are currently being finalized. In the interim, interested contributors are encouraged to contact iDigBio Project Manager Jason Grabon at jgrabon@flmnh.ufl.edu for blog submission instructions.

iDigBio Wiki

Resource: https://www.idigbio.org/wiki
Purpose: A component of the iDigBio website. Community-maintained resource for display of information related to iDigBio, Thematic Collections Networks, digitization issues, solutions, continuation of dialog from workshops/working groups, and any other relevant content. Contribution of new content and updates to content is welcomed from all (iDigBio site registration required).

iDigBio Forum

Resource: https://www.idigbio.org/forum
Purpose: A component of the iDigBio website. Discussion and collaboration environment for all topics related to iDigBio, Thematic Collections Networks, digitization issues, solutions, workshops, working groups, and any other relevant topics. Initiation of new forum topics and contribution to ongoing discussions are welcomed from all (iDigBio site registration required).

iDigBio Listserv

Resource: idigbio-L@lists.ufl.edu
Purpose: Community communication and announcements. To add yourself to the list, email listserv@lists.ufl.edu with the following command in the body of the email:
subscribe idigbio-L first_name last_name
e.g. subscribe idigbio-L Jane Doe

ADBC IT Listserv

Resource: ADBCIT-L@lists.ufl.edu
Purpose: ADBC Information Technology community communication. To add yourself to the list, email listserv@lists.ufl.edu with the following command in the body of the email:
subscribe ADBCIT-L first_name last_name
e.g. subscribe ADBCIT-L Jane Doe

Note that the ADBCIT-L listserv should not be used to submit User Services tickets to iDigBio (e.g., site down, authentication issues). A separate list has been created for that purpose. A form on the iDigBio.org website accepts User Services issues related to the site or digitization questions, and routes the issues to iDigBio support personnel.

Community Announcements

Resource: https://www.idigbio.org/community-announcements
Purpose: A component of the iDigBio website. A resource for the community to share announcements related to biodiversity, digitization, and biological collections (iDigBio site registration required).

iDigBio Project Scope

The following scope statement outlines the activities that are both within and outside of scope for iDigBio, and casts the roles of iDigBio and the TCNs in the larger context of the Network Integrated Biocollections Alliance (NIBA) report and the Strategic Plan for Establishing a Network Integrated Biocollections Alliance final report, available in a text-only version as well as a brochure version.

iDigBio Scope Statement

Integrated Digitized Biocollections (iDigBio) is the national resource for digitized information about existing, vouchered natural history collections within the context established by the community strategic plan for the Network Integrated Biocollections Alliance (NIBA) and is supported through funds from the NSF program Advancing Digitization of Biological Collections. As such, iDigBio serves as the administrative home for the national digitization effort; fosters partnerships and innovations; facilitates the determination and dissemination of digitization practices and workflows; establishes integration and interconnectivity among the data generated by collection digitization projects; and promotes the uses of biological/paleontological collections data by the scientific community and stakeholders including government agencies, educational institutions, NGOs, and other national and international entities to benefit science and society through enhanced research, educational, and outreach activities. iDigBio provides these services to all stakeholders with clarity, simplicity, transparency, intuitive methodology, and intuitive design.

To accomplish these objectives, iDigBio is responsible for the following specific in-scope activities:

Activity #1 - Implement a scalable and secure cloud-based infrastructure and web portal to enable the storage, integration, search, and retrieval of existing biological/paleontological specimen data, images, and other media files contributed by Thematic Collections Networks, other networks, resources, and collaborating institutions.

Activity #2 - Deliver appliances that integrate and package existing digitization technologies in a manner that enhances and/or simplifies the user experience. Appliances are intended to improve the deployment and interoperability of digitization tools, and to simplify integration with the iDigBio specimen database and storage infrastructure.

Activity #3 - Provide user services to support interaction with both specimen databases and with appliances. User services will support both data/appliance contributors and data/appliance consumers. User services are provided in the form of a ticket submission and tracking system for requests and problems, telephone support, email support, user documentation, and site visits.

Activity #4 - Research, evaluate, benchmark, integrate, and disseminate digitization methodologies, end-to-end processes, tools, recommended standards, and workflows that improve the efficiency and scalability of digitization.

Activity #5 - Provide user services to support efficient, scalable and effective digitization of specimen images, media, and specimen data. User services are provided in the form of a ticket submission and tracking system for requests and problems, telephone support, email support, user documentation, and site visits.

Activity #6 - Coordinate and fund workshops and working groups to:

- a. Foster partnerships and collaboration within the collections community, as well as to connect to stakeholder organizations external to the collections community.
- b. Conduct training related to digitization, technology, workflows, and other applicable fields.
- c. Recommend standards, common practices, guidelines, workflows, and optimal digitization tools and software for use by ADBC participants.
- d. Foster innovation related to bio/paleo-collections digitization and imaging. The outputs of these innovation workshop sessions may include:
    - i. Specific application and/or hardware development requirements that are assigned to existing organizations funded for, and tasked with, tool development.
    - ii. Documentation of challenges and proposed solutions that may lead to proposals to obtain funding for separate projects to deliver required technologies.
    - iii. Creation or improvement of digitization, imaging, and databasing tools resulting from “hackathons”. These sessions bring together skilled session participants to deliver a specific functional product during the workshop. Tools created in “hackathons” must be delivered with pre-conceived strategies for maintenance and sustainability.

Activity #7 - Facilitate the development of standards for digitization, technology, and process training.

Activity #8 - Coordinate and execute iDigBio Education and Outreach activities. Provide advice to and coordination with other digitization projects regarding the integration of outreach activities.

Activity #9 - Provide opportunities and technologies that encourage communication, collaboration and status reporting among members of the ADBC community.

Activity #10 - Oversee development of a community implementation plan to accomplish digitization of existing biodiversity collections in the US, and establish the long-term sustainability of the ADBC data and related infrastructure, and for iDigBio user services operations.

Activity #11 - Establish an iDigBio Internal Advisory Committee that 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. Also establish an External Advisory Board that meets annually to provide advice regarding project activities, the integration of digitization projects, research, education and outreach activities, strategic direction, and management policies.

Activity #12 - Track research outcomes, the results of outreach activities, and innovative discoveries related to the project.

iDigBio Out of Scope Activities

In order to reduce uncertainty in the scope of iDigBio's mission and to prevent scope creep as project requirements are evaluated, the following specific activities are defined as outside the scope of iDigBio:

1. Direct development of new tools (hardware or software), or improvements to existing software tools intended to enhance the digitization of existing, vouchered biological/paleontological collections. iDigBio is not funded or staffed to execute hardware or software development; exceptions are the creation and maintenance of the core iDigBio.org website, the portal/database designed to integrate digitized specimen data, and integration via appliances with existing tools that support digitization.
2. Collections in institutions physically located outside the United States will not be included in the iDigBio collections integration portal. Specimens collected outside the United States but housed within a US collection/location are within scope. *Federally owned collections will be integrated with iDigBio through the federal data center when it is established.*
3. Occurrence records and media not supported by voucher specimens (e.g., bird sightings without a collected specimen) will not be included in the iDigBio collections integration portal. *However, as other resources for these data are established, appropriate links to specimen ancillary data will be created.*
4. iDigBio is not responsible for the acquisition, data curation/management, and quality control of data provided by TCNs and other collaborating collections. However, as part of the execution of in-scope Activity #1 and Activity #2, iDigBio will endeavor to provide tools, features, error-checking, historical record tracking, and feedback mechanisms designed to simplify data curation/management, fitness for use tracking, and quality control by TCNs and other contributing institutions.

Summary of Input, Needs, Requirements, Questions, and Responses

Summit sessions and post-summit survey feedback highlighted a number of needs, requirements and questions from the community. The following section summarizes this input and provides responses and plans designed to address these issues. Issues without an existing response plan include follow-up activities and target dates for next steps. Action items are summarized at the end of this document.

1) Define an appliance and the purpose of appliances.

An appliance may take two forms:

- A package of existing digitization software tools pre-loaded and pre-configured on a virtual machine. The virtual machine may be downloaded and instantiated on a local computer or server at any institution. The intent of a virtual machine appliance is to simplify the initiation of a digitization effort at an institution.
- A web service that a digitizing institution can utilize to improve digitization or integration with the iDigBio specimen database and cloud computing infrastructure.

The creation of appliances should not be confused with new tool development. Appliances enable improved deployment of existing tools and simplified integration with the iDigBio specimen database and cloud computing infrastructure.

2) What is meant by "Integrated Data"?

"Integrated" is the preferred term for the manner in which iDigBio's portal will collect and manage data, distinct from an "aggregator" (e.g., GBIF) that may not maintain persistent relationships.

3) Define the roles of the iDigBio HUB and the TCNs.

iDigBio roles are summarized in the iDigBio Project Scope section of this document. iDigBio is a facilitator for collaboration, communication, dissemination of standards and practices, and an integrator of digitized specimen data, images and media. iDigBio is not a curator of the quality of data shared with the iDigBio specimen database; however, iDigBio will assist in enabling the use of tools, features, error-checking, historical record tracking, and feedback mechanisms designed to simplify data curation by TCNs and other contributing institutions.

TCNs are responsible for specimen image and media capture, for local databasing of the digitized specimen record, and for transfer of this information to the iDigBio integrated specimen database. TCNs are also responsible for collaboration with iDigBio to develop consensus on both curatorial and technical standards, processes and practices.

Clear and consistent communication between iDigBio personnel and the TCNs is required to continuously clarify roles and to add value to ADBC. Refer to Action Item #1.

4) Enable communication between iDigBio and the TCNs, and among TCNs.

Collaboration and communication resources in the form of a Forum, Wiki, Blog and Listservs have been established. Additional technologies to enable virtual meetings, training, and ad-hoc collaboration within the community are being evaluated at this time for effectiveness and cost vs. benefit.

The iDigBio HUB will communicate news, activities, plans, developments, work products, and status on a regular basis via various communication channels.

- Regular blog posts will provide public updates regarding biodiversity, digitization, and project status.
- Blog posts and other key items will be summarized and communicated via an eNewsletter.
- Standing monthly iDigBio / TCN virtual meetings will be held to communicate status and ongoing issues, workshop and working group status, and to encourage collaboration. (Refer to Action Item #1).
- Ongoing activities and discussions that occur within working groups will be accessible via the iDigBio Forum.
- Work products from iDigBio staff, workshops, and working groups will be disseminated on the iDigBio Wiki when ongoing public input/modification is appropriate.
- Codified standards and process documents that require formal publication will be released on the iDigBio website (e.g., a process for managing GUIDs to be utilized by ADBC participants). Unlike the Wiki, these documents cannot be directly modified by the public, but are open for public comment.

5) Address GUID persistence and tracking.

Survey results indicate this topic requires high-priority attention within the next three months, but should not require a workshop for resolution. iDigBio does not intend to impose a community-wide GUID standard, but will produce a standard for interoperability with iDigBio services. Ongoing work on a community standard is being undertaken by GBIF and other organizations. However, a general policy statement from iDigBio is required to ensure that data providers understand how GUIDs will be utilized by iDigBio, and to address the implications of a GUID implementation at any institution integrating with the iDigBio specimen database. Refer to Action Item #2.

6) Address the issues of data storage locations and data backups/redundancy.

iDigBio will provide a storage architecture that can be divided into two main modules: digital object storage and metadata storage. For iDigBio's purposes, metadata is defined as all textual data that relates to a digital object. This includes attributes typically referred to as data in specimen databases as well as attributes about the digital file more commonly known as metadata like camera or lighting information for images.

Each part of the storage is being treated separately to take advantage of the best software available for each task. Storage of large un-searchable binary objects and small highly-searched text are different challenges and there is no one single product that does both well. In order to provide a unified biological object storage system, an additional API and middleware layer will be placed on top of the best object and metadata stores available. An additional advantage of this additional layer will be the flexibility to replace underlying technologies without disturbing applications developed on top of the iDigBio API.

The current object storage system is OpenStack Swift. This is a cloud object store designed for large (1 MB+) objects that are stored and retrieved as one whole unit via a HTTP API. Objects are chunked and stored on slow disks on a cluster of commodity machines. Various replication options are available and the initial set up will be a single cluster of nodes connected by gigabit networking; the expected total storage available in the initial setup will be 50 TB after system and redundancy overhead.

The current metadata storage system is Riak. This is a key value clustered storage system that will hold JSON (a textual description of a JavaScript object, ex. variables, hashes, and arrays) that represents the metadata for objects. Riak stores this data in a distributed manner across multiple nodes. Riak has native map-reduce query functionality that allows arbitrary searches across any piece of metadata within a defined time period (usually, seconds).

For cross-site replication, there are three main approaches under consideration (these follow a primary/backup model and focus on disaster recovery): replication at service API layer, at the Swift/Riak middleware layer, or at the file layer. Replication at the API layer creates replicas at the time they are inserted into the cloud; middleware-layer across sites in a primary/backup fashion is supported by OpenStack Swift, as well as in the Enterprise (not open-source) version of Riak. Swift/Riak data objects are stored in file systems, so approaches that utilize well-known tools (such as based on rsync) are possible. The choice of a cross-site replication approach will be to a large extent dictated by costs and availability of off-site resources that iDigBio can secure.

7) Produce a guide for data standardization and data quality.

The specific request is related to data standardization that goes beyond Darwin Core, the establishment of standards for data quality, and ultimately how collaborators and iDigBio will exchange data using common standards/protocols. Other related topics include the storage of raw data vs. transformed data, a bidirectional interface between iDigBio and collaborators to synchronize updates of data and annotations (curatorial feedback), and the specific feedback pathways. The resolution of these topics will require a workshop that will translate into one or more ongoing working group(s). Refer to Action Item #4.

Current tools and standards must be leveraged whenever possible. iDigBio has published to the Wiki a Glossary of Tools, Glossary of Terms, and Glossary of Projects and Organizations to help identify and utilize existing resources. The community is encouraged to assist with the expansion and maintenance of these Glossaries to ensure that relevant items are made known to iDigBio staff and other organizations that are conducting digitization.

8) Establish Service Level Agreements (SLAs) for data accessibility and requirements for data access capabilities.

SLAs for data accessibility (allowable downtime, site responsiveness, client types supported, etc), specifications related to what type of data should be accessible, and the requirements related to accessing that data from the iDigBio specimen database are topics for a workshop. These topics can be combined with the workshop specified in the previous item summary, as well as addressed through subsequent working groups. Refer to Action Item #4.

9) Develop a plan for sustainability of the iDigBio specimen database and user support services.

Sustainability is a critical yet complex issue that will require a creative and complex solution. However, significant insight can be gained from existing initiatives that have successfully achieved sustainability. A Sustainability Working Group will be established to address this issue. Summit feedback places this as a High Priority issue that needs to begin to be addressed within the next 12 months.

10) Establish plans to help those TCNs that “retire” from NSF funding remain sustainable and part of the iDigBio portal.

Plans to support “retired” TCNs are highly correlated with broader sustainability planning. This issue will be addressed by the Sustainability Working Group.

11) Collaborate regarding a solution for authority files.

The primary authority file requirements from the summit are related to interaction with:

1) Taxonomic authority sources
2) Collector names authority sources
3) Geographic names sources

Authority files were combined with “standards” in the post-summit survey during the evaluation of follow-up workshops and working groups, however this is a unique issue requiring unique and ongoing attention by an Authority File Working Group. Refer to Action Item #5.

A forum topic for authority files has been established on the iDigBio website to enable the community to begin to document issues and discuss shared problems among TCNs. Engagement through this forum is encouraged in advance of the initial Authority File working group's workshop.

12) Clarify the specific roles of iDigBio staff and personnel.

The iDigBio staff directory will be improved to include all staff members, photos, brief biographies and specific project responsibilities. Refer to Action Item #6.

13) Assist with technical training of TCNs and other organizations interested in conducting digitization of their collections.

iDigBio digitization specialists Gil Nelson and Deb Paul are currently conducting research, evaluation, benchmarking, and dissemination of digitization methodologies, end-to-end processes, tools, recommended standards, and workflows that improve the efficiency and scalability of digitization. Five site visits are planned within the first two months of 2012 to gather information and begin to evaluate recommended processes. Publication of findings will be posted on the iDigBio website and Wiki, forming a "toolkit" of best practices, standards, and techniques. Glossaries of digitization Tools, Terms, and Projects and Organizations have been seeded on the iDigBio Wiki; community review and enhancement of these wikis are encouraged. An annotated bibliography of many important web-accessible documents will be posted to the iDigBio website in January 2012.

Documentation is backed by User Services. The iDigBio website will be enhanced to enable the submission of issues/questions regarding digitization and technical training needs directly to User Services. In the interim, User Services may be engaged via the "Website Feedback" form on the iDigBio website.

Additional specific technical training needs should be discussed in the recurring TCN / iDigBio monthly conference calls (refer to Action Item #1) so that these issues can be addressed.

14) Identify end-user Stakeholders and obtain input from this group.

There are a significant number of potential end-users who operate outside of the ADBC community. Identification of, and engagement with, this vast audience will require ongoing activities best served by a working group. The Stakeholder Inclusion, Community Building & User Support Working Group identified within this document will serve as the driver for this ongoing effort. Appendix C contains a preliminary list of stakeholder groups.

15) Research and publish the results of potential integration with data management software and hardware tools.

The process of identification and selection of applications and services in iDigBio is driven by input gathered from the community and by technology and resource availability assessments made by the iDigBio team. iDigBio will host virtual machine servers in its compute cloud, and will package and disseminate virtual appliances integrating tools that are of general use and applicability to the community. Open calls will be published on the iDigBio Web portal outlining requirements and the process of selection for joint development of appliances and services.

Virtualization technologies (such as provided by commercial and open-source products including VMware, VirtualBox, and Xen) will be used as a basis to integrate software into ready-to-use appliances, and their deployment may target either end-user environments (where appliances run on a user's workstation or local server) or a hosted server environment (where appliances run on cloud infrastructures such as the iDigBio cloud).

The tools integrated into iDigBio are typically tools that have been developed and are actively in use by the community. The role of iDigBio is not to create new tools, but rather to integrate them and help reduce barriers to their deployment. In addition to the development of appliances, an important role of iDigBio is to provide hosting resources and best practices to foster the development and facilitate the dissemination and sharing of appliances that can be developed and managed by members of the community.

Given the limited personnel resources of iDigBio and the effort needed to create and maintain sustainable appliances, collaboration and consultation with the community (including tool developers and users) are keys to successful design, implementation and dissemination. The iDigBio team seeks to team up with developers of tools with broad applicability to develop appliances that integrate such tools. The collaboration process will be described in the iDigBio Web site as an open call for developers interested and willing to work with the iDigBio team to create an appliance for purposes of dissemination and/or hosting by iDigBio.

Refer to Action Item #7.

16) Avoid recreating the wheel, particularly concerning GBIF, TDWG and SPNHC.

Existing solutions, standards, practices, and techniques will be utilized by iDigBio to the fullest extent possible. iDigBio does not intend to generate new standards or best practices that supersede the publications of these well-established organizations. Workshops and working groups will seek to include members of organizations that have addressed (in full or in part) applicable issues in the past.

17) Establish plans for iDigBio to keep those outside the NSF-funded efforts informed about the project.

iDigBio currently offers several Resources for information, which were previously described in this document. iDigBio will continue to evolve methods to inform those outside of the NSF-funded efforts by utilizing those resources, as well as the use of social media and an eNewsletter. iDigBio has a presence on Twitter and Facebook that will be utilized for project updates. The iDigBio.org website will also soon be enhanced with the capability to “Like” iDigBio.org, as well as the capability to “Share” individual articles posted to the site. Engaging content that is “Shared” by one Facebook user can be seen by all of their contacts, and will serve as another mechanism to inform the broader community.

18) The expertise associated with collections is as important as the collections. What will iDigBio do to help sustain the community of scientists involved with collections?

iDigBio can assist scientists involved with collections through training, documentation, and User Services support related to digitization. Collaboration tools such as the iDigBio Forum and iDigBio Listserv can also be utilized to connect scientists in need to those individuals who have solutions. To be successful, broad and active community participation is required; this activity is expected to grow organically as tools are utilized and information is populated on the iDigBio website. Education and outreach activities, as well as advertisement of beneficial research results derived from collections information, can help to inspire support for the collections community.

19) Provide opportunities to collaborate with engineers regarding significant technical issues. This would include Google, Microsoft, and university engineering departments.

iDigBio has engaged large corporations and engineers regarding support and collaboration for significant technical issues. As relationships are formed, members of these groups may be included in appropriate technical specification development sessions, workshops and working groups. However, it is important to note that significant new tool development is outside the scope of the iDigBio project.

20) Establish licensing covenants, copyright and attribution policies, and mechanisms for data, image and media protection and attribution.

This is an issue best addressed early in the process by a workshop to establish formal policies and mechanisms. An Image, Media, and Data – Rights, Licensing Covenants, and Attribution workshop is required to draft policies, as well as to identify potential functional requirements for the iDigBio portal to protect media and data.

21) How can iDigBio data be used to draw meaningful scientific conclusions?

A well-constructed set of functional requirements for the search, display, and extraction of specimen records is required to adequately satisfy this question. A working group tasked with Serving the Research Community / Informatics will be established to identify the functional needs for the iDigBio portal from the perspective of a research scientist.

22) How can contributors present content for publication on the iDigBio blog, such as the sharing of experiences, innovations, and methods?

Blog contributions are welcomed from all sources. Formal editorial guidelines are currently being finalized. In the interim, interested contributors are encouraged to contact iDigBio Project Manager Jason Grabon at jgrabon@flmnh.ufl.edu for blog submission instructions. Any user (following registration) also has open access to initiate discussions in the iDigBio Forum, or to post content to the Wiki that can benefit from community input. Community Announcements may be published by any registered user; top announcements will be featured in the monthly eNewsletter.

23) How can iDigBio assist with training people on proper geo-referencing techniques and procedures?

iDigBio will engage the geo-referencing community to produce and aggregate documentation, host web-accessible training modules, host onsite training workshops, and produce short course content. Geo-Referencing is an ongoing activity that will require frequent attention, and is best addressed through a working group. A Geo-Referencing Working Group will be established to provide this attention.

24) How can iDigBio provide details regarding Broader Impacts?

Site utilization statistics will be maintained on the iDigBio website to track usage and impact. End-user surveys will be conducted to gauge impact, and to identify opportunities for improvement. Furthermore, iDigBio will maintain relationships with downstream data partners (e.g., EOL) to receive and track utilization statistics from downstream portals.

25) What are iDigBio's policies and procedures related to locality data abstraction, data removal, and/or image screening of endangered, rare, or protected species records?

A Data and Metadata Requirements and Standards Workshop will be required to answer these questions. Several possibilities exist; data contributors may not publish this data to iDigBio, iDigBio may remove locality information from records marked as sensitive by the data contributor, or iDigBio may publish all data regardless of sensitivity. A firm position on this issue is a required deliverable from the workshop or subsequent working group, and should reflect prior analysis from other organizations (e.g., TDWG) regarding this concern.

26) Can the HUB build onto "georeferencing.org" (owned by VertNet) and provide a site of central information for geo-referencing to the community?

http://www.idigbio.org should serve as the central site for this type of information. Documentation contributors are encouraged to utilize the resources of the iDigBio Forums, Wiki, or submission process for Blog publications in order to publish content to the community.

27) Can the HUB make GEOLocate a software appliance?

GEOlocate is being evaluated as a potential software appliance. Please refer to Action Item #7. iDigBio will host open calls for collaboration with any application developers interested in appliance development.

28) Is there money to build new tools?

iDigBio is not funded to build new tools. However, iDigBio can apply resources to bring developers together for improvements to existing tools, to integrate and deploy existing tools as appliances, to conduct workshops focused upon documentation of challenges and proposed solutions that may lead to funding proposals to deliver required technologies, and fund workshops and working groups to generate technical requirements for other organizations that are funded to build tools.

29) What is the short-term product roadmap for storage?

Q1 2012: Provide a Version 0 iDigBio Demonstrator backed with a prototype storage solution using native system APIs.

Q3 2012: Provide production iDigBio storage services via iDigBio API.

Short-Term Action Items

Action Item #1: Evaluate, and with budget approval implement scalable technology to enable multi-site communication between iDigBio personnel, TCNs, and other interested parties. Schedule recurring meetings with a core group of iDigBio, TCN, and other collaborative participants, beginning in January 2012. Produce a preliminary agenda for the first meeting to establish appropriate attendees for future meetings, as well as a consistent meeting format that addresses status reporting and issue discussion. Clearly communicate project timelines, benchmarks and milestones. Post meeting summaries on the iDigBio website following each meeting.
Due Date: January 2012 (the kickoff meeting was held on January 11th)
Responsible: Cathy Bester, Jason Grabon, TCN PIs, iDigBio Senior Personnel
Accountable: Jason Grabon
Desired Outcome: Implement regular communication and collaboration opportunities between iDigBio, TCNs, and other interested parties. Continue to define and clarify roles in the early conversations. Share best practices, project status, and recommendations for issue resolution. Initiate workshops and working groups as needed to address important issues. Improve transparency to the broader community by publishing meeting summaries.

Action Item #2: iDigBio will draft a flexible workflow for implementation of GUIDs at TCNs and collaborating institutions. The workflow will describe the actions that occur at the TCNs and at iDigBio, and will clarify important touchpoints between organizations. The draft will be sent to TCNs and Summit participants as a Request for Comments, with the expectation to finalize the practices for GUID implementation by mid-February.
Due Date:
1/26/2012 – iDigBio will provide a GUID implementation RFC to the TCNs and broader community via the iDigBio Forum.
2/17/2012 – Finalized guidance regarding GUID implementation will be published to the TCNs and collaborating institutions via iDigBio's website.
Responsible: Greg Riccardi, iDigBio IT Personnel, TCN PIs
Accountable: José Fortes and Greg Riccardi
Desired Outcome: Publication of a flexible method to implement GUIDs at participating institutions, as well as explanation of the future interaction between collections databases and the iDigBio specimen database based upon these GUIDs.

Action Item #3: Implement a ticket submission and tracking system for "Digitization" and "iDigBio Portal/Website" User Support Services. Define functional requirements and use cases to implement this capability.
Due Date: March 2012
Responsible: Kevin Love, Alex Thompson, Jason Grabon
Accountable: Kevin Love
Desired Outcome: Provide a form or other mechanism on the iDigBio website to enable question/issue submission and tracking related to: 1) The iDigBio.org website; 2) Digitization; 3) The iDigBio specimen database portal (future).

Action Item #4: iDigBio will host workshops to enumerate specific requirements related to data standardization, data logistics, data quality standards, data exchange standards/protocols, SLAs for data accessibility, requirements for access to the data in the iDigBio specimen database, storage of raw unaltered data vs. transformed data, data quality measures, a bidirectional interface between iDigBio and collaborators to synchronize updates of data and annotations (curatorial feedback), messaging infrastructure, data object versioning, and recording of additional data such as phenotype statements on specific specimens (extending DarwinCore). Based upon planning conversations, these workshops will likely be conducted in single a 3-day session, with follow-up activities to be completed by topic-focused working groups.
Due Date:
Data and Metadata Requirements and Standards – March 2012 (3/28 – 3/30)
Image and Media Requirements and Standards – March 2012 (3/28 – 3/30)
Data Storage, Curation, and Transfer Standards for Specimen Data and Media – March 2012 (3/28 – 3/30)
Responsible: Reed Beaman, Greg Riccardi, iDigBio IT Personnel at ACIS, TCN IT Personnel, Kevin Love, Cathy Bester, identified members of the community.
Accountable: Reed Beaman
Desired Outcome: Produce functional requirements and use cases, including cost/time estimates where appropriate, for the identified workshop topics. Functional requirements will be translated into technical requirements for system design. Agreed-upon standards will be documented and published to encourage interoperability with other organizations interested in publishing digitized data to iDigBio. TCNs will provide a preliminary set of digitized data to test the mechanisms to share, store, and synchronize data. Establish oversight for these deliverables and future related activities by creating one or more Technology Working Group(s) as an output of this session.

Action Item #5: Establish an Authority File Working Group to evaluate needs, issues and potential paths for resolution related to authority files. This includes authority files for taxonomic names, collector names, and geographic locations names. EoL may be a partner for taxon name resolution within various naming schemes.
Due Date: Establish working group and select a Lead by February 2012
Responsible: iDigBio IT Personnel, TCN PIs & Senior Personnel, TCN IT Personnel, iDigBio PIs
Accountable: Greg Riccardi
Desired Outcome: Establishment of authority files to meet the needs of TCNs and other collections data providers. The components required for a solution may necessitate development of iDigBio database requirements, requirements for interface with existing authority sources, identification of requirements for new authority sources, and/or a “hackathon” to develop and deliver a workable, sustainable solution.

Action Item #6: Improve the iDigBio staff directory to include all staff members, photos, brief biographies and specific project responsibilities.
Due Date: February 2012
Responsible: Kevin Love, Alex Thompson, Jason Grabon, Cathy Bester
Accountable: Kevin Love
Desired Outcome: Increase stakeholder understanding of, and access to, iDigBio project staff.

Action Item #7: Develop and publish on the iDigBio Website a technology roadmap outlining the deployment of iDigBio specimen database capabilities and its service application programming interfaces (APIs), and open calls describing the requirements and processes for server hosting and integration of appliances.
Due Date: Initial version with partial information by February 2012; updated version by September 2012
Responsible: Andréa Matsunaga, Renato Figueiredo, Gil Nelson, Deb Paul
Accountable: José Fortes
Desired Outcome: Improve transparency into iDigBio activities to assess interoperability with existing technologies. Produce a community dialog around identification of APIs that show promise and should be evaluated for availability in the iDigBio specimen database and appliances.

Action Item #8: Establish and maintain a Biologist FAQ on the iDigBio Wiki.
Due Date: Ongoing
Responsible: All iDigBio and TCN Personnel
Accountable: Cathy Bester
Desired Outcome: Produce a clearinghouse with answers to archived Biologist questions from the Forum, Listserv, Blog, User Services, and Contact Forms.

Action Item #9: Establish and maintain an Information Technology FAQ on the iDigBio Wiki.
Due Date: Ongoing
Responsible: All iDigBio and TCN Personnel
Accountable: Cathy Bester
Desired Outcome: Produce a clearinghouse with answers to archived IT questions from the Forum, Listserv, Blog, User Services, and Contact Forms.

Action Item #10: Schedule and conduct a workshop to produce policies and requirements related to Image, Media, and Data – Rights, Licensing Covenants, and Attribution. This workshop may be held virtually via video/tele-conferencing.
Due Date: April 2012
Responsible: iDigBio IT Personnel, TCN PIs & Senior Personnel, TCN IT Personnel, iDigBio PIs, Cathy Bester
Accountable: Gil Nelson
Desired Outcome: Produce policies and functional requirements related to image and data rights. Produce Intellectual Property Agreements with the TCNs and other institutional participants, and a Memorandum of Understanding with data end-users.

Action Item #11: Schedule and conduct a Tool Innovation Workshop aimed at better understanding and addressing technology needs from the digitization program.
Due Date: Approximately July 2012
Responsible: iDigBio IT Personnel, TCN PIs & Senior Personnel, TCN IT Personnel, iDigBio PIs, Chris Norris, Jim Beach, Cathy Bester
Accountable: TBD. This is a joint effort between iDigBio and the steering committee from S2I2.

Desired Outcome: Produce deliverables related to specific technology challenges that will feed into the various software innovation programs being developed in the context of NSF's Cyberinfrastructure Framework for 21st Century Science and Engineering.

Action Item #12: Schedule and conduct a Paleocollections workshop.
Due Date: Workshop to be held on April 27th and 28th
Responsible: Bruce MacFadden, Pam Soltis, Jason Grabon, Cathy Bester
Accountable: Bruce MacFadden
Desired Outcome: Allow a forum for dialog and communication among the paleontological collections sub-disciplines (vertebrate, invertebrate, paleobotany, and microfossils). Present examples of on-going digitization initiatives among the neontological collections communities and examples of currently-funded ADBC digitization (e.g., existing TCNs) and related programs. Facilitate dialog between the paleo and neontological collections communities.

Action Item #13: Schedule and conduct a Digitization Workflows workshop.
Due Date: May 2012
Responsible: Gil Nelson, Deb Paul, Austin Mast, TCN PIs & Senior Personnel, TCN IT Personnel, iDigBio PIs, Cathy Bester
Accountable: Gil Nelson
Desired Outcome: Standardize workflows and train personnel regarding specimen and label digitization procedures and workflows. Record pertinent training presentations for publication and re-use by other collections and future TCNs. Publish all training documentation and materials prepared for this workshop for public consumption.

Action Item #14: Establish a Geo-Referencing working group.
Due Date: April 2012
Responsible: Andréa Matsunaga, TCN PIs & Senior Personnel, Cathy Bester, identified experts in the field
Accountable: TBD – Established Experts in the Geo-Referencing Field
Desired Outcome: Provide an ongoing focus on geo-referencing training, improvements, collaboration, and integration with tool providers.

Action Item #15: Establish a Citizen Science / Crowdsourcing working group.
Due Date: March 2012
Responsible: Bruce MacFadden, Betty Dunckel, Cathy Bester, Tom Nash, TCN PIs & Senior Personnel
Accountable: Austin Mast
Desired Outcome: Provide an ongoing focus on utilizing existing citizen science and crowdsourcing applications, collaboration and integration with tool providers, and training.

Action Item #16: Establish a working group focused on Augmenting Optical Character Recognition (OCR) and Natural Language Processing (NLP).
Due Date: July 2012
Responsible: iDigBio IT Personnel, Cathy Bester, TCN PIs & Senior Personnel
Accountable: Deb Paul

Desired Outcome: Identify opportunities to leverage tools and technologies that are successful both within and outside of the biology digitization domain. Identify opportunities to integrate these tools, or to seek funding for tool development.

Action Item #17: Establish a working group focused on Sustainability.
Due Date: January 2013
Responsible: iDigBio PIs, TCN PIs & Senior Personnel, Cathy Bester
Accountable: Larry Page
Desired Outcome: Establish and implement plans for iDigBio and TCN sustainability, including support for "retired" TCNs.

Action Item #18: Establish a working group focused on Serving the Research Community / Informatics.
Due Date: July 2012
Responsible: Pam Soltis, TCN PIs & Senior Personnel, iDigBio IT Staff, TCN IT Staff, Cathy Bester
Accountable: Pam Soltis
Desired Outcome: Identify needs and functional requirements related to utilizing iDigBio data to serve the research community.

Action Item #19: Establish a working group focused on Outreach and Education.
Due Date: October 2012
Responsible: Bruce MacFadden, Betty Dunckel, TCN PIs & Senior Personnel, Cathy Bester
Accountable: Betty Dunckel
Desired Outcome: Coordinate outreach and education activities between iDigBio and the TCNs, and produce innovative outreach and education deliverables.

Action Item #20: Establish a working group focused on Stakeholder Inclusion, Community Building & User Support.
Due Date: August 2012
Responsible: Kevin Love, Gil Nelson, Deb Paul, TCN PIs & Senior Personnel, Cathy Bester
Accountable: Larry Page
Desired Outcome: Identify, engage, and build relationships with the broader stakeholder community. Acquire iDigBio Portal functional requirements from this broad end-user group.

Workshops

Tool Innovation Workshop

The Scientific Software Innovation Institutes (S2I2) has proposed to contribute funding toward an iDigBio-hosted workshop aimed at better understanding the technology needs of the national digitization program, with deliverables feeding into the various software innovation programs being developed in the context of NSF's Cyberinfrastructure Framework for 21st Century Science and Engineering. Specific objectives and organizational details are currently under development, and may include:

- Develop and integrate a crowdsourcing application for large-scale geo-referencing (high priority for TCNs)
- Providing cost-effective technologies for automated workflows
- Improve OCR technology for specimen labels (make it more accurate, with the capability to parse out data into appropriate database fields)
- Augment OCR and label digitization with a common crowd-sourcing platform/website
- Provide services that interface with definitive authority files (for collector names and taxonomy)
- Improve the capabilities of current geo-referencing tools
- Provide mobile access to digitized records
- Improve tools for storage, transfer, and management of data

Lead: TBD – Discussions are ongoing
Timeline: TBD – Approximately July 2012. Workshop must occur before November 2012.

Image, Media, and Data – Rights, Licensing Covenants, and Attribution

A workshop will be held to address policies and functional requirements for the portal related to rights, licensing, and attribution of media and data. The group will explore subjects such as acceptance of creative commons without restriction of derivatives, public domain usage, best practices, fair use, attribution requirements, and iDigBio portal technical requirements needed to enforce any restrictions. A comprehensive set of policies and functional requirements should be the final output of the workshop and post-workshop activities. Personnel from the Encyclopedia of Life have been identified as an excellent resource for this workshop due to their experience with licensing requirements. Due to significant existing experience within the community regarding media and data rights, a small group meeting virtually via video/tele-conferencing may suffice for this workshop.
Lead: Gil Nelson
Timeline: April 2012

Data and Metadata Requirements and Standards

Establish data elements that must be captured and maintained by the TCNs and the iDigBio specimen database. Utilize existing standards to the greatest degree possible, such as TDWG standards. Key personnel and resources identified during the iDigBio Summit include: John Wieczorek, James Macklin, Dave Remsen (GBIF), Greg Riccardi, and representation from the library community. Address subjects including:

- File management issues (e.g., naming, organization). Leverage standards such as www.archives.gov.

- Tracking of actual label data vs interpreted or modified data.
- Identify standards and extensions required (DarwinCore, ABCD, Audubon Core).
- Inclusion of Stratigraphic data.
- Data quality/accuracy standards, including capability for the contributor to identify records with suspected quality/accuracy issues. Identified records could be filtered from research-related search results, or specifically included in searches for crowdsourcing input via annotations.
- Tracking of annotations.
- Notation of the GPS spatial reference system.
- Georeferencing and locality data, including county, township, elevation, etc.
- Masking of sensitive data.
- Selection of existing digitized specimen information to provide test data sets in advance of TCN digitization. What are important databases that we might consider? Examine characteristics of the data (from each) that should be considered.

Lead: Reed Beaman
Timeline: March 2012

Image and Media Requirements and Standards

Establish image and media standards that must be adhered to by the TCNs and the iDigBio specimen database. Utilize existing standards to the greatest degree possible, such as TDWG and Morphbank standards. Key personnel and resources identified during the iDigBio Summit include: Bob Morris, Gregor Hagadorn, imaging experts, Morphbank personnel, and biologists/paleontologists representing various organism types with extensive imaging experience. Address subjects including:

- Standardization (e.g., image quality, resolution, metadata).
- File management issues: (e.g., naming, organization).
- Dealing with composite/aggregate images and how to separate out distinct images from the larger whole (including labels).
- Balance between image quality and utility vs. cost.
- Small specimen capture needs (i.e., microscopic preparations). Apply different standards based upon quality and purpose of use?
- Conversion of analog images to digital format, including metadata capture.
- Archival standards (e.g., DNG) vs. web presentation standards (e.g., JPG – at what quality, thumbnail images)
- Standards related to image processing, enhancement and manipulation, and identification of modifications within the record or image metadata.
- Accommodation of multiple images and multiple specimen labels from a single specimen.

Lead: Reed Beaman
Timeline: March 2012

Data Storage, Curation, and Transfer Standards for Specimen Data and Media

Enumerate specific requirements related to data exchange standards/protocols, SLAs for data accessibility and website performance, requirements for access to the data in the iDigBio specimen database, storage of raw unaltered data vs. transformed data, data quality measures, a bidirectional interface between iDigBio and collaborators to synchronize updates of data and annotations (curatorial feedback), messaging infrastructure, data object versioning, and recording of additional data such as phenotype statements on specific specimens (extending DarwinCore). Key personnel and resources identified during the iDigBio Summit include: Casey McLaughlin (FSU), Nahil Sobh, Alex Thompson, Edward Gilbert, Katja Seltmann, TCN IT staff, Michelle Butler (NCSA), Chris Jordan, DataOne. Address subjects including:

- Large file transfers from collections with limited bandwidth capacities.
- Archival of original image files.
- Reconciliation of differences in identification histories (e.g., various specimen annotations).
- Treatment of geo-reference data modifications (annotations, justification or reference to source in the annotation, maintenance of history, ultimate authority for determination of the current geo-reference).
- Assessment of a modular approach (provider takes on some responsibilities, HUB takes on others), a comprehensive package of HUB responsibilities, or a mixed model based upon institution needs.
- Are label images stored at the HUB, or does the HUB maintain only digitized data from the label?
- Potential interfaces with:
    - SGR (Scatter Gather Reconcile)
    - Filtered Push
    - SYMBIOTA
    - SPECIFY
    - GEOLocate / BioGeomancer

Lead: Reed Beaman
Timeline: March 2012

Paleocollections

Address the digitization needs, opportunities and Grand Challenges of the US paleontological collections community throughout the US. The specific goals of this workshop are to:

1. Allow a forum for dialog and communication among the paleontological collections sub-disciplines (vertebrate, invertebrate, paleobotany, and microfossils).
2. Present examples of on-going digitization initiatives among the neontological collections communities and examples of currently funded ADBC digitization (e.g., existing TCNs) and related programs.
3. Facilitate dialog between the paleo and neontological collections communities.

This workshop will include about 30 invited participants representing non-federal collections throughout the US. Participants are encouraged to participate in other iDigBio working groups that address issues that impact both the neontological and paleontological communities.
Lead: Bruce MacFadden
Timeline: April 27th and April 28th, 2012

Digitization Workflows

A workshop will be held to standardize, assist and train personnel regarding specimen and label digitization procedures and workflows. Topics would include equipment calibration, hardware, color bars, scales, consistency, training of future staff, basic decisions about what to image,

how to prep and stage the materials for imaging, integration of iDigBio HUB appliances and file/image transfer within the workflow (including republishing to data aggregators and image/metadata repositories such as Morphbank, GBIF, EoL), and data/image backup and archival at the HUB. Key personnel and resources identified during the iDigBio Summit include: Representatives of Tritrophic and Bryo-Lichen TCNs, Nicole Tarnowsky, Michael Bevans, an industrial workflow specialist, someone from the Jardin Botanique in Paris where they have employed an industrial approach to specimen digitization, Chris Norris of Yale Peabody who previously held very effective sessions on workflow, Jim Beach, Linda Ford, Rod Eastwood, Vince Smith, and representation from SPNHC.
Lead: Gil Nelson
Timeline: Approximately May 2012

Portal Requirements

A combined workshop involving the Scientific Research Community and broader end-user Stakeholder groups will be held to continue to enhance and define requirements for the iDigBio portal.
Leads: Larry Page and Pam Soltis
Timeline: October 2012

Working Groups

Authority File Working Group

A working group will be formed to evaluate needs, issues and potential paths for resolution related to authority files. Initial scope includes authority sources for taxonomic names, collector names, and geographic names. The group will explore the possibilities of building authority files vs. integrating with existing authority sources. The summit identified this as a top-priority issue that requires resolution in the immediate future. Key working group personnel and resources identified during the iDigBio Summit include: David “Patty” Patterson (taxonomy expert); Richard Pyle (Hawaii Biological Survey); David Remson (GBIF); the existing efforts of the “Global Names Initiative”; Bill Piel (Yale Peabody Museum); Stinger Guala (USGS); John Wieczorek (UC Berkeley); Matt Yoder; ITIS; Catalogue of Life
Lead: Greg Riccardi
Timeline: Initial members of the working group will be identified in February 2012. A kickoff teleconference will be scheduled by the end of February 2012 to establish priorities, expectations for deliverables, the need for a workshop, workshop planning and a workshop date (if required).

Technology Working Group(s)

One or more IT working groups will be formed to codify iDigBio technology standards and technical requirements, and to provide input into portal design and development activities. The working group(s) will be formed from the technology standards workshop to be held in March, and will be engaged in providing input into ongoing iDigBio portal and appliance development.
Lead: Reed Beaman
Timeline: March 2012

Digitization Workflows

A working group will be formed to review, document and improve digitization workflows, including manual processes and technologies.
Lead: Gil Nelson
Timeline: May 2012

Geo-Referencing

Geo-referencing has experienced significant advancement in recent years. The conclusion is that leading experts in this field will be invited to constitute and lead a working group on this topic. The working group will be formed primarily to address the significant needs of coordinating and providing geo-referencing training, with resources for onsite and/or virtual training provided by iDigBio. The working group will also identify geo-referencing issues, licensing concerns, quality control issues, and will produce appropriate documentation. Key working group personnel and resources identified during the iDigBio Summit include: Nelson Rios, John Wieczorek, Carol Spencer, and Andréa Matsunaga. Issues identified during the Summit that may be explored by the working group include:

1) Determine if multiple existing geo-referencing services can be integrated to provide more comprehensive solutions (i.e., GEOLocate in coordination with Google Maps).
2) Opportunities for partnerships between iDigBio and geo-referencing service providers, including the use of iDigBio compute resources to improve processing capacity, and workshop funds to conduct geo-referencing training.
3) Overcome problems related to the interpretation of legacy localities.
4) Engagement and involvement of local communities to geo-reference specific areas.
5) Identify appropriate existing geo-referencing documentation and training materials for publication, in an aggregated and annotated format, on the iDigBio website (e.g., GBIF best practices, GML (Geographic Mark-up Language), European Petroleum Survey Group (EPSG) Codes, GEOLocate, BioGeomancer, Google Maps, Specify, Arctos, Georeference resources from HerpNet (GIS/Georef Resources), optimal geo-referencing workflows).
6) Investigate and potentially leverage NEON's tools and processes at iDigBio.
7) Determine if existing problems with geo-referencing of Marine data in existing tools can be resolved via a "hackathon", grant proposal, or other action.
8) Determine if multi-language support in existing tools is a significant need, and if this can be added via a "hackathon", grant proposal, or other action.
9) Determine if GEOLocate support for county polygons is a significant need, and if it can be added via a "hackathon", grant proposal, or other action.
10) When existing materials do not exist, develop instructional videos, online modules, basic GIS courses, and advanced GPS training courses for release on the iDigBio website.

Leads: TBD – Established Experts in the Geo-Referencing Field
Timeline: April 2012

Citizen Science / Crowdsourcing

A working group will be formed to identify specific opportunities to use existing citizen science resources for digitization. Opportunities for partnerships between iDigBio and crowdsourcing service providers will be explored, as well as methods to centralize/simplify crowdsourcing, crowdsourcing validation, and advertisement of crowdsourcing needs and citizen science opportunities to the general public. Key working group personnel and resources identified during the iDigBio Summit include: Representatives of TCNs, the Citizen Science Alliance, Afron Smith (Zooniverse), Vizzuality, eBird, Nathan Wilson (representing EoL as well as Mushroom Observer), Michael Giddens (SilverBiology), , WikiSpecies, Herberia@Home, Bruce McFadden, Vince Smith (Vibrant), John VanDyke (Bugguide), Earthwatch, ReCaptcha, the Field Museum, Rob Guralnick, Zack Murrell.

A workshop may be appropriate to kick off this working group, with a keynote speaker/expert on crowdsourcing in general (Astronomy, Google, etc).
Lead: Austin Mast
Timeline: March 2012

Augmenting Optical Character Recognition (OCR) and Natural Language Processing (NLP)

A working group will be formed to explore potential opportunities to enhance OCR via "Hackathons", integration with and application of existing OCR and NLP resources outside of the traditional biodiversity domains to enhance quality and capture rate, and functional requirements for projects devoted to improvements in OCR and NLP technologies. iDigBio will not develop new tools directly. Key working group personnel and resources identified during the iDigBio Summit include: Chris Freeland (Biodiversity Heritage Library), representatives should include Tritrophic and Bryo-Lichen TCNs, Stephen Gottschalk, NYBG project manager for the currently supported Plants and Fungi of the Caribbean project that is breaking new ground in the approach to OCR management, representatives of the Apiary Project, the Salix Project, Tesseract, ABBYY, representatives from the Royal Botanic Gardens Edinburgh (Elspeth Halston) and/or Kew (Anna Saltmarsh), Read Beaman, Ed Gilbert, Jason Best, the US Postal Service, CIA, banks, John Hart, Xerox, IBM, Google, library community, LoC, Nathan Wilson (Mushroom Observer).
Issues to be explored by the working group include:

- Collaborative meeting opportunities co-located at other events.
- Value of pursuing technologies to capture data on handwritten labels.
- Challenges parsing data into fields from OCR (i.e., Natural Language Processing).
- Evaluate and potentially leverage existing practices and technologies such as CAPTCHA, APIARY, SALIX, ABBYY OCR, Tesseract OCR, and Adobe OCR.

Lead: Deb Paul
Timeline: July 2012

Sustainability

A working group will be formed to develop and implement plans for iDigBio and TCN sustainability. Objectives will include establishing funding and infrastructure to enable continued maintenance of digitized specimen records in the iDigBio specimen database, maintenance of the iDigBio website and portal, maintenance of iDigBio appliances, continuation of iDigBio User Services, and continuation of interactions with de-funded TCNs in future years. This working group should include representation from other collaborative projects that have achieved sustainability (e.g., Dan Stanzione, NESCent, DataOne, XSEDE, DataNet, TerraGrid).
Lead: Larry Page
Timeline: January 2013

Serving the Research Community / Informatics

A working group will be formed to identify needs and functional requirements related to utilizing iDigBio data to serve the research community. Analytics, informatics, dataset download and attribution options, and other key topics will be addressed. Key working group personnel and resources include: Andrew Hipp (The Morton Arboretum), Pam Soltis, Austin Mast.
Lead: Pam Soltis
Timeline: July 2012

Outreach and Education

A working group will be formed to coordinate outreach and education activities between iDigBio and the TCNs, and to produce innovative outreach and education deliverables. Objectives include:

- Introduce the public to local biodiversity.
- Introduce the public to collections (and historical components).
- Produce and deliver a Biodiversity/Informatics short course.
- Train IT-capable Biologists for future generations.

Key working group personnel and resources identified during the iDigBio Summit include: Carolyn Lewis, AMNH, Philadelphia Museum, Gaye-Lynn Clyde Milwaukee Public Museum, Joe Cook, Carolyn Ferguson, Anna Monfils, Jill Holliday (UF), and representation from each TCN.
Lead: Bruce MacFadden
Timeline: October 2012

Stakeholder Inclusion, Community Building & User Support

A working group will be formed to identify, engage, and build relationships with the broader stakeholder community (i.e., groups outside of the ADBC initiative). This group should be consulted for input regarding portal requirements for the specimen database (e.g., data search needs, data and media display needs), identification of User Support and training needs, and a general user needs assessment. The broader stakeholder community may also provide assistance to the iDigBio and TCN projects via production environment bug detection, feature requests, and user acceptance testing. A list of identified stakeholder organizations is available in Appendix C of this document. Key working group personnel and resources identified during the iDigBio Summit include: Tom Nash; Austin Mast; Michael Gibbons; Betty Dunckel; Citizen Science Alliance; Rick Bonney (citizenscience.org)
Lead: Larry Page
Timeline: August 2012

Procedural/Administrative Lessons Learned

- Clearly establish mission, purpose and scope in advance of breakout sessions.
- Provide more opportunities for informal interaction between participants.
- Provide more opportunities for Q&A with presenters.
- When providing guest WiFi connection credentials, include the SSID for the network in addition to the access id/password.
- Use tiny URLs if presenting Google document links for online collaboration in written documentation.
- Use larger fonts for participant names on name badges.
- Avoid evening presentations.
- Scale back the quantity of breakout session deliverables to allow more time for thorough discussion and thoughtful answers.
- Conduct a more in-depth preview and training session with breakout session facilitators to ensure full clarity and understanding of the session’s objectives, as well as a review of effective facilitation techniques.
- In future workshops, seek out participation from subject matter experts in external domains (i.e., outside of the biology community)

Appendix A: Identified Specimen Imaging Tools

Specimen Imaging and Post-Processing Tools	Explanation of Selection / Details
Open Zoom	Tiled image processing and display standard with many implementations
GIMP	Open source “Photoshop”
Computed tomography (CT for 3-D)	
Lightfield image	Uses directionality of light, retains more information and allows focal point to be changed in software
LEAF system - camera, super high resolution (for specimens with depth of field issues)	Global Plants Initiative project
E-Box (for herbarium specimens)	Light box system (MK direct); allows for standard imaging, standard positioning of all elements, can change out cameras. Best for flat Herbarium specimens.
SAT-scan, to be replaced by a more affordable option (for Invert specimens)	Used in Australia and in London (insect collections); camera mounted to a robotic arm, tilts, best for insects (InvertNet solution is more affordable).
Gigapan	Robotic arm that allows panorama shots. Stitch images together to yield higher quality image overall.
Herbscan (for herbarium specimens)	Lifts specimen up to scanner
Automontage (for specimens with depth of field issues)	Stacking software (other similar options are available for free). Best for large specimens and depth of field, also very small specimens and long distance microscope lenses. Free/open source options are Combine-Z or Image J.

Appendix B: Identified Specimen Imaging Tool Gaps

Gaps, Issues and Opportunities for Improvement
No known solution for getting good depth of field and image quality, combined with rapid capture.
There is no system for efficient capture of specimens on microscope slides, or other very small specimens.
An automated process to capture data from labels is needed.
Technician training time can be significant. Training materials should be standardized and produced for mass consumption. Effective training should enable understanding of tool usage and imaging workflow.
Fully-automated specimen / label image capture systems do not exist.
Jarred / pickled specimens can be difficult to capture without extensive specimen preparation (i.e., removal from the jar).

Appendix C: Identified Stakeholder Organizations

Stakeholder Organization	Key Individual(s) Within the Organization
AAM (Association of Museums)	
Academia at all levels (college university and K-12)	
AFS (Field Stations)	
ALA	
All biological societies	
ASTC	
BIEN (part of iPlant)	Brian Enquist
Biodiversity Researchers	
Biology Database	Shannon Peters
BLM	
Bureau of Land Management (Interior)	
Bureau of Reclamation (Interior)	
CBOL	
CCH/SEINet -- Large-scale data shareholders	Dick Moe; Les Landrum/Tim Lowry
COL	
Commercial Service Providers (eg. web consultants)	
Consortia: taxon-specific (have portals, concerned with data/best practices; can be used as regional organizing pools)	
CSA (Citizens Science Alliance)	
Discover Life	
DNR (State specific)	
DOD - Department of Defense	
DOE - Department of Energy	
DOI Climate Centers	Damien Shea; Douglas Beard
EoL	
EPA	
ExEP (Exotic Plants)	
GBIF	Donald Hobern
GNA	
GSCG (UN, CBD)	
Homeland Security	
IALE (International Assn Landscape Ecology)	
IAWGSC (inter agency working group on scientific collections)	
Informaticians	
Interagency working group on scientific collections	Scott Miller
iPlant	
ITIS	Stinger Guala
IUBIO	
IUCN	

Land Grant	
Lifewatch/ Framework 7	
Local level: Parks, Cities, LEAs (local education agencies)	
LTER	
MorphBank Standards	Greg Riccardi
Museum Curators	
NABT (National Assn. Bio Teachers)	
Nat. Phenology Network	
National Park Service (NPS)	Anne Hitchcock
Nature Serve	
NEON	
NEOTOMA	Russ Graham
NGOs - Nature Conservancy (and other conservation organizations), Nature-Serve, Botanical Gardens and Museums, Cultural Institutions, Audobon, World Wildlife Fund, Sierra Club, Defenders of Wildlife	
NIH- NCBI	
Non-TCN organizations (e.g. VertNet);	David Bloom
NSF	
Politicians, public officials, policy-makers, lobbying groups	
Private citizens, citizen scientists, local and regional societies, clubs and centers (Bug guide, Wikispecies)	
Professional Organizations: e.g., ABLS -- American Bryophyte and Lichenological Society	Committee
Professional Societies (BSA, ESA, AIBs et al)	
SCICOL	David Schindel
Species file	Matt Yoder
SPECIES LINK ("Brazilian GBIF")	
SPNHC	Tim White
State - AID programs	
State Heritage Programs	
State level: Departments of Natural Resources; Conservation; Transportation; Wildlife; Natural Heritage; Division of Parks and Forests	
TDWG (taxonomic database working group)	Chuck Miller
Teachers and Educators - e.g., NSTA	
TNC (The Nature Conservancy)	
UNESCO	
United Nations	Edward Morton
University Administrators	
US Fish and Wildlife Service	
USDA ARS	
USDA-APHIS	Ann Bartuska
USDA-NRCS	
USDA-CSREES	

USFS (Forest Service)	
USGS	Lucy Edwards
USGS -- Core Science Analytics and Synthesis	Stinger Guala
USUH	Zack Murrell
Volunteer based organizations - (e.g., AARP, Americorps)	
WHO	
Zooniverse (Crowdsourcing)	
Zoos	