iDigBio 2012 External Advisory Board Report

NSF Grant Number: EF-1115210
Award Type: Cooperative Agreement
Award Title: Digitization HUB: A Collections Digitization Framework for the 21st Century
Award Dates: 7/1/2011 – 6/30/2016

iDigBio PIs: Lawrence Page (PI), Pamela Soltis (co-PI), Bruce MacFadden (co-PI), José Fortes (co-PI), Greg Riccardi (co-PI)

External Advisory Board Members: Stanley Blum, Vincent Smith, Donald Hobern, Karen Francl, Gerald Guala

Overview:
The External Advisory Board met with the iDigBio PIs and Project Manager on April 5, 2012 via a virtual conference. Presentations were provided by iDigBio personnel. Inquiry and recommendation periods were exercised by the External Advisory Board. Recommendations, concerns and ideas are categorized in the following document within sections for broad topics. Each section contains a summary of the specific area of discussion, and where appropriate we have included iDigBio's current or planned response to address the item.

Table of Contents

General Comments

The role of iDigBio

The EAB questioned the expected relationship between iDigBio and the TCNs - Is the expectation that the TCNs exist as coordinating networks for institutions that hold data with iDigBio as a cloud integration layer above what they are doing, or are each of the TCNs expected to have their own portal with iDigBio underpinning that with shared storage infrastructure, or both?

iDigBio has been obligated by circumstances and TCN funding models to serve both roles. There are TCNs that do not have the infrastructure to host their own data and/or portal, while others will maintain their own data and/or portal while also exporting that data into the iDigBio specimen database. Mechanisms are currently in place for TCNs without sufficient resources to store data or to host portals to request hosting services from iDigBio.

Regarding other support activities, iDigBio is not staffed or funded for extensive tool development. As a result, it is recommended that future reviews of TCN proposals should ensure that any key, unsolved portion of the TCN workflow is not expected to be resolved by iDigBio without explicit confirmation of that planned collaboration prior to proposal submission. Many TCNs had expectations regarding iDigBio tool development that would solve their workflow issues (e.g., Optical Character Recognition and Natural Language Processing). iDigBio is investigating digitization optimization opportunities in terms of both process and technology, including relationships with partners such as Xerox, but iDigBio may ultimately be unable to provide the “hoped for” solutions.

ADBC-Funded TCN and PEN adaptability and capability

ADBC TCNs and PENs must demonstrate in their proposal their adaptability to the evolving nature of the iDigBio infrastructure. This requires a demonstration of a strong IT component within the TCN that can work with iDigBio IT personnel to craft solutions and resolve problems. When determining funding, the scientific value of a TCN should be weighed along with the technology capabilities described in the proposal in order to increase the likelihood of a successful outcome from the TCN and better interaction and integration with iDigBio.

Sustainability in regard to long-term infrastructure

iDigBio was not funded for storage. The current storage resources procured by iDigBio from other initiatives are not expected to support the data storage requirements of TCNs beyond a projected timeframe of 3 years. Additional funding sources to be described as part of iDigBio's required sustainability plan will be needed in order to continue to support the ingestion and storage of data generated by TCNs. This is a critical issue that should be addressed via NSF funding, exploration of third-party provider donations, and provision of access to community storage resources.

Overall Project Management and Administration

iDigBio interaction with potential TCNs

The EAB questioned if iDigBio is sufficiently involved with potential TCNs that are submitting new proposals under the ADBC program.

In Year 1, there were a number of email questions and visits that focused on roles and interaction with iDigBio. Questions were addressed with a form response, as well as a posting of frequently asked questions. While these mechanisms still exist, as iDigBio matures and continues to develop policies and procedures the solicitation process will become more detailed. For example, TCNs are now aware of the requirement to provide data systems that can produce a persistent GUID, and an Intellectual Property Policy has been published. Data standards, capability to adhere to data ingestion protocols, consistent efficiency measurements and modeling of the digitization process are also details that are being formalized and will be communicated to potential TCNs. iDigBio has a Documentation section on the website for publication of official policies, and personnel are always willing to speak with a potential TCN to ensure that their questions are resolved.

Information Technology

Minimum Information Standards may by excessively limiting

A standard that dictates that only records that include a certain minimum level of information will be made discoverable through the network will limit creative and useful applications of data. For example, if only an image of a specimen is available, that data object may still have utility in terms of

including those objects in services that enable the public to assist with the transcription of incomplete information. There are advantages in supporting discovery for all records while providing services that return only those records that meet the minimum requirements.

iDigBio understands that the discoverability of all records is very important. There are methods in place to encourage assistance from the public for digitization transcription and georeferencing, and iDigBio should enable the capability to expose data to these methods. As a result of several meetings and a Workshop that addressed this subject, iDigBio will adopt the approach of accepting for ingestion any record that contains only the minimum requirement of a GUID. However, minimum information standards that are fit for use for specific research purposes will be clearly defined. These Use Case specifications will help to: 1) Allow an end-user to filter out records that do not contain data required for the research activity; 2) Deliver statistics to the collections community and funding organizations on where data gaps exist that are reducing the scientific value of data exported to the iDigBio specimen portal.

Fill the technical gap for institutions with minimal IT staff

The EAB commented that the use of IPT, a proven tool for specimen data exchange, is a good choice with the capability of rapid implementation. Community adoption of new standards and technologies is extremely slow, therefore project progress will advance more rapidly by implementing standards such as Darwin Core and tools such as IPT. However, the chasm between having an Excel spreadsheet and getting specimen data from the spreadsheet online is a task that appears insurmountable for most small institutions. Stinger recommends working with GBIF to produce improvements that get data into IPT easily and quickly for smaller institutions. Working with smaller institutions will contribute to Broader Impacts.

iDigBio can address this recommendation by:
1) Utilizing in-place technical staff to provide support and training to smaller institutions as they bring collections online or make technology investment decisions.
2) Finalizing an internal evaluation of the limitations of IPT, Dublin Core, Darwin Core, and Audubon Core that present roadblocks for success and collaborating with the standards bodies to produce resolutions that overcome those roadblocks.

Regarding standards limitations, iDigBio will participate in a workshop in May in Kansas to begin outlining key gaps. Also, as iDigBio identifies gaps in Darwin Core, there is an existing mechanism to place requested enhancements into the GBIF repository. If resolved in this manner, the improvements are made through community consensus and become available to all collections using Darwin Core Archives. This is a preferred approach compared to implementing iDigBio-specific extensions to the standards.

Help the community to solve problems that stem from GUIDs

The EAB agrees that a service to impose stable GUIDs on a large section of the community's collections will be a gigantic win for everyone. Functionality recommended by the EAB involves a service to upload a set of known data records and receive feedback as to whether additional related data exists in the iDigBio data store (e.g., additional specimen data for a Genus of interest).

iDigBio has published simple guidelines for establishing unique GUIDs and will release complementary services to provide a proxy resolution service and object services to return information about requested records. The recommended GUID functionality will be considered and compared to the existing plans for services to determine if additional value can be added by expanding the service.

Rapid application of data for research and outreach

The EAB expressed concern that there appears to be a divide between the delivery of exciting science questions and outreach activities as a result of bringing this content together, and the timing of the delivery of some of the IT functionality. This may result in community discord due to the reality of the IT implementation timelines failing to match the level of expectation. A small demonstration of end-to-end process, (IT to outreach for example), would be valuable early in the project.

The project may also benefit from an early win by "solving" record relationship complexities related to one specific science question that already exists within ADBC. This would show that iDigBio is providing functionality not currently offered by GBIF, ALA or BISON. One example is for iDigBio to build a demonstration as early as possible (even if is with sample data) to demonstrate science questions that could be answered related to the Tri-Trophic TCN data. This demonstration would send strong messages to other communities not yet represented by TCNs of the potential benefit of the iDigBio infrastructure.

iDigBio will have post-doctoral fellows and graduate students in 2012-2013 who will be able to work with collections data and to demonstrate what gaps exist in workflows and processes. Use Cases will be developed for potential research projects that will use iDigBio specimen portal data to enable appropriate planning for the development of the cyberinfrastructure. This will take time, but will ultimately result in a solution that serves research needs rather than acts as a simple data store.

Useful georeferenced taxonomic data from TCNs are likely to be delayed for a long time. In the meantime, iDigBio is gathering existing data sets of high quality to allow people to access data directly from the iDigBio specimen portal in the short term through API's.

To address the need to produce some demonstrable product early in the project, iDigBio will release a simple query interface that interacts with the existing (non-TCN) data sets that iDigBio is able to ingest. The cyberinfrastracture group will also show examples of what might be possible in the future with rich data sets via prototypes. Demonstration of the usefulness of certain tools and functionality

will lead to buy-in by the community. Also, release of an iDigBio implementation plan to the public will help to establish expectations and a common understanding within the community.

During cyberinfrastructure development, grad students and post-docs will be recruited to establish education and outreach activities that leverage the cyberinfrastructure. Post-cyberinfrastruture development outreach activities will include engaging school children and other downstream users. Education and outreach opportunities are expected to keep pace with the functionality of the cyberinfrastracture that will support that education and outreach.

iDigBio data model and strategy for IT functionality

The EAB recommended that an early quick win for iDigBio is to demonstrate that iDigBio is not ALA or BISON, but rather a collections-based cyberinfrastructure enabling improved access to new data, and improvements to the quality of existing data. Exposure of iDigBio data to BISON (or others) only via APIs (to the exclusion of the development of an iDigBio user interface) could produce immediate portal capability and enable iDigBio resources to focus on other activities.

Several members of the EAB expressed agreement, citing the fact that existing portals can serve up the iDigBio data and provide good visualizations of the data. However, the capability that is underserved is the delivery of a virtual natural history collection with services, data access, and tools that can help to transform the taxonomic process and the data curation in ways that are enabling and powerful to the institutions themselves. Understanding what the TCNs are promising, in particular to their own taxonomic communities, is absolutely critical to what iDigBio should be focusing on.

iDigBio will need to consider this approach as it is a departure from the original vision of the project, which is a combination of a specimen portal and data services provided through APIs. There are different expectations from each TCN, as each is pursuing different research questions and has different scientific goals. iDigBio will persistently integrate (rather than merely aggregate) information that are produced by the TCNs. As currently planned, to distinguish iDigBio in the early phases the portal will demonstrate the ability to integrate across databases with persistent, versioned records that include images.

Complement rather than duplicate efforts of other entities

In line with the item above, the EAB expressed concern that iDigBio's technology plans may duplicate significant portions of the effort from organizations like GBIF, which already has an elaborate infrastructure for data cleaning, a data portal, and integration infrastructure. Instead, iDigBio could focus resources on locating and mobilizing existing but unconnected data (e.g., from small institutions). This may expose a large amount of existing content for the first time to other portals. Implementation of iDigBio appliances and training focused on data export from these institutions would be a non-trivial effort, but would result in a significant impact to the research community.

iDigBio is looking to collaborate and leverage existing technologies as much as possible, including GBIF data cleaning procedures and the use/extension of the IPT infrastructure. Appliances that are adopted will have the potential to enable currently "dark" collections to easily export data into iDigBio. This is more of a collaborative approach than a duplication of effort or product.

In terms of value-add from the iDigBio specimen portal, one of the current differentiators is the development of data-cleaning tools that iDigBio can push out to the providers for use prior to data export, to improve the likelihood of receiving high quality data. iDigBio is also seeking out and looks forward to identifying and mobilizing "dark" digitized collections. The sum total of these services and training efforts is a specimen database that will enable records to be persistently referenced, annotated, massively horizontally scalable, a source for newly integrated but existing digitized data, with progressively cleaner data fields as cleaning tools improve over time.

Community and Public Outreach

iDigBio should engage the community through openness and information sharing

iDigBio will provide Workshop meeting presentations and meeting summaries to the public via the iDigBio social portal (www.idigbio.org) . In order to engage the public, iDigBio will also live-broadcast and record Workshop sessions whenever feasible, in order to open the content to a greater audience. To ensure that new voices are heard within the community, future Workshops will include openings for applicants rather than strict creation of invitation-only participant lists based upon decisions by iDigBio staff and/or Workshop committees. The presentations that were provided to the External Advisory Board are also available on the iDigBio website.

iDigBio expects these actions to address the External Advisory Board's recommendations to focus on outreach to the ADBC community and to provide appropriate initial and ongoing insight into the goals and activities of the project.

Learn from prior successes in other projects to demonstrate quick wins

Based upon experience at the Atlas of Living Australia, the EAB advised that it is much easier to demonstrate rapid value and interest to citizen scientists and the public than to genuinely support taxonomic research or curatorial activity. Showcasing where specimens came from was a quick win at ALA and generated a lot of excitement. The much harder task is to convince key stakeholders in ADBC institutions that the infrastructure will be significantly beneficial to them rather than just another way to share their data more widely. As iDigBio finalizes the 10-year implementation plan, the potential to dedicate resources to this type of effort will be prioritized and included in the plan. A demonstration of end-to-end use, from data ingestion to outreach, is recommended by the EAB.

Outreach to participation from small collections

In an effort to recruit participation from small collections, iDigBio has awarded the first of several Visiting Scholar awards to Dr. Anna Monfils, who is from a small collection in Michigan. Upcoming workshops including Paleocollections and Botany 2012 include participants from different types and sizes of institutions. About half of the botany workshop participants are from small institutions who want to learn how to use some of the digitization tools that have been developed. A series of training workshops will be held in Year 2 to address specific needs of small institutions and to catalyze digitization efforts that implement optimal processes at these institutions. NSF has also started to award digitization funding to PENs – most of which will likely be awarded to small institutions.

Integration of data from existing small collections that are currently not online

Institutions with digitized specimen data that are not available online are an opportunity for iDigBio. This would require an initial assessment, or "State of the Union" of collections and how much is digitized. Progress toward digitization (and indirectly outreach progress and research benefits) can then be quantitatively measured.

iDigBio is working with collaborators including Hank Bart, Karen Francl and David Schindel to establish baseline biorepository data to provide the most accurate data possible toward this end. Deb Paul and Gil Nelson are also visiting mature and new collections to contribute a qualitative analysis of the state of collection digitization in many institutions across the country. The EAB recommended several other resources that iDigBio will pursue to quantify collections and digitization for benchmarking purposes, including:

1) A NSCA survey conducted in conjunction with NSF (within the last 2 years) on digitization. The survey included several data points such as the size of collection and how much is digitized.
2) Elizabeth Martin at USGS is creating a survey for BISON that may yield insight.
3) USGS has Greenbook which assesses the state of federal collections.
4) The GBIF registry.
5) A federal metadata clearinghouse is being developed.
5) Bar Code of Life Registry (Schindel), Biodiversity Collections Index (BCI), Index Herbariorum, and other existing resources that can be consolidated in some way. BCI and Biocollections.org (Bar Code) clearly need to be merged, and there seems to be some readiness to merge the data into the GBIF registry. GBIF is interested in working with different communities in order to understand what additional features would be appropriate within the GBIF registry to make it a much more general purpose data resource and collection discovery source, with a focus on simplicity.

The EAB recommends that iDigBio take an active role in developing a registry, in coordination with the entities named above, rather than a passive role with the expectation that another group will solve the problem. Resource constraints at several of the other initiatives may inhibit progress toward the creation of a central repository.

Presentations

Role of the External Advisory Board

Jason Grabon – Introductions, Meeting Overview and Expectations

iDigBio Welcome and Overview

Larry Page – iDigBio and ADBC Overview

iDigBio Information Technology

José Fortes – iDigBio Technology, Cloud Computing and Appliances

iDigBio Digitization

Greg Riccardi – Support for Digitization and Informatics

iDigBio Research

Pam Soltis - Research Coordination and Scientific Community Outreach

iDigBio Education and Public Outreach

Bruce MacFadden – Education, Outreach and Evaluation

Project Background

José Fortes – The HUB: Then and Now

Community Engagement

Jason Grabon – Social Media and Community Engagement