Meeting Information

Meeting Date/Time: March 22, 2013 / 1:00-3:00 PM USA Eastern Daylight Time

External Advisory Board Members: Vince Smith, Stan Blum, Stinger Guala, Karen Francl (present but commented via email), Donald Hobern (absent but commented via email)

iDigBio PIs: Larry Page, Pamela Soltis, Bruce MacFadden, Greg Riccardi, José Fortes (absent but was represented by Renato Figueiredo)

iDigBio Staff: David Jennings, Joanna McCaffrey, Cathy Bester, Shari Ellis, Betty Dunckel, Kevin Love

Executive Summary

The External Advisory Board (EAB) members met with the iDigBio PIs and project staff on March 22, 2013, via an Adobe Connect virtual conference. Presentations were given by iDigBio personnel to outline iDigBio's activities, progress, challenges, and opportunities faced by iDigBio during the past year. Following the presentations, an open discussion period allowed iDigBio to respond to questions, comments, and recommendations from the EAB members. In general, the EAB members were quite pleased with iDigBio's progress. The EAB was impressed with the quantity of work that iDigBio has produced while meeting its goals, staying on track, and with only a few items on the timeline being delayed.

Table of Contents

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (#EF1115210)

iDigBio Presentations

The following are the presentation topics given to the EAB by iDigBio personnel during the meeting. Each topic is linked to the specific presentation.

Topic	Presenter
Welcome, Introductions, and Role of EAB	David Jennings
iDigBio Overview and Relationship to Other Initiatives	Larry Page
iDigBio Technology, Cloud Computing and Appliances	Renato Figueiredo
iDigBio Optimization of Digitization Workflows and Processes	Greg Riccardi
iDigBio Research Coordination, Scientific Community Outreach	Pam Soltis
iDigBio Education and Public Outreach	Bruce MacFadden
iDigBio Implementation Plan and Project Progress	David Jennings

Project Management and Administration

Overall Progress

In general, the EAB members were quite pleased with iDigBio's progress. The EAB was impressed with the quantity of work that iDigBio has produced while meeting its goals, staying on track, and with only a few items on the timeline being delayed. The EAB congratulated iDigBio on its efforts and activities.

Workshops and Training

The EAB remarked that all of the iDigBio Workshops and Symposia have been extremely popular within the community. In addition, the EAB recognized the significant value that iDigBio is able to provide by sponsoring these training activities. The EAB expressed appreciation for iDigBio's efforts.

Role of the EAB

The EAB noted the obvious constraints with personnel and resources in trying to link together so many activities and infrastructures and questioned the role of the EAB with regard to the prioritization process. The governance structure of iDigBio utilizes the EAB as "community advisors" to guide the strategic activities and directions of iDigBio, whereas iDigBio's internal Steering Committee is the governance body responsible for prioritizing resources and personnel.

Internationalization

With regard to the significant collaborations that iDigBio seems to be making within the U.S., the EAB inquired about any iDigBio efforts towards internationalization. Specifically, the EAB wondered if iDigBio was involved in the European funding opportunity called Horizon 2020. iDigBio stated they were not involved in Horizon 2020, but would speak with the NSF program officers about the extent to which NSF is working with the European Community (EC) towards building larger global frameworks for funding (i.e., getting beyond national boundaries in terms of funding). On another internationalization

front, the iDigBio digitization group has launched an International Whole-drawer Digitization Interest Group, in co-coordination with CSIRO and membership from the UK, The Netherlands, and Germany, as well as U.S. institutions and TCN members, and a member from the Smithsonian. Since whole-drawer digitization of insects is more widely practiced outside of the U.S., this interest group brings together international representatives who are implementing whole-drawer digitization strategies.

Biodiversity Collections Index

The EAB asked if iDigBio had any contact with the Biodiversity Collections Index (BCI) people regarding their notion of a registry of national resources. iDigBio indicated that we have been communicating with them, primarily though David Schindel, and to a lesser extent with Barbara Thiers (regarding a particular merger of Index Herbariorum and BCI). The EAB indicated that BCI has a prototype implementation up, and suggested that some coordination or rallying from iDigBio could help set it on its way.

Cyberinfrastructure

MISC

The EAB asked what "MISC" stood for. MISC stands for Minimum Information Standards for Scientific Collections. The iDigBio MISC Working Group serves as an advisory group to iDigBio as it develops plans and strategies for accepting and ingesting data provided by contributors.

Sustainability

The EAB remarked that it was not clear from the presentations whether the fundamental data architecture is a relational database, or a "big table" model. The EAB noted that these decisions will interlock with whether this is an archival repository or a discovery tool and expectations of scalability and performance for data publishers and for data users. The iDigBio specimen portal data architecture combines several technologies to allow iDigBio to be highly-available, scalable, fault-tolerant, flexible and monitorable. At the core, the textual data is maintained in a NoSQL/"big table" distributed database while the media objects are stored in an object store. On top of that, full-text and domain specific indexes are created to speed up searches; relational databases are used to facilitate the navigation through relationships and to deal with the internal structured data; and external services are used for the mapping component. All iDigBio public programmatic interfaces are REST based.

Archival Storage

The EAB remarked that it was unclear from the presentations whether or not iDigBio's responsibilities include long-term primary archival storage for data or simply hosting of web-accessible, interconnected copies, or both. iDigBio responded by stating that it hosts, maintains, and archives web-accessible copies of data that are linked back to the provider for annotation purposes. The data provider is responsible for maintaining the primary data set and for permanent archival storage. iDigBio will maintain record versioning to ensure a historical perspective on the records that iDigBio maintains.

iDigBio is trying to position itself as the focal point for data/image storage and is working on long-term storage solutions, recommendations, and guidance. iDigBio also desires to work collaboratively with image providers, providing for storage of web presentable images in iDigBio's infrastructure or in non-iDigBio image repositories.

Web Site (www.idigbio.org)

The EAB commented that they have been following the growth of iDigBio's website over the past year, and recognize the significant impact of the information being disseminated, particularly the digitization resources. iDigBio encouraged the EAB members to visit the newly redesigned website and provide any feedback or comments.

Specimen Data Portal (portal.idigbio.org)

The EAB asked whether or not any sort of user testing and/or experience has been performed on the data portal. iDigBio stated that the feedback on its V0 demonstrator portal was used to guide the design and implementation of the recently released V1 portal. In addition, iDigBio has met with several curators and collections managers from the Florida Museum of Natural History to get their input. Most recently, the ACIS group (cyberinfrastructure) has hired a new person who will have responsibility for the user interface design. iDigBio has been encouraging the community to experiment with the portal and give us feedback, which iDigBio records as "tickets" in Redmine so we can review each request and make a decision whether or not to move forward on a particular item. However, iDigBio recognizes that we have very little feedback on the newly released V1 so far and encouraged the EAB to provide input.

Data Ingestion

The EAB observed that the specimen data portal currently contains about 1 million records and inquired as to how many of these have been mobilized through the activities of the TCNs. iDigBio stated that the current data sets in the portal are from the Florida Museum of Natural History and are not related to any TCNs. iDigBio has a plan for ingestion of the TCN data, which will be mobilized in the very near future. iDigBio stated that getting TCN data into the portal is a top priority.

Collaboration with Collections Management Software

iDigBio remarked that one of the things we actively working on is ensuring that all of the major collection management software tools can send data to iDigBio in a straightforward manner. In particular, iDigBio has been working with KE EMu, Specify, Symbiota, SilverBiology, Arthropod Easy Capture and other tools to develop packages that can seamlessly resolve a lot of the issues with exporting data.

Getting Data from Small Institutions

The EAB observed that although iDigBio's focus is getting TCN data first, their secondary focus appears to be examining how to bring smaller institutions into the iDigBio effort. The EAB appreciates iDigBio's "no museum left behind" mentality, but questioned how the recruitment of these institutions would

occur. iDigBio noted that it is actively seeking collaborations and is making headway within the broader collections community, which we believe will help drive these collaborations. Due to existing relationships between iDigBio staff and several small, non-TCN but NSF-supported collections, we have accessed test data and reviewed methods for bringing these data into the iDigBio portal.

Getting Data from Large Institutions

The EAB commented that the whole program of iDigBio, ADBC, and NIBA is about getting all of the nation's resources up and as visible as possible. Along that vein, although small collections are certainly important, the EAB indicated that iDigBio and ADBC could harness much more by making overtures to larger museums and institutions that are not part of TCNs, noting that many of these institutions are mounting digitization efforts, have crowdsourcing initiatives in place, and have active education programs. The EAB observed that many of the larger institutions currently publish their data via IPT, so iDigBio could potentially tap into that data stream. iDigBio has existing relationships with several of these institutions through working group and workshop activity and agrees that it will continue with communications regarding data acquisition from these institutions.

The EAB also suggested reviewing a survey that was done by CollectionsWeb, which presented the potential scope of digitization and what the space looked like. The EAB suggested comparing iDigBio's efforts with that space to give us a good indication of where we stand. This could also be used to demonstrate to science administrators and the community the potential of the digitization effort if everyone were involved.

API

One of the EAB members indicated that he would like to interact with the iDigBio API and provide feedback. iDigBio indicated that read APIs are readily available to anyone, and write APIs are available to users registered on the iDigBio website with an API key that can be obtained contacting the cyberinfrastructure group by e-mail. iDigBio provides documentation of the public API through the cyberinfrastructure working (CYWG) group wiki (https://www.idigbio.org/wiki/index.php/Cyberinfrastructure_Working_Group), and encourages interested developers to join and participate in the group to provide feedback.

GBIF

The EAB asked how iDigBio was going to interact with GBIF, particularly with getting its data into the GBIF ecosystem. In addition, the EAB noted that many of the TCNs are serving data independently to GBIF and wanted to know iDigBio's plan for preventing duplicate data. iDigBio's position is that the ideal scenario is for iDigBio to act as a transfer hub to GBIF. iDigBio considers their position not as a mandatory requirement, but rather as a potential high-value service to the TCNs and GBIF. This is clearly an area where iDigBio, GBIF, BISON, and other initiatives will need to work together to develop a sustainable solution. We all have the same goal of making data on biodiversity accessible to everyone, and we can benefit from a better understanding of how best to work together to achieve that goal.

TCN Support

The EAB observed that most of the TCNs have their own portal, infrastructure, databases, etc. As such, the EAB questioned how much the TCNs are committing to using iDigBio resources and developments wherever possible, and whether or not iDigBio would have any ability to enforce and/or ensure efficiency on this front. iDigBio responded that many of the TCNs use iDigBio computing resources, such as virtual machines, temporary storage, etc., in the progress of doing their work, but that these resources are not outwardly visible. The TCNs are routinely relying on iDigBio for support and expertise.

Digitization

Name Resolution

The EAB asked how far iDigBio had come on its name resolution work. iDigBio indicated that we are asking our data providers, TCNs in particular, to give us their organized names, which are mostly coming in as taxon concepts rather than just names. iDigBio noted that the Global Names Architecture (GNA) is claiming they will be able assign a taxon concept id to all of these names. Then, when iDigBio put them into Taxonomic Name Resolution Service (TNRS), we will be able to get concept level information back out again. iDigBio's plan is to use, rather than create, name resolution services that will allow data contributors to use taxonomies that best represent each contributor's taxonomic world view. Taxonomic records submitted to iDigBio in support of collection object records will be stored in an iDigBio-related instance of TNRS, with taxon concept identifiers extracted from GNA.

OCR

The EAB asked if there was a particular group within iDigBio that was taking the lead regarding the development of Optical Character Recognition (OCR). iDigBio responded that there is an active iDigBio working group on this topic (https://www.idigbio.org/wiki/index.php/Augmenting_OCR), with the personnel from Florida State University (FSU) taking the lead. The EAB commented that setting up OCR services would be very helpful to the digitization community. iDigBio stated that this is already being discussed as a potential outcome of iDigBio's recent OCR hackathon. iDigBio's OCR working group is isolating and testing methods and services for effectively incorporating OCR technologies into digitization practices.

Research

The EAB was impressed with the kind of research possibilities that might come out of having all of the collections data mobilized. The EAB speculated that it could be worthwhile talking with the data resources experts within iDigBio to create "services" that could enable research. iDigBio responded that discussions are currently ongoing with cyberinfrastructure group about mechanisms where it could be very easy to pull specimen data, genomic data, the bioclimatic data, and whatever else is associated with that specimen together, and then also have accessible (API or web interface) ways to run different types of analyses. The idea is that if we can identify all of those components, we can make this easily

repeatable by other people across disciplines and research interests. We will try to do the same kinds of things for a number of research problems.

Data Linking

The EAB asked if iDigBio was participating in the effort to link DNA collection and sequencing and genomics to specimens, specifically with TDWG's Genomics Biodiversity Working Group (GBWG). iDigBio responded that collaboration efforts are well underway, including iDigBio's hosting of a common portal to various genomics resources. iDigBio will also be participating in the GBWG workshop at the Smithsonian later in April.

Outreach

The EAB applauded the outreach efforts and activities of iDigBio. The EAB suggested looking into the various Research Experiences for Undergraduates (REU) projects that are hosted at natural history museums and at universities with natural history museums. The EAB further suggested looking at ways to assess student engagement (similar to the work being done by Anna Monfils), and disseminating this to the broader community.