The iDigBio External Advisory Board Recommendations for 2017
July 28, 2017

This document contains the latest set of recommendations provided by the External Advisory Board (EAB) to iDigBio. The current board operates under a revised structure implemented in 2016 that created two committees focused on **data usage** and **sustainability**. The formation of the data use committee resulted from a request by NSF to focus the last five years of ADBC funding on efforts to promote the use of ADBC data for research as well as broader impact activities. The sustainability committee's role is to help iDigBio develop a sustainability plan for post ADBC funding. The EAB submitted its 2016 recommendations in December, 2016 to iDigBio, who in turn submitted the report to the National Science Foundation (NSF).

The iDigBio responded to the 2016 EAB recommendations in March, 2017. Both the EAB and iDigBio agreed to maintain an ongoing dialogue in preparation for the in-person EAB meeting October 31, 2017. To that end, this document summarizes the latest recommendations by the EAB for data management-usage and sustainability. Although the EAB continues to operate as one working group, we have maintained a separation between these two areas of focus.

Role of the EAB - "An external advisory board, whose membership will be subject to approval by NSF's cognizant program official, will meet at least once a year and provide written and verbal advice to iDigBio on its activities, including progress and integration of digitization projects, research, education and outreach activities among all funded institutions, and provide advice on strategic directions, management policies and sustainability. It is anticipated that board members will serve for 2-3 years and have expertise in scientific research and/or sustainability of biological infrastructure." (iDigBio EAB website)

Chair

- Neil Cobb, Northern Arizona University

Data Use Committee

- Paul Kimberly, National Museum of Natural History, Washington
- Linda S. Ford, Harvard University
- Vince Smith, Natural History Museum, London
- Jason Knouft, Saint Louis University (Chair)

Sustainability Committee

- Mary Klein, Former President and CEO of NatureServe
- Donald Hobern, Global Biodiversity Information Facility
- Barbara Thiers, New York Botanical Garden
- Eva Huala, The Arabidopsis Information Resource (Chair)

External Advisory Board Recommendations

iDigBio Data Use

Data Use

The EAB has recommended that the first priority for iDigBio should be to document basic data use metrics and effective ways to communicate data use summaries to increase the use of the data by a broad and growing community that includes a variety of research disciplines. Once iDigBio develops a solid understanding of who uses the data and how they use it, they can better promote the use of iDigBio data.

.

Discussion Point 1.a., Pages 1-4

iDigBio requests that the EAB provide a specific suite of suggested metrics for iDigBio to use. While we believe it is ultimately iDigBio's responsibility to identify the most appropriate metrics to report, we can share some thoughts based on our own experiences. The EAB notes that it is important to distinguish between providing metrics regarding individual record sets and portal-wide metrics. Statistics reported by iDigBio for record sets are relatively complete; but there is more that can be done to improve the usefulness of portal-wide metrics. One option is to consider the statistics displayed on the NHM website, which is a good model to initiate a discussion. Several additional suggestions gathered from the EAB are provided below.

- Data Trends: What metrics ought to be highlighted and what statistics should be provided on data-use metrics?
    - The most useful metrics should be monthly data presented to show trends in data use, and how they are changing over time. Graphical visualizations would be very helpful and easy to understand for the casual user and/or administrator. We recommend dropping the cumulative metrics. Ideas for specific metrics – some of which are already collected by iDigBio - include[1]:
    - Measures of the overall size of the user community (e.g., unique visitors per month)
    - Measures of how useful the product is to the user community. These estimates could be presented as number of repeat users, publication references, ratio of records in search results to records in downloads (or records viewed), or number of records downloaded on a monthly basis.
    - Estimation of which records appear in the most search results. This may give some idea of which data are most in demand, and then these data could be promoted and supported more heavily.
    - Statistics on location of users, U.S. state or country.
    - Documentation of the state of and improvements in standard (emerging in TDWG) data quality tests on biodiversity data.
    - A goal suggested by the EAB is to measure trends, and reduce "noise" in the trends from events that cause spikes in usage. For example, it is not clear whether there is an upward trend in usage in the graph of monthly, aggregated data usage on page 3 of the iDigBio response.

[1] Note: these ideas are not listed in priority order.

- It also would be interesting to quantify the usage impact of events (e.g., WeDigBio). This information could help with estimating the potential return on investment of future events.

- Where should usage statistics be presented on the portal?
  - We encourage iDigBio to consider creating a "Metrics Dashboard" so all search and download data are exposed to the publishers who can then filter and query the data however they wish (e.g., discipline-specific stats, downloaded records for Madagascar, all of a publisher's raw search/download metrics). This tool would allow publishers to query the metrics, instead of iDigBio trying to figure out all of the fixed statistics that we should be seeing. This would also help determine what are the most common metrics used and, then, these can become the fixed statistics for the less experienced publisher. We understand this may be outside of the scope of iDigBio's current work plan, but it is something to work towards that would distinguish iDigBio's portal as an exceptional tool for data publishers.

Discussion Point 1.b., Pages 4-12

- iDigBio welcomed the development and application of common data usage metrics, tracking and communication with GBIF and others. "iDigBio representatives are committed to leading the community discussion around harmonization of data use statistics with TDWG's Biodiversity Services and Clients interest group.
  - The EAB feels that it is very important for iDigBio to continue collaborating with other data portals such as GBIF to develop shared standards for reporting data usage, and to develop a division of labor among organizations that makes sense for the future. Likewise, it will be important to continue working with the larger biodiversity informatics community (e.g., TDWG) to create synchronized data across data aggregation platforms that will leverage iDigBio data for end users. It is likely that there are lessons to be learned from the business community, where they are explicitly developing "collective impact" strategies, including when and how to structure collective impact activities.

- Which metrics collected and provided by GBIF are particularly valuable and useful?
  - Metrics that aggregate usage data (across all collections) and that focus on who is using the data would be helpful.
  - The following metrics from GBIF are useful:

    -The list of peer-reviewed articles using GBIF-mediated data (mediated by GBIF's use of DOIs for data sets and for query results), although the EAB recognizes that there are challenges associated with gathering this information.

    - Documentation of the mobilization of occurrence records over time and the number of species represented in the mobilized occurrence records (change over time in records about biodiversity).

    - Documentation of the temporal distribution of occurrence records (when specimens were collected).

Discussion Point 1.c., Pages 12-13

- In its response, iDigBio focuses on the challenges of attribution that arise when multiple data initiatives openly share data.
    - This is an important problem, and the EAB is pleased that iDigBio is playing an active role in addressing the problem. It will be very helpful to the individual data providers as well as the business sustainability planning effort to be able to quantify, at some level, the extent of usage by other portals of data that originated from iDigBio.

Discussion Point 2.b., Pages 16-17

- Publishing journal papers
    - The EAB was pleased to see the statistics on publications.
    - The EAB felt that publication data should be part of the iDigBio usage metrics dashboard – easily and prominently accessible from iDigBio's home page.
- iDigBio Newsletter
    - The Research Spotlight section of the newsletter is a very good idea. It raises the question of how widely the newsletter is distributed and read? Could iDigBio share statistics about who is being reached by the newsletter?

Discussion Point 2.e., Page 19

- Should the iDigBio portal evolve to include research tools that can engage a broader community of scientists?
    - Again, the need for identifying who the broader community is crucial. One needs to know who that end user is before developing research tools to meet their needs.
    - One caution with limited resources is not to "over-tool" – research and development can be an expensive funding section within itself and the maintenance cost to stay at the forefront seems outside the scope of iDigBio. iDigBio should try to be an excellent data portal that interacts well with externally developed research tools.

Discussion Point 2.g., Pages 20-21

- Tools for researchers to easily integrate their data into the iDigBio portal.
    - This is extremely important and even more important than developing research "tools." Anything that facilitates researchers getting their research and data more publicly known should promote voluntary and core use of iDigBio.

iDigBio Sustainability

The primary role of the Sustainability committee is to prepare for iDigBio operations after ADBC funding. The National Science Foundation does not typically provide long-term support for operations, the ADBC program is intended to establish digitization programs throughout the United States (e.g., TCNs and PENs) and central organization to coordinate a national effort (iDigBio).

Discussion Point 1.a., Pages 21-22

Original EAB Comment: The recommendations above on diversifying the user community are also relevant to sustainability because a diverse user base creates a broader range of possibilities for seeking support (funding, partnerships, and collaborative tool development) that is the foundation of a sustainable business model.

EAB Response to iDigBio: The EAB is satisfied with the ongoing effort to promote broad use of iDigBio data and encourages continued efforts in this area. However, additional effort is needed to measure the use from different scientific fields or end uses such as conservation vs basic research. We encourage the iDigBio leadership to give some thought to how this could be accomplished.

Discussion Point 1.b., Pages 22-23

Original EAB Comment: Usage tracking should include the number of unique visitors per month and the geographical location and institutional affiliation of users. This information can be gathered through Google Analytics.

EAB Response to iDigBio: For this year, we would like separate reports for the website and portal containing a table of monthly unique visitors, sessions and page views beginning with the earliest available data. If possible, we would like usage from iDigBio internal staff to be excluded.

After this initial report is provided, future usage reports to the EAB should include monthly unique visitors, sessions and page views for the past year, together with a year over year comparison to the previous year. We recommend dropping the cumulative reports and the 'searched' and 'seen' categories for the portal usage data as we don't view these as good indicators of user interest in the data.

Alternatively, the EAB could be provided with direct access to the Google Analytics interface for the website and portal by adding the email address of one board member with view and analyze privileges to the Google Analytics accounts.

Discussion Point 1.c., Pages 23-24

Original EAB Comment: Periodic online surveys can identify the research domains and research questions of interest to the community of data users. By identifying the communities that most value iDigBio/TCN data, tools and expertise, we can gain a better understanding of

how those user groups might be able to help sustain key elements of iDigBio and the broader digitization effort into the future. An annual survey should be sufficient.

EAB Response to iDigBio: We support the idea of including one or two simple questions on research domains and research questions. We think this can be at a high level that would not trigger any privacy Discussion Points, such as pick lists that includes some common research areas (e.g. systematics, evolution, ecology, developmental biology, etc) and some downstream uses of iDigBio data (basic research, conservation, …). This could be tracked over time to provide a metric on the diversity and breadth of the user community.

Discussion Point 2.a., Pages 24-25

Original EAB Comment: Coordination with GBIF and other data providers will be essential in defining the unique role of the iDigBio cyberinfrastructure that will need to be preserved after the current funding period. A deep discussion with these groups on current areas of strength and future plans will help clarify where iDigBio's future focus should be. This should be an ongoing process involving regular meetings that include GBIF, iDigBio and other groups, and collaboration on areas of mutual interest.

EAB Response to iDigBio: Please provide a synopsis and list of action items from the Germany meeting, and any other relevant concrete steps planned to address this Discussion Point.

Sustainability Planning will need to consider each of iDigBio's four core areas of activity (engaging the collections community, digitization, database/informatics, and research/education) to determine what must be sustained in each area and develop appropriate plans.

Discussion Point 3.a., Pages 25-27

Original EAB Comment: iDigBio should identify specific metrics to assess progress within each of the four activity areas and set achievable and measurable goals for the remainder of the funding period. This will have a dual benefit of focusing efforts for the remainder of the funding period and forming the basis for an estimate of the work within each area that is not likely to be completed under the existing funding.

EAB Response to iDigBio: The strategic plan contains categories in which numerical metrics could be developed but does not have specific measurable goals. We encourage the iDigBio leadership to develop a small number of metrics for each major activity area that can provide a numerical assessment of progress, and provide regular updates to the EAB on progress in achieving these numerical targets. We would like to see clear goals for the number of collections and individual records to be included in iDigBio by the end of each remaining year of funding, and similar concrete and measurable goals in other key areas.

Discussion Point 3.b., Pages 27-28

Original EAB Comment: Sustainability planning should focus on each of the four activity areas of iDigBio. A draft plan for each area should include; 1) assessment of the need for continued activity in each of the four areas after the end of the current funding period; 2) how the activity

might change assuming the proposal goals are met; and 3) a set of ideas including both federal grant funding and other revenue streams that could contribute to sustaining the anticipated future activity (based on perceived value to users, as derived from data use metrics). Assessment of need in each area should include discussions with a range of stakeholders including data providers and aggregators and a variety of data users.

EAB Response to iDigBio: We consider the sustainability plan a good starting point but a good deal more work will be required before this can be translated into a robust plan of action. The EAB can help guide iDigBio in this area but the ownership of this task rests on iDigBio. One of the key messages we would like to convey is that each activity area within iDigBio will need to be separately assessed for community need beyond the current funding period, and the potential for user-derived revenue, grant support or other funding sources in each activity area.

A good next step to develop a realistic sustainability plan will be for the iDigBio leadership to develop a rough conceptual budget for years 11-15 that shows both costs and potential revenue (broken out by source) for each major activity (e.g. workshop expenses, estimated cost of attendance and number of attendees, cost of portal maintenance and data loading and amount of institutional support available for these expenses, and so on). That document can serve as a guide to what needs to be done before the end of the current funding period to make sure those funds will be available (e.g. write a grant, establish pricing and begin sales for downstream user accounts, etc). This can be further developed into a timeline of tasks to make sure the sustainability planning stays on track.

Discussion Point 3.c., Pages 28-31

Original EAB Comment: In order to develop better metrics on digitization progress, iDigBio should obtain more accurate estimates of numbers of specimens and data providers. There appear to be three classes of data that iDigBio ingests 1) TCN and PEN data from US institutions, 2) data from individual collections and non-ADBC funded programs in the United States, and 3) data from larger museums and programs outside the United States. Estimating the digitization trajectories for #1 and #2 (within reasonable margins of error) are critical in assessing the success of the ADBC program and for planning a strategy to implement beyond 2020. Understanding how many collections and how much data has been produced by #1 versus #2 is very important in knowing the direct impact of ADBC funding and who has been included in TCN funding and who is not likely to receive ADBC funding. Additionally, knowing the progress for the different major taxa is important in establishing sustainability plans with different priorities. For example, if 75% of vertebrate specimens will be digitized in US collections by 2021 but only 10% of arthropod specimens, then devising two separate plans might be appropriate. For arthropods, planning would focus on ways to greatly increase the rate of digitization of existing collections, whereas the priority for vertebrates would be to focus on ensuring that new specimens are routinely digitized and promoting use of the digitized data in a range of research areas. There should be a separate assessment for digitization by museums focused on paleo versus modern specimens as well as TCN efforts digitizing non-specimen records (e.g., vocal recordings). Finally, the vast majority of specimens exist within a small subset of museums, and so another key metric would be to assess the relative progress of efforts within each of the larger institutions and develop strategies to ensure digitization efforts are maximized at each institution.

EAB Response to iDigBio: We appreciate the data provided in response to this discussion point. A further step in this area would be to set some concrete goals for the number of new collections that can realistically be brought into iDigBio during the remaining years of funding, based on the annual rate to date, and by subtracting this number from the list of all collections not yet in iDigBio, estimate the number that will remain unincorporated at the end of funding. This number can be used as the basis for planning the next phase of digitization efforts after the current funding ends.