iDigBio: Integrated Digitized Biocollections
University of Florida

Site Visit Report – 4-5 April 2013

Executive Summary

Intellectual Merit

Progress on the project. The iDigBio project has made much progress since the initiation of funding. Highlights of the achievements thus far include: 1) The project has defined and begun implementation of a software and data storage system that is responsive to community needs, and has given every indication to date that it will meet the project goals. 2) The project has dedicated significant resources towards integrating with the TCNs and understanding their needs, and has taken note of existing projects that can be usefully integrated into iDigBio. 3) The project has created significant synergies with existing TCNs, and has greatly reduced the barrier to digitization by other network projects. 4) The project has initiated and fostered a vast array of partnerships. Including partner representatives in the Internal Advisory Committee is an effective method to enhance these collaborations, which are essential to the goals of the project.

iDigBio leadership. The Leadership Team has developed a detailed project implementation plan that is a highly effective means to communicate among the project team the expectations for each project component, as well as providing clarity for expectations from partners. The project team has also shown appropriate flexibility to move the project forward in partnership with diverse institutions and collaborations. The dedication of the iDigBio leadership team to the success of the project is highly evident, as is that of the very competent staff. A highly collaborative environment has been developed among the staff and Leadership Team, fostered by the open approach to management by the director. We cannot overstate how impressed we are with the highly productive team effort that has occurred to date. We encourage discussions with NSF to enhance the support for this project, in particular for additional personnel, to ensure the continued progress and build upon the success to date.

Suggestions

A strategic plan for iDigBio should be developed. It is now appropriate to construct a prioritized listing of potential future options for sustainability and incorporate these into the plan, along with potential means to obtain financial support. We encourage clarity as to who on the project team serves as the lead for contacts with the large number of partner institutions. We suggest that some method to maintain a communications database that is accessible to all project personnel could reduce the potential for misunderstandings as the project expands. Additional release time for the PI and other members of the Leadership Team is appropriate and we encourage the PIs to discuss this in conjunction with NSF and their institutions.

The role of the External Advisory Board should be further clarified, a plan for transitions for members of this Board established, and effort to enhance the diversity of the Board should be

taken. Broadening participation in iDigBio activities is essential. This project should be providing national-scale leadership in developing methods to enhance the cultural and ethnic diversity of participation in the science of collections.

An explicit evaluation plan, linked to metrics of success for the various activities, is essential to provide evidence of project impacts. Better methods need to be put in place to obtain demographic information from participants as well as to track products, including new collaborations, which have arisen from the project. A leadership evaluation plan should be developed, built in part around input from the variety of constituencies with whom the Leadership Team interacts, with clarity as to who conducts the evaluation.

Mission creep remains a potential risk. Keeping a strict focus on the job of getting more deposited data and providing access to data by more mechanisms should trump any peripheral tool development. The project is currently providing a central repository for all TCNs. This simplifies the process of data acquisition, but the possibility of intractable data storage costs remains a risk. We encourage the team to develop plans to ameliorate this concern. There is risk associated with the overhead for capturing data from each of the many different partnering collections. The team should consider methods to alleviate the need for adaptations to be done on a case-by-case basis, particularly for the smaller, resource-starved collections.

Broader Impact

Scientific outreach activities. The iDigBio team has been very successful at engaging the TCNs and collections community via the establishment of working groups and workshops. The cyberinfrastructure plans will enable a large community of users to access the information provided by the TCNs and, eventually, the international collections community. Results of efforts from the first two years of the grant are being disseminated via publications, presentations, and newsletters published on the website.

Educational outreach activities. The iDigBio project will facilitate enhanced outreach activities that will engage the public and highlight the uses of the data provided by the TCNs. It is still early days for these plans, but the team has set in motion the means for implementation.

Human resources. Mentoring for postdoctoral researchers, graduate students, and staff members include professional development and interactions across disciplinary boundaries. Communication among all participants in the project is evident as is the commitment to achieving the project goals.

Suggestions

The iDigBio team must enhance diversity in all its aspects for working group, workshop, and the external advisory board. Formal cross-training among the graduate students and postdocs associated with the project needs to be emphasized. We encourage open access for all workshop proceedings.

The educational outreach activities need to be ramped up during the second half of the grant period. Specifically, the team needs to develop educational modules and mechanisms to facilitate dissemination of outreach activities by TCNs.

The website needs a major redesign and testing by focus groups. It is strongly recommended that the IT team engage focus groups to periodically test and review the website. This could easily be done as part of the workshops. Community feedback would be helpful.

There needs to be some careful consideration regarding the scope of the iDigBio project prior to engaging research communities that are not focused on collections digitization.

General report of site visit

Vision and Strategic Plan

Implementation plan. The iDigBio cooperative agreement programmatic terms and conditions has served as the overall planning document concerning the major objectives of the project. Details of planning efforts to meet the project objectives have been codified in an Implementation Plan, version 3, (IP3) that has been revised significantly from earlier versions. IP3 includes (i) statements of the vision and mission for the project, (ii) a listing of potential components of the project noting carefully which are considered in the scope of the project and which are not, (iii) description of the governance structure, (iv) statements of the key responsibilities for various project personnel, (v) methodology and roadmaps for project management, education and outreach, serving the research community, digitization, and cyberinfrastructure, (vi) communication, change and risk management, and (vii) project closure procedures. IP3 also includes a highly detailed timeline for project progression noting estimated timing for various components as well as estimates of the status of progress on each component.

IP3 is a highly effective means of communicating among the project team the expectations for each project component, as well as providing clarity for expectations that is particularly of significance for the TCNs. Detailed elaboration of the in-scope and out-of-scope activities helps to ensure that the large number of partner institutions in the TCNs are informed about the limits of what IDigBio should be expected to provide.

Data management strategy. Given the lack of knowledge early in the establishment of the project regarding what TCNs would be supported, and their capabilities regarding cyberinfrastructure and other key aspects of the digitization project, the project team has shown appropriate flexibility to move the project forward. Evidence for this is the capability the project has developed to maintain data in the iDigBio facilities while building the middle-ware appropriate for eventual cloud-distribution, the rapid development of Working Groups, Workshops and other activities to meet project objectives, and the development of a rational governance structure that has ensured true partnerships with TCNs rather than a top-down structure. Including TCN representatives in the Internal Advisory Committee is an effective method to enhance the partnerships that are necessary to success of the project as well as ensure that an appropriate level of input from institutions external to those of iDigBio leadership team is obtained.

Project partnerships. The project has rightly focused its limited resources over the first two years of the project on building the partnerships with TCNs, constructing the cyberinfrastructure needed to get the major digitization effort initiated, and coordinating the initial collection of data for the project. There has been relatively little effort devoted yet to obtaining additional resources for project components that require supplementation beyond the support available in the base award, though there has been at least one funding request submitted. The lack of emphasis to date on sustainability of the project beyond the initial period of support is appropriate. We suggest that it is now appropriate to construct a prioritized listing of potential future options for sustainability and incorporate these into a strategic plan rather than as an addendum to IP3. This could include plans for how support for additional needs might be obtained, and assignments for these to various project personnel.

The project has initiated, and in some cases highly developed, a vast array of partnerships. We suggest that as these continue to expand, it may be appropriate to increase the clarity as to which project personnel have responsibility for partnership management. This is one component of risk management for the project since it only takes one "dropped-ball" to derail what would otherwise be a continuing successful collaboration. It is possible that it may be more efficient to assign an individual on the project team as the lead for contacts with the continually increasing number of partner institutions, acting as the interface to other project personnel and to maintain a communications database that is accessible to all project personnel to reduce the potential for misunderstandings.

Value Added Nature of Integrated Data

Integration with awarded TCN projects. The iDigBio project has dedicated a significant amount of their effort towards obtaining a good integration with the Awarded TCN projects. As part of this effort, they have taken several useful and proactive steps to create social and technical integration of the funded first and second year TCN projects. They have met with these driving user communities to gather requirements and have set expectations of these groups for the role of iDigBio as a coordinator and distributor of the TCN data products. They have used a number of workshops to foster a spirit of cooperation and excitement about the integration of digital data across the TCNs, and to train others in digitization of collections.

While it is hard to evaluate how tightly integrated the existing TCNs are with each other at this point (and given how little time has elapsed in the project, it is unreasonable to expect a high level of integration), it is clear that these groups are engaged with the iDigBio team, and that the iDigBio team is taking the kinds of steps that will build trust with these groups, and will ultimately bring about a close integration.

The iDigBio team has defined a data model that will both meet the current needs/abilities of the TCNs, and will be sufficiently flexible to permit new data items to be added should the community requirements change. They have created a database architecture that will meet the known requirements of the project in a highly scalable way. They have begun to receive and publish the data that are deposited, and are working to help each TCN project provide the data in a format that is maximally useful for distribution to iDigBio users. Although the actual data

collection is just beginning, the iDigBio project has offered infrastructure services to other projects that will help them be seen as a valuable infrastructure by the community, including VM support for projects including the Vertnet project and community projects such as Symbiota and Filtered Push, and offering teleconferencing support via Adobe connect which has been used extensively by the TCN groups.

Synergies with TCNs. One of the most obvious things about the iDigBio project is that it has created enthusiasm within the collections community and among its constituent TCNs. This is clearly a high priority activity of the project, and pervades their communications. Of particular value in the work to date are the training sessions, which have engaged the constituents in overcoming technical hurdles to digitization of specific types of specimens. There is little doubt that this kind of activity breaks down the silos that would exist among a set of disconnected digitization projects, and ensures that activities that are necessary for each TCN do not have to be re-created individually by each of the TCNs. The activity also places the iDigBio project firmly in the role of an enabling hub for a large community. The synergy created by the project is expected to grow and find other forms of expression as the project progresses, and will catalyze a rapid cross-fertilization of ideas and methodologies among both the TCNs and other stakeholders as the community's understanding of the need for digitization grows. Without the iDigBio project, the overhead for each individual TCN would be much higher. The main risk at this point is the overhead for capturing data from each of the many different partnering collections. Currently, adaptations must be done on a case-by-case basis, often with smaller, resourced starved collections that may not have the dedicated IT specialist needed to assist with the transfers.

Adaptivity of the iDigBio project. It is still early days for the iDigBio project to expect rapid addition/adoption of new technologies, as the project must remain highly focused on developing and becoming production ready with its core technologies. However, it is clear that the project has been open to a wide variety of technologies that have been developed by existing projects. Their plans for Year 3 include enabling data access for EOL, BiSciCol, and GBIF, and integration with tools such as Filtered Push into data management applications (e.g., Specify, Arctos, Symbiota, Arthropod Easy Capture Database, KE EMu). This development is important and valuable, but must be conducted in a measured way so as to focus on making each added feature production ready before deployment.

Education and Human Resources

Scientific community interactions. Educational outreach activities for iDigBio have ramped up as the TCNs became active. Most of the efforts have focused on the establishment of collaborative working groups based on themes representing the interests and needs of the TCNs (e.g., Augmenting OCR, Biodiversity Informatics Management, Cyberinfrastructure, Developing Robust Object to Image to Data, Georeferencing, etc.). Concurrently with the establishment of working groups, iDigBio has offered, and is developing, a series of workshops aimed at the stakeholders from the TCNs and from the collections community as a whole. These workshops have focused on digitization techniques, workflows, IT standards, educational outreach, georeferencing, and software development (e.g., the Augmenting OCR Hackathon). Workshop content is available to participants and nonparticipants via the iDigBio website via written

summaries, audio/visual recordings, and pdfs of powerpoint presentations. To access the a/v content, one must have an account and be logged into the site. We encourage open access for all workshop proceedings.

Workshop topics and working groups can be proposed via the iDigBio website. In addition, workshops have been organized for national meetings of scientific societies (e.g., Botany2012, Botany2013, Biodiversity Information Standards (TDWG) 2013 Annual Conference, etc.), and symposia focused on iDigBio efforts and TCN interests have been organized or presented (e.g., 2013 SPNCH, Botany2013). Ideas for K-12 outreach activities are in the development stages, but could be patterned after the efforts of the University of Florida Museum of Natural History program for California Teachers (Fossils in Panama).

In reviewing the demographics of workshop participants, one of the major challenges that emerged is the lack of diversity for underrepresented minority groups. Changing this pattern will require active recruitment efforts, and the PIs and senior personnel need to develop mechanisms to engage members of these communities.

K-Gray interactions. Educational outreach activities, in terms of the development of learning modules for K-Gray, have not yet been started. Similarly, the incorporation of outreach products from the TCNs has not yet been initiated. These activities will need to ramp up in the second half of the funding period.

Educational outreach activities for the general public are in the planning or beta-test stage. A workshop on engaging the public (IDigBio Public Participation in Digitization Workshop) was offered in 2012 with numerous opportunities for engaging the public in activities related to iDigBio identified. Similarly, scientific outreach activities based on the use of the iDigBio resources are in the planning stages with case studies under development as examples of the utility of the resource.

Human resources. Postdoctoral and graduate student mentoring activities have included opportunities for interactions among personnel with different areas of expertise. An example is the monthly journal club where students, postdocs, staff, and faculty meet together to discuss topical papers. Postdoctoral associates have participated in professional development workshops, iDigBio workshops, presentations, and have been actively engaged in outreach activities. Graduate students have been encouraged to participate in professional development workshops, present papers at scientific meetings, and to engage in outreach activities. Formal cross-training among the graduate students and postdocs associated with the project needs to be emphasized.

Website. A major challenge identified for iDigBio is the website. The panel recognizes that the website is under development, but there are some basic issues that need to be resolved before it will be useful to the systematics community, policy makers, or the general public. Part of the problem is information overload and the difficulty in finding material quickly. There is a lot of excellent content on the website, but it needs to be organized in a user-friendly format. A beta site would be advisable before offering messy data to the broader research community.

Suggestions for the website:

On the front page there are boxes highlighting the number of specimen and media records as well as record sets. It would be useful to use these boxes as hyperlinks that would take the user to a summary page of what is included in each of these categories (alphabetized and with the ability to organize by other taxonomic categories).

The portal search needs to be organized by taxonomic categories first, keywords secondarily. When there is no entry, a page should come up stating so, and then give an option to search on keywords associated with the taxon name.

There should be a search function on tags on the first or second page clicked on in the site. Tags should be listed somewhere so the user knows what content is accessible via those hyperlinks. A sitemap should be provided via a hyperlink from each page of the website.

Content organization is in need of improvement. For example, the bibliography page default should be an alphabetized list of products, and everything on the page should have content. The sort functions should be in larger font – they are currently easy to miss.

Active indexing for each page with numerous entries needs to be done – possibly in a sidebar. Content should be alphabetized as the default with options to organize it differently via sort functions.

It is strongly recommended that the IT team engage focus groups to periodically test and review the website. This could easily be done as part of the workshops. Community feedback would be helpful.

Collaboration: Other Institutions; International: other End Users

Collections community. The iDigBio team has actively engaged the collections community and enabled collaborative efforts among the TCNs for solving technical challenges in digitization, workflow, and other activities via the working groups and workshops. Plans for engaging the international collections community are underway, but at the early stages of development. The iDigBio team will need to be careful to balance the needs of the TCNs with those of the international collections community, specifically by truly functioning as leaders for best practices in data access, aggregation, and standardization. The recent formation of a working group that includes international members is a good step toward building a broader community. Organizing workshops for the international community would also be good for building these relationships.

Other end users. It is not yet clear how the team will engage policy makers, government agencies, and NGOs in the iDigBio project. Case studies are being developed to demonstrate the utility of digitized collections for addressing challenges of major impact such as disease

outbreaks, the effect of global change on biodiversity, etc. The case studies presented to the review team could be published as a brochure to deliver to relevant agencies in addition to organizing targeted presentations or workshops to engage members of these communities.

Suggestions. The enthusiasm of the team for establishing collaborative efforts is laudable, but there needs to be some careful consideration regarding the scope of the iDigBio project prior to engaging research communities that are not focused on collections digitization. Facilitating the development of research tools via workshops, publications, etc. may be sufficient for enabling collaborative efforts in contrast to establishing workbenches for doing data analysis such as ecological niche modeling and data mining of genomic data. However, the iDigBio site may be a good repository for some of the tools developed by the broader research community.

Leadership and Management

Leadership. The Site Review Team notes that the devotion of the project Leadership Team to the success of the project is highly evident. The project staff, despite the diverse nature of their disciplinary backgrounds, functions highly effectively together and the project leadership has clearly fostered this through their management style, mentoring of personnel, and openly collaborative view of decision-making. The ability of the Directors to adjust as needed to changing personnel circumstances is evident from the successful transition to a new project manager at the end of the first year of operation. A collaborative environment has been developed among the staff and Leadership Team, which has allowed the project to move forward across very diverse responsibilities over the initial stages of the project and the Leadership Team repeatedly remarked on the importance of this collaborative team effort in building their success to date. This is a significant achievement particularly in light of the highly dispersed nature of the physical locations (both on the UF campus and between the UF and FSU institutions) of many of the project leadership team and staff.

The structure established to obtain input from stakeholders both internal and external to the project staff has no doubt been a major contributor to the success of the project in fostering the communication and collaboration necessary between the IT and biological science personnel. We were also impressed by the evidently seamless collaboration between the personnel at the two institutions involved, UFL and FSU. The success of the project in setting up the initial cyberinfrastructure, fostering partnerships and initiating a host of activities is evidence that the project Directors are highly capable leaders. We cannot overstate how impressed we are with the highly productive team effort that has occurred to date, and the Leadership Team should be proud of their efforts. The PI has been a highly effective motivator for the project, with a leadership style that encourages collaboration. We applaud his efforts to develop a highly successful Leadership Team and provide management oversight while maintaining his other responsibilities at UF. As this project further expands, we suggest that some additional release time for the PI and other members of the Leadership Team is appropriate and we encourage the PIs to discuss this in conjunction with NSF and their institutions.

External advisory board. The variety of input mechanisms for guidance and advice that have been established through the governance plan have evidently been quite effective as the project team noted in particular that they modified their approach to standards for data in response to

partner input. We see no reason to suggest any major changes to the governance structure. However, the role of the External Advisory Board should be further clarified and a plan for transitions for this Board established. Having a very small initial Board of individuals highly cognizant of the initial project needs was appropriate, but we encourage the leadership team to consider an expansion of the Board to meet the changing needs for advice. In particular, we urge that the Board be expanded to provide viewpoints from diverse perspectives, to encourage suggestions of activities that could broaden participation in project activities from individuals in underrepresented groups, and addition of individuals with expertise in education and outreach as these components of the project expand.

Project evaluation. While there has been considerable thought given to project evaluation, particularly in developing assessment tools for individual participant reactions to iDigBio activities, we suggest that development of an explicit evaluation plan, linked to metrics of success for the various activities, would be beneficial. This likely will lead to an enhanced effort from the evaluation coordinator, but it is extremely important for the project to have clear evidence of success (quantitative as well as qualitative) on appropriate metrics (these include products such as publications, usage of the cyberinfrastructure developed, and contributed digitizations by the partners) by the time of the next Site Review. One component of this involves more complete collection of base demographic information on participants in project activities. The response rate and associated acquired information for participants to date is not adequate, and mechanisms should be developed to improve this, certainly for those individuals who receive support to attend project activities and also for participants who do not receive direct support from the project. We encourage the evaluation team to consider new evaluation methods, such as network analysis, to quantify new partnerships/collaborations which have arisen due to project activities, particularly between the individuals from different collections and IT communities.

Leadership evaluation. In response to questions from the Site Review Panel, the leadership team outlined a plan for leadership evaluation built in part around surveys for input from the variety of constituencies with whom the Leadership Team interacts. While there is evidently no push from the university administration structure for a leadership evaluation, we suggest establishment of a clear process of leadership evaluation that can be provided both to appropriate university officials and to future external reviewers. Clarifying who is responsible for collecting survey results and carrying out the evaluation of senior project leadership is essential to this process. For some projects this is done by an external advisory board, and in others by independent components of the base institution's administration.

Expansion. Broadening participation in iDigBio activities is essential. As the major US infrastructure project in the collections arena, this project should be providing national-scale leadership in developing methods to enhance the cultural and ethnic diversity of participation in the science of collections. We strongly encourage the leadership team to accept this responsibility, to reach out to individuals and organizations which have been successful in broadening participation (e.g. the ESA SEEDS project, SACNAS) and to consider fostering additional partnerships themselves and in conjunction with the TCNs with minority-serving institutions. We were particularly concerned regarding the lack of mention in the roadmap in IP3 for education and outreach of any plans to reach out to underrepresented groups. Although it may

be a difficult task, it is essential that the Leadership Team devote effort to enhancing diversity in all its forms among the community of researchers involved in collections science. We encourage the leadership team to consider assigning responsibility for this to one of the personnel on the project, or bring in expertise from UF or FSU to assist in this effort.

Cyberinfrastructure and Data Management

Software Architecture. The project has created an architecture that seems suitable for web based access to large data stores, and for viewing/retrieval of images and media. Development of this architecture is currently on schedule in the development plan, and many new features are planned. Over the next two years, the project will continue developing tools for ingesting data, enabling data access mechanisms for related projects, and integration with community tools. The plans to integrate existing tools, such as taxon name registries and Globally Unique IDentifiers (GUIDs) should be very useful additions to the services offered by iDigBio. The main concern currently is whether or not Drupal will be a sufficiently flexible platform for the web presence that iDigBio requires.

Priorities for the software development seem to be set by the Steering Committee, and are further evaluated through discussions with the TCNs, the IAB, and the EAB, as appropriate. This should give sufficient flexibility to adjust priorities and make changes in the CI development plan as new issues arise over the course of the next two years. It will be important for the steering committee to manage expectations and the desire for new features with the desire to harden production features for release, while at the same time being as responsible as possible to the community. Given that current activities are already limited by budget, addition of new features such as tool/workbench implementations should be delayed until the existing web site has been shored up and the data resources available through the web site become richer and more accessible. A more general recommendation would be to focus energy on the task of serving data programmatically to other workbench developers who have the necessary developer and computational resources.

The iDigBio project has also used the workshop forum to promote integration and problem solving outside the iDigBio project. For example, they hosted an OCR hackathon to promote solutions to some difficult problems in this area.

Data storage. The project plans to integrate data derived from all the TCNs initially. The data at present consists of digital images with a small set of metadata. There current plan is to advocate for depositions that meet the criteria of the Darwin Core, while permitting/accommodating submission of additional metadata. The project will also support the submission of digital media including video and sound recordings, with appropriate metadata.

At present, collections data is made available and searchable through the iDigBio web site, although the amount of data deposited to date is relatively small. The architecture includes a cloud-based data storage system that should be scalable, and permit data access from a distributed set of storage sites. This opens up the possibility of having individual TCNs assume responsibility for their own data storage, which seems to open one way of distributing the costs of long term data storage. At present, however, all data storage is taking place at UF, and is

under the control of the iDigBio project. This may be preferable for performance (speed of query and availability), although apparently was not originally part of the presumed responsibilities of the iDigBio project. From the standpoint of making sure the investment in the iDigBio project come to fruition, housing the data for the project for the near term at any rate seems a very prudent decision until such time as the costs become unmanageable. The data are stored in a geographically redundant way.

At present the estimated data needs of the community are modest, amounting to tens of terabytes. The iDigBio staff projects that growth over the lifetime of the project might reach 100s of Terabytes for digital images, but digital media may reach the 10 PB level. At present the iDigBio storage resources (~100 TB) are adequate for near term storage needs. The project data storage architecture is implemented on a cloud model, which should provide the scalability necessary for the project for the foreseeable future. Software solutions that improve performance and manage failures when searching over large data stores will no doubt be incorporated as required. However, there is no strong plan for who pays for the storage when the need exceeds the capacity of the existing servers, nor is there an obvious way to shift storage costs back to the TCNs (though we believe the initial intent of NSF was for the TCNs to be responsible for the storage costs, and that a distributed storage model was an important feature of the initial iDigBio proposal). While adding data to the resource is correctly on the front burner today, the issue can become a critical one if the project is successful. Developing plans for how data costs will be managed in the long term should be an important activity by Year 4.