TCN / iDigBio Progress Update
March 2013

Contents

InvertNet Update (by Chris Dietrich)

Digitization of slide-mounted and ethanol-preserved specimens is ongoing at several collaborating institutions. To date > 2,500 images have been uploaded, representing >1600 slide boxes and >300 vial racks, together comprising >20,000 individual arthropod specimens.

Most recent activities at the lead institution continue to involve upgrading the project web platform and development and testing of the hardware and workflow for digitizing drawers of pinned specimens.

Student programmers continued working on a image segmentation/annotation tool and customization of the image ingest form. They also began work on linking the InvertNet portal to BugGuide.org, a popular crowd-sourced online field guide to North American insects and other arthropods.

Our new drawer imaging prototype system, based on a 4-arm delta robot design, is up and running and co-PI Chris Taylor described this system as part of his presentation at the recent iDigBio-

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

sponsored Wet Collections Digitization workshop. We are continuing to test the system to streamline and automate the workflow and optimize the number of images obtained for each drawer. We plan to bring the prototype to the Field Museum for a demonstration at the April workshop on pinned insect digitization.

Computer science graduate student, JonMark Lau continued work on improved approaches to 3-D reconstruction.

Sustainability. INHS biodiversity informaticians Phil Anders and Matt Yoder (members of the permanent INHS IT staff) have been attending the biweekly InvertNet technical team meetings (that also include our collaborators from engineering and computer science departments) since early in our project. This will familiarize INHS IT staff with the InvertNet cyberinfrastructure and allow them to assume responsibility for maintaining the HUBzero cyberinfrastructure platform that supports the InvertNet portal and content management system beyond the period for which we have NSF funding. In our original data management plan, submitted with the TCN proposal to NSF, we stated that INHS would form an InvertNet steering committee consisting of scientists and IT support staff to provide support for TCN cyberinfrastructure beyond the 4 years for which we have been funded by NSF. Our main goals with regard to sustainability are to provide a cyberinfrastructure platform that is suitably robust and user friendly that users of the infrastructure, including various collections that join our network, will incorporate the workflows and infrastructure we develop into the routine, day-to-day operations of their collections.

North American Lichens and Bryophytes Update (by Corinna Gries)

Digitization progress in numbers:

Lichens- http://lichenportal.org

- Herbaria: 34 (up by 6 since 12/2012)
- Label images on iDigBio server: 253018 (up by 55315 since 12/2012)
- Specimen records in database: 922383 (up by 66403 since 12/2012)

Bryophytes - http://bryophyteportal.org

- Herbaria: 25 (up by 4 since 12/2012)
- Label images on iDigBio server: 186259 (up by 76616 since 12/2012)
- Specimen records in database: 1327062 (up by 54446 since 12/2012)

A number of images have not been uploaded by collaborating institutions and are expected in the near future, as we are assured they have been captured and are managed locally. Specimen records are increasing through imaging of labels and through integrating existing databases, which explains

that more specimen records have been added to the lichen portal than images. One specimen record may be connected several label images, which explains why there were more images than records added to the Bryophyte portal.

Digitizing smaller collections:

Both NY and WIS have completed or are in the process of digitizing collections from smaller herbaria (VT, MONT, NHA) and WTU has provided access to records from their collaborators. ILLS is in the process of receiving their first shipment and F has been digitizing specimens at MOR.

Outreach:

The citizen science interface for LBCC http://lbcc1.acis.ufl.edu/ is under intensive development. The content from the old LBCC community site has been moved and work in underway on the crowd sourcing application. Currently the emphasis is on the integration with the Symbiota data entry application, which needs to be simplified and made more accessible to the non-expert user.

Several Symbiota trainings are planned for this year. Upcoming in the near future (March/April) are for bryologists at the 'so-be-free' meeting in California and for lichenologists at the Tuckerman lichen workshop in Georgia.

An abstract for the presentation on LBCC has been submitted to the Mycological Society Meeting by Andrew Miller.

Symbiota developments:

The data feed from Symbiota to iDigBio is almost complete. Only minor adjustments are expected to be needed at this point and then all new data from Symbiota portals will be available to the new iDigBio interface.

Ed Gilbert was involved in organizing and running the first hack-a-thon. As a result of the workshop several new digitization approaches are being integrated into Symbiota.

Tri-Trophic Project Update (by Toby Schuh)

a. Progress in digitization efforts

We have digitized just shy of 300,000 insect specimens; these are novel records, all digitized since the beginning of the TCN award. The tritrophic TCN has captured images for more than 260,000 plant specimens.

All collaborating institutions, entomological and botanical, are actively capturing images and specimen data and are on track with the progress projected in our original proposal to the NSF.

b. Share best practices and standards

We have prepared and posted on our TCN website manuals for best practices to be used in the execution of our digitization activities (http://tcn.amnh.org/documents) . These documents are updated based on feedback from project participants, with the aim to constantly improve the quality of our data capture. We have participated in the MISC Working Group, Cyberinfrastructure Working Group, and the DROID Working Group. We also participated in the Wet Collections Digitization Workshop recently held at the University of Kansas. With regard to the last, it is our view that many issues associated with wet collections were left partially addressed, or not addressed at all in Kansas, and that another such workshop should be a priority.

c. Identify gaps in digitization areas and technology

We look forward to a more final decision on the issue of UUID structure for our project and for the overall biological collections digitization community. Resolving this issue will put to rest one of the few remaining technical issues confronting our TCN.

It is our view that pre-curation and verification of identifications are important aspects of all TCNs and we encourage continued support of and emphasis on these issues.

In relation to sustainability (see below) we want to emphasize the need for an identifier resolver at the record and institution levels. The absence of such a resolver has been a problem for GBIF, the community in general, and will doubtless be a problem for iDigBio.

d. Report on collaborations with other TCNs, institutions, and organizations

We are developing a multi-trophic level collaboration with the NSF-funded Bee Databasing Project to apply methods of data-mining to our TCN database and to produce comparative analyses of trends in phenology and the effects of environmental change on host specificity.

e. Ideas for sustainability of the digitization efforts and the collaboration networks

We argue that one model of sustainability is supporting a distributed network of easily customizable, lightweight applications such as Arthropod Easy Capture (AEC). It is for this reason that we have submitted to the code for AEC to Source Forge as a mechanism for getting the software into the open-source community. We argue for the inclusion of software support in NSF-grant applications as a cost effective alternative to the continued de novo development of new applications. Such an approach allows for continued development and improvement without the need for repeated development of core elements of the application. One possible

and effective extension of this concept is the development of networks based on taxon expertise, as opposed to always centralizing data aggregation around institutional ownership. With the incorporation of identifier resolvers we believe the taxon-focused approach will be increasingly viable.

PALEONICHES Update (by Bruce Lieberman)

Here is where things stand with our Paleoniches TCN. Regarding the University of Kansas, we currently have 41,128 collection objects, with accurate, verified localities, databased. Further, 10,431 of those collection objects are georeferenced - that equals 853 georeferenced localities. We have also hired a new post-doc, Dr. Michelle Casey, who received her Ph.D. from Yale University and she will begin working on the project at KU June 1st. We are very excited about adding her to our team. Finally, applications from both the Yale Peabody Museum and the University of Texas were submitted to form a PEN with our TCN and each of these did get funding so those are two new collaborations that we will soon be initiating as part our project.

Since the last update, PI Hendricks (San Jose State University; SJSU) and his undergraduate student assistant have put the "Digital Atlas of Ancient Life" online at http://www.geosun.sjsu.edu/~jhendricks/AtlasTemp/ (this website location is temporary). Other project participants have commented on the website and their suggested changes have been incorporated. The website has also incorporated a Twitter feed. Additionally, two groups of Neogene gastropods have been added to the website: the Conidae and Fasciolariidae. In the next month, attention will shift to adding a third group of Neogene gastropods (the Murcidae) to the webpage and initial construction of the Pennsylvanian period component of the site will begin.

At Ohio University, PI Alycia Stigall reports 632 unique species including a total of 2,406 individual specimens were cataloged and digitized during February. This included all steps from specimen ID and labeling to Specify entry and collection organization. The digitization of basic data (specify entry of basic data and georeferencing of localities) for the Jack Kallmeyer collection is on target for completion by the end of April 2013.

For other participants working under her, the Invertebrate Paleontology Department at Cincinnati Museum Center has hired 2 undergraduate students from the University of Cincinnati. They are currently photographing Ordovician fossils and entering collection records into the KeEmu database. Efforts are underway to format the catalogue record data in order to georeference localities for Ordovician taxa. At the Miami University Limper

Museum, Kendall Hauer and two student workers have corrected preexisting errors in the database's descriptions for all of the Shideler localities in Ohio, Indiana, and Kentucky (the core of the collection). In addition, approximately 450 of Shideler's 1,040 Ohio localities have been georeferenced.

Macrofungi Update (by Barbara Thiers)

a. Digitization Activities:

Assembling existing records to share through the MycoPortal: During February 2013, approximately 424,00 existing records were uploaded to the MycoPortal. As of 1 Feb 2013, approximately 890,000 specimen records from 22 institutions are being shared through the MycoPortal. We still have perhaps another 100,000 existing records to add from participants who will join the project this coming year. We currently have 20,928 images on line, including specimen, label, and living organism images.

New Digitization (all Participants)

1) Skeletal records (specimen label image plus data record consisting of taxonomic name, collector and collecting number) created: 18783

2) Label images captured: 19554

3) Full records captured: 3544

4) Records completed in the NYBG Project Center: 2371

5) Specimens imaged: 0

6) Fieldbook records created (database record consisting of locality, collector, number, and date): 7723

7) Ancillary items digitized: 1400 (Hesler notebook pages digitized and saved as individual species-based files

b. Project Management:

Participant Interaction: Preliminary work began on initiation of subawards for year 2 participants; additional correspondence gave more details about participation to upcoming participants.

Project Training: Work to revise and update the project procedures manual continues

Project Communication: Plans were finalized this month for presentations to be given about the project at summer meetings:
Society for the Preservation of Natural History Collections (SPNHC) meeting: Presentation and poster about the project by B. Thiers and S. Acensio.

Mycological Society of America (MSA) meeting: Abstracts for four presentations have been submitted: Oral presentations by Dr. Andrew Miller (ILLS) and Matthew Foltz (MICH); poster presentation by B. Thiers and R. Halling and Bates; poster presentation on Hesler notebooks by K. Hughes and R. Petersen.

c. Education and Outreach:

B. Thiers was invited to write an article for the amateur mycology magazine McIlvainea about the MycoPortal and how to collect and preserve scientifically valuable specimens.

d. Research:

Nothing to report.

e. Identify gaps in digitization areas and technology:

Natural language parsing—this is a serious gap for us. Since we use Symbiota, it sounds as though soon there will be additional tools that were developed by Darryl Lafferty embedded in the system. However at the moment we are having to spend more time keystroking than would be ideal.

f. Report on collaborations with other TCNs, institutions, and organizations:

Nothing to report at this time.

g. Ideas for sustainability of the digitization efforts and the collaboration networks:

All of our participants have their own imaging equipment and have (or will) receive training in all the techniques needed to capture and manage their data. We anticipate that the MycoPortal will be key to some major research efforts in the mycology community, notably the effort to

create a Mycoflora of North America. We are hoping that the use to which the MycoPortal data can be used by both professional and amateur mycologists will be incentive for institutions to continue to add data to the Portal.

New England Vascular Plants Update (by Patrick Sweeney)

Hardware & Software Development

The engineering group at the University of Oklahoma have been focused on testing the conveyor system and refining the controller station Server-Client application. Below are some details:

Hardware development:

1) Ran integration test on 18ft conveyor system in OU laboratory. All major components worked together; however, there are some minor software bugs to be resolved.

Software development:

1) Continued making refinements to entry system UI-Interface to key-in specimen meta-data to the database.

2) Continued time studies to bring down time taken by the entry user to key-in all relevant specimen meta-data using the UI interface.

3) Continued to modify database to incorporate image, specimen and process level meta-data.

The Hardware/Software development team at OU anticipates that the conveyor belt apparatus can be deployed in June or July.

Other activities:

- Draft versions of rdf/xml format for data exchange files are complete and are awaiting feed back from FilteredPush team. Currently working on code to generate data exchange files from digitizing station databases.
- Moved portal database and Symbiota instance to iDigBio server. Currently testing deployment before making the production instance.

Digitization

Preparing collections for digitization and capturing collection level-information (i.e., “pre-capture”) are still the the primary activities. Collections-level data capture is occurring at most of the budgeted

digitizing institutions with some institutions exceeding 50% completion of collections-level data capture. At institutions where sorting is completed, pre-capture rate is proceeding as fast as expected or faster. At institutions where considerable sorting and barcoding are happening in conjunction with pre-capture, the pre-capture rate is considerably slower. Primary digitization, that is imaging of specimens and capture of specimen-level data, will begin no later than July, when the conveyor belt system and light-boxes are due to be deployed.

Collaborations

Collaborating with Ed Gilbert (LBCC project) and Paul Morris (FilteredPush) to develop mechanism to transfer data from digitization stations to Symbiota.

Southwest Collections of Arthropods Network Update (by Neil Cobb)

The Southwest Collections of Arthropods Network (SCAN) is currently focused on activities in five areas, 1) finalizing draft protocols, 2) soliciting participation by non-SCAN museums, 3) digitizing specimen label data,. 4) Developing Filtered Push, and, 5) enhancing Symbiota for entomological collections.

1. We have four draft protocols in prep or finished. The first protocol for entering, editing, and searching for data in SCAN Symbiota is located on the project website http://scan1.acis.ufl.edu/?q=content/protocols . We have four more in preparation 1) Unique identifiers, 2) Using SCAN data portal, 3) Imaging, and 4) Quality Assurance.

2 & 3. One of our broader impact activities is to host specimen records from collections that are not funded by the NSF-ADBC program. We have just developed a search function to easily parse out records from NSF-funded SCAN collections and other museum data that are have alternate sources of funding to upload their data to SCAN Symbiota . The five non-SCAN collections we have included are listed below. Table 1 below shows statistics compiled from all the records on the SCAN Symbiota portal and only records from museums receiving NSF-ADBC funding. We are digitizing SCAN specimens at our projected rate although we need to continue georeferencing since only 50% of our records are georeferenced. We have started to image but will not plan to go into production mode until summer, 2013.

Non-SCAN Collections

Entomology Collection at the Natural History Museum of Utah

Gregory P. Setliff Collection – Kutztown University

Museum of Comparative Zoology, Harvard University

Colorado Plateau Museum of Arthropod Biodiversity National Park Collections

Scarab Central: World Scarabaeoidea

Table 1. Summary statistics for digitizing and georeferencing specimen label data as of March 12, 2013. First row includes all records in SCAN data portal and the second row only includes data from NSF-ADBC participating museums.

	Specimens	**Families**	**Genera**	**Species**	**# Georeferenced**	**# Specimens Identified to Species**	**# Georeferenced and Identified to Species**
All Specimens	345,502	1,004	7,505	15,393	43%	64%	38%
SCAN	211,257	512	4,852	9,178	56%	73%	50%

4. FilteredPush has reconfigured the SCAN network instance to use the W3C Open Annotation Community Group's Open Annotation Ontology Specification. Filtered Push has demonstrated harvest of a taxonomic authority file from Symbiota into an ontology, and the use of reasoning on that ontology to notify parties who express interests at higher taxonomic ranks of annotations making assertions about included taxa.

(A) Set up an OAI/PMH harvesting of the taxonomic authority file from SCAN Symbiota into the network, along with conversion to an ontology.

(B) Set up demonstration for using the ontological representation of the taxonomic hierarchy to match an annotation that asserts an identification using a lower taxon with an interest expressed on a higher taxon.

(C) We are working on productizing (1) and (2) for deployment in the production SCAN system in the next month.

(D) We have started discussing with Ed Gilbert the development of user-interface elements within Symbiota to notify taxonomists of inadequately determined specimens with images that they may be able to put finer-grained identifications on.

(E) We've arranged a mechanism with the Specify-6 team for coordinating releases of Specify-6 upgrades with releases of the Specify Driver that we use to help ingest annotations into Specify.

(F) We are (this week) discussing the integration of annotations with future Specify versions with the Specify/Dina collaboration that is developing a web-based next-generation Specify.

Requests from FP for SCAN taxonomists:

(A)Provide a list of taxonomists involved in the project and the higher taxa that they are responsible for providing taxonomic determinations?

5. Symbiota now has all of the functionality that is essential for entomological collections and we will continue to add more options in 2013 as time permits. We have created a summary statistics function so anyone can obtain summary data for any set of collections. This currently includes # specimen records, # georeferenced, # identified to species, and the # identified to species and georeferenced. The statistics are provided for an entire collection and also by family. By April, 2013 we will have a these same summary statistics displayed by country.

iDigBio Update (by David Jennings)

- NSF will be conducting a site visit at iDigBio on Thursday, April 4, 2013. We don't anticipate the need for TCN participation directly, but need other support, such as completion of this month's progress report. We also need copies of all previously submitted TCN annual reports to NSF.
- iDigBio recently released and update to the Website and Portal. The website was redesigned to focus on making it easier to understand and to use, and to be more approachable to a lay visitor. The Portal and APIs were redesigned to focus on correcting shortcomings identified in the technology demonstrator and in completing the foundation for a system that will serve the community for years to come, including user interface improvements, improved stability and flexibility of the API, and due diligence to create requirements for data providers that are minimal but sufficient to ensure the smooth operation of the system.
- Joanna sent out a questionnaire to all TCNs to assess the readiness of TCN data for ingestion into the iDigBio portal. This is allowing us not only to plan our ingestion process but also to better assist the TCNs with any issues.
- Shari sent out an assessment survey to the TCNs, which will be sued for the site visit.
- iDigBio created a Drupal portal for SCAN to facilitate an internal network of users and make data accessible to the public. This technology is available to all of the other TCNs and they should contact Kevin for more information.

- IDigBio uses Adobe Connect for the steering committee meetings etc. This collaboration software is available for use by the TCNs as well. Contact Kevin for help using/setting up Adobe Connect software.
- Many workshops and other events are being planned. Please check on the iDigBio website for details.