Symposium:
Workflows and Challenges in the Digitization of Biological Specimens

8:30 am – noon Friday, April 12th Parlor D
Chair: Ashley Morris

ASB and iDigBio are proud to co-sponsor this symposium and Saturday's workshop on the digitization of biological collections. Integrated Digitized Biocollections, or iDigBio, is the national resource for Advancing Digitization of Biodiversity Collections (ADBC) funded by the National Science Foundation. While the effort to digitize museum collections has gone on for quite some time, many of us are either just being exposed to this or are working with collaborators to develop new Thematic Collections Networks (TCNs) or Partner to Existing Networks (PENs) in our own disciplines. The methodology used and technologies employed to digitize a collection vary widely across groups, and the reasons for choosing one over another may not be obvious to those of us who are new to the effort. This symposium featuring speakers from existing TCNs and future TCNs, followed by Saturday's workshop on workflows, will inform a large number of individuals from diverse disciplinary backgrounds in a short period of time. Our goals are to inform and educate both novice and experienced digitizers, while building a support network such efforts.

Scheduled speakers:

8:30 AM - *Brief remarks*

Ashley B. Morris (Organizer), Department of Biology, Middle Tennessee State University

8:35 AM - Introduction to iDigBio

Gil Nelson, iDigBio and Robert K. Godfrey Herbarium at Florida State University

8:45 AM - InvertNet: A new paradigm for digital access to invertebrate collections

Chris Dietrich[1], Johh Hart[2], David Raila[3], Umberto Ravaioli[4], Nahil Sobh[5], Omar Sobh[6], Chris Taylor[6]

[1] Illinois Naural History Survey, Prairie Research Institute, University of Illinois; [2] Department of Computer Science, University of Illinois; [3] Department of Computer Science, University of Illinois; [4] Department of Electrical and Computer Engineering, University of Illinois; [5] Beckman Institute for Advanced Science and Technology, University of Illinois;[6] Illinois Natural History Survey, Prairie Research Institute, University of Illinois

9:15 AM - Plants, herbivores, and parasitoids: Tri-trophic digitization strategies

Kimberly Watson[1], Robert Naczi[2], Melissa Tulig[1], Randall Schuh[3], Katja Seltmann[3]

[1] The William and Lynda Steere Herbarium, The New York Botanical Garden, Bronx, New York; [2] Institute for Systematic Botany, The New York Botanical Garden, Bronx, New York; [3] Division of Invertebrate Zoology, American Museum of Natural History, New York, NY

9:45 AM - 10:30 AM *Break for Poster Session*

10:30 AM - Efficiencies and challenges of organizing an ADBC TCN project on southeast freshwater macrofauna

Hank Bart
Tulane University Biodiversity Research Institute, Belle Chasse, LA

11:00 AM - So many herbaria, so little time: Challenges and opportunities in biodiversity informatics

Zack E. Murrell

Department of Biology, Appalachian State University, Boone, NC

11:30 AM - Collaborative digitization workflows with Specify 6

Andrew Bentley

Biodiversity Institute, University of Kansas, Lawrence, KS

Abstracts

Chris Dietrich[1], Johh Hart[2], David Raila[3], Umberto Ravaioli[4], Nahil Sobh[5], Omar Sobh[6], Chris Taylor[6]

InvertNet: A new paradigm for digital access to invertebrate collections

[1] Illinois Naural History Survey, Prairie Research Institute, University of Illinois; [2] Department of Computer Science, University of Illinois; [3] Department of Computer Science, University of Illinois; [4] Department of Electrical and Computer Engineering, University of Illinois; [5] Beckman Institute for Advanced Science and Technology, University of Illinois;[6] Illinois Natural History Survey, Prairie Research Institute, University of Illinois

InvertNet, one of the three Thematic Collection Networks (TCNs) funded in the first round of the U.S. National Science Foundation's Advancing Digitization of Biological Collections (ADBC) program, is tasked with providing digital access to ~60 million specimens housed in 22 arthropod (primarily insect) collections at institutions distributed throughout the upper midwestern USA. The traditional workflow for insect collection digitization involves manually keying information from specimen labels into a database and attaching a unique identifier label to each specimen. This remains the dominant paradigm, despite some recent attempts to automate various steps in the process using more advanced technologies. InvertNet aims to develop improved semi-automated, high-throughput workflows for digitizing and providing access to invertebrate collections that balance the need for speed and cost-effectiveness with long-term preservation of specimens and accuracy of data capture. The proposed workflows build on recent methods for digitizing and providing access to high-quality images of multiple specimens (e.g., entire drawers of pinned insects) simultaneously. Limitations of previous approaches are discussed and possible solutions are proposed that incorporate advanced imaging and 3-D reconstruction technologies. InvertNet couples efficient digitization workflows with a highly robust network infrastructure capable of managing massive amounts of image data and related metadata and delivering high-quality images, including interactive 3-D reconstructions in real time via the Internet.

Kimberly Watson[1], Robert Naczi[2], Melissa Tulig[1], Randall Schuh[3], Katja Seltmann[3]

Plants, herbivores, and parasitoids: Tri-trophic digitization strategies

[1] The William and Lynda Steere Herbarium, The New York Botanical Garden, Bronx, New York; [2] Institute for Systematic Botany, The New York Botanical Garden, Bronx, New York; [3] Division of Invertebrate Zoology, American Museum of Natural History, New York, NY

Integrated data on insect herbivores, their plant hosts, and their insect parasitoids are currently not accessible, nor are comprehensive data on their relationships available online. The primary goal of the Tri-Trophic Digitization TCN (Thematic Collections Network), funded in July 2011, is to digitize, integrate, and make accessible online ±4 million specimen records (±2 million de novo) representing three major groups of organisms from the North American biota: the phytophagous and economically important insect order Hemiptera, 20 of their target host plant families, and their insect parasitoids in the Hymenoptera. With the combined expertise and resources of 32 collaborating museums (14 herbaria and 18 entomological institutions), we are implementing streamlined workflows for specimen data capture, imaging, and georeferencing in order to maximize digitization efficiency. Complete insect specimen data is captured in a centralized database via a web portal hosted by the American Museum of Natural History, with representatives for each species imaged. Conversely, all herbarium specimens are imaged and minimal data captured initially, with remaining data captured later from the images through the use of OCR software, duplicate matching, or manual entry. All data and images will be made available through web portals including the TCN project website, GBIF, iDigBio, and Discover Life, the latter providing animal-plant data integration and mapping tools. By assembling and integrating data on geographic distributions, host associations, and phenologies, our tri-trophic approach will benefit a wide range of research questions and practical applications in such fields as agriculture, systematics, conservation, ecology, climate change studies, and biogeography.

Hank Bart

Efficiencies and challenges of organizing an ADBC TCN project on southeast freshwater macrofauna

Tulane University Biodiversity Research Institute, Belle Chasse, LA

This presentation reports on an effort that is underway to organize an ADBC TCN project on freshwater macrofauna of the southeastern US. The project would involve collections of fishes, crawfishes and aquatic mollusks at institutions in Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Texas and Virginia. Digitization efforts would focus mainly on crawfishes and mollusks, as fish collections of the southeast are largely digitized and networked. The project would benefit from the availability of digitized and georeferenced SE fish collection localities and the locality matching features of GEOLocate's Collaborative Georeferencing Platform. The project will also benefit from records of SE macrofauna being digitized in other TCN projects (e.g., InvertNet). The resulting integrated resource on freshwater macrofauna of the southeast would be used to explore the biotic impacts of river modification. However, the project is not without challenges. A number of the collections lack permanent curators, a situation that poses issues for institutional leadership. An approach to the project is presented that capitalizes on the efficiencies and attempts to address the challenges.

Zack E. Murrell

So many herbaria, so little time: Challenges and opportunities in biodiversity informatics

Department of Biology, Appalachian State University, Boone, NC

Major issues facing the museum informatics community are 1) the cost/benefit of obtaining specimen data from small collections and 2) the data quality needed to be useful to scientists. The SouthEast Regional Network of Expertise and Collections (SERNEC) was developed from the southeastern herbarium community. With six years of funding from the NSF Research Coordination Network program, we developed an inventory regarding the number, size, distribution and curatorial expertise of the 232 southeastern herbaria. We also organized curators, through workshops and symposia, to develop a community well-versed in current biodiversity informatics methods. We are now poised to address "dark data" and the "long tail of science", as we gather metadata and specimen data from smaller regional collections. Additionally, we are working to access the collective expertise of the regional curatorial community in order to accomplish the higher-end specimen processing of georeferencing and concept mapping. The goal of our efforts is to generate a research dataset at a fine scale, by accessing all the collections in the region, and at a scope that is large enough to address significant biogeographical, conservation and climate change questions. We face substantial challenges in this effort. We plan to meet these challenges by using tiered expertise of citizen scientists, students, and herbarium professionals, coupled with cutting-edge efficiencies in data acquisition, to generate high quality specimen metadata. Through these efforts, we intend to provide information to the greater museum community regarding the value of small collections and resident curatorial expertise to global biodiversity informatics efforts.

Andrew Bentley

Collaborative digitization workflows with Specify 6

Biodiversity Institute, University of Kansas, Lawrence, KS

The Specify Software Project supports more than 400 museum and herbarium collections with open source software and technical support services for digitizing and mobilizing specimen data. Most collections utilizing Specify are repositories at universities or smaller colleges with holdings in the tens to hundreds of thousands of collection objects, but Specify also supports collections with over a million data records, as well as large multinational projects.Specify aligns with goals of the NSF ADBC Program, iDigBio HUB, and the Thematic Collection Networks to promote collaborative cataloging, by offering interfaces which can be individualized to local collection data workflows, as well as functions associated with collaborative data entry, duplicate specimen discovery, data reconciliation, and publishing records to species occurrence data aggregators.We will touch on Specify 6's capabilities for data entry, bulk data uploading and validation, as well as the integration of specimen images and labels into digitization workflows. We will nimbly show Specify technologies for image, label and OCR data archiving, duplicate data discovery and re-use (Scatter Gather Reconcile), and for data publishing. We will preview a new generation of Specify which moves specimen data management to the web. Finally, we will re-affirm our commitment to support curation and research with biological specimen data, and to collaborative development and co-ownership of open source software for biodiversity collections.

Workshop:
Workflows and Challenges in the Digitization of Biological Specimens

8:30 am – 5:00 pm Saturday, April 13th Parlor D
Chair: Ashley Morris

Saturday's workshop is modeled after the iDigBio DROID (Developing Robust Object-to-Image-to-Data) workshop. Developing an efficient workflow that works across collaborative institutions seems to be one of the biggest challenges to scaling up of digitization efforts. The target audience will be one of diverse taxonomic backgrounds, such that the workshop will emphasize best practices that are broadly applicable. Speakers from Friday's Symposium will also participate in the workshop, providing insight into workflow design and implementation. Gil Nelson and Deb Paul (iDigBio) will work with Ashley Morris (MTSU) and Hank Bart (Tulane University) to lead participants in this effort.