Advancing the Accessibility of Digital Media for Biological Research in the 21st Century

May 7, 2015

A meeting presented by the National Resource (Home Uniting Biocollections, or HUB, at the University of Florida and Florida State University) for Advancing Digitization of Biodiversity Collections (ADBC) in collaboration with the Cornell Lab of Ornithology, Cornell University

The digital era has moved the biological sciences into the era of "big data" by creating new opportunities for data collection, storage and access, and leading to an expectation for open data sharing that will benefit the broader scientific community. Indeed, funding agencies and many journals now require that the research they support make data broadly accessible in an appropriate data archive, which in turn is creating enormous potential for new discovery. The biological research community has risen to this challenge in a number of key ways, for example through creation of key standards and best practices, data ontologies, and specialized data archives where researchers can easily deposit or access data of various types (e.g., GenBank, MorphBank, Dryad).

But these opportunities also are creating enormous challenges and a growing need for "biocuration" – practical and powerful approaches to managing data and making them accessible to the broader research community. For example, "biodiversity media" – audio and video recordings of living organisms in nature – capture the behavioral phenotype of individuals in ways that traditional museum specimens (e.g., "study skins") simply cannot, and various scanning technologies are unlocking hidden information from physical specimens. At the same time, new technologies have increased the quality and decreased the expense of digital recording in the field and the lab. Today it is common for researchers to collect many hours of media, for example in projects aimed at studying individuals within and across populations, or acoustic monitoring of particular regions. Importantly, these same approaches are also becoming popular with the general public, and a large number of nonprofessional "citizen scientists" are collecting and sharing biodiversity media through traditional recording efforts or by using camera traps and "nature cams". As a result, biodiversity media are being collected at global and rapidly expanding scales, and are a valuable resource for scientific discovery, yet the strategies and approaches used to preserve and make accessible other data formats or specimens are not suitable for digital media; a different approach is needed.

The growing prevalence and volume of digital media data raises a number of key questions that have not been addressed. Which of these data should be preserved and for how long? What standards should be adopted, and what are the best practices for preserving these data? How can the data be organized, and the associated metadata enriched, to make the recordings more accessible to those who need them? How should cost-effective storage (e.g., cloud storage) be balanced with ease of access? And how can these media be linked to associated specimens, data, and publications archived elsewhere? Some of these questions are unique to audiovisual data, whereas others apply to a spectrum of research data types. These questions and others must now be addressed for audiovisual data, informed by advances made for other data formats

but brought into the realm of audiovisual, and done so within the broader context of advancing research in the biological sciences.

This meeting will bring together museum professionals, researchers who collect and use biodiversity media in various ways, and others familiar with digital media and databasing approaches to discuss the above questions. The main deliverable of this meeting will be a white paper that addresses explicitly the challenges of making audiovisual data broadly accessible. This report will not provide answers to all of the issues raised, but rather will develop a framework that identifies and prioritizes the key issues and make recommendations for the most important next steps. This report will be a key step toward finding sustainable solutions. Moreover, although some of the issues addressed will be unique to audiovisual data, others will be more broadly relevant to other data types, such that solutions for audiovisual data can help inform approaches to archiving other data types and vice versa. This meeting also will lead to one or more review papers for publication in peer-reviewed journals, where they will help inform the broader research community of both the power and the challenges of working with audiovisual data. The publications that result from this meeting will serve as a guide for future development of audiovisual data archival capacity, and as such will have clear and long lasting impacts on the broader research community.

Schedule: May 7 (Thursday)

Time	**Topic/Activity**	**Responsible**
	Breakfast (on own)	
7:30 a.m.	Depart hotels, gather @ CLO	
8:00-8:05 a.m.	Welcome & logistics	Mike Webster, Cornell
8:05-8:25	Biodiversity media in the 21st Century	Mike Webster
8:25-8:45 a.m.	Fleshing out the agenda (group discussion)	Mike Webster
8:45-8:55	Introduction to Discussion 1	Gil Nelson, iDigBio
9:00-9:40 a.m.	Discussion 1: What should be preserved? • Are media specimens? • Which media to preserve? • Digitization of legacy media? • Levels of accessibility?	Breakout groups
9:40-10:10 a.m.	Discussion 1 synthesis	Gil Nelson, iDigBio
10:10-10:40 a.m.	Coffee Break	

10:40-10:50	Intro to Discussion 2	Ed Scholes, Cornell
10:50-11:30	Discussion 2: How to preserve it? • Standards and best practices? • Should these apply across the board? • Impacts on sustainability?	Breakout groups
11:30-12:00 p.m.	Discussion 2 synthesis	Ed Scholes
12:00-1:00 p.m.	Lunch (Catered, on-site)	
1:00-1:10 p.m.	Intro to Discussion 3	Greg Riccardi, iDigBio
1:10-1:50 p.m.	Discussion 3: Maximizing scientific impact • Metadata quality • Tools/incentives to facilitate archival? • Crowd-sourcing approaches?	Breakout groups
1:50-2:20 p.m.	Discussion 3 synthesis	Greg Riccardi
2:20-2:40 p.m.	Coffee break	
2:40-2:50 p.m.	Intro to Discussion 4	David Winkler, Cornell
2:50-3:30 p.m.	Discussion 4: Databasing and linking • Linking media to specimens & other data • Publication data archiving • Sharing/linking digital media	Breakout groups
3:30-4:00 p.m.	Discussion 4 Synthesis	
4:00-4:20 p.m.	Break	
4:20-5:00 p.m.	Discussion: Overall summary and key points	Mike Webster
5:00-5:30 p.m.	Next steps	Mike Webster
5:30 p.m.	Break for dinner	