Inaugural Digital Data in Biodiversity Research Conference
Plenary and Capstone Sessions Abstracts
5-6 June 2017
Co-sponsored by the University of Michigan and iDigBio
Hosted at Michigan League, University of Michigan

Bentley, Andrew
Collections Manager, Ichthyology
KU Biodiversity Institute
abentley@ku.edu

Natural History Data Pipelines: The Good, the Bad, and the Ugly

Collections, aggregators, data re-packagers, publishers, researchers, and external user groups form a complex web of data connections and pipelines that form the natural history knowledge base essential for collections use by an ever increasing and diverse external user community. We have made great strides in developing the individual parts of this knowledge base and we are now well poised to integrate these capabilities to address big picture questions. Although we need to continue work on the individual pieces, the focus now needs to be on integration of these disparate sources of data that create the pipeline.

In order for the system to function efficiently and to the benefit of all parties, information, data, and resources need not only to be linked efficiently but flow in both directions. Data and resource flow in the reverse direction is the focus of this talk with the collections supplying the data for these pipelines not benefiting from those above them in the chain. There are benefits to collections from inclusion by aggregators and subsequently use by researchers and publishers that are not being realized. This presentation will focus on some of the important elements that need further development in order for all to succeed.

Brainerd, Beth
Professor, Biology and Medical Science
Brown University
elizabeth_brainerd@brown.edu

Video Data and Motion Analysis in Comparative Biomechanics Research

Film or video recordings have long been important primary data for research in comparative biomechanics. Innovations have included the use of two or more cameras to capture 3D motion, and the use of two X-ray video cameras (fluoroscopes) to capture 3D motion of bones in vivo. Over the past decade we have developed X-ray Reconstruction of Moving Morphology (XROMM), which combines dual-fluoroscopy with bone models from CT scans to produce accurate animations of 3D bones moving in 3D space. Radio-opaque markers are also placed in muscles to measure length changes (fluoromicrometry). XROMM and fluoromicrometry have led to many discoveries, such as the catapult mechanism used by frogs to jump farther than their muscles should allow, the use of 'swimming muscles' in fish to power rapid expansion of the head for suction feeding, and the interactions of bird feet with sediment to produce models for fossil trackway formation by non-avian dinosaurs. Digital video data sets pose many challenges for data archiving and sharing, including large individual file sizes

and the need for extensive metadata and file linkages for effective reuse. I'll describe database projects for video data management (xmaportal.org, zmaportal.org) and for developing community standards for video data management.

Contreras, Dori
PhD Candidate
University of California Berkeley
UC Museum of Paleontology
dorilynne@berkley.edu

Field Collections to Digital Data: A Workflow for Fossils and the Use of Digital Data for Reconstructing Ancient Forests

The integration of curation and digitization with project-focused data collection is a key component to performing time-efficient studies from new fossil collections. Standard workflows for processing fossil specimens starting from initial field collection and continuing through digital analysis/measurement are not widely established. Here I present my workflow for reconstruction of a diverse Late Cretaceous flora from plant macrofossils preserved in an extensive recrystallized volcanic ashfall deposit. I have established 26 quarries spanning the >1.2 km length of exposure and performed census counts of leaf specimens at each. Additionally, over 2000 fossiliferous slabs with voucher specimens have been collected for further study, curation at the University of California Museum of Paleontology, and for measurement of leaf characteristics from digitized specimens. I will discuss how various tasks have been worked into a pipeline to streamline data collection and curation, including both specimen imaging, generation of geo-referenced individual specimen records, and the use of digital data in research. I will also delve into parts of the workflow that have caused processing bottlenecks, and what measures have been taken to resolve them, especially how physical re-organization of space and both people and physical resources have been key to developing the workflow.

De Palma, Adriana
Natural History Museum, London
A.De-Palma@nhm.ac.uk

The PREDICTS Project: Projecting Responses of Ecological Diversity in Changing Terrestrial Systems

PREDICTS is a collaborative project that aims to produce global models of how local biodiversity responds to land use and related human impacts, in order to make projections under possible future scenarios. A necessary first step was to collate from published studies a large, reasonably representative database of biodiversity data from multi-site surveys whose sites differed in the nature or intensity of human pressures (Hudson et al., 2017). With this task came a number of challenges; in particular, how can we take hundreds of relatively small datasets after corresponding with hundreds of authors, and curate them in a coherent and comparable way? Thanks to the generosity of many researchers and the efforts of collaborators, the PREDICTS database is now the largest of its kind and has allowed us to provide further insights into how different species and regions respond to local land-use change across the globe (De Palma et al., 2016; Newbold et al., 2016).

Hobern, Donald
Executive Secretary
GBIF Secretariat
dhobern@gbif.org

Preserving Evidence of Biodiversity Patterns: GBIF and Persistent Biodiversity Data Management
Abstract: Our understanding of biodiversity patterns rests upon many streams of field observations, in particular natural history specimens, published literature, ecological research, citizen science, remote sensing, and genomics. Despite the wide variety represented by these activities, they generally retain a shared information core that serves as evidence of the occurrence of particular species in time and space, often with associated measurements of the relative abundance of species in a particular sample. GBIF focuses on standardized mobilization and organization of these core elements. Existing standards support a scale of detail from simple checklists through atomic occurrence data to richer co-occurrence data from field samples. The goals of the GBIF implementation plan include simplifying and supporting data publishing and assisting with delivery of the most detailed version possible for each data source. Shared effort toward these goals will allow us to overcome information loss and preserve the richest possible evidence-base on species distributions and biodiversity patterns.

Kearney, Maureen
Associate Director of Science
Smithsonian
National Museum of Natural History
kearneym@si.edu

Expanding the Power of Natural History Knowledge: Research and Collections at the Smithsonian's National Museum of Natural History
A major responsibility for natural history museums in this era of rapid global change is to mobilize our collections data and natural history knowledge for science and society. Humans have collected and documented the natural world for centuries—compiling big planetary data in the process—yet this information is fragmented and still largely inaccessible. The untapped dark data in natural history collections can help us answer recalcitrant questions about our changing natural world if we illuminate it.

Collections-based research outcomes must also be more effectively mobilized and leveraged. Natural history scientists help us comprehend the fundamental nature of the planet, of organisms (including humans), and of evolutionary and ecological interactions throughout the history of life on Earth. And natural history museums support the precise disciplinary components — biological, geological, and anthropological sciences — that are increasingly recognized as the fundamental, linked systems necessary to understand our changing planet and its inhabitants.

Taken together, the unique data and scientific expertise residing in natural history museums makes them rare and deep reservoirs of knowledge. Enormous potential exists for natural history museums in the 21st century if they highlight their unique niche as irreplaceable research and data centers for the study of global change. This can only be realized, though, if museums build large-scale pipelines and open-source, dynamic platforms to digitize, structure, link, and share our natural history data and knowledge. We are at an exciting inflection point technologically in our ability to rise to this demand —

to transform natural history collections into research infrastructure and knowledge that can play a significant role in scientific and societal issues. At the Smithsonian's National Museum of Natural History, we are embarking on a new decadal science strategy inspired by this challenge.

McCartney, Peter
Program Officer
National Science Foundation
PMCCARTN@nsf.gov

A Vision for a National Cyberinfrastructure for Biodiversity Research and what NSF can do Enable it
NSF investments in digital data for biodiversity cover a broad range of award topics including informatics research, software development, digitization, databases, ontology development, training, and community engagement. While some of these are large centrally coordinated activities, many awards are to individuals or small teams. This poses challenges in promoting the development of a national infrastructure for biodiversity through the collaborative efforts of independent projects. This talk presents an overview of the programs and sample awards made by NSF in recent years to characterize the overall portfolio of NSF investments in biodiversity data and how they are positioned within the larger environment of international, government, and non-government supported activities. Emergent opportunities for synergy, standardization, and coordination are discussed along with potential mechanisms for stimulating and enabling them.

Smith, Stephen
Assistant Professor
Dept. Ecology and Evolutionary Biology
University of Michigan
eebsmith@umich.edu

The Utility of Large-scale Phylogenetic Analyses for Understanding the Evolution of Biodiversity

The promise of a comprehensive view of the tree of life, whether for a particular clade or the entire tree of life, has been a major motivation of the systematics community for decades. Broader and more complete phylogenies allow for evolutionary, biogeographic, and ecological questions that cannot be addressed with smaller phylogenies. There have been several efforts to build large and comprehensive phylogenies for particular clades and these have resulted in phylogenies for thousands of species. The Open Tree of Life project aimed to construct a comprehensive draft tree of all life using published phylogenies. While this effort succeeded in the construction of a tree and many tools to support this effort, these have limitations. Here, I will describe new efforts and new ways for combining the resources from the Open Tree of Life with other phylogenetic analyses to construct a dated and comprehensive tree. Specifically, I will discuss our construction of a comprehensive tree for seed plants containing 80,037 taxa from GenBank and 356,807 total taxa. I will describe some of the challenges presented and ways for the community to move forward. I will also discuss how these large phylogenies have been used to address questions about the origins of biodiversity.

Soltis, Pamela
Professor
University of Florida
psoltis@flmnh.ufl.edu

Linking Heterogeneous Data in Biodiversity Research
Emerging cyberinfrastructure and new data sources provide unparalleled opportunities for mobilizing and integrating massive amounts of information from organismal biology, ecology, genetics, climatology, and other disciplines. Key among these data sources is the rapidly growing volume of digitized specimen records from natural history collections. With nearly 100 million specimen records currently available online, these data provide excellent information on species distributions and changes in distributions over time. Particularly powerful is the integration of phylogenies with specimen data, enabling analyses of phylogenetic diversity in a spatio-temporal context, the evolution of niche space, and more. However, a major challenge is the heterogeneous nature of complex data, and new methods are needed to link these divergent data types. Ongoing efforts to link and analyze diverse data are yielding new perspectives on a range of evolutionary and ecological problems. We will present multiple case studies that address different aspects of ecology and evolutionary biology that have been addressed using specimen data and related heterogeneous data sources. Although many specific hypotheses may be addressed through integrated analyses of linked biodiversity and environmental data, additional value of such data-enabled science lies in the unanticipated patterns that emerge.

Summers, Adam
Associate Professor
University of Washington
fishguy@uw.edu

Big Data, Museum Specimens, Access and Archiving - Lessons from #scanAllFish
We have begun a multi-institution effort to CT scan all of the vertebrates. This requires a different approach to CT scanning than has previously been common. By scanning many specimens at the same time, in the same interrogation space, large data sets that require extensive post-scan analysis. Free, open source software (FOSS) available on multiple platforms is used for this analysis step. This means that a single CT scanner can produce data that will keep many off-site data analysis sites productive. A typical, high resolution scan can image 20 species with sufficient resolution for every common role of CT. The raw data for this scan is a series of projections that are not usually archived. These projections are transformed into slices, which are what most people think of as CT data. However, it is important to consider that there are many parameters involved in producing the slice data and in our experience it is not uncommon to return to the projections to recompute the slices. The reason this is an issue is the associated data is of staggering size. Over 2000 species of vertebrates we have an average data file size of 14GB. Large genome data sets are an order of magnitude smaller. We expect to store over half a petabyte of data for 30,000 vertebrates. Storing and backing these data up is an issue. It is also interesting to consider what collections plan to do when these data are returned to them with the specimens.

Webster, Mike
Robert G. Engel Professor of Ornithology
Director, Macaulay Library
Cornell University
msw244@cornell.edu

Using "Digital Specimens" to Explore the Behavioral Phenotype

Biological research has relied on study specimens for centuries, and today emerging new technologies and analytical techniques have increased the scientific value of these specimens dramatically. But today we can also collect a new type of specimen, the digital "media specimen", which is an audio or video recording of a bird in nature. These recordings capture key aspects of wild birds — their acoustic signals, physical displays, and other important behaviors – in ways that traditional physical specimens simply cannot. This talk will introduce the concept of digital media specimens and illustrate their use in recent research. Properly curating such media and making them accessible, including connecting them with associated physical specimens, will require partnerships between traditional natural history collections, media libraries, and data aggregators.