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

Beach, James
Assistant Director Biodiversity Institute
University of Kansas
beach@ku.edu

Supporting the Biological Collections Community with Specify Software for the Long Run

The Specify Software Project produces and supports Specify, a biological collections data management platform for collection curation, digitization, and data publishing. The Project is a descendent of the MUSE Project which together have been funded for 30 years by the US National Science Foundation. About 500 collections around the world use Specify 6 for specimen data processing.
The latest generation is Specify 7, a web application hosted by the Specify Cloud server or by institutions for themselves. All Specify 6 and 7 software is currently open source licensed.

During 2017, we are undertaking a process to identify a sustainable revenue model and non-grant sources of financial support, for the ongoing software engineering of Specify as well as associated helpdesk and data management services. The development of a sustainable revenue model for Specify software is an open community process. We will present various options for collections community governance and support of the software in the future. We invite suggestions and feedback to the organizational options which will be discussed.

Bemmels, Jordan
University of Michigan PhD student
jbemmels@umich.edu
Co-authors: S. Joseph Wright, Smithsonian Tropical Research Institute; Nancy C. Garwood, Southern Illinois University; Simon A. Queenborough, Yale University; Renato Valencia, Pontificia Universidad Católica del Ecuador; Christopher W. Dick, University of Michigan & Smithsonian Tropical Research Institute

Biogeographic Filtering and the Assembly of Neotropical rainforests: Insights using Ecological Traits Derived from Digital Biodiversity Data

Despite the barrier to intercontinental dispersal posed by the northern Andes region and adjacent dry habitats, numerous Neotropical rainforest species are distributed in both Amazonia and Central America. To investigate the potential role of biogeographic filtering across the northern Andes region, we examined life-history traits, environmental tolerances, and biogeographic distributions of >1,000 woody plant species locally co-occurring in forest plots in Panama and Amazonian Ecuador, and tested whether certain trait features are associated with widespread (i.e., spanning both sides of the Andes) vs. regional (cis- or trans-Andean) geographic distributions. As we expected, a widespread distribution is predicted by a distinct suite of traits, including high drought tolerance, broad elevational range, and pioneer traits associated with high dispersal-colonization ability. Our results highlight how biogeographic

filtering across the northern Andes region has mediated floristic assembly of Neotropical rainforests, especially in Central America where biotic interchange has heavily influenced regional community composition. The presentation will emphasize advances and challenges of processing >300,000 occurrence records from GBIF to define geographic ranges and estimate environmental tolerances of each species. Leveraging digital biodiversity data to characterize ecological traits allowed us to address novel macroecological questions across a range of otherwise poorly characterized rainforest species.

Boyer, Doug
Assistant Professor, Database director
Duke University
douglasmb@gmail.com
Co-authors: Julia M. Winchester, Gregg F. Gunnell, Seth Kaufman, Timothy M. McGeary

MorphoSource: A Virtual Museum and Digital Repository for 3D Specimen Data
Improvements in 3D scanning technology and computing power over the past several decades have led to increased use of 3D data and more powerful, data driven approaches in comparative biology and paleontology. The transition to a digital approach brings up questions about best practices and requirements for data archiving. Ideally, all studied 3D datasets of natural history objects should become freely accessible in flexibly searchable databases in order maximize the reuse potential of these rich resources. Achieving this goal poses challenges concerning implementation, governance, and participation. We describe MorphoSource, a web-based virtual museum and 3D data repository that adheres to standards of quality and file format recently established by the community of scientists working with 3D data. Users can upload and download high fidelity 3D renderings of specimens derived from a variety of scanning modalities. Each 3D media file is associated with a specimen record (the basic unit of organization) and assigned a DOI as well as a globally unique identifier. Archived data can be easily searched and downloaded, and downloaded data can be used by researchers, educators, and others. MorphoSource is increasingly endorsed by major museum collections and journals as an appropriate solution for hosting researcher-generated 3D data.

Butts, Susan
Senior Collections Manager
Yale University
susan.butts@yale.edu
Co-authors: Talia Karim, Gil Nelson, Christopher Norris, Mark Uhen, Jocelyn Sessa, Dena Smith

ePANDDA: enhancing Paleontological and Neontological Data Discovery API
The paleontology community has exemplary data repositories for research on climate and biodiversity through time, including: iDigBio, the national hub for neontological and paleontological specimen data, iDigPaleo, a product of the Fossil Insect Collaborative TCN built for educational use and a prototype for aggregating collections data with Darwin Core paleo data, and Paleobiology Database, a heavily-cited database of fossil occurrences built from primary scientific literature. The sources are excellent on their own, but integrating data from them is laborious and problematic, as the connectivity between modern and fossil, and specimen and literature-based resources does not currently exist. Funded by the NSF EarthCube initiative, the ePANDDA project seeks to correlate and combine data from those three data providers via an API, and provide the user with a single data set that provides utilities for clustering by variables. This talk will discuss climate and biodiversity research use cases of the ePANDDA API based on

the research interests of the ePANDDA principal investigators and input from core data users who attended a 2016 ePANDDA workshop. It will also discuss how ePANDDA can collaborate with other existing geological and biological resources.

Clites, Erica
Museum Scientist/EPICC Project Manager
University of California Museum of Paleontology
eclites@berkeley.edu
Co-authors: Liz Nesbitt, University of Washington Burke Museum; Peter Kloess, University of California Museum of Paleontology; Austin Hendy, Los Angeles County Museum of Natural History

Supporting Research Pipelines through the Creation of Stratigraphic and Taxonomic Concordances
The institutions of the Eastern Pacific Invertebrate Communities of the Cenozoic Thematic Collections Network (EPICC TCN, http://epicctcn.org) are actively digitizing 1.6 million marine invertebrate fossils from the past 66 million years of Earth's history. To aid in interpretation of this occurrence data, we are compiling concordances of stratigraphic and taxonomic data. These concordances will support research pipelines by standardizing stratigraphic and taxonomic data across nine institutions. Stratigraphic concordances are being prepared for stratigraphic units found in British Columbia, Washington, Oregon and California within the TCN's fossil collections. The basis of these concordances are the USGS GeoLex resource (https://ngmdb.usgs.gov/Geolex/search) and unpublished USGS Geologic Names Committee Archives. These are reviewed and updated following more recent published literature. Stratigraphic concordances will include script-ready data sheets published through Data Dryad as well as short, written descriptions of problematic formations. A taxonomic concordance is also being compiled based on taxonomic authority papers. This concordance is based on an assembly of taxonomic data for Recent and Quaternary mollusks (our largest taxonomic group), which provides a robust taxonomic framework for the introduction of taxonomic names from older fossil literature. Other taxonomic groups and reconciliation of outdated taxonomic concepts are being undertaken with the collaboration of experts in those taxonomic groups.

Damerow, Joan
Postdoctoral Fellow
Field Museum of Natural History
jdamerow@fieldmuseum.org
Co-authors: Patina K. Mendez, Petra Sierwald, Rudiger Bieler, Matthew J. Yoder, R. Edward DeWalt

Taxonomic data quality in GBIF: a case study of aquatic macroinvertebrate groups
Addressing data quality is a primary concern for biodiversity data creators, users, and aggregators. In the case of taxonomically diverse groups, such as insects and mollusks, digitized species names taken from specimens may be outdated or incorrect. Global indices, such as the Global Biodiversity Information Facility (GBIF) Backbone Taxonomy, are constructed by aggregating taxonomic sources of varying quality into a synthesized index of species names. A synthesized backbone integrates datasets and should improve data quality by facilitating automatic corrections of synonyms to accepted names. However, it can also compound problems as taxonomic specialists who contribute new and updated species occurrence data may find their nomenclature reverted to the synthesized name used by GBIF. Here we present a case study evaluating the prevalence of outdated names in GBIF records for several groups of aquatic macroinvertebrates in the USA. We trace the workflow of how community-accepted species

lists for these groups are updated and eventually integrated into aggregators (if they are integrated). We discuss current barriers and potential improvements to the system from the perspective of data contributors and users.

Delaunay, Mathilde
PhD student
Muséum national d'Histoire naturelle, Paris, France
mathilde.delaunay@mnhn.fr
Co-authors: Régine Vignes-Lebbe, Romain Nattier

How do People see Biodiversity? Using a Digital Identification Key in a Citizen Science Program
"Spipoll'' is a French citizen science program about pollination. To assist the volunteers, a multi-access digital identification key for the insects has been created with the Xper3 platform. The pictures, identifications and series of steps followed by the participants have been recorded since September 2015 thanks to the Xperience system.

The recorded data allow to study the behaviour of the citizens when they observe an insect, and to deduce the taxonomic confusion and the misunderstanding of character states. The identification paths give elements on how the entomofauna diversity is perceived. Which morphological traits are chosen most frequently? Are the most noticeable characters selected to the detriment of those which need advanced entomological skills?

Here we will present the Xperience system. The first analysis of the database shows that some morphological parts are perceived more easily than others, and that people are sensitive to the quality of character descriptions in the keys. These elements must be taken into account in order to improve digital identification tools, in particular those used by the general public.

Esquivel, Alicia
Garden NDSR Resident
Chicago Botanic
esquivelndsr@gmail.com
Co-author: Constance Rinaldo, Ernst Mayr Library Museum of Comparative Zoology, Harvard University

Using Statistical Analysis to Calculate the Size of Biodiversity Literature
To address gaps in the corpus and focus digitization efforts, the Biodiversity Heritage Library continuously refines their content and collection analyses. As a National Digital Stewardship Resident, I am working with BHL to conduct a collection analysis and have been exploring the use of statistical models to scope the amount of biodiversity literature in the public domain. The capture-recapture method has been used in ecology to determine the population size of specimens without indexing each member of the population. The statistical probability is calculated by capturing a sample, marking each specimen in the sample, releasing sample back into the full population, capturing an additional sample, and measuring the amount of marked, or recaptured, specimen compared to specimen caught for the first time. This methodology has been applied to other data to estimate sizes of uncountable populations. For example, this method is used in epidemiology to calculate patients with particular diseases, in human rights to calculate amount of human rights violations, and more recently in literature

reviews to calculate the amount of published material on a given subject. Using a mixture of online databases and hardcopy bibliographies, we will calculate the size of biodiversity literature based on a capture-recapture methodology.

Fisher Daniel
Professor
University of Michigan
dcfisher@umich.edu

3D Surface Models in Paleontology and Archaeology

Digital data on the external form of specimens are central to many paleontological and archaeological analyses. Digital models minimize handling of fragile and/or heavy specimens, facilitate access and collaboration, allow complex measurements, enhance visualization of surface topography, and simplify inspection of multi-object assemblies. In our UMORF (University of Michigan Online Repository of Fossils) project, we produce high-resolution surface models and host them through a custom web application (http://umorf.ummp.lsa.umich.edu) offering control of viewing parameters, measurement capability, labeling of elements, and toggle-control of visibility and color information. Two exemplar applications illustrate important features of our approach. First is the BonePicker interface, designed to facilitate access to data on the osteology of the American mastodon (*Mammut americanum*) using a complete, interactive, 3D skeleton. The BonePicker supports precise identification of serial elements that differ only subtly from one another (ribs, vertebrae, podials), as needed in taphonomic studies documenting patterns of carcass processing that demonstrate human association with late Pleistocene mastodons. The second exemplar involves models of fragments of cortical bone from mastodon long-bone diaphyses. Disabling specimen color enhances visualization of subtle features of fracture topography and allows us to recognize fractures as having formed when the bone was fresh, demonstrating human association.

Flannery, Maura
NY Professor of Biology
St. John's University
flannerm@gmail.com

iDigBio and the Digital Humanities

The use of herbarium specimens in the humanities is in its infancy; the availability of specimens online means that historians, artists, and others can now mine these resources in their work. The iDigBio community can encourage links to these other communities, and there are models available. The Botanica Caroliniana project has produced a website where a Mark Catesby's herbarium specimen is displayed alongside the plate from his book depicting it, as well as the accompanying description. There is also a website for the specimens and writings of a later Carolina botanist, Henry Ravenel. The Reconnecting Sloane project in Britain seeks to link his herbarium specimens with his books, correspondence, and other materials. Also in Britain, the Royal Botanic Garden, Kew has a site dedicated to the work of the 19th-century botanist Nathaniel Wallich: search for a species and find related herbarium specimens, manuscripts, and illustrations. The Missouri Botanical Gardens has created something similar for George Engelmann. As digitization of herbarium collections and those of related libraries continue, more attention should be given to linking them. The Engelmann site is available through the Biodiversity Heritage Library and might therefore be a model for future endeavors.

Hammock, Jennifer
Scrum Master, EOL
Smithsonian
hammockj@si.edu
Co-author: Katja Schulz, Smithsonian

The Encyclopedia of Life v3: constructing a linked data model

Biodiversity data holdings in the Encyclopedia of Life include taxa, linked to each other by relationships of lineage and ecology, and linked also to habitats, ecological categories, management categories, and other attributes. These data are best modeled as a graph, which makes these relationships explicit, and available for reasoning and searching across.

This case study will describe how these data are being modeled for EOL v3, how quantitative and categorical attribute values are treated, how provenance and other metadata are recorded, what kinds of identifiers are used, from which sources, and how this graph intersects with other realized graphs and potential graphs, including hierarchically organized gazetteers and ontologies, occurrence data, literature, photos, and other resources.

Measured in number of nodes, EOL hosted data is a small piece of the Biodiversity Knowledge Graph, but it has high connectivity and holds the potential to support powerful searches within biodiversity data and connected graphs throughout the natural sciences and beyond. Additional areas of high connectivity that could further enrich the graph include: attributes linked directly to specimen or observation records; geographic coordinates relative to named locations; connections among people such as researchers and wildlife observers providing occurrence data.

Haselhorst, Derek
Graduate Research Assistant
Program in Ecology, Evolution and Conservation Biology, University of Illinois
dhaselh2@illinois.edu
Co-authors: Shu Kong, School of Information and Computer Sciences, University of California, Irvine; Charless C. Fowlkes, School of Information and Computer Sciences, University of California, Irvine; J. Enrique Moreno, Center for Tropical Paleoecology and Archaeology, Smithsonian Tropical Research Institute; David K. Tcheng, National Center for Supercomputing Applications, University of Illinois; Surangi W. Punyasena, Department of Plant Biology, University of Illinois

Automating Tropical Pollen Counts using Convolutional Neural Nets: from Image Acquisition to Identification

Pollen types are identified according to a suite of distinctive morphological characters, including shape, size, and ornamentation, which are understood to capture the phylogenetic relationships of pollen taxa. However, parsing finescale morphological variation among hyperdiverse tropical pollen samples represents a challenge for the human analyst.

Here, we present an automated workflow for pollen analysis, from the automated scanning of pollen sample slides to the automated identification of pollen types using convolutional neural nets (CNNs). We work with aerial pollen samples from lowland Panama with >115 pollen types. To train our CNN, pollen

types were tagged using a virtual microscope; metadata was digitally recorded for each imaged pollen grain, including its identification and coordinates.

We compare the results of our analysis on three levels. We: (1) compare pollen counts made with a standard microscope to those made on our virtual microscope; (2) measure identification accuracies of the CNN classification system; and (3) measure the ability of an expanded automated system to simultaneously segment and classify pollen types from scanned slides.

Our preliminary results (>80% accuracy) demonstrate the ability of next-generation deep learning tools like CNNs to consistently classify pollen and lay the groundwork for a fully automated big-data pollen classification system.

Knouft, Jason
Associate Professor
Saint Louis University
jknouft@slu.edu

Integrating Relevant Hydrologic Measures with Digitized Biodiversity Data to Investigate Climate Change Impacts on Freshwater Fishes

Streamflow, water temperature, and sediment are primary factors influencing the traits, distribution, and diversity of freshwater taxa. Ongoing changes in climate are causing directional alteration of these variables, which can impact local ecological processes and result in extirpation of populations. Accurate estimation of these environmental variables at watershed (and larger) scales is critical for predicting the response of species to human-induced changes in freshwater habitat. Nevertheless, access to data sets characterizing spatial variation in these variables is not available for many regions. Considering the vast amount of digitized biodiversity data that are available for freshwater taxa, development and application of appropriate hydrologic data are critical to the advancement of our understanding of freshwater systems. These hydrologic data can be integrated with digitized biodiversity data to not only investigate the distribution of suitable habitat for a species, but also address trait-based ecological and evolutionary responses to rapid changes in climate. I will discuss development and application of these types of environmental data as well as provide two examples of how these data can be integrated with digitized freshwater biodiversity data to investigate the potential impacts of climate change on freshwater taxa.

Kost, Michael
Assistant Curator
University of Michigan, Matthaei Botanical Gardens and Nichols Arboretum
michkost@umich.edu
Co-authors: David Michener, Maricela Avalos, and Adam Hulyksmith, University of Michigan, Matthaei Botanical Gardens and Nichols Arboretum; Jason Tallant, University of Michigan Biological Station; Peter Knoop, University of Michigan, College of Literature, Sciences and the Arts, IT Advocacy and Research Support; Shannon Brines, University of Michigan, School of Natural Resources & Environment

Developing an Enterprise GIS to Support Collections Management, Teaching, and Research

We are developing an Enterprise Geographic Information System (GIS) to collaboratively map and manage data on our living collections of plants, animals, and natural communities, with a primary goal of making the data readily available for teaching and research. The scope of the GIS includes numerous University of Michigan properties managed separately by the Matthaei Botanical Gardens and Nichols Arboretum, Biological Station, School of Natural Resources & Environment, and Ecology and Evolutionary Biology. The data for these properties are currently dispersed across a variety of paper and electronic formats, including 16,000 plant records in a Microsoft Access database, and much of the data lacks digital location information. Migrating the data into the GIS and enforcement of a standardized metadata format will enable students and faculty to easily access and utilize the data, as well to contribute new data, through ArcGIS software, Google Earth, Excel, and other desktop, mobile, and web applications. The GIS will also make it easy for researchers and instructors, at U-M and elsewhere, to discover our data through ArcGIS Online, The Big Ten Academic Alliance Geoportal, and other data hubs.

Mayer, Paul
Fossil Invertebrate Collections Manager
The Field Museum
pmayer@fieldmuseum.org

How Digitizing and Tagging Helped Solve the Tully Monster Mystery

The Tully monster (*Tullimonstrum gregarium*) is a 307 million year old problematic fossil known only from northeastern Illinois. Since it was first described in 1966 there has been debate about where it fits on the tree of life. Scientists have assigned it to various phyla including: Mollusca, Annelida, and Chordata. In 2015 Yale University scientists, teaming up with Field Museum and Argonne National Laboratories scientists, examined over 1,200 Tully monster specimens from the Field Museum's collections to search for rare or overlooked morphologic clues that might solve this 50-year mystery. To locate specimens efficiently, 1,305 Tully monster specimens were digitized in three weeks, creating a total of 4,441 images including images of each part and counterpart in low-angle and cross-polarized light. Each specimen was also examined for eight morphologic traits. All traits observed in the specimen and the lighting techniques were incorporated in the file naming protocol developed for this project. All images were uploaded to our EMu database and a total of 12,400 keywords derived from the file names were created and attached to both the catalog and multimedia records. These tags allow researchers to quickly and efficiently search for specimens by their key morphological traits.

Metz, Mark
Research Entomologist
USDA ARS SEL
mark.metz@ars.usda.gov

Open Source Tools for Digitization Workflows

Free, open source, non-proprietary, cross-platform tools are available to enhance digitization workflows. Many of these techniques are great time savers, especially for large projects that otherwise would have required a considerable amount of keystrokes and/or proofing for standardization and integrity. Others, provide automation that a computer can do faster and more accurately. I will share in a demonstration format some of my favorite ways to capture and manipulate taxonomic data that were not born digital, not readily accessible, or unstructured. My hope is that you will be able to apply these techniques or to stimulate creativity for similar techniques in your workflows.

Pearson, Katelin D.
Graduate student
Florida State University
kds15e@my.fsu.edu

Hole-y Plant Databases! Understanding and Preventing Biases in Botanical Big Data

The recent mass digitization of biodiversity specimen records has enabled researchers to investigate biological questions on an unprecedented scale. However, the accuracy of these analyses relies largely on the quality of the data, which may be influenced by spatial, temporal, taxonomic, and other biases that arise from the unsystematic nature of collecting. In this study, we sought to determine: 1) what biases have been identified within herbarium specimen databases, 2) how have researchers corrected for these biases, and 3) what are the potential effects of failing to account for biases when using collections data? Furthermore, we examined herbarium specimen databases for previously untested biases that may have significant effects on existing and future collections-based research. Our results underscore the need for researchers to assess and understand common and dataset-specific biases when leveraging these powerful sources of data. Understanding biases in biodiversity collections data is critical not only to ensuring accurate analyses of natural phenomena, but also for the community of specimen collectors as we strive to close the gaps in our understanding of Earth's biota.

Punyasena, Surangi
Associate Professor
University of Illinois
punyasena@life.illinois.edu
Co-authors: Shu Kong, Charless C. Fowlkes, and Stephen T. Jackson

Reconstructing the Extinction Dynamics of *Picea critchfieldii* – The Application of Computer Vision to Fossil Pollen Analysis

The spruce *Picea critchfieldii* is the only tree known to have gone extinct during the last deglaciation (20,000 – 11,000 years BP). It is known from fossil needles and seed cones from Louisiana, Georgia, and Tennessee. However, our previous work shows that *P. critchfieldii* pollen is morphologically distinct from other spruce and can be used to track changes in its geographic range and abundance preceding extinction.

Here, we reconstruct changes in population abundance at Anderson Pond, Tennessee, and Cupola Pond, Missouri. We employ a recently developed method of classifying pollen images that incorporates feature learning and patch representation, dictionary construction, and location-aware sparse coding for classification. The approach represents a breakthrough in automated fossil pollen identification, emulating human analyst capabilities; the best models can transfer training from modern reference material to the classification of taphonomically altered fossil material. No published study has been capable of this before.

Our preliminary results indicate that *P. critchfieldii* was the dominant spruce at Cupola and increased in abundance at Anderson during deglaciation. This extends the known range of *P. critchfieldii* and suggests that rapid decline of spruce in the southern US during the last glacial may be tied to extinction of *P. critchfieldii*.

Reznicek, Anton (Tony)
Curator
University of Michigan Herbarium
reznicek@umich.edu

Illustrating Value Added in Databasing Historical Collections: Entered, Proofed, and Done (or Not!)
With limited data on labels, transcription can be viewed as an afterthought, with the difficulties to be overcome merely updating/checking names and determinations, and georeferencing, with the science framed by analyses of the transcribed data, utilization of specimen images, etc. However, this overlooks many issues. Digitizing of the unique First Geological Survey plant collections from Michigan (1837-1840), numbering nearly 1200 sheets, provided a rare opportunity to integrate species level information with pre-settlement vegetation data, adding a valuable baseline to studies of range changes, extirpated species, and shifts in community composition. The original data, however, presented obstacles, extending even to the discovery of the specimens in the collection. Larger scale geography was often missing (no state or country), leading to mis-attribution of localities. Collections were frequently missing the year, and localities were often cryptic and abbreviated. A primary driver here is that with handwritten labels, the drive to abbreviate was substantial. Linked with curatorial practices of the 19th century such as re-copying labels into a standard format, often with discarding of the original labels, produced often impenetrable labels requiring extensive research to make the data accessible. This brings to the forefront the differing standards needed for working with historical collections.

Sandall, Emily
Graduate Student
Frost Entomological Museum, Penn State University
els22@psu.edu
Co-author: Andrew R. Deans, Penn State University

Importance of Life Stage Capture in Dragonfly Specimen Digitization
Life stage is apparently rarely captured during the digitization of natural history collections, but it can highlight sampling biases, life history changes, and gaps in taxonomy and systematics. Through the digitization of non-adult specimens of clubtail dragonflies, for example, it is possible to identify the taxa that lack diagnostic keys for all life stages. While likely only a small proportion of an entomological collection includes all life stages for a single taxon, the larvae and exuviae (the skin shed upon emergence from the aquatic larval stage to the adult terrestrial stage) can provide a more comprehensive representation of morphological characters for species' identification and evolutionary questions. Furthermore, recording the life stage of dragonflies in the digitization process enables analysis of the distribution of a dragonfly taxon throughout its life history, which is both aquatic and terrestrial. By digitizing the Gomphidae specimens in collection at the Frost Entomological Museum at Penn State University, proportions of life stages were compared to other digitized collections. This analysis reveals gaps in both keys and collections, as well as the need for increased diagnostic and natural history data for this family.

Schulz, Katja
Program Coordinator
Smithsonian National Museum of Natural History
SchulzK@si.edu

Co-author: Jennifer Hammock, Smithsonian National Museum of Natural History

Encyclopedia of Life Version 3: New Tools for the Exploration of Biodiversity Knowledge
Biodiversity informatics is focused on mobilizing and inventorying primary biodiversity data (names, classifications, occurrences), but efforts to integrate, contextualize, and synthesize this information are taking shape. As the volume of structured data about organisms increases, semantic approaches facilitate access to and reasoning over linked information from heterogeneous sources. The Encyclopedia of Life (EOL, eol.org) is developing a new software architecture that takes advantage of this emerging infrastructure. EOL V3 will support complex queries across a comprehensive, semantically enriched data set of taxon attributes managed in a redesigned TraitBank framework that is scaffolded by a dynamic, curated reference taxonomy. Other new features include autogenerated natural language descriptions of taxa, species number estimates and coverage/knowledge indices for groups across the tree of life, data visualizations, and knowledge-based recommender systems for the exploration of content along multiple axes, including phylogeny, ecology, life history, relevance to humans and other characteristics derived from structured data. Many of the primary data and data products needed for a comprehensive, interconnected biodiversity knowledgebase are not yet shared freely in machine-readable formats. To encourage the proliferation of data-driven applications like EOL V3, progress towards accessibility and interoperability of large scale biodiversity data sets is essential.

Uhen, Mark
Assistant Professor
George Mason University
muhen@gmu.edu

Paleobiology Database: A Community Based Data Service for Research, Education, and Museums
Paleobiology Database (PBDB, paleobiodb.org) is an open resource of global paleontological data on all types of fossils through all time periods. Data types include references; taxonomic names, opinions, and classifications; fossil collection data; occurrences; and geologic time scales all in a relational database. PBDB currently contains over 1.32 million occurrences of over 350,000 fossil taxa and is expanding every day as over 380 researchers contribute data. These data have been used for studies of paleobiodiversity, paleogeography, macroevolution, history of science, and many others. The PBDB Navigator interface plots fossil collections on interactive paleogeographic maps using GPlates paleo plate position data. PBDB also has an API data service that allows other cyberinfrastructure resources to automatically query the database and use the data in for their own purposes. Users can also use the data service to query the database and run any kind of analysis that they can conceptualize. PBDB is working to incorporate specimen level occurrence data, and is also building a clearinghouse for lesson plans that use PBDB data and visualization tools. PBDB strives to be a comprehensive resource on macrofossil data for deep time, and an open platform for storing these data and sharing them with the world.

Urban, Michael
Postdoctoral Researcher
University of Illinois - Urbana Champaign
urban1@illinois.edu
Co-authors: Mayandi Sivaguru, Ingrid Romero, Glenn Fried, Charless C. Fowlkes, Washington Mio, Carlos Jaramillo, Surangi W. Punyasena

The Application of Optical Supperresolution Microscopy to the Study of Pollen Morphology
Airyscan confocal superresolution and structured illumination superresolution (SR-SIM) microscopy are powerful new tools for visualizing taxonomically important features on and within the pollen wall. Optical superresolution microscopy can resolve features below the diffraction limit of light (these two techniques achieve 100-140 nm resolution), without the laborious and destructive imaging of electron microscopy.

We compared the two forms of superresolution using three modern pollen types with distinct morphologies (*Croton hirtus*, *Dactylus glomerata* and *Helianthus* sp.). Both proved to be viable techniques. However, the microscopes do not have identical performance. The SR-SIM performs best when pollen is thin-walled with simple surface structures and a high signal-to-noise (*D. glomerata*), while the Airyscan excels with thicker-walled pollen possessing more complex morphology (*Helianthus* sp. and *Croton* sp.).

Optical superresolution has the potential to expand the research questions that can be addressed using fossil pollen. We preview three projects that use superresolution: (1) assessing the taxonomic affinity/affinities of the putatively pantropical *Striatopollis catatumbus*, a biostratigraphically important Cenozoic palynomorph; (2) applying superresolution and machine learning to the reconstruction of grass pollen diversity spanning a 25,000 year sediment core from Lake Rutundu, Kenya; and (3) the recovery of phylogenetically significant features from modern and fossil Bombacoideae.

Winchester, Julie
Postdoctoral Associate
Duke University
julia.m.winchester@gmail.com
Co-authors: Doug M. Boyer, Department of Evolutionary Anthropology, Duke University; Maureen A. O'Leary, Department of Anatomical Sciences, Stony Brook University; Jocelyn A. Sessa, Department of Paleontology, American Museum of Natural History

The Importance and Challenges of Database Integration: MorphoBank, MorphoSource, and the Paleobiology Database
Placing scientific research products (matrices, media, stratigraphy) in a digital, web-based, public repository is increasingly recognized to be a standard 'best-practice' for science. As specialized databases grow in size, communication among them has clear benefits for research. We describe our current efforts to integrate three existing databases: MorphoSource (3D media), MorphoBank (2D, 3D media and matrices), and the Paleobiology Database (PBDB; stratigraphy, geography, synonymy). All of these databases contain taxonomic data relating to living and extinct species. Using primarily taxonomy, we propose to link these databases via APIs to provide data access from and deposition to these three sources within each database. This will enable research that includes accessing 1) imagery when building matrices; 2) the phylogenetic context of species when collecting 3D data; 3) extensive comparative 3D anatomical data when collecting new fossils of related species; and 4) geological and temporal information for the above research possibilities. None of these tasks are currently possible in an automated fashion. In addition, links to iDigBio's specimen archive are a natural fit to this integration, and are currently being explored by linking MorphoSource specimen and media records with iDigBio, and by ePANDDA, an NSF-funded project connecting iDigBio, iDigPaleo, and the PBDB.