BCoN workshop abstracts

Attribution patterns in mammal collections before and after the digitization age

Cody Thompson
University of Michigan, Museum of Zoology, Ann Arbor, MI. cwthomp@umich.edu

U.S. mammal collections began computerizing their catalogs in the late 1970s, and most large institutions were fully digitized by the early 1980s. In 2001, mammal collection data were mobilized online via the NSF-funded Mammal Network Information System (MaNIS) project. MaNIS was very successful in facilitating and enhancing research, education, and the conservation of mammals by providing a single data portal to aggregate high-quality georeferenced specimen records from U.S. mammal collections. Subsequent data aggregators, such as VertNet, iDigBio, and GBIF, have now replaced MaNIS and other taxon-specific portals and have expanded digitization efforts to include images and other digital media. However, no data aggregator to this point has fully addressed attribution concerns for participating collections. Although most aggregators have attempted to provide search and download information, these data are not easily accessible and are not comparable across different portals. In addition, this limited view of collection use often is difficult to grasp by administrators and does not replace the tangible benefit of cited collection use in publications. To determine whether data aggregation has affected collection citations, I attempt to examine publication data from several major U.S. collections before and after online data aggregation.

Linking Across Collections: Host/Parasite Relationships

Mariel Campbell[1], Dusty McDonald, Eric Hoberg, Carla Cicero[2], Joseph Cook[1].
[1] *Museum of Southwestern Biology, University of New Mexico, Albuquerque;* campbell@carachupa.org
[2] *Museum of Vertebrate Zoology, University of California, Berkeley.*

One of the primary goals of the natural history museum community has been to collectively document the diversity, abundance, and distribution of living organisms. In archiving physical voucher specimens and metadata, museum collections provide verifiable records that an individual organism belonging to a specific taxon existed at a specific place and time. Increasingly museum specimens are also being recognized as an important nexus between distinct data streams as biological samples are sourced for investigations that produce large volumes of data (e.g. genomes, isotopic data, CT scans). Developing research-grade databases that track and integrate these derived data streams is challenging. Yet another challenge is tracking data on ecological, evolutionary, and life histories of associated organisms (e.g., host-parasites, mutualists, plant-pollinators, predator-prey). In the case of parasites and hosts, these organisms typically belong to widely divergent taxa that have historically been collected and maintained in separate institutions with little to no overlap in data management and curation. This results in loss and divergence of critical data on their shared ecological and evolutionary relationships. We present here extensions to Arctos developed to create and query reciprocal linkages between parasite records and their associated host records in multiple Arctos collections and external institutions. These extensions have been applied to integrate data representing other complex species interactions across collections.

Integrating source modifiers with sequence data through a new GenBank submission module in Symbiota

Andrew N. Miller[1], Phil Anders[1], Neil Cobb[2], Ben Brandt[2], and Ed Gilbert[3]
[1]*Illinois Natural History Survey, University of Illinois, Champaign, IL; amiller7@illinois.edu*
[2]*Merriam-Powell Center, Northern Arizona University, Flagstaff, AZ*
[3]*School of Life Sciences, Arizona State University, Tempe, AZ*

The Symbiota biodiversity data management system is utilized by hundreds of natural history collections worldwide to serve specimen-based content online. These collections, which encompass algae, bryophytes, fungi, invertebrates, lichens, plants, and vertebrates, are organized into 38 portals which serve ~22 million specimen records. Specimen records can be linked to images, tissues, DNA sequence data, and species information. Symbiota allows the documentation of species occurrences (specimens and observations) and provides the tools for data packaging (checklists), visualization (mapping and images), educational material (keys, identification quizzes), and easy access to data sets as downloadable csv files. The system is designed with robust import, export, and publishing tools that utilize Darwin Core as the data exchange standard. Although specimen metadata have been assembled in several collection management systems, integration into pipelines for serving source modifiers of DNA sequence data does not exist. A majority of GenBank sequence records lack the most basic source modifiers. A new GenBank sequence submission module is being developed to automatically populate source modifiers during sequence submission, thus greatly improving the integration and completeness of specimen and genetic data.

Proposed redevelopment of GBIF data integration pipeline

Donald Hobern
GBIF Secretariat, Copenhagen, Denmark. dhobern@gbif.org

The Global Biodiversity Information Facility, GBIF, operates systems to harvest, aggregate and normalize evidence of species occurrence from many sources, including natural history collections. During 2018, GBIF plans to re-engineer this pipeline as an open collaborative activity in partnership with other international biodiversity data networks and national GBIF nodes. One of the goals behind this plan is to support richer indexing of associated data associated with natural history specimens (including images, sequences and morphometric data) in a modular fashion based on standards and practices followed by curators and taxonomists working with different groups. For example, for insects, life stage and COI sequence could be treated as standard elements to aggregate and index wherever possible. Within insects, caste and dorsal/lateral/frontal image could be treated as standard elements for ant specimens. Processing pipelines may differ for zoology, botany and paleontology. In order to achieve such a model, GBIF will need efficient ways to work with taxonomic communities to determine achievable data profiles for different taxa.

Designing, Implementing, and Benefiting from a Collections Attribution Channel: the view from iDigBio

Alex Thompson, Deb Paul, Gil Nelson
iDigBio, University of Florida, Gainesville, FL. godfoder@acis.ufl.edu

In this talk, we will focus on those attribution issues that are most relevant to aggregators, highlighting steps iDigBio has already taken or is considering. These issues include data requirements, data

standards, metadata standards, API development, an identifier schema that promotes linkages and automated citation processing, the challenges of tracking and ensuring fidelity of derived datasets against those downloaded, and implementations that make citations as durable as the objects cited.

Portals Statistics: Biting the hand that feeds you

David Bloom
Vertnet, University of Berkeley, Berkeley, CA. dbloom@vertnet.org

Data aggregators, often referred to as data portals, are a key player in the discoverability and delivery of data for research, teaching, policy-making, and general exploration. Most, but not all, portals provide basic usage statistics to data publishers, but those statistics are not often standardized leaving data publishers with the unenviable, and unnecessary, task to bring these statistics into concert with each other. There are standing requests from data publishers for aggregators to standardize their outputs and to provide the same basic statistics of use. While the portals may have other requirements to meet and the need to produce their own statistics to make placate their funders, it is important to remember that without data from data publishers, portals will cease to exist. Thus, it is in the best interests of all data portals to provide a standardized set of statistics for the benefit of data publishers in order to support those publishers in their efforts to seek funding and maintain the support of their local institutions.

Biodiversity Information Serving Our Nation (BISON): five years later

Anne Simpson
BISON, US Geological Survey, Reston, VI. asimpson@usgs.gov

In the NIBA implementation plan report published in 2012, the US Geological Survey (USGS) was reported to be "developing a program—Biodiversity Information Serving Our Nation (BISON)—that will contribute significantly to the implementation of NIBA... it will serve as an integrated resource for biological occurrence data from the United States and will function as the US node in GBIF. [I]t will [also] mobilize and integrate environmental data for sustaining the nation's environmental capital." Since publication of the report, the BISON application has grown from 110M to 375+M species occurrence records displayed on 50+ possible map layers. This presentation will briefly discuss some challenges as well as lessoned learned in the large-scale aggregation of diverse species datasets, from attribution to attrition. By implementing a name lookup using the Integrated Taxonomic Information System, BISON also enables a broad species lookup function to improve species search results among its diverse datasets.

DataONE Snapshot

David Vieglais
DataOne, University of New Mexico, Albuquerque, NM. vieglais@ku.edu

The National Science Foundation supported Data Observation Network for Earth (DataONE) has implemented infrastructure supporting a global, federation of interoperable Earth Science data repositories known as Member Nodes. DataONE provides core services including search, replication, authentication, usage metrics and a consistent application programming interface for participating Member Nodes. These core services help ensure users find relevant research data products and those products remain resolvable, citable, and accessible over the long term. The project also has a significant

outreach component providing community resources to support a broad spectrum of research data stakeholders. DataONE has been in production operations since mid 2012 and currently supports 43 Member Nodes and 360,000 diverse data sets on earth science topics. A brief overview of DataONE will be provided with emphasis on data integration and attribution characteristics of the infrastructure.

Data integration and provenance in the Encyclopedia of Life

Katja Schulz
Encyclopedia of Life, National Museum of Natural History, Smithsonian Institution, DC.
eolspecies@gmail.com

As a large-scale biodiversity aggregator, the Encyclopedia of Life (eol.org) harvests information from diverse content providers. This includes taxon names and hierarchies, text, media, and structured data records (traits, occurrences). Data integration relies on names reconciliation algorithms that leverage taxonomic information from many different resources (e.g., Catalogue of Life, ITIS, WoRMS, NCBI). A variety of community standards are employed by content providers (e.g., Dublin Core, Darwin Core, Audubon Core), and data are further harmonized through post-hoc semantic annotations from domain ontologies and controlled vocabularies (e.g., ENVO, Uberon, PATO). EOL data are disseminated through human and machine accessible interfaces, and each data record is accompanied by available metadata on provenance and other relevant parameters. Due to the great diversity of attribution models communicated by content providers, attribution data are heterogeneous and often complex, making compliance with attribution requirements difficult for data users. To facilitate and track data use and redistribution, automated solutions are needed based on globally unique identifiers for individuals and institutions. Since different aspects of provenance will be significant depending on the intended data use, better standardization of contributor roles (e.g., author, compiler, publisher, funder) is desirable, as well as more detailed attribution guidance for data users.

Big Data Needs Little CRUD - data co-creation and integration for a comprehensive taxonomic framework

David F. Mitchell and Thomas Orrell
ITIS, Smithsonian Institution, DC. *mitchelld@si.edu*

The Integrated Taxonomic Information System (ITIS – www.itis.gov) maintains a regularly updated database of global scientific names and their hierarchical or synonymic relationships. ITIS partners with Species2000 to create the Catalogue of Life (CoL), a source of quality taxonomic data that allows scientific names strewn through literature as far back as 1753 to be conveniently found. Big biodiversity data is underpinned by scientific names, which means how we create, read, update and delete (CRUD) scientific names and their associated data deserves careful attention. To meet the objective of providing a complete taxonomic database ITIS is developing a new data-editing platform for taxonomic data - Taxonomic Workbench 6.0. The new collaborative software will allow data stewards and taxonomic experts to actively participate in the ITIS data quality process. Simultaneously ITIS seeks to develop services within the Taxonomic Workbench to integrate nomenclatural content within CoL, with the objective of broadening name coverage to fill known taxonomic gaps and adding missing content from covered taxonomic groups such as synonyms, subsequent combinations, and orthographic variants.

Biodiversity Heritage Library – User-informed priorities for improving research efficiency in open science

Carolyn Sheffield
Biodiversity Heritage Library, Smithsonian Libraries, Smithsonian Institution, DC. sheffieldc@si.edu

With over 53 million pages of open access literature and archives covering 500 years of biodiversity research, the Biodiversity Heritage Library (BHL) serves as a critical resource for studying life on earth. Included in the BHL corpus are species descriptions, distribution records, historic climate records, records of scientific exploration and discovery, and related data. Within these pages, over 180 million instances of species names have been indexed and made searchable to help facilitate research. With the growing need to store, access, preserve, and understand ever quantities of biodiversity data, BHL recognizes several key opportunities for further improving research efficiency to the literature and archives that make up this corpus. Results of recent user needs analysis and best practices research will be covered, along with top priorities for technical development identified through that research. Specifically, updates on BHL's top two priorities – implementation of full text search and incorporation of available transcriptions—will be covered. Examples of anticipated implications for enhancing research efficiency include improved access not just to species level descriptions but to specimen level descriptions as museum codes mentioned in the text become searchable. This talk will also provide a framework for discussing how BHL might best intersect with other developments underway or in planning across the larger biodiversity community.

The challenges of attribution: Insights from one biological collection

Michael Webster
Cornell Laboratory of Ornithology, Ithaca, NY. msw244@cornell.edu

Proper attribution is critically important to the long-term maintenance and growth of biological collections, and yet has become increasingly challenging as digitization of specimens and widespread data sharing become commonplace. In this talk I will illustrate both the benefits and challenges of proper attribution with respect to a very specific type of collection: one that curates "biodiversity media" that capture the behavioral phenotype of individuals (e.g., audio recordings of acoustic signals). Such recordings are often collected by a single individual who is sometimes part of a larger research group, are contributed to a hosting research collection which is sometimes associated with a federated group of such collections, and the data are then passed on through to one or more data aggregators, where it is often discovered by a researcher wishing to use the recording for a project. Proper attribution is important to each of these links in the chain of data-delivery. Other types of collections have slightly different, though generally very similar, needs for data handling and attribution. The intent of this presentation is not to provide easy answers, but rather to clearly illustrate the issue to facilitate discussion of possible solutions.

Linking specimen data to GenBank records

John Sullivan
NCBI/Genbank, Bethesda, MD. john.sullivan@nih.gov

GenBank (NCBI/NLM/NIH) is an indispensable resource for vast domains of medical, biological and ecological research. Sequences derived from biological specimens, hosted on GenBank, are among most valuable and visible research products of biodiversity collections. Hence, these institutions have an

interest in receiving attribution when these products are used or cited and in tracking their usage. A first step towards attribution is linking specimen data in online biocollection databases to GenBank records. I will describe the two methods by which supplementary specimen information from external providers can be linked to individual GenBank records. The first method is via a source modifier ("specimen_voucher," "culture_collection," or "bio_material") submitted concomitantly with the sequence data. Specimen metadata submitted in Darwin Core triplet format, e.g. "/specimen_voucher="USNM:MAMM:602070"" will automatically generate a hyperlink for institutions that have registered a formula for mapping specimen IDs to URLs. The newly public Biocollections Database is a registry for these Darwin Core institution and collection codes. The second method is via LinkOut, a service that creates sidebar links on individual database records to related resources on external websites. Among the advantages of LinkOut is that providers can be unaffiliated with the submitter of the GenBank data and LinkOuts can be added or changed at any time by the provider without assistance from NCBI staff.

Avifaunal Change Over Three Decades in North America Detected Via Integration of Specimen and Observational Data

A. Townsend Peterson
Biodiversity Institute, University of Kansas, Lawrence, KS 66045. town@ku.edu

Although change and turnover of faunas and floras are important topics in ecology and evolutionary biology, studies over long timespans and broad areas are rare. We explored historical (i.e., before 1980) specimen data for North American birds, and identified 139 sites with demonstrably complete avifaunal inventories. We compared those inventories with recent (i.e., after 2010) data for the same sites deriving from observational datasets: of the 139 sites, 108 had demonstrably complete avifaunal inventories for the recent period as well. Patterns of avifaunal turnover had significant spatial structure, meaning that there were broad spatial patterns and not just random gains and losses. We then explore and analyze species-level patterns of range expansion or collapse, and relate these patterns to environmental changes occurring across the study area and across the species' geographic ranges. Via broad integration of diverse biodiversity data sources, we were able to assemble a view of avifaunal turnover that is broad-scale in both space and time and that is not otherwise achievable.

Case Studies in Simplifying Attribution Across Internet-Scale Workflows

Rob Guralnick
University of Florida, Gainesville, FL. rguralnick@flmnh.ufl.edu

Aggregators serve as a critical lynchpin between data providers and consumers, and provide services to both. In previous talks, my colleagues from VertNet have discussed tools such as migrators, which are meant to support standardization of data both at the level of aggregators and for potential improvements at the source. They have also covered provider statistics, which support a much-needed function to support provider's needs about digitized collection use. Aggregators can also help assemble data demanded by users, but not easily consumable, as I will briefly mention related to traits such as body size. Full attribution, however, needs to be baked into the whole process of developing information and knowledge products, downstream from where providers and aggregators "sit". I will describe work we are undertaking to help automate production of provider citation reporting and creating community metadata for tracking provenience related especially to niche modeling. These tools are being baked into the R statistical software set. While it is tempting to think that it is APIs or

GUIDs that might hasten better data integration and attribution, but rather it is standards and tools built on top of them that are required here too.

Tracking Data from Download to Publication – and Back?

Pamela S. Soltis
Florida Museum of Natural History, University of Florida, Gainesville, FL. psoltis@flmnh.ufl.edu

Digitized natural history data are enabling a broad range of innovative studies of biodiversity. Large-scale data aggregators such as GBIF and iDigBio provide easy, global access to millions of specimen records contributed by thousands of collections. A developing community of eager users of specimen data – whether locality, image, trait, etc. – is perhaps unaware of the effort and resources required to curate specimens, digitize information, capture images, mobilize records, serve the data, and maintain the infrastructure (human and cyber) to support all of these activities. Tracking of specimen information throughout the research process is needed to provide appropriate attribution to the institutions and staff that have supplied and served the records. Such tracking may also allow for annotation and comment on particular records or collections by the global community. Detailed data tracking is also required for open, reproducible science. Despite growing recognition of the value and need for thorough data tracking, both technical and sociological challenges continue to impede progress. In this talk, I will present a brief vision of how application of a DOI to each iteration of a data set in a typical research project could provide attribution to the provider, opportunity for comment and annotation of records, and the foundation for reproducible science based on natural history specimen records. Sociological change – such as journal requirements for data deposition of all iterations of a data set – can be accomplished using community meetings and workshops, along with editorial efforts, as were applied to DNA sequence data two decades ago.

Publishing and Markup of Collection Data

Teodor Georgiev, Donat Agosti, Lyubomir Penev
Pensoft Publishers, Sofia, Bulgaria. preprint@pensoft.net

Taxonomic practice dictates that authors cite the primary biodiversity data from collections for the species used in their analyses. Usually this kind of data is cited as a list of species occurrences in the respective taxonomic treatment section of the article. Pensoft and Plazi have developed a workflow for tagging of various Darwin Core compliant data. While Pensoft is focusing on markup during the publication process and visualizes the data in both the human-readable (HTML) and machine-readable version (taxPub JATS XML) of the articles, Plazi has developed technologies for post-publication text and data mining of the same types of data. These technologies altogether allow for assembling interoperable data from various sources including the historical literature. At the end, data extracted from literature may be put together by various aggregation engines, such as GBIF, Plazi TreatmentBank or the Linked Open Data based OpenBiodiv knowledge graph. Information on species occurrences, including raw data from biological collections, are stored in different biodiversity databases such as GBIF, PlutoF, iDigBio and others. To facilitate the use of this data in publications, Pensoft developed an API-based import from the above-mentioned resources direct into manuscript developed in the ARPHA Writing Tool and intended for publication in the Biodiversity Data Journal. Our workflows may also act as a curation filter for occurrence data as, once data are imported into a manuscript in the publication pipeline, their accuracy is expected to be vetted by authors and reviewers. Tracking the usage of collection data in publications is important for authors and collection managers and it is possible by using the

standardized collection and institutional codes (currently we use GRBIO for that) but far better through persistent identifiers of individual collection objects.

What Publishers Can Do (and Are Doing) to Facilitate Data Integration and Attribution

David Parsons
Elsevier, New York, NY. *d.parsons@elsevier.com*

Elsevier has initiated a number of projects to assist the academic community in pursuing an agenda of better-documented elements of the research process. Four "research elements" journals – *Data in Brief, MethodsX, HardwareX,* and *SoftwareX* – are allowing authors to co-submit information for peer-review which would typically be hidden in supplementary files at the back of the original research paper. Co-submission of datasets without peer review is also now possible, through Mendeley Data integration directly within our submission portals. Aside from building this seamless co-submission workflow, Elsevier is progressively implementing the Transparency and Openness Promotion (TOP) Guidelines across all eligible journals, as well partnering with external repositories to allow for dynamic linking between research paper and datasets.

Collections support through NSF as it applies to data integration and attribution

Reed Beaman
National Science Foundation, Washington, DC. *RSBEAMAN@nsf.gov*

No abstract provided

Data Integration and Attribution: A Library's Perspective

Scott Hanrath, Ada Emmett, Jamene Brooks Kieffer
KU Libraries, Lawrence, KS. *shanrath@ku.edu*

Libraries bring long experience with harmonizing disparate sources of information in order to represent, locate, and engage with the scholarly record. As the scholarly record has evolved from print to digital, and begun to incorporate scholarly products beyond books and articles, libraries have adapted traditional roles and workflows. Integrating and attributing specimen data into biodiversity research presents core issues that are familiar to libraries, while introducing nuances that are specific to relevant disciplinary communities. Librarians from the University of Kansas will discuss challenges and opportunities with regard to scholarly identifiers, data management and discovery, and institutional and discipline-specific repositories.

Whole Tale: The Experience of Research

Bertram Ludaescher
University of Illinois at Urbana-Champaign, University of Illinois at Urbana-Champaign, Champaign, IL.
ludaesch@illinois.edu

In this talk, I will present on the Whole Tale project, which is a multi-discipline, multi-year effort to encapsulate the full story of digital scholarship by reducing the barriers between scholars and the computational instruments they utilize. In this talk, I will describe our approach to cyberinfrastructure, our utilization of existing tools, and the new components we are developing.