Biodiversity Data: Refinement of Technology and Implementation Methods

Tim Diamond

Johns Hopkins University: Museum Studies Digital Curation Certificate Program

AS.460.674.81. FA22: Digital Curation Research Paper

Jocelyn Boigenzahn

12/18/22

Abstract

Biodiversity data consists of taxonomic specimens and information that inform our interpretations of ecosystems and life on Earth. Museum projects, exhibitions, and research utilize biodiversity data to construct answers and educational programming for staff and visitors. Cleaning and maintaining biodiversity data, however, is a difficult challenge that involves moderation and refinement of data entry, inventory, workflows, and protocols. Creating an ideal framework that involves the utilization of technology and the management practices of data standards will help in developing baseline recommendations for institutions struggling to maintain their biodiversity collections.

Surveys were sent to listservs and museum professionals to acquire interpretation and data surrounding biodiversity data practices. From survey results, three interviews/case studies were performed with one staff member, respectively, from the University of Wyoming Museum of Vertebrates, Bernice Pauahi Bishop Museum, and the Smithsonian National Museum of Natural History. These interviews and surveys, in conjunction with a literature review, were conducted to explore processes and strategies currently being utilized to develop biodiversity data frameworks. Results indicate a strong desire for customizable and malleable databases that integrate institutional-level decision-making and preventative error protocols. In addition, thorough documentation and active engagement with staff and volunteers contribute to long-term benefits to data management standards.

Table of Contents

Introduction and Problem Statement

Biodiversity data describes the taxonomic and informational data relating to species and the ecosystems (such as locality and occurrences) that inform our interpretations of current and prehistoric life on Earth. Museums and cultural institutions make a concentrated effort to digitize, store, and share biodiversity data through their museum databases, as well as online aggregators such as Integrated Digitized Biocollections (iDigBio), VertNet, and the Global Biodiversity Information Facility (GBIF). However, effectively disseminating biodiversity data of high quality has numerous challenges. Data transcription, inventory, and workflows result in many local procedures for the handling of biodiversity collections that have not been standardized. This leads to messy and inconsistent data within and across biodiversity collections. Creating standardized frameworks that involve technology and workshop/training protocols are necessary to mitigate these data concerns and improve the future viability of said data.

Maintaining biodiversity data quality is necessary as Nelson and Ellis's *The Impact of Digitization and Digital Data Mobilization* (2019) details that there is an increasing trend of scientific publications utilizing digitization project data. New research and projects/initiatives from institutions stem from this data. One example from Nelson and Ellis's (2019) study combined museum record data with historical precipitation data to track the emergence of cicada populations based on predicted rain or weather patterns. Data and projects build on each other from thorough and accurate specimen data. Nelson and Ellis (2019) outline problems that persist in data accuracy due to the resolution and accuracy of taxon names and geospatial coordinates lacking consistency across the field. According to Nelson and Ellis (2019), several institution projects have tried to correct inconsistencies by addressing improper usage of the Darwin Core

standard or antiquated synonyms, but comprehensive solutions are not yet available that can be applied to all institutions.

The following research question is addressed in this paper: What developmental frameworks are most efficient in cleaning and refining biodiversity data? The two supplementary questions include: What functions and technology are helpful in cleaning biodiversity data? What are efficient ways to enforce data consistency and quality standards across biodiversity collections? A literature review will help assess the frameworks that have been effective to maintain biodiversity data quality across different institutions. Three case studies, consisting of survey responses and interviews from the Bishop Museum, University of Wyoming Museum, and National Smithsonian Museum of Natural History, will pinpoint the more intricate processes that occur in formulating these frameworks, both from a technology and a staff development perspective. These assessments combined will formulate the future trajectory for museum institutions in assembling biodiversity data.

Notable Terms:

Biodiversity: Biological diversity among and within plant and animal species in an environment.

Informatics: Interpreting, transforming, and assessing how to use sets of data.

Controlled Lists: Vocabulary entered within the database that is restricted within formatting rules such as search retrieval functions and pick lists.

Taxonomies: Vocabulary that is classified and organized into a hierarchical structure, appropriate for specimen classification.

Literature Review

The literature review section Functions and Technology encompasses quantifiable mechanisms and programs that enhance biodiversity data validity. The *Identifying Criteria* subsection outlines principles and mechanisms that institutions could utilize and some decision-making parameters associated with them. The *Frameworks, Tools, and Prototypes* subsection describes more specific frameworks from one (or multiple) institution(s) such as prototype transcription services, decision-making initiatives, and the benefits (or potential ones) that resulted from them.

Functions and Technology

Identifying Criteria

As museums and cultural institutions develop technology-based frameworks, outlining the processes and standards from a holistic perspective is necessary. Chapman's *Developing Standards for Improved Data Quality* (2020) discusses that the museum data will define to the data user what criteria and protocols are needed for improving data quality. Furthermore, Chapman (2020) mentions concrete applications to those policies. Before deciding on or developing the technology function, the institution should assess its data weaknesses from both a content and a staff mindset. Chapman (2020) acknowledges the process could be considered a fitness-for-use framework providing a structure for data use and context. While general and open-ended, the article's content emphasizes this assessment as a good brainstorming practice for considerations in implementing technology. For example, Chapman (2020) states that in the data quality solutions section, for a validation test, you should have parameters and specifications for the test. This helps narrow down the code needed to create it and the role that technology contributes.

Veiga's *A conceptual framework for quality assessment and management* (2017) defines data quality planning and user/staff specifications in the form of conceptual frameworks. A conceptual framework assesses data quality (DQ) processes including needs, solutions, and data quality reports (2017). Veiga (2017) explains that once the data quality measurements are specified, the institution can outline proper assertions for the data. A specific mechanism mentioned by Veiga (2017) is the Application Programming Interface (API) which implements data enhancements to mass biodiversity data. Users can copy and paste pieces of code provided by aggregators and set them in pieces of open-source software or host programs to extract data from one host into the institution's data. An institution could pull higher taxon data from GBIF's API to fill in the classification data of the animal in a quick and consistent manner.

Anderson's *Optimizing biodiversity informatics* (2020) highlights limitations and improvement measures different aggregators can take to improve biodiversity data quality. It is another problem altogether when the databases that institutions are extracting from have data quality problems of their own. Anderson (2020) discusses that while the Darwin Core standard is used to facilitate data limitations, it lacks functionality in assembling data suitable for analysis. This leads to inconsistency based on differences in data needs and quality between institutions. Furthermore, Anderson (2020) mentions data access lacks accessibility and transparency, creating misleading impressions of data quality. Anderson (2020) proposes stable unique identifiers as a solution as they provide a way of indicating record lists and facilitating qualitative information surrounding biodiversity statistics. Anderson (2020) enumerates that automated cleaning systems, quality flags, and institutional collaboration are needed to create strong and widely available biodiversity data addressing scientific questions.

Oritz-Troncoso's *Data management aspects of public engagement* (2015) describes the awareness problem with digital and technology-related frameworks. One broad framework Oritz-Troncoso (2015) outlined is the Data Management Association (DAMA) creating the Data Management Body of Knowledge (DMBOK). The main requirements Ortiz-Troncoso (2015) outlined for this framework are developing the data policy, an appropriate data integration architecture, making information accessible to different audiences, training volunteers, and implementing quality control measures. One term Oritz-Troncoso (2015) noted is quality awareness, which involves the processes of training on transcription, annotation, and geo-referencing by combining scientific and public outreach. Outlining collaboration initiatives within a broad definition article indicates that staff and audience integration is commonly used for developing these digital frameworks.

Frameworks, Tools, and Prototypes

Hill's *The notes from nature tool* (2012) highlights transcription services and transparency methods in utilizing biodiversity data to its fullest potential. A prototype transcription service from Hill (2012), titled Notes from Nature, organizes data and images into projects, collections, and missions. Hill (2012) observes that the system gives a sense of purpose and reward to users as the users transcribe record/ledger pages, which will then be displayed as part of their profile pages. Hill (2012) explains reward badges are earned for different users and collective map processes show how volunteers are filling in information for staff to utilize. While in its infancy, this template for reward and transparency will help users and staff feel motivated in catching errors and doing their part in proper research and quality assurance with biodiversity data.

A potential digital infrastructure framework was outlined in Dhindsa's *A Novel Model for Building Digital Infrastructure* (2021). This framework promotes collaboration between citizens and professionals regarding biodiversity data. Dhindsa (2021) describes the proposal itself as a Biological Informatics system that utilizes visual and imaging tools connected to computation tools and cloud networks. The expected results from Dhindsa (2021) include participation and engagement with multiple types of staff and collaborators. Dhindsa (2021) describes how the system facilitates species matching and name parsing along with community-based validation systems with scientific advancements. The level of infrastructure and quality assurance measures, combined with staff/volunteer relationships, could create a strong template for re-evaluating and improving biodiversity data.

Moritz's *Towards mainstreaming of biodiversity data* (2011) describes the GBIF Data Publishing Framework Task Group that was developed to inquire about and recommend improvements for biodiversity data quality and access. The Task Group from Moritz (2011) concluded that extensive and elaborate metadata standards will fail because they require excessive resources. Moritz (2011) added that any museum/institution process should include a metadata operation that is collaborative and continuous, as well as data quality measures prioritized according to data demand. This describes a process that is intuitive but fluctuates according to current interests or projects being developed. The article also lists recommendations directed toward GBIF. One recommendation from Moritz (2011) is to strengthen its trusted digital repository network. Combining applicable standards as much as possible with other aggregators like VertNet and iDigBio would help in data unification and resource management, as they would collaborate as one aggregation system.

Peterson's *Data Leakage and Loss* (2018) discusses the inherent problem of the lack of data standardization. Peterson (2018) describes that data leakage occurs when records are compromised by incomplete and unshared data. Broad solutions from Peterson (2018) are outlined including the tool and community-based georeferencing initiatives. An example cited within Peterson's (2018) article is Wieczorek's *The point-radius method for georeferencing* (2010), which describes localities as coordinate pairs and distances, encompassing uncertainty and the locality information. Peterson (2018) also emphasizes that data should be placed online with documentation as soon as possible as a practical measure to expand and refine the data before it gets too unwieldy. Incorporating pieces of data into one (like the point-radius method) will only help in solidifying data creation and dissemination processes, putting data online at a faster rate.

Data Consistency and Quality Standards

This section of the literature review addresses logistical, training, and staff standards/protocols that museums utilize to enforce data quality standards on biodiversity data. The *Templates and Workflows* subsection outlines methods such as brainstorming sessions and work modules used to facilitate data standard enforcement. The *Acknowledging Stakeholders* subsection describes practices for collective community engagement and morale initiatives, particularly among volunteers, to maintain biodiversity data.

Templates and Workflows

Having the tools and technology for data quality assurance is necessary, but quantifying the how and the why of enforcing those data practices is also needed for improved data quality. Sutter's *Practical guidance for integrating data management* (2015) outlines a long-term ecological monitoring project (LTEM) involving data having sufficient qualities for

interpretation by different users. Sutter (2015) outlines broad-scale data quality practices, with completing a regular data quality assessment being a key component in the process. Sutter (2015) describes the LTEM project as consisting of regular postseason review meetings that instilled a learning perspective into the project. Sutter (2015) elaborates that this perspective extends into protocols and procedures reflecting changes in current practice. The review meetings present an ongoing and evolving relationship between constituents and the data itself, creating accountability and transparency at a constant rate. These meetings build relationships between partners and staff, creating a responsible museum ecosystem that can adapt and evolve as needed.

Karim's *Digitization workflows for paleontology collections* (2016) outlines how digital workflow modules can increase data quality standards. Workflow modules can highlight notable mistakes museum staff make while adhering to digital curation. Examples from Karim (2016) include a pre-digitization setup such as assessing priorities, databases, and updating taxonomies within the said database. Karim (2016) describes that data entry itself is another module where, for quality control standards, students or interns prepare data spreadsheets, and supervisors check the data prior to being uploaded. A newer module Karim (2016) describes is proactive digitization, where digitization is done by scientists during fieldwork. This article emphasizes preventative measures to ensure data is accurate. The proactive digitization component is compelling as it could be refined to where the professional field workers perform entry on one specimen/piece of data, establishing a specimen template for further entries. This relationship serves as a framework for museum staff and interns to revisit if they have data quality concerns or questions.

Wieczorek's *Meeting Report: GBIF hackathon-workshop* (2014) discusses the refinement of the Darwin Core standard. Wieczorek (2014) explains that Darwin Core is a uniform data standard for biological specimen information that can and should be translated to different institutions as they work with data. Wieczorek (2014) outlines new Darwin terms, mechanisms, and improvements to the system including sanitizing values for fields, verifying the integrity of data rows and controlled vocabulary, and identifiers being used as expected.

Discourse also surrounded the Integrated Publishing Toolkit (IPT) in conjunction with Darwin Core. Wieczorek (2014) examines the IPT as part of the Darwin Core Archive in which data files are organized as star schema and core data files are attached to extensions, where the extensions point back to the core file, creating a more robust record. Wieczorek (2014) discusses a dispute in the workshop between using the two terms occurrence and event. Karim within Wieczorek's (2014) study mentions the dispute progressed into whether the word occurrence was related to the term's organism or taxon and how occurrence differed from the word event. Wieczorek (2014) states there was no resolution (at the time of this article), and the discussion will continue in future meetings. This level of debate for every term reinforces the notion of producing a proper agreement or consensus measurement.

Nelson's *Digitization workflows* (2015) outlines a process for different plants, algae, and fungi. Nelson (2015) explains that the conceptual development of these (and other) workflows was devised by the Developing Robust Object-to-Image-to-Data Workflows Workshop (DROID) in Gainesville, Florida. Nelson (2015) describes disparities and refinements for workflows including digitization tasks, protocols, infrastructure, and practices. Nelson (2015) mentions that through these meetings, workflow modules were created that are best viewed as malleable templates. Many of the modules contain preparation-based procedures for new or

emerging museums. For example, according to Nelson (2015), Module 10, *Selecting a database*, addresses NSF-funded natural history databases and what they may offer to institutions. More networking and resource pinpointing to foundational modules such as these will give museums and institutions a head start in developing their own foundations and/or communicating with different institutions about refining data standards and models.

Nelson and Paul's *Five task clusters* (2012) outlines a practice and questionnaire carried out across different natural history institutions to develop effective digitization practices for biological specimens and data. According to Nelson and Paul (2012), of the nine institutions assessed, five task clusters were commonly identified including pre-digitization and staging, specimen image capture, specimen image processing, electronic data capture, and georeferencing specimen data. The analysis portion from Nelson and Paul (2012) highlights general procedures that were commonly practiced. Nelson and Paul (2012) details camera settings such as the shutter speed and focus point as a starting point to direct institutions toward further reading. Alternative strategies were also articulated for practices that failed. One problem discussed by Nelson and Paul (2012) was the barcode extraction failure rate. As a solution, the institution designed new software to integrate camera filenames with database entry names.

Acknowledging Stakeholders

The National Research Council (NRC) of The National Academies Book *Preparing the Workforce for Digital Curation* (2015) posits that the role of digital curation does not just belong to curators. This role stems from all fields and departments that have a stake within the museum. The NRC (2015) describes that trusted and accredited standards require governance, succession plans, and management policies. The NRC conducted a National Science Foundation (NSF) workshop outlined in the book. The NRC (2015) mentions that with the number of stakeholders

involved, the following challenges emerged: selection strategies (what to preserve), how much content to include (usability for the future), tools and techniques/technologies to implement, as well as cost and accountability models. The NRC (2015) pushes for greater accountability and transparency from all communities to help with digital curation initiatives.

Wiggins *Mechanisms for Data Quality and Validation* (2011) writes about the dichotomy between data quality and staff and volunteers within citizen science. A table from Wiggins (2011) summarizes data validation methods and their effectiveness measurements from questionnaires, with expert reviews carrying the highest average of effectiveness with 77%. Elaborating on the expert review statement, Wiggins (2011) pinpoints different data validations that may be used instead of expert opinion from data validation and that quality management techniques should be based on the growth and size of the data set. Wiggins (2011) states that focusing on refined data entry forms and data mining techniques for large-scale data is necessary for improved data interpretation. This analysis mirrors the proactive digitization angle from Karim in that through expert reviews, volunteers can assess the data "template" for incoming specimens, and in conjunction with refined data forms, this should help mitigate resource and time constraints for these digitization workflows.

Turnhout's *Citizen science networks in natural history* (2016) discusses interviews and documentation that work to refine citizen science and digital data. Turnhout (2016) mentions how validation and quality control are commonly debated as part of the discussion on ensuring volunteer data reliability. The issue concerning community belonging is discussed within validation procedures and the relationship between experts and volunteers. An anecdote in Turnhout's (2016) article describes that volunteer members want to understand staff and protocol, not just be vessels that enter data for projects. The conclusion from Turnhout (2016) is

that data validation and collecting are distributed and collective, comprising interactions with recorders, field guides, and organizations. This ties into the GBIF hackathon article, emphasizing collaboration and consensus and leaving contentions between different communities and groups at the door to deliver the best possible digital data.

Research Design

The quantitative data research design of this paper consists of the examination of three case studies. The data gathered for these case studies were completed through a survey and an interview with a select member from each museum. The research began with a general survey distributed to the digitization listserv group at the Integrated Digitized Biocollections (iDigBio) National Resource Site (iDigBio, n.d.). iDigBio's working groups have created forums for discussion and analysis of digitization-related solutions, such as workshops to enhance and refine these best practices for biodiversity data. In addition to the listserv, the researcher shared the survey with data curators, managers, and collection staff outside of the listserv, for responses.

The survey itself consisted of a questionnaire with seven main questions (see appendix for a complete question list) that respondents were asked to answer. The survey's content includes pick lists and short, two to three-sentence free response sections intended to answer the seven main questions, along with sub-questions. The seven main questions included the following: What software do you currently use? What metadata schemas are currently being used? What specific controlled vocabularies are used to enforce data standards? What quality assurance measures have been put in place? Does the institution utilize volunteers? What methods are the most helpful in improving data management standards? Are they using/considering strategies to improve resource/training standards?

Respondents were asked to identify their institution type (Collection Holder, Data Contributor, Data Aggregator, or Other) and include their email and name within the survey section if they were willing to respond to an in-depth interview over Zoom and further email inquiries. The members (and their institutions) who self-selected for the interview project were the following: Dr. Elizabeth Wommack from the University of Wyoming Museum, Dr. Richard Pyle from The Bishop Museum, and data manager Adam Mansur from The Smithsonian National Museum of Natural History.

The main topics covered in the in-depth interviews included: their role at their institution, whether they use commercial or open-source platforms, technical limitations with software, revision/upgrade recommendations, methodology to improve data quality, challenges enforcing data standards, challenging moments in staff conversations, and steps/improvements being considered to further refine the data process. Depending on how the interviewees responded to those questions, some additional questions were asked. These additional questions included the following: use of the Darwin Core standard, utilizing transcription services, collaboration frameworks, governance with financial stakeholders, workflow module templates, and preventative/proactive measures to enhance data integrity. These case studies, developed through current data gathered via surveys and interviews, in conjunction with the literature review, provide a layered comparison and analysis through which answers to the proposed question: What developmental frameworks are most efficient in cleaning and refining biodiversity data? and sub-questions: What functions and technology are helpful in cleaning biodiversity data? What are efficient ways to enforce data consistency and quality standards across biodiversity collections? can provide an assessment of the technology available and data standard frameworks being employed now that could be considered for future deployment elsewhere.

Case Study 1: University of Wyoming Museum of Vertebrates

The University of Wyoming Museum of Vertebrates (UWYMV) is a museum located in Laramie, Wyoming. The museum was founded in 1890 and has since moved into the University of Wyoming's Berry Biodiversity Conservation Center in the spring of 2011, hosting an assemblage of vertebrates and preparatory labs for skeletons, skins, and fluids (Home, n.d.). The mission of the UWYMV is to "document and understand regional and global biodiversity through acquisition and investigation of collections to advance academic knowledge and public appreciation of the natural world" (Home: University, n.d.). The UWYMV houses over 11,800 specimens using the Arctos Collaborative Collection Management Solution online collection database including mammals, birds, reptiles, amphibians, and fish (UWYMV Collections, n.d.). The Arctos database itself functions as a community center, a collection management information system, and a research-grade data provider all in one, providing fundamental infrastructure for natural history, cultural, and art institutions (About Arctos, n.d.). Survey and interview data gathered from UWYMV was completed by Dr. Elizabeth Wommack, the staff curator and collection manager of the institution. Wommack's responsibilities cover collecting and processing specimens, teaching students and volunteers, and curating the material itself.

Case Study 1: Survey and Interview Data

In response to the questions about cleaning/refining software, Wommack responded to none of them, but instead focused on the Arctos collection database, as it also helps with those cleaning functions. She expressed extreme satisfaction (rating level 5), emphasizing that the database itself is collaborative. In addition, a level 4 (high) satisfaction rate was selected for the Darwin Core Metadata schema. Controlled lists and taxonomies were selected for controlled vocabularies, citing that Arctos manipulates these flexible terms across multiple institutions,

creating a more holistic data practice (E. Wommack, Personal Communication, Oct 27, 2022). Arctos is accessible online; therefore, no specific installation procedures are strictly tied to local computers, helping with infrastructure, data management, and institution costs. While institutions do pay to use Arctos, there are no tier payments, increasing viability in terms of features and compatibility regardless of collection size.

Most importantly, Arctos is run and managed by the museums that use the database. Every institution that joins gets a seat on the management team and a co-chair role in the development processes. Meetings with the Arctos staff are held monthly to discuss and hold tutorials; each member also may reciprocate by sharing outside training to the Arctos community, making fast and responsive timetables toward change possible as the staff is all working with the same data (E. Wommack, Personal Communication, Oct 27, 2022). Wommack identified workshops and training sessions as important ways to improve data standards. With Arctos, these initiatives are fulfilled and broaden the importance of different types of data. These collections then connect for research and educational development or programming.

One example of these meeting results included designating unique agent names. Agents in Arctos have additional information tied to them such as ORCID (a persistent digital identifier), publications, and relationships with other Agents or institutions. If staff cannot provide this information, Arctos creates a Verbatim Agents field, which allows use of the information for the individual name in any form, while still searchable throughout the database. Wommack stated that, despite Arctos's many features, it can still be challenging to keep up with communication and meetings (E. Wommack, Personal Communication, Oct 27, 2022). Furthermore, Arctos has a lot of available tools, potentially increasing the complexity of the database. Wommack discussed and shared the Arctos wiki page during the interview, citing the

many resources and tutorials to help with the initial challenges of the system (E. Wommack, Personal Communication, Oct 27, 2022).

Quality assurance was defined by Wommack as having data fields filled out properly to create a sense of completeness, or set core values. These are complemented by Arctos's code tables, which outline specific and accepted terms that staff and volunteers can fill in. In her educational programming, Wommack trains students by having them perform data entry, starting with simple accessions and progressing to more complex work such as catalog bulk-loading (E. Wommack, Personal Communication, Oct 27, 2022). The students are monitored with the guidance of the instructors. Wommack chose asking for or researching different institutions for advice as an additional improvement measure in the survey. Also, a desire for digitization of Archive collection types within Arctos was noted; Wommack stated they do not have archival documents stored there and are currently only in hard copy.

Concerning volunteer practices, Wommack stated that while they actively enforce data standards on volunteers, they do not use software packages to do so, instead teaching the standards through actions they undertake. All volunteers are assigned to curatorial data work with the option to work in the lab. The volunteers are also guided through a list of different museum tasks such as tying tags, georeferencing, and data entry (E. Wommack, Personal Communication, Oct 27, 2022). In short, data standards are used as a teaching tool to educate volunteers on quality data and gain a more holistic perspective of the institution's tasks.

Case Study 2: Bernice Pauahi Bishop Museum

The Bernice Pauahi Bishop Museum (BPBM) is located at 1525 Bernice Street in Honolulu, Hawaii. BPBM was founded in 1889 and has progressively expanded into a central

resource for museum objects related to Hawaii and other Pacific cultures (About, n.d.). It also has one of the largest natural history collections in the world, supporting the native population, and is going through a ten-year strategic revision to reenergize the engagement of both staff and visitors (About, n.d.). The mission of BPBM is it "inspires our community and visitors through the exploration, celebration, and perpetuation of the extraordinary history, culture, and environment of Hawai'i and the Pacific" (About, n.d.). BPBM includes specimens, documents, and biodiversity databases that have allowed staff to engage in research for over 100 years (Explore, n.d.).

Survey and interview data was completed by Dr. Richard Pyle, the Senior Curator of Ichthyology and Director of the Center for the Exploration of Coral Reef Ecosystems (XCoRE) at the BPBM. He has worked at BPBM for 36 years; for the first 25-30 years he worked as an impromptu database coordinator as the museum could not initially afford an informatics team. Pyle and the museum are now in the process of cultivating an actual team and transferring data management responsibilities to others.

Case Study 2: Survey and Interview Data

Pyle did not list any cleaning/refining software as an option, instead indicating custom in-house tools optimized for different purposes in digitization and collection management. These were defined as case-by-case parsing, pattern matching, and scripts performed in Microsoft Access Visual Basic and later SQL. An example described was Agent name parsing, where the database was populated with multiple fields, and the task was to search all the database names and create a huge data index. From there, all the values would be parsed and condensed to the actual first-person names and abbreviations (R. Pyle, Personal Communication, Oct 25, 2022). Pyle indicated a level 3 (moderately satisfied) under the satisfaction question in the survey but

selected all the above for improvements including reporting systems, error-catching mechanisms, improved layout, and sharing mechanisms.

Darwin Core and other were selected in the survey for metadata schemas. The custom option was an in-house tool for taxonomic data modeling, contributing to Darwin Core and Taxon Concept Schema (TCS) development. Pyle described it as one of the four authority databases relating to their core content including taxonomy, geography, Agent names, and literature/citations. The greater plan is to collapse taxonomy, Agent names, and literature into one data model known as the Global Names Usage Bank (GNUB), intending to standardize and unite all the pieces of data into a central data hub (R. Pyle, Personal Communication, Oct 25, 2022). The data model and parsing techniques show that Pyle expressed a desire for these schemas to fulfill robust normalized data to improve data consistency.

The controlled vocabularies Pyle uses included subject heading lists, controlled lists, authority files, taxonomies, and thesauri. Pyle described that they strive to leverage and learn about the vocabulary they don't use to further improve data consistency (R. Pyle, Personal Communication, Oct 25, 2022). Completeness, consistency, and uniqueness were selected for enforced quality assurance measures. Consistency was emphasized in his survey elaboration; most of the time spent cleaning data is transforming values to a consistent form.

For improving general data standards, workshops and training sessions were selected. Pyle elaborated that workshops yield more tangible benefits as opposed to broad conferences. Meetings and workflows can be helpful but are considered malleable to a fault as biodiversity data constantly evolves and changes based on the current work demands of the institution. For example, a series of SWAT analyses and meetings were conducted with one (undisclosed) data organization, and it led to nothing valuable as the company was in over its head regarding

biodiversity. This was not anyone's fault, as Pyle reiterated, but it shows the complexity of collaboration measures (R. Pyle, Personal Communication, Oct 25, 2022).

Regarding volunteer use, Pyle described the main challenge as enforcing data standards on volunteers. Pyle maintained that most of them were retirees that invested themselves in the data entry, and in turn, Pyle wanted to retain them as they became accomplished in the process. The ideas for incentives described included dashboards that showed progress (e.g., you did the 10,000th data clean up, here is a gift card). Timestamps and activity logs were mentioned as well, to increase a sense of awareness of their contributions, as well as recognition from staff (R. Pyle, Personal Communication, Oct 25, 2022).

Discussing the history of the BPBM, Pyle stated that the museum was in "survival" mode for most of its existence regarding data standards with its parsing methods and lack of financial backing. Because of the pandemic, there was a robust shift in IT investment and informatics research. A formal informatics department is currently being established, including the hiring of three or four more staff and acquiring data storage centers and new servers. The intent is to have the staff and resources in place by March 2023, when the staff will explore new software tools (R. Pyle, Personal Communication, Oct 25, 2022).

Pyle gave more historical context to data standards, stating most data practices were done in-house. BPBM adapted informatics standards from the Pacific Information Center around the 1970s, where punch-card systems were purchased. In addition, they acquired a multitude of Apple 2 computers and siloed databases in the 1980s. This created a hodgepodge of data with no underlying structure. The initiative was to then keep primary labels, nudge them to the backend, and assess comparable tables and fields and merge them into a single system (R. Pyle, Personal Communication, Oct 25, 2022).

An exhaustive analysis of a future database began. The choices were to continue this homegrown parsing method (not sustainable) or purchase specific natural history collection databases such as EMu or Specify. The staff, however, were acclimated to custom-designed data forms and data practices, particularly grid data views and forms, which (at least at the time) were not supported by EMu or Specify. Pyle mentioned that integrating data to these commercial natural history databases would create a step backward on the taxonomic standards and capabilities they performed in the homegrown method. Pyle then discussed EarthCape, a biodiversity data platform they are currently investigating (R. Pyle, Personal Communication, Oct 25, 2022). EarthCape is optimized for custom data entry methods and taxonomic data modeling, making it an ideal fit for the staff at BPBM. Nevertheless, this investigation is still in the prototype phase and is ongoing.

Case Study 3: Smithsonian National Museum of Natural History

The Smithsonian National Museum of Natural History (NMNH) is located on Constitution Avenue NW, Washington, D.C. The museum was founded in 1910 and stewards a collection of over 147 million specimens and artifacts (About, n.d.). NMNH's mission is to "promote understanding of the natural world and our place in it" (About, n.d.). Major disciplines that comprise their research and digitization programs include Anthropology, Botany, Entomology, Invertebrate Zoology, Mineral Sciences, Paleobiology, and Vertebrate Zoology (Our Research, n.d.).

Survey and Interview data for NMNH was completed by Adam Mansur, the IT Specialist and Data Manager within the NMNH Mineral Sciences Department. His tasks include working in the EMu database, importing and cleaning data, enforcing data, and strategizing with staff for museum-wide data initiatives. While mineral sciences departments study rocks and minerals as

opposed to biodiversity, Mansur mentioned that his department's general data practices, standards, and workflows are like biodiversity practices. Information such as localities, geography, name classifications, and interpretations/workflows are commonplace in all large-scale museum digitization programs regardless of the exact discipline.

Case Study 3: Survey and Interview Data

For cleaning/commercial software options, Mansur selected other and named Python with a level 5 (extremely satisfied) rating. Mansur described Python as an open-source computer language that creates easily readable data and has a strong community of developers tied to the software, making it more flexible than EMu (the natural history collection database used by NMNH). Features such as web APIs and the Jupyter Notebook computing platform support Python's features including formatted text, annotated code/documentation, and improved flexibility with the data itself (A. Mansur, Personal Communication, Nov 1, 2022). Python's weaknesses were described as it being slow in utilization, referring to readability over the speed of the functions. While it can be a friendly code language, there is still a steep learning curve for those unfamiliar with coding programs.

Dublin Core, Darwin Core, and Biological Collection Access Service (BioCASe/ABCD) were selected as metadata schemas. Mansur added that they share data using discipline and project specific bespoke schemas provided by geoscience organizations such as the System for Earth Sample Registration and NOAA's Index to Marine and Lacustrine Geological Samples. They metadata schemas are developed from XML or CSV scripts and are mapped into EMu for specific custom schemas (A. Mansur, Personal Communication, Nov 1, 2022). Mansur gave the schemas a satisfaction rate of level 5, noting that geoscience schemas are not as widely used and are ideal for the smaller Mineral Sciences department.

For vocabulary, controlled lists and authority files were chosen. Mansur explained that controlled lists help prevent errors in data entry. Authority files help them integrate additional information without managing it in EMu. Mansur expressed interest in working more with thesauri and ontologies, but the NMNH EMu implementation does not allow them to use these vocabularies. Completeness and uniqueness were selected for quality assurance measures. Mansur described how these measures are easy to evaluate, but accuracy checks require additional steps making them rather time-consuming. He expressed interest in performing more QA measures but is at a loss for how to manage them on a large scale (A. Mansur, Personal Communication, Nov 1, 2022).

Mansur selected staff meetings and training sessions as the most important ways to improve data management standards. Training sessions and meetings were described as specific and specialized toward staff, making more of an impact on the data quality. Conferences and workshops were noted as good ideas but impossible to execute in a timely manner. He wants to ask for or research different institutions for advice, as the Mineral Sciences department does not have much data expertise (A. Mansur, Personal Communication, Nov 1, 2022). Consulting with other institutions means he can learn about the processes of other organizations and develop structures to handle data consistently between these organizations.

Mansur discussed volunteers, emphasizing the importance of making sure the volunteers are interested in working with data and have a background in that type of work. Citing past projects with previous volunteers, Mansur said you can’t pigeonhole volunteers into something they don’t want to do (A. Mansur, Personal Communication, Nov 1, 2022). Providing those willing with detailed instructions and proper work review measures is also a standard in volunteer interaction.

Discussing the Smithsonian itself and data challenges, Mansur stated that data standard methodology has been a consistent problem (A. Mansur, Personal Communication, Nov 1, 2022). The wide variety of departments and staff inevitably leads to different standards. With work standards changed due to the pandemic, switching to more remote work and a greater reliance on database work, more communities have been developed (e.g., around geography and georeferencing fields) with an emphasis on cleaning up that data. Many demographics that work in the institution are not overly familiar with databases and vocabulary protocols. They may not enter the exact terms or values and enter new ones. Mansur cleans up the data on his own time after the fact (A. Mansur, Personal Communication, Nov 1, 2022). Discrepancies with country and state names/classifications come up all the time based on the ISO Country List of Names database. For example, North and South Korea have different classifications with one being a republic and another a democratic republic. Therefore, enforcing proper standardized values for individual data pieces becomes increasingly challenging.

In data management logistics, collection managers do most of the cataloging. For rock specimens that have detailed metadata, the collection managers will email back and forth with Mansur, deciding on the data entry together. These emails are handled case-by-case, and Mansur discussed the lack of regular/weekly meetings to settle terms and data logistics (A. Mansur, Personal Communication, Nov 1, 2022). Lots of cataloging is also done by contractors positioned into specific data projects. While contractors help with data entry, they may not have much experience either. Some contractors make decisions that turn out to be the wrong one, creating systematic problems within six-month to year-long projects. While not currently implemented, Mansur discussed more regulations and documentation in independent contracts and more consistent data quality checks with contractors, such as sending in a data spreadsheet

weekly to mitigate data issues moving forward (A. Mansur, Personal Communication, Nov 1, 2022).

Mansur concluded he would like to do more templates and noted the NMNH is moving in that direction. In June 2022, a task force charged with planning the future of informatics and data science at the museum performed surveys for the NMNH science community, and the results expressed interest in sharing and the consolidation of workflows across all departments (A. Mansur, Personal Communication, Nov 1, 2022). This result also stems from highly specialized individuals retiring from their jobs, leaving confusion and ambiguity on specific data practices. Small and intrinsic in-house development support was another suggested template for improving data on large-scale projects. These would include two-week sprints to build and test features, technology, or work processes to create more data standard efficiency measures (A. Mansur, Personal Communication, Nov 1, 2022).

Analysis

Based on the data collected, it is possible to pose answers to the main and sub-questions this paper aims to address for the field of biodiversity data management. Those questions are again stated here: What developmental frameworks are most efficient in cleaning and refining biodiversity data; what functions and technology are helpful in cleaning biodiversity data; and what are efficient ways to enforce data consistency and quality standards across biodiversity collections? According to the data collected, customizable databases, programs, schemas, and controlled list vocabularies provide optimal solutions and tools to clean biodiversity. The accuracy quality assurance, volunteer engagement, training sessions, and consultation measures are the primary methods in enforcing data quality standards. The detailed benefits to these initiatives are described below.

Software across the three institutions presents the trend that museums want to maintain their data and protocol while making the manipulation, entry, and parsing of data more efficient and flexible. Anderson's *Optimizing biodiversity informatics* (2020) addresses this desire, as many databases lack functionalities to improve or manipulate problematic records. The UWYMV shows the greatest extent of this trend, with the Arctos Online database curating a staff-centered consensus strategy across data entry and methodologies that are easily translatable. BPBM and NMNH show a desire to modify or bypass their initial database strategies. Mansur's Python is made to enhance data flexibility and readability absent from the restrictive EMu database. BPBM is investigating EarthCape to easily translate their taxonomic data models into quantifiable data. This translation, transition, and customization leads to data that is more usable, as Peterson's *Data Leakage and Loss* (2018) states that data cannot be just integrated but should be checked, enriched, and documented.

Complementing the flexibility trend, all three institutions selected Darwin Core as the primary metadata schema as the starting point for their data development process. Nelson and Ellis's *The history and impact of digitization and digital data mobilization* (2018) reaffirm the schema as the default framework for biodiversity data from varied sources, so it is the logical starting point for all three departments. The basis for greater data customization, however, is also showcased in Wieczorek's *Meeting Report: GBIF hackathon-workshop* (2014) where Darwin Core is described as a tool to build on or refine more complex data or interoperability formats. Pyle extends Darwin Core's features through in-house taxonomic data modeling, combining their separate internal databases for beneficial data connections. Mansur extends the schema with explicit bespoke schemas related to geology projects to maximize standard results. Wommack,

meanwhile, extends and disseminates policies and definitions through the staff decision-making process within Arctos.

While vocabulary choices vary, Wommack, Pyle, and Mansur all selected the survey choice controlled lists as the vocabulary preference. As Mansur described, controlled lists narrow down data entry into choices, mitigating potential errors. Controlled lists are a tool and concept that help fulfill data quality (DQ) enhancement measures. Veiga's *A conceptual framework for quality assessment* (2017) lists the processes in DQ enhancement as prevention, correction, and recommendation. These processes described by Veiga (2017) include suggesting similar terms, filling in fields, and recommending coordinates based on locality info. Wommack solidifies and maintains these controlled lists within Arctos. Pyle did not elaborate on them as much, but the consolidation point he made regarding vocabulary highlights the efficiency and mitigation angle that are central to controlled lists.

Quality assurance through database software comes less defined. Wommack and Mansur emphasized completeness, while Pyle talked more about consistency. In elaboration, Wommack discussed code tables and student training in fixing poor data. Mansur expressed a desire for accuracy (despite the current conditions), while Pyle maintained consistency because it is easy to fix the mistakes of others if those errors are repeated. However, templates for uniformity and standardization naturally come from clear rules and advice/guidance. Wiggin's *Mechanisms for Data Quality and Validation* (2011) suggests multiple methods of quality assurance, but it includes expert review along with some form of documentation as a guide. Furthermore, in the mechanism's combination table from Wiggins (2011), the highest percentage, 23%, reported that photos along with expert review were the most used methods. While the answers are different,

the end results from Wommack, Pyle, and Mansur allude to accuracy. Maintaining guidance and visual aids could comfort individuals learning data standard protocols.

Volunteer assessments were consistent; Wommack, Pyle, and Mansur all concurred that volunteers cannot be taken for granted. Mansur alludes to an invested interest by the volunteer members as a key factor, as well as not pigeonholing them into a specific role. Pyle leaned toward the recognition of volunteers with activity logs, dashboards, and timestamps documenting their progress on different projects. Wommack proctored an active understanding of volunteers with the volunteer programs encompassing how the institution works and what volunteers can contribute to the museum. Transparency, knowledge-building, and contribution attributes permeate Hill's proposed *The notes from nature tool* (2012) as it shows a built-in discussion platform for volunteers to actively engage with scientists for enrichment and clarification. Professionals engaging with volunteers provides them with a sense of validation and comfort in their efforts. Turnhout's *Citizen science networks* (2016) cites a lack of response from a Scotland museum staff member regarding a volunteer's comment, leading to more disinterest in the project.

Training sessions were universally selected from all three interviews. Pyle and Wommack also selected workshops, while Mansur chose staff meetings. Nelson's *Five task clusters* (2012) indicated protocols and specialized training were the answer to technical specimen curation difficulties. All three respondents described that training sessions provided tangible results and improvements to staff members qualified to perform these tasks. On a more holistic scale, Sutter's *Practical guidance for integrating data management* (2015) mentions that strong training includes quality assurance methods and logistical/safety training, in addition to data collection procedures. Wommack, once again, mentioned Arctos and how training sessions

were performed and brought back to the collaborative meetings within the database, creating more data unification measures and likely more cumulative protocol.

Consultation and refinement are the ongoing themes shown in the improvement measure responses. Wommack reinforced Arctos's value within the decision-making process of the museum. Pyle outlined activity measures (e.g., logs and timestamps) to keep volunteers feeling valued in the workplace. Mansur's task force survey results led to ideas for workflow consolidation and test sprints to create new and simplify old and existing technology and initiatives. Chapman's *Developing Standards for Improved Data* (2020) adds to the test sprints idea, stating the test needs to be clearly defined in their outcomes and implemented across the whole community for improved transparency and communication measures. Supplementing the survey and community/volunteer engagement idea, Moritz's *Towards mainstreaming of biodiversity data publishing* (2011), indicates surveys prove value and impact for specific data, and the investment will have a probative impact on data integrity and community/staff relationships.

Conclusions, Recommendations, and Next Steps

As seen in the collected data and analysis, the most efficient technology and functions used to clean biodiversity involve databases and programs that have improved flexibility in manipulating and editing data. These databases include functions that dictate museum decision making and protocol, such as the Arctos database. Customized metadata schemas and controlled vocabularies extend data parameters and mitigate risk in data errors. Additional programming from data aggregators, such as GBIF, provide further measures in technologies that correct and streamline data.

Expert review, visual documentation, training sessions, and an emphasis on data accuracy constitute the solution to enforcing data consistency and quality standards across institutions. Dialog, acknowledgement, and engagement measures, such as activity logs, support volunteer enthusiasm and accomplishments and create a foundation for continual, and quality, data entry. With these practices in mind, here are some implications and recommendations to take away from these findings.

The most effective technology frameworks involve products and services that include all-in-one features. If willing, natural history museums and institutions should consider investigating and possibly joining the Arctos online database system. The database system itself is intuitive with records and multimedia information easily uploaded online. Once an institution signs on, they also become participants in shaping the metadata schemas and protocols through regular meetings, providing uniformity, not just in their collections, but across multiple like-minded natural history institutions as well. Under the Join Arctos tab on its website, the first step is filling out a collection form so that an Arctos Working Group officer can discuss and arrange for an institution's specific needs (Join, n.d.). Taking the time to have that conversation, regardless of whether they sign up, should at least give the institution a reassessment of its needs before choosing the next course of action.

With the emphasis on controlled lists and error-catching mechanisms, in-house development or extension tools should be implemented to improve data consistency and/or metadata schemas. Examples include APIs from GBIF that allow users to input a sequence/code to fill in taxonomic data such as family, genus, and species with one command. The Species API webpage on GBIF benefits from the integrity of the GBIF website as it works against data kept within the indexes, at least ensuring consistency when performed (GBIF Species, n.d.).

Regardless of the extension/development tool, rigorous testing practices, as alluded to by Mansur, are needed to increase transparency and training development for staff/volunteers utilizing the programs. These programs and mechanisms should be developed incrementally to avoid long-term problems that could permeate different digitization projects.

To enforce efficient data quality standards, enriched documentation and community-building measures are necessary. If not already done, institutions should create workflow templates that include detailed instructions and photos/visuals for every data entry and digitization process step so that volunteers, staff, and contractors feel secure in what they are doing. Pictures, arrows, and case-study-like process descriptions will mitigate the overwhelming nature of these complex database systems. If meetings are not regularly held, staff should sit down with volunteers/contractors to establish set dates for progress check-ins and be physically available for questions and comments. Recognition systems such as activity logs, timestamps, and guided museum procedure workshops should be developed to make volunteers feel like they are contributing and understand the bigger picture of the museum's work.

Collection and data managers across museums and natural history listserv groups are the intended audiences that may benefit from these findings. All parties mentioned are in constant discussions with peers and perform extensive research to maintain, manipulate, and refine biodiversity data. These baseline recommendations such as Arctos, extension features, and logistical practices, combined with insights from Mansur, Pyle, and Wommack, give or enhance potential ideas and solutions these audiences may implement.

That said, there is still more research to be done. Data interpreted from this paper comes from only three institutions. Performing more case studies and surveys across different institutions should yield more varied practices and interpretations. The initial survey listed

commercial software options such as OpenRefine that can help clean and refine data. Most of the initial replies are customized extensions to regularly installed natural history databases. Investigating different institutions that use the commercial software listed in the survey could provide more robust recommendation measures for institutions depending on their size and needs.

Other obstacles in these practices not addressed in this paper are financial barriers, diverse stakeholders, and governance structures that can potentially hinder biodiversity data progress. Pyle alluded to this with a reinvention plan during the start of the pandemic. Mansur also briefly discussed time and financial constraints in relation to developing new technology and standard-based practices. Performing interviews with institutions discussing financial constraints, successful grant proposals, and appealing to multiple stakeholders could help institutions stimulate ideas and break these barriers that hinder their progress in biodiversity data refinement.

References

About Arctos. Arctos. (n.d.). Retrieved November 28, 2022, from https://arctosdb.org/about/

About Us. Bishop Museum. (n.d.). Retrieved October 20, 2022, from https://www.bishopmuseum.org/about/

About the Museum. Smithsonian National Museum of Natural History. (n.d.). Retrieved October 27, 2022, from https://naturalhistory.si.edu/about

Anderson, R. P. et.al. (2020). Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society. *Frontiers of Biogeography, 12*(3), 4,6,9. https://doi.org/10.21425/f5fbg47839

Chapman, A. et.al. (2020). Developing Standards for Improved Data Quality and for Selecting Fit for Use Biodiversity Data. *Biodiversity Information Science and Standards*,4–*5*, 7, 8. https://doi.org/10.3897/biss.4.50889

Dhindsa, A., Bhatia, S., Agrawal, S., & Sohi, B. S. (2021). A Novel Model for Building Digital Infrastructure for Biodiversity Studies. *Journal of Physics: Conference Series, 1714*(1), 5–6. https://doi.org/10.1088/1742-6596/1714/1/012011

Dictionary.com. (n.d.). *Biodiversity Definition & meaning*. Dictionary.com. Retrieved June 15, 2022, from https://www.dictionary.com/browse/biodiversity

Explore. Bishop Museum. (n.d.). Retrieved October 20, 2022, from https://www.bishopmuseum.org/explore/

GBIF Species API. GBIF Species API. (n.d.). Retrieved November 10, 2022, from https://www.gbif.org/developer/species

Hill, A. et.al. (2012). The notes from nature tool for unlocking biodiversity records from museum records through citizen science. *ZooKeys*, 209, 224–225,227–228. https://doi.org/10.3897/zookeys.209.3472

Home:: University of Wyoming Museum of Vertebrates. (n.d.). Retrieved October 18, 2022, from https://uwymv.wyomingbiodiversity.org/

Join Arctos. Arctos. (n.d.). Retrieved November 6, 2022, from https://arctosdb.org/join-arctos/

IDigBio Home. iDigBio. (n.d.). Retrieved September 26, 2022, from https://www.idigbio.org/

Karim, T. S. et.al. (2016). Digitization workflows for paleontology collections. *Palaeontologia Electronica*, 6,8,10. https://doi.org/10.26879/566

Moritz, T. (2011). Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF Data Publishing Framework Task Group. *BMC Bioinformatics, 12*(S15), 3–4,6. https://doi.org/10.1186/1471-2105-12-s15-s1

National Research Council of The National Academies. (2015). Chapter 2: The Current State of Digital Curation. In *Preparing the Workforce for Digital Curation* (pp. 25,27,39). essay, National Academies Press.

Nelson, G. et.al. (2015). Digitization workflows for flat sheets and packets of plants, algae, and fungi. *Applications in Plant Sciences*, *3*(9), 2–3,7. https://doi.org/10.3732/apps.1500065

Nelson, G., Paul, D., Riccardi, G., & Mast, A. (2012). Five task clusters that enable efficient and effective digitization of biological collections. *ZooKeys*, *209*, 19,23,29,33. https://doi.org/10.3897/zookeys.209.3135

Nelson, G., & Ellis, S. (2018). The history and impact of digitization and digital data mobilization on biodiversity research. *Biodiversity Information Science and Standards, 2*, 3–5. https://doi.org/https://doi.org/10.1098/rstb.2017.0391

Ortiz-Troncoso, A. (2015). Data management aspects of public engagement with biodiversity documentation. *PeerJ PrePrints,* 3,6. https://doi.org/10.7287/peerj.preprints.992v2

Our Research. Smithsonian National Museum of Natural History. (n.d.). Retrieved October 27, 2022, from https://naturalhistory.si.edu/research

Peterson, A. T. et.al. (2018). Data Leakage and Loss in Biodiversity Informatics. *Biodiversity Data Journal*, *6*, 3,9-11. https://doi.org/10.3897/bdj.6.e26826

Sutter, R. D. et.al. (2015). Practical guidance for integrating data management into long-term ecological monitoring projects. *Wildlife Society Bulletin*, *39*(3), 451, 454, 457, 461. https://doi.org/10.1002/wsb.548

Turnhout, E., Lawrence, A., & Turnhout, S. (2016). Citizen science networks in natural history and the collective validation of biodiversity data. *Conservation Biology*, *30*(3), 533,536–538. https://doi.org/10.1111/cobi.12696

UWYMV Collections. Collections:: University of Wyoming Museum of Vertebrates. (n.d.). Retrieved October 18, 2022, from https://uwymv.wyomingbiodiversity.org/index.php/collections

Veiga, A. K. et.al. (2017). A conceptual framework for quality assessment and management of biodiversity data. *PLOS ONE*, *12*(6), 4, 8–9. https://doi.org/10.1371/journal.pone.0178731

Wieczorek, J. et.al. (2014). Meeting Report: GBIF hackathon-workshop on Darwin Core and sample data (22-24 May 2013). *Standards in Genomic Sciences, 9*(3), 585–586, 588, 594–595. https://doi.org/10.4056/sigs.4898640

Wieczorek, J., Guo, Q., & Hijmans, R. (2010). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, *18*(8), 748. https://doi.org/10.1080/13658810412331280211

Wiggins, A., Newman, G., Stevenson, R. D., & Crowston, K. (2011). Mechanisms for Data Quality and Validation in Citizen Science. *2011 IEEE Seventh International Conference on e-Science Workshops*, 3,6,16. https://doi.org/10.1109/esciencew.2011.27

Appendix A: Survey

Section 1: Data Survey Scope

Problem Statement Data transcription, inventory, and workflows result in many local procedures for the handling of biodiversity collections that have not been standardized. This leads to messy and inconsistent data with biodiversity collections. Creating standardized frameworks that are widely implemented across biodiversity data collections could improve data quality and increase efficiency. This research will focus on identifying challenges and possible solutions to this issue. These include technology (e.g., software improvements), wider implementation of different tools, such as community adoption and promotion of metadata schemas and controlled vocabularies (e.g., through workshops or other training), or other methods.

Question: What developmental frameworks are most efficient in cleaning and refining biodiversity data?

- What functions and technology are helpful in cleaning biodiversity data?
- What are efficient ways to enforce data consistency and quality standards across biodiversity collections?

Research Design: Literature review will help assess the frameworks that have been effective to maintain biodiversity data quality across different institutions. A short survey will be sent out to different Listservs within the iDigBio page. Survey respondents have three weeks to complete the short questionnaire. If willing, survey respondents will be followed up on with three in-depth survey interviews based on a selection of biodiversity aggregators and museum staff that will pinpoint the more intricate processes that occur in formulating these frameworks, both from a technology and staff development process. These assessments combined will formulate the future trajectory for museum institutions in assembling biodiversity data. More specifically, the research design will highlight the best approaches to improve data quality. These include:

1 Improved Software

- What software is used in cleaning biodiversity data?
- How satisfied are users with this software?
- What improvements, if any, would be useful?

2 Use of Existing Tools

- What metadata schemas are currently used or preferred?
- What controlled vocabularies are currently used?

3 Training Resources

- What training resources or workshops currently exist?
- How can training be improved or more widely offered and used?
- Are there workflows being used as training resources for staff/volunteers?

Target Audience: The following include the Digitization Listserv from iDigBio and other museum professionals. https://www.idigbio.org/wiki/index.php/IDigBio_Listservs The target audience includes natural history museum professionals that seek to improve and refine their workflows and methodologies for digitization and museum workflow initiatives.

Note: By completing this survey or questionnaire, you are consenting to be in this research study. Your participation is voluntary, and you can stop at any time.

Section 2: Survey Questions

Please respond to the following questions as applicably as possible.

Institution Type: ☐Collection Holder ☐Data Contributor ☐Data Aggregator

☐Other (Please Explain):

Email Address and Respondent Name (if willing to clarify and/or respond to further inquiries by an interview via Zoom, phone, or further email responses):

1 What software are you currently using to clean and edit biodiversity data?

☐OpenRefine

☐Easy Data Transform

☐Trifacta

☐Data Wrangler

☐ZoomInfo OperationsOS

☐DemandTools

☐DataCleaner

☐Melissa Data Cleansing

☐Other (Please Explain:)

1.1 On a scale of 1 to 5, how satisfied are you with using your specific software?

Not Satisfied Extremely Satisfied

☐1 ☐2 ☐3 ☐4 ☐5

1.2 What improvements do you desire the most out of your current software system based on what you are using? If you have no desire/ideas for improvements, choose N/A

☐Better reporting systems (e.g., integrated software to reduce redundant data entry)

☐Error catching mechanisms (programs to correct hand made errors)

☐Improved system layout/format (easier usability and/or personalized workflows)

☐Improved data sharing mechanisms (sharing online or with other institutions easier)

☐Other (please explain):

☐N/A

2 What metadata schemas are currently being used at the institution (check all that apply)?

☐Dublin Core (DC)

☐Metadata Object Description Schema (MODS)

☐Darwin Core

☐Integrated Taxonomic Information System (ITIS)

☐Directory Interchange Format (DIF)

☐Ecological Metadata Language (EML)

☐Geospatial Interoperability Framework (GIF)

☐Biological Collection Access Service (BioCASe/ABCD)

☐Other (please explain):

2.1 How useful do you perceive the metadata schemas to be (can be based on one or multiple schemas being used)?

Not Helpful Extremely Helpful

☐1 ☐2 ☐3 ☐4 ☐5

2.2 If below a 4, what are the prevailing challenges with these schemas? (2-3 sentences)

3 What specific controlled vocabularies have you been using to enforce data standards? (Check all that apply)

☐Subject heading lists

☐Controlled lists

☐Synonym ring lists

☐Authority files

☐Taxonomies

☐Alphanumeric classifications

☐Thesauri

☐Ontologies

☐Folksonomies

☐Other (please explain):

3.1 Explain why these forms of controlled vocabularies have been the most helpful (2-3 sentences)?

4 What quality assurance measures (using the technology/software) have been put in place to ensure data consistency across the museum staff and/or volunteers? (Check all that apply)

☐Accuracy

☐Completeness

☐Consistency

☐Validity

☐Uniqueness (not duplications or overlapping of values)

☐Timeliness (e.g., data available when required)

☐Other (Please Explain):

4.1 Which quality assurances measures do you believe are the most important for ensuring data standards and why? (2-3 sentences)

5 Does your institution use volunteers to fulfill any of this type of work? If so, are there specific challenges with enforcing data standards and quality assurance measures with volunteers? (Check all that apply)

☐Enforcing data standards on volunteers (quality assurance checks)

☐Lack of motivation/drive from volunteers

☐Lack of communication between both staff and Volunteers

☐Other (please explain):

☐Institution does not utilize volunteers

5.1 How (or would) have you addressed these challenges? (If volunteers are used) (2-3 sentences)

6 Which of the following have been the most helpful in reassessing and improving data management standards? (Check all that apply)

☐Conferences

☐Workshops

☐Meetings (staff only)

☐Building (and refining) workflow templates

☐Stakeholder meetings (e.g., staff discussions with financial members and/or volunteers)

☐Training sessions

☐Other (please explain):

6.1 Explain why these methods you selected have been the most effective (2-3 sentences)

7 Have you been using (or are considering using) any of the following to improve these resource/training standards? (Check all that apply or choose N/A)

☐Asking for or researching different institutions for advice

☐Employee one-on-one meetings for transparency improvement measures

☐Incorporating surveys post meetings for feedback

☐Incentivize or reward members for work/accomplishments during these meetings/training sessions

☐Other (Please Explain):

☐N/A

7.1 If you have **not** selected N/A, name two to three benefits these improvement measures have given to the data standard work environment.

Appendix B: Interview Questions

1 What is your role at your institution? What are your day-to-day tasks?

2 Does this institution have experience with commercial and/or open-source platforms? How would you compare them in strengths and weaknesses?

3 Were there any technical limitations or problems in utilizing the technology/software? What were the steps taken to fix or adjust the scope of the product to help with maintaining collection content?

4 If you could recommend revisions or upgrades to the software you're using, what would they be?

5 Can you give me an overview of a notable methodology the institution has used to help measure, validate, and improve the quality of data?

6 Concerning enforcing data standards, what challenges are you aware of from your experience? How would you recommend overcoming those challenges?

7 Can you discuss any challenging moments in projects where staff have had to reconvene and reassess how to maintain these data standards more efficiently? (e.g., go to more workshops, weekly check-ins, a reinvention of the project plan?)

8 Are there steps/improvements you and the staff are considering implementing to address these challenges further?

Specific Prompt Questions (To Be Brought up Based on Interview Answers - If Applicable)

1 Does the institution utilize the Darwin Core Standard? If so, have you revised/modified it in any way to suit your data preferences? Have there been instances where there was conflict concerning definitions and terms during meetings surrounding Darwin Core and how (or were) they resolved?

2 Have you (or the institution) utilized transcription services? If so, do you create incentives or goal measures to keep volunteers or staff invested in the data transcription? For context, the source below describes the Notes from Nature tool that emphasizes volunteer engagement and incentives.

Hill, A. et.al. (2012). The notes from Nature Tool for unlocking biodiversity records from Museum Records through citizen science. *ZooKeys*, *209*, 224–225, 228. https://doi.org/10.3897/zookeys.209.3472

3 Many literature pieces have cited collaboration frameworks between citizens and professionals to facilitate species matching, name parsing and other biodiversity data quality measures. Have you created workshops or training sessions (or related types of programming) that combine the public and professionals together to enhance data validation?

4 Literature cites governance structures and financial stakeholders as a common challenge in growing biodiversity data. Have there been any approaches done to help these stakeholders “see” the value of the data from your perspective?

5 Institutions have created workflow modules that are used as templates and guidance for new institutions pertaining to data quality. Has the institution created similar workflow modules? If so, what were the challenges in making them?

6 An article from Palaeontologia discussed a data entry system where a piece of digitization is completed by a qualified supervisor prior to upload, emphasizing a preventative or proactive measure of data quality. Has the institution considered or utilized preventative or proactive measures to ensure there is a template for good digital records?