



National Digital Stewardship Residency | New York
Final Report
Vicky Steeves

Host: American Museum of Natural History

***Preservation of Scientific Research and Collection Datasets at the American
Museum of Natural History***

**NDSR FINAL REPORT:
AMERICAN MUSEUM OF NATURAL HISTORY**

Table of Contents

Description	Page 1
Overview	Page 1
Project Partners	Page 1
Project Execution	Page 3
Project Activities	Page 3
Project Alterations	Page 5
Project Accomplishments	Page 5
Outreach and Dissemination	Page 6
Analysis and Execution	Page 6
Project Results	Page 6
Next Steps	Page 8
Project Deliverable	Page 9

Description

Resident: Victoria Steeves

Project Title: Preservation of Scientific Research and Collection Datasets at the American Museum of Natural History

Overview

The project at the American Museum of Natural History resulted in a survey of the digital research and collections data generated in the Science departments. Science at AMNH incorporates the Research Library and five scientific divisions and their staff, including: researchers, curators, students, fellows, postdocs, associates, and collections staff. To properly assess the status and extent of digital data in Science, I created a structured interview guide and interviewed every Museum curator and other relevant science staff. These interviews centered on three main topics: storage, management, and preservation of digital data, focusing on research workflows, volume of data, management practices, and preservation as a basis to provide recommendations to the Museum on best steps moving forward in stewardship of that data.

The interview process occupied the first four months of my Residency. The next four months were devoted to the analysis of the survey results and research into industry standards for the storage, management, and preservation of scientific data and the development of reports on these topics. The last month was dedicated entirely to the final report, which included all previous reporting in the residency as well as final survey results and recommendations to the Museum.

An analysis of this nature has never been conducted at the Museum before, and served, as the first needs assessment survey for digital data. The results of this project are not only a final report which details the survey results, recommended best practices moving forward, and a comparison to industry standards to current trends in Science, but also several reports detailing projections of growth in data generated, an environmental scan of other natural history institutions and their solutions, and a digital asset retention recommendation plan. Additionally, the structured interview guide will be retained as a result of the project, along with the survey results in raw and structured formats.

These results were all achieved throughout the course of the Residency. The final report has been going through an ongoing editing process in the month of June to ensure complete compliance with Museum needs. An unintended result of the project has been a shift in the institutional culture in the Museum. I have noticed that from my first week at the AMNH to my last, there has been an increase in support for digital stewardship initiatives. This bodes well for the future of the incredibly important digital data created daily at the Museum.

Project Partners

Barbara Mathé, Museum Archivist and Head of Library Special Collections. Barbara was my day-to-day supervisor and oversaw the development of the survey and reviewed my analyses and reports. She helped plan every stage of the project with me, such as project materials and timeline. Barbara was also deeply committed to making sure I adapted to the Museum in every way: work environment expectations, organizational hierarchies, and the ins and outs of each division and department in Science and beyond. She ensured I knew as much as possible about the Museum to make my interviews successful. Barbara also accompanied me to the first of my interviews in order to provide feedback about the process including my demeanor and body language, possible question confusion, etc. Her continual mentorship surrounded not only my project, but also my

personal development as an LIS professional.

Tom Baione, Harold Boeschenstein Director of the AMNH Research Library: Tom was largely responsible for getting the resources I needed to complete the project. He also played a large role in opening the doors for advocacy work at the Museum by allowing me to sit in on meetings with Museum administration. Tom was also very involved in acclimating me into the work environment within the Library, and was an invaluable resource when navigating the Museum's many divisions and departments. In the beginning of the interview process, Tom accompanied me and provided great feedback as a third-party observer to my interviews. He was also a great reviewer for my reports, blog posts, presentations, etc. and always made time when I needed to meet with him.

Scott Schaefer (Ph.D.), Associate Dean of Science for Collections and Curator in the department of Ichthyology in the division of Vertebrate Zoology. Scott provided the necessary reinforcement within Science that gave this project validity with Science staff. By providing his seal of approval, he ensured that we got a higher level of cooperation when scheduling interviews and speaking with staff. Scott was also a great resource when it came to dissemination of project results within the Museum. At the beginning of my Residency, both the project and I were formally introduced at the open session of the Science Senate, a monthly meeting of curators and science staff at the Museum. At the end of the Residency, I was able to present more fully the results of the interview process, which was wonderful because most of the people in the final Senate meeting were in my dataset. Scott was also great at playing Devil's advocate when I began developing the survey, forcing me to think more deeply about how to explain questions, which seem obvious to the LIS professional, but less obvious to others. It was great to have the independent review from someone outside the Library.

Library Staff, the AMNH Research Library: I was constantly supported by every Library staff member I met. After describing my project to each new staff member I was introduced to, they immediately understood and offered explicit support, usually saying, "This is so important. Let me know if I can do anything to help!" A few staff members in particular helped me make great headway in the project. Rebecca Morgan, CLIR Project Archivist, was an incredible resource as she had done previous needs assessment initiatives within Science for their physical archives. She had some contacts in Science that I utilized during my interviews and generally helped me understand the layout of each department in Science. For instance, in one department it was the collections manager who delves into digital data collection and organization, while their job description only mentioned the physical assets they deal with. Becca was able to give me that insider knowledge when she realized I had missed interviewing that staff member. Additionally, Jen Cwiok, the Library's Digital Lab Manager, was especially instrumental in understanding the Library's technical landscape while also bridging the gap in my understanding of the push and pull of digital data in and out of the Library.

I was also invited to participate in a working group created by Library staffers shortly after my arrival to the Museum, called the "Digital Asset Working Group," inspired they told me by what my project represented at the Museum as the next concrete steps in institutional digital data management (currently a department by department effort). Being invited into this group meant a lot to me in terms of both bolstering my professional skills as well as integrating myself into the Museum. For other library staff members to think to include me meant quite a bit; it meant that they believed that I could help them in their own projects, which was the best compliment they could have paid me.

Science Staff, including my interviewees and others, weren't identified as project partners in the

grant proposal, but rather as my dataset. Interestingly enough, they were both. Each new staff member I interviewed or even just met in Science offered me a new perspective on how to approach the project from a purely institutional standpoint, and it was immensely helpful in forming my final recommendations. It was above and beyond anything I required of them. Not only that, but some of my staunchest defenders and advocates were rooted in Science. These staff members would promote the aims of the project within their own department, and in some cases, even beyond to other divisions. The up swell of support from Science really was the catalyst for making this project as successful as it was.

Information Technology Staff, AMNH IT, with special thanks to Michael Benedetto, Deputy CIO. While all the IT staff supported my needs (in terms of both software needed and bolstering my understanding of the technological landscape of the Museum and IT's position within the AMNH), my weekly meetings with Mike Benedetto were some of the most helpful of my project. Each week, I discussed my project trajectory and the results I had gotten from my interviews so far. In these meetings Mike not only allowed me to bounce ideas off of him for the project, but he also provided me with invaluable mentorship around project management, Museum technical landscape, and professional development. We had a free exchange of ideas, which was helpful in assessing my next steps at the Museum and beyond. He was also a great ally in outreach activities within the Museum, often advocating for digital preservation initiatives at high levels of Museum administration while citing my work.

Project Execution

Project Activities

Project Phase: Needs Assessment | September 2014-January 2015

Tasks:

- Create a semi-structured interview guide to direct interviews with relevant science staff.
- Meet with Information Technology staff members to gain an understanding of the technological landscape of the Museum.
- Begin an outreach campaign with Science staff to underscore importance of project.

Deliverables:

- A semi-structured interview guide for the assessment of AMNH digital assets.
- A draft of the detailed survey of AMNH digital assets.
- Meet with the IT department to get a baseline understanding for the technical landscape of AMNH.
- Interview the data creators and managers in the Science divisions to identify:
 - Data formats
 - Quantity of data
 - Type of software and hardware used to support data creation and management
 - Current storage locations (geographic, physical, and virtual)
 - backup schedules
 - How data is transformed/basic data workflows
 - Long-term usefulness of data

Project Phase: Analysis & Research | February 2015 - April 2015

February 2015 Tasks:

- Apply sociological methods for transforming qualitative interview data into quantitative

- data.
- Research current methodology on preservation of scientific databases.

February 2015 Deliverables:

- Compile the results from the interviews in one comprehensive document.
- A report on guidelines for evaluation of policies for:
 - appraisal for determining the length of retention of digital assets
 - data management plans for federally funded grant proposals
- A report that contextualizes AMNH scientific databases within the larger digital preservation community.

March-April 2015 Tasks:

- Research long-term preservation solutions for AMNH's scientific digital assets.

March-April 2015 Deliverable:

- A report comparing long-term preservation options for AMNH based on survey, including a cost estimate and five year projection.

Project Phase: Recommendations | May 2015

Tasks:

- Compile the final report as a resource for ongoing digital preservation planning at AMNH, comprised of:
 - previous reports, and
 - guidelines for local best practices for digital asset management and preservation.

Deliverable:

- Final Report

Project Alterations

All of the project deliverables were completed. There were only slight alterations to the project and only in terms of the timeline and scheduling. Interviewing close to 60 people and working around their schedules meant that there were some inevitable missed deadlines, and subsequent work had to be pushed back. For instance, I was supposed to have finished interviewing all the relevant staff members by the end of January. At that point, about 95% had been interviewed, with the other 5% missing due to fieldwork schedule, injury, or other timing conflicts. The last interview didn't occur until April 1st.

Project Accomplishments

My most valuable accomplishment from the Residency was the interview process and its development and completion. Previous to arriving at the Museum, I had never done any work analyzing qualitative data or performing a needs assessment. This position would have me doing those two jobs primarily. I drew heavily from outside resources, but in particular an asset from Purdue University called the "Digital Curation Toolkit." This included not only a semi-structured interview guide but also instruction on best practices and use cases. I relied on the instructional aspect of the Toolkit in my first interviews, and then began to feel comfortable as I continued through the needs assessment phase. The semi-structured interview guide that I developed underwent nine separate revisions throughout the cumulative interview process, largely in refining the wording or order of questions based on initial interviews. By the end of my term at the Museum,

I interviewed close to 60 staff members in Science including curators, curatorial associates, postdocs, scientific illustrators, database administrators, and more.

A tangential accomplishment was the build-up of excitement around this project at the AMNH. I have received nothing but support from the staff I've spoken to, with only a few voicing concerns related to privacy and access of their data. When I interviewed a new staff member, they often expressed happiness or relief that the Museum had taken the initiative in assessing and preserving their digital data. Many currently don't have the time or resources for proper storage, backup, and management of their digital data. This project gave them an opportunity to become more aware of the importance of preserving data and an opportunity to express what they'd like to see come from this project. Both have been useful for the interviewee in most cases, and helpful for me in writing my final recommendation.

The final report, which I wrote at the end of my Residency, was also an incredible accomplishment. At its second draft, it is a little over 100 pages long, including appendices. This is the cumulative effort of my entire Residency and includes all previous reports as well as the complete analyses of the data and recommendations for local changes that could be made at the Museum. This is my first needs assessment report and it feels significant, especially since this has been the beginning of my career.

Outreach and Dissemination Activities

The most notable outreach activity I have engaged in since becoming a resident has been conference attendance and participation. During my nine-month residency, I presented on either my project specifically or NDSR at large during the METRO Annual Conference, American Library Association Midwinter Conference, the Mid Atlantic Regional Archives Conference (MARAC), and the Preservation and Archiving Special Interest Group (PASIG) Conference, and attended code4Lib's annual conference. Additionally, I have presented at a meeting for the Art Libraries Society of North America, New York Chapter (ARLIS/NY) and participated in the Archive-It NY User Group meeting held in New York City.

I have also been digitally published, writing about my project and NDSR at large, which has been disseminated extensively through the use of social media and listserv outreach. My NDSR-NY cohort maintained [a blog collaboratively](#), with each resident about posting twice a month depending on the rotation. These posts centered on our projects directly or topics closely related to our projects, and gave the residents a great platform to update other information professionals on our projects' developments and other subjects of interest. Furthermore, each resident was published once on the Library of Congress's digital preservation blog, the SIGNAL, writing a project status update or in my case (since I was the first resident to be published on the SIGNAL) [a report on the value of NDSR](#) and programs like it. METRO, the supervising organization for NDSR-NY also published two articles in which a METRO staff member interviewed the five residents about our experience in the program. Lastly, I have also been published in the National Museum of Natural History's Field Book project blog as a guest writer, [reporting on my experience](#) with the importance of field books in the digital age.

With the LIS field so active on Twitter, I've also found great opportunities in social media to engage with other professionals about NDSR (with the #NDSR), digital stewardship, and even my project specifically, especially during conferences. The active conversations during conferences (using standardized conference hash tags to track them) have enabled me to network beyond my immediate circle and create professional relationships, some of which bore fruit for my project.

Being a part of the NDSR community has afforded me these incredible opportunities to showcase both my own work, the work of my cohort, and the program at large.

Analysis and Execution

Project Results

The AMNH has the unique opportunity to play steward to some of the most important scientific data of our time, however with most of it digital and with no institutional data management or digital preservation system, the data is constantly at risk. However, this was largely unknown and underrepresented at the Museum previous to my Residency. One measurable impact of this project has been the change in institutional culture surrounding digital preservation. At the beginning of my residency, I was hard pressed to find someone who acknowledged digital preservation as needing immediate attention, aside from collections managers within Science who worry about the status of their digital objects.

At the end of my residency, there was recognition amongst many that this was important. I'm hoping that this positive momentum coupled with my comprehensive needs assessment report will be used by the Museum to establish a sustainable digital preservation program, or at the very least begin to provide institutional support for data management and preservation initiatives. The final report has been requested by many Museum administrators, in and out of Science, and serves as the chief deliverable of the project, as it outlines all the results from the interview process. The report, totaling over 100 pages, is comprehensive in its approach, outlining the current state of digital data safety at the Museum, and strategies the Museum can employ to minimize risk.

The NDSR project at the American Museum of Natural History has been successfully completed; all project deliverables have been finished and data delivered. However, there were a bevy of administrative challenges at the Museum, most notably scheduling with relevant staff. This was problematic because of the incredible work ethic of AMNH scientists: they are constantly at work in the labs, teaching courses at the Museum's graduate school and elsewhere, collaborating with colleagues at other institutions, or out on fieldwork. Fitting into their schedule was difficult, and caused a slight problem in the original project timeline. I ended up finishing my last interview about three months after my scheduled deadline. This was anticipated, and so while there was not much that we as project partners could do about this, there was time scheduled in the grant proposal to allow for interviews to come in easily after its marked end.

One other significant challenge was answering the bias against librarians. A comment I've received several times during my interviews is "Why is this coming out of the library? Isn't this an IT problem?" While there has been great support for my project, there has been a consistent ignorance of library services, both existing and potential. Many staff members did not even conceive that the information professionals in the Research Library could assist them in their data needs, whether it is storage, management, and especially preservation. However this was a great opportunity for the AMNH Research Library to advocate for itself and its professionals. The questions above gave me a great platform to explain the role of information professionals in the digital age, and why the Research Library in particular is best suited to be the home of this type of digital data analysis. This was an unexpected challenge to overcome as well as a great opportunity to discuss how the Library can help Science.

This project has also been impactful for AMNH Information Technology in particular. During weekly

meetings in IT, it was often discussed how this project is a great way to bridge the communication gap between AMNH IT and Science, as well as the fact that this project also provides concrete reasons to expand the IT infrastructure at the Museum, making it easier for IT projects to get traction at higher levels of Museum administration. Although it was an unexpected consequence, I'm happy to contribute to this effort to grow IT and its relationships within the Museum.

I hope that I will be able to submit my compilation of research to the larger digital preservation community and contribute to the niche of natural science data management and preservation. This is to be determined, but I believe that it will be helpful to at least Museum LIS staff nationally. I hope that my work here will illuminate strategies other scientific research institutions can utilize to preserve their data. I am also looking forward to contributing to the many groups, subcommittees, and other professional organizations I met during conferences.

Personally, this project has helped me develop a wealth of professional skills that were previously underdeveloped. This was my first real "office job," as my work previous to this was mainly technological instruction and support positions, a lot of it self-directed and individual (in that I did not work with more than two other people). This was a great opportunity to learn how to integrate myself into an institution's culture. It was my first real foray into a professional setting and so I've developed a wealth of "soft skills" that I will carry forward into my next "office job," most notably project management and communication skills.

Additionally, I've learned qualitative data analysis and needs assessment planning and reporting, quantifiable skills which I had no experience in prior to the Residency. Through the interviewing of staff members over a five-month period, I was able to continually eliminate ineffective questions and refine my interviewing skills (including reading body language in myself and others and interpersonal skills) as I talked to curators with a range of personalities. Because of my background in computer science, learning the qualitative data analysis (QDA) software was not so difficult; however learning the principles and best practices behind QDA, and then applying them, was a skill that required constant practices and reapplication to understand fully and apply successfully.

The most impactful personal response to this project has been my career trajectory. Previous to the Residency, I had no real direction within LIS. I thought I would be better suited to special libraries with interesting collections, which is what drew me to the AMNH. Now after experiencing the challenges and opportunities in working with science data, I am captivated. I will most definitely be continuing on in LIS through working with the management, preservation, and reproducibility of scientific data. This Residency has given my career a definite direction, for which I am immensely grateful.

Next Steps

The American Museum of Natural History now has a document that outlines its needs in regard to digital preservation. I am hoping that this document will be used moving forward with a digital preservation program within the Museum that starts in Science and moves beyond to incorporate all the many important, institutional-wide digital assets. I'm hoping that the positive momentum surrounding digital preservation coupled with my comprehensive needs assessment report will be used at the Museum to establish a sustainable digital preservation program.

I will continue working on the digital preservation and reproducibility of scientific research and collections data at NYU Libraries and the NYU Center for Data Science in direct support of the Moore/Sloan Data Science Environment Team moving forward. I have accepted a position there as

a “Research Data Management and Reproducibility Librarian.” In this role, I will refine existing best practices for reproducible and open data, provide instructional and consultation services in Research Data Management (including how to best satisfy federal grant requirements and open data requirements) to faculty and advanced students, and I will be involved in the efforts to design a data repository and storage infrastructure for researchers. I will also be conducting ongoing assessment and monitoring of researcher needs, much like what I accomplished in the first half of my Residency. This is a perfect hands-on extension of the work I did at the AMNH--whereas at the Museum I performed a needs assessment and produced a document outlining recommendations for digital data stewardship, at NYU I will be working closely with researchers doing hands-on work with data management and reproducibility.

Project Deliverable

For my product to display from the project I have added my interview guide. The questions highlighted in grey represent the questions that I sent to interviewees ahead of time. Please find the guide below:

NDSR @ AMNH: Interview Guide

Name:

Date of Interview: Email:

Department: Phone Number:

BACKGROUND *A place to write in a bit about the interviewee’s background, their primary research focus, and anything else of interest.*

INTRODUCTION

Would it be ok for me to record our session as a form of audio note taking?

- *I will be the only one listening to it, it not shared with anyone*
- *It will be destroyed after I am done with it*
- *WHY?*
 - *fill in gaps with my own note-taking to ensure accuracy*
- *I will send you all notes for you to verify and amend as needed*

The goal of this interview is to better understand your data storage and curation needs. Overall, it will take up to an hour of your time. Using the results from these interviews, I will write a report that I hope will inform the way the Museum handles digital assets in Science, making it easier for you to store and access your data.

SURVEY QUESTIONS

1. **Part 1: Preliminary Data Description and Context** *Starting a Conversation*
 1. I know that your research centers on XYZ, but can you tell me what you are doing in a little more detail? Example: multiple projects, routine operations, etc. *[to provide some context for the data we will be discussing]*
2. **Part 2: Data Lifecycle** *I’m going to ask you a couple of questions about the lifecycle of your data, just to give me an overview of the range of data your work generates. [This will help me understand the transformations your data go through, which affects storage and management.]*
 1. If you could describe data creation as a series of stages, what would those be?

For example: raw data to refined data to analyzed data to final data.

1. For each stage:
 1. Do you know how much data is generated? How much?
 2. Do you know what format are the data in? Which ones?
[May or may not be relevant, given answer to pre-interview questions]
 3. What processes are the data put through? Examples: run through statistical software, transformed through image processing.
 4. What tools are used to generate this data? Examples: digital imaging, CT scanners, statistical software, POY or Malign or other specialized software for a RDB, etc.
 1. Is it necessary for others to have this tool to make use of your data or does it stand on its own?
 1. Are there alternatives to these tools?
3. **Part 3: Data Storage** *These next questions deals with specifics on storage and resource allocation in regards to data storage. ~flag if they don't know the answer to the first 3 questions~*
 1. Do you know many GB or TB of data do you currently have stored? How many?
 2. Do you know how much data do you generate annually? How much?
 3. Do you know what are the file formats your data are in? Which ones?
 4. How is your data stored?
 1. Where is it physically located?
 5. What resources are available to you for data storage?
 6. Are you able to allocate resources (both institutional and otherwise) to cover your storage needs entirely?
 1. What resources would help the most to answer your data storage needs? Example: additional storage media, additional space on the IT server, additional staff to manage your data.
4. **Part 4: Data Management** *I'm familiar with the NSF data management requirements, but I was wondering...[important because it ensures authenticity and integrity, ensures data lasts as long as it needs to, compliant with grants]*
 1. Do your main funding sources require any data management?
 1. If yes: What are the requirements?
 2. Are there other ways that you currently manage your data?
 1. What software do you use to manage your data? *Some examples could be Microsoft Access or Windows drag and drop.*
 2. Do you have a data manager on staff?
 3. Do you make back-ups? Backups could be a copy you never look at again on a hard drive you update every month or so, or a folder of all your work in dropbox that you update once a week.
 1. How often?
 2. Where are they stored?
 4. Do you take security measures to protect your data? Example: encrypting a drive, making backups regularly, making restore points on your computer.
 5. Do you have version control for your data? If so, how?
5. **Part 5: Access** *These questions are meant to give me a better idea of what kind of storage you need moving forward, dealing specifically with access of data you put into storage.*

1. *[can you tell me]*How often do you need access or go back to:
 1. Raw data
 2. Refined/analyzed data
 3. Final production-ready data
 2. *[I want to hear about if you use repositories]* Is your data in a repository?
 1. If yes: is this based within the Museum or through an outside service?
 3. Is it important to you to have the ability to restrict access to datasets or certain information within datasets to authorized individuals at will?
- 6. Part 6: Data Description** *These next questions deal with access in more specific terms, of organization and description. [ensures that the data are accessible and understandable to users]*
1. How is your data organized? Do you have data in multiple places? Example: raw data on an external hard drive, analyzed data on a computer, publication materials on IT server.
 2. How is your data described? Are there standards, such as Darwin Core, that you use? Do you create keys for your data fields?
- 7. Part 7: Discovery** *These questions get into the specifics of use of your data and how much you want others to find your data.*
1. Do you know or can you imagine what kind of data could be most useful to others: raw data, refined/analyzed data, or final, production-ready data?
- 8. Part 8: Data Preservation**
- We want your research to be available in perpetuity, and these questions deal with that preservation aspect of data management. We want people to be able to use and reuse your data as the field continues to move forward. [digital objects and tools change quickly so it's important to make priorities, want your data to be safe (if AMNH blew up), EX: there are emulation software to get data off storage media--what if that was necessary for your data but we didn't know the tools you used to save your data?]*
1. Which stage in the data lifecycle contains the most important assets to preserve, manage, and maintain over time? Raw data, refined/analyzed data, production-ready data?
 2. Is there anything that needs to be preserved alongside your data to make it useful? Example: tools, keys, explanatory or descriptive documents.
 3. How long do you think your data will be useful or have value for you or others were it to be preserved?

CONCLUSION

Do you have any questions for me at this time? Thank you so much for taking the time to meet with me, I appreciate it! Best of luck with your work, and I look forward to seeing you again and hearing more about your research!