AMERICAN MUSEUM OF NATURAL HISTORY

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

---

Goal Summary

To obtain a broad overview of the extent and status of American Museum of Natural History (AMNH) digital assets pertaining to Science. A survey will be developed to measure and describe scientific digital assets resulting in a metric to predict ongoing data curation needs as a baseline to review options for long-term digital preservation. It will also identify existing practices and policies for integrated data access and management at the AMNH.

Specific Objectives

To develop and implement a survey of existing digital assets in AMNH Science Departments that include the Research Library and other administrative units administered by AMNH Science. Among the data to be gathered will be the current storage location, uses, and administrative management of the assets, the status of associated metadata, the software applications used to access the data, and the file format along with a metric for storage requirements and associated costs specific to that format, as well as projections for growth. The survey will also include notes about the present lifecycle of each digital asset and the elements needed for the development of an institutional appraisal policy to determine the duration of data retention and which data will be selected for long term preservation.

To research the issues surrounding the preservation of scientific databases like those used by AMNH Science, e.g., current work in searchable long-term preservation of scientific data through semantic web representations.

To chart a comparison of the various options for long-term digital preservation including cloud and collaborative possibilities coupled with broad estimates of present and anticipated costs.

All of the above will be combined into a report as a basis to review the status of AMNH scientific data that may be used as a model for similar research-based natural science museums.

Timeframe & Deliverables

Overall – 9 months

Months 1 through 4 — The resident will be introduced to the primary project team and to key personnel within the Museum. After initial planning discussions, the resident and mentors will meet with the IT department to determine the overall technological landscape. Initial group meetings will be held with members of each Scientific Department. The survey will be in development during this initial process and ready to be used, and modified as needed, during individual interviews with the AMNH curators and relevant scientific staff in charge of collections (approximately 50 individuals). In addition to identifying data formats, the amount of data in each format, and current storage and backup schedules, discussions will be held about how the data is used and how long it is useful, to distinguish working data from final data, to determine how to evaluate and choose data for long term digital preservation. These parameters will vary based on the processes of different scientific disciplines and individual scientists.

Deliverable: Draft of detailed survey of AMNH scientific digital assets, subject to revision as the project proceeds, and beyond. Guidelines for evaluation of policies for: appraisal for determining the length of retention of digital assets and data management plans for federally funded grant proposals.

Months 5 through 6 — Research on current methodologies regarding the preservation of scientific databases. Based on information and context gained from the initial survey and interviews, the resident will research and document current methodologies that address the management of scientific and collection databases.

Deliverable: A report that contextualizes AMNH scientific datasets within the larger digital preservation community. General and some specific recommendations will be made, including identification of issues that might require further evaluation.

Months 7 through 8 — A comparison of long-term preservation solutions. Based on the resident's accumulated knowledge of the institution and its priorities, s/he will review various options for long term digital preservation including cloud and collaborative opportunities, incorporating current research reports and developments in the field of digital preservation.

Deliverable: A comparison of possible long term preservation options for AMNH based on the digital assets survey which will include current cost estimates and a five year projection.

Month 9 — Final Report compiled and submitted.

Deliverable: The final report will be a compilation of all previous deliverables into a comprehensive document in a form, outlined as a resource for ongoing digital preservation planning at the AMNH including guidelines for local best practices for digital asset management and preservation. It should be noted that the structure of the report will be complete but allow for easy identification of areas where more research is needed, either institutionally or within the broader digital preservation community.

Resources Required

2 mentors (primary mentor, Barbara Mathé, MSLS, Museum Archivist and Head of Library Special Collections, and Scott Schaefer, Ph.D., Associate Dean of Science for Collections), 1 resident

Access to AMNH staff in Science Departments including library, archives and museum (LAM) collection staff and AMNH IT staff.

An office and a laptop computer will be provided in the Library for the resident.

Context

Science at the AMNH administratively incorporates 5 scientific divisions, their researchers, students, fellows, associates and collection staff. Collections are also included in the Research Library, the Center for Biodiversity & Conservation, the Sackler Institute for Comparative Genomics, which includes the Ambrose Monell Cryogenic Collection (AMCC). The Science Computer Cluster Facility is a major resource used by museum research scientists, postdoctoral fellows, graduate and undergraduate students whose work relies heavily on high-end capability computing in areas of biology, genomics, astrophysics, and anthropology. The Microscopic Imaging Facility contains state-of-the-art imaging instruments and image analysis software for use across departments. Other sections under Science administration include the Hayden Planetarium, the Southwestern Research Station, the Richard Gilder Graduate School, an innovative Ph.D. program in comparative biology, and Natural Science Collections Conservation which oversees the conservation of the biological specimens in collections and on exhibit. The data in these sections will overlap with—or be minimal as compared to— the Science divisions and computational and imaging facilities.

AMNH scientists are at the forefront of developing and utilizing cutting-edge approaches in computing paradigms to address problems of broad application in the biological and physical sciences. For instance, researchers in Invertebrate Zoology have developed and implemented phylogenetic algorithms used by scientists around the world. While those in Astrophysics, in collaborations with scientists world-wide, use high-resolution numerical simulation techniques in research and also in the Hayden Planetarium. Anthropological research employs an anthro informatic approach, utilizing computational and phylogenetic tools to build on databases of human kinship variants. The NSF recently funded a 10 year effort to create a shared portal for digital data and images of biological collections. iDigBio https://www.idigbio.org/ promises to integrate access to information about collections held across the country, but the preservation of assets in the collaboration is dependent upon its individual partners. AMNH is a federally-funded partner in this initiative. AMNH Science has few peers in the museum community in terms of the breadth and scale of its digital collections; however, it is similar to many research museums in lacking a unified strategy to manage its research data. Except for efforts at the Smithsonian, no other major museum has delved into this issue. At the same time, funders, including the NSF and NIH, are increasingly requiring researchers to produce data management plans (DMPs) that describe how the results of grant-funded research will be preserved. Research universities have responded with a variety of solutions but this project will articulate the challenges of research data management in a museum-specific environment.

Required Knowledge and Skills for Resident

This position will involve interactions with scientists across many disciplines. No individual can be expected to have a comprehensive understanding of these many and varied areas of research but an interest in science, intellectual curiosity, and the willingness to respond to the challenge of determining different individuals' methodologies and needs is essential for this project. Specific requirements include: A graduate degree in Library/Information Science, Archival Studies, Information Science or equivalent.

- Deep interest /understanding of the issues pertaining to management of digital resources
- Awareness of/ability to research current relevant efforts in digital preservation
- Strong organizational skills and the ability to shift focus from big picture infrastructure issues to the details of a unique database design
- Ability to communicate clearly and effectively in writing, meetings and interviews and to explain complex technical ideas in a simple and concise way to others with less technical knowledge.
- Ability to work with others as part of a team
- Facility with MS Office, particularly MS Excel

Preferred Knowledge or Experience

- Knowledge or experience in current and emerging best practices, tools, principles and standards for digital preservation and curation
- Familiarity with descriptive, technical and preservation metadata standard application and use
- Conceptual familiarity with XML encoding and Linked Open Data within the semantic web
- Understanding of database technologies and server environments
- Knowledge of survey and information gathering technologies
- Knowledge of digitization guidelines and parameters (color space, file formats, resolution
- Conceptual/practical knowledge concerning the cost environment of managing and preserving digital assets
- Experience in a digital library, archive or related heritage environment