

Corresponding Evidence:
The Use Of Museum Correspondence Files In Support Of Collection Records

Bernadette G. Callery
University of Pittsburgh, School of Information Sciences
Pittsburgh, Pennsylvania, USA
bcallery@sis.pitt.edu

Keywords: Museum Archives, Digital Curation, Administrative Archives, Collection Records, Collection Management Systems

Introduction

While the correspondence files of a natural history museum can be readily seen as valuable sources of information about the institution's own administrative history, they frequently contain information about the objects acquired by the museum. Collection information in museums is seldom centralized, as the organizational structure, and subsequent recordkeeping systems, tend to be divided along disciplinary lines, with individual departments frequently maintaining their own records. It is also the nature of natural history collections that the accession record, typically the initial record describing individual specimens or artifacts, is frequently incomplete at the point of creation and that information about the individual specimen or collection grows by accretion over time, as the collections are curated, used in research, or prepared for exhibition. In this intellectual environment, information about the identity of a given object and its circumstances of acquisition must be assembled from a variety of internal sources, carefully weighing the relative merits and authority of each source. Maintaining the understanding of these often fragmentary records and their interrelationships with other institutional records is a delicate matter, requiring familiarity with individual as well as institutional memory.

This paper will discuss the challenges of adding correspondence files to this fragile information ecosystem, using as a case study a collection of 30 letterbooks containing copies of the outgoing correspondence of the director's office of the Carnegie Museum of Natural History, Pittsburgh, Pennsylvania. These letterbooks cover the period of 1896-1913, which were the critical early years of the Museum's existence. The contents of

these letters include, among many other subjects, instructions and encouragement to field collectors in South America, Africa, and the American West, negotiations with donors for major gifts to the Museum, and specific construction details on the Museum's 1907 expansion building which was specifically created to appropriately house the growing number of dinosaurs and other paleontological specimens, insect, bird and mammal collections, and archaeological and anthropological artifacts. Even in the 1907 letterbook volume which is almost totally taken up with the many details of construction and exhibit installation associated with the opening of the new building, the then director, the noted entomologist W.J. Holland, had time to write to Richard Rathbun, Assistant Secretary of the Smithsonian, and Caspar Purdon Clark, director of the Metropolitan Museum of Art, on behalf of an unnamed local collector who was looking for someone to "arrange for him a collection of ancient time pieces which he has been accumulating for a number of years past."¹ Shortly thereafter in a letter to Clark,² Holland revealed that the donor was H. J. Heinz, Pittsburgh food purveyor and philanthropist, who gave his major collection of watches to the Museum later that year, indicating that Holland's attentions to him were repaid. On the other hand Holland could be quite short with potential donors whose collections he did not want for the Museum, as he indicated in his refusal of a bullet mold, "It is a waste of postage stamps on your part to offer the object which you say you have at the price that you ask for it."³

There are approximately 15,000 individually numbered items in the letterbook collection, with an average of 500 items per letterbook. The content is largely typescript, and appears to have been produced on thin paper by using single-sided carbon paper, one of the many wet or dry copying techniques so lovingly detailed in the 1999 work by Rhodes and Streeter, *Before Photocopying: The Art & History of Mechanical Copying, 1780-1938*.⁴ Each letter book is arranged chronologically and contains a handwritten name index of the correspondents, clearly constructed as the letterbook was assembled, as there is little alphabetical arrangement beyond that of the original letter in the index entries. There is no subject access to the content of the letters beyond what contextual knowledge an individual researcher might bring to a search of the name index. The inherent vice of both the paper used for the copies and the insertion technique of pasting individual tissue

sheets onto gummed stubs, which are then stab-sewn into the expansion folds of the binding structure, has resulted in a very unstable format, which has further degraded over time and inadequate housing. Not all the manuscript indexes are complete, missing parts of pages or entire sections of the index. Additionally, while access to the letterbooks is limited, both physically and intellectually, they tell only part of the story. Incoming correspondence is scattered across collections in the institutional archives as well as files maintained in the individual museum departments. A variety of additional record types, including the primary accession record, exist in multiple – and variant – versions – across the informational structure of the museum. Part of the challenge of managing museum archives is that collection records are always current, regardless of when they were created or amended, and information dealing with an individual object could include correspondence dating from the earliest years of the museum up through the present.

Networked museum collection management systems, particularly those which allow the capture of the form and content of such supporting, but often invisible, documentation as deeds of gift and collection permits, in addition to the fielded data derived from ledger books and other item-level cataloging tools, increasingly successfully mimic institutional memory. In discussing how institutional museum archives are used to bridge these gaps between the museum's various – and occasionally conflicting – recordkeeping systems, museum archivist Deborah Wythe specifically mentions the use that registrars make of institutional archives, noting that they “are able to supplement the basic object information in their files with more detailed information from records in the archives, documenting credit lines, donor information, and exhibition history.”⁵ Recognizing the tension between the institutional archives and the curatorial departments for responsibility for these institutional records, she acknowledges that “records relating to the collection are always permanent and logically a part of the museum archives, but many such records are permanently active and needed in the originating department on a daily basis.”⁶ The contents of the letterbooks, which represent the outgoing correspondence of the director's office, comprise a reasonably complete record of this phase of the ongoing activity of the institution at the highest administrative level.

Digitization as a strategy for access

One approach to simultaneously create access copies of the letterbooks and capture their content for potential integration with the Museum's other resources would be to digitize the text and then convert those page images to searchable text using Optical Character Recognition (OCR) software. While commercially available OCR packages are much touted in the business and records management communities as the perfect solution to the problems of management of legacy document collections, our initial experiments in applying OCR to the typescript letterbook content have not produced encouraging results. In this preliminary stage of our investigations, it is necessary to contain costs and therefore we have not purchased any of the high-end products. It is certainly possible that they would yield better results. However, the fact remains that due to the nature of the copying process that produced the letterbooks, individual characters in the text are not well-formed and there are considerable background artifacts such as print-through, smudges, and torn and soiled edges, all of which contributes to the difficulty in obtaining a clear image for the OCR process.

Given the poor condition of the letterbooks, some preservation reformatting will have to be done as the material cannot continue to be used safely in its present form. If the OCR is not successful and users will not have access to full-text searching of the content as a means of locating information on a particular object or event, one option would be to transcribe the existing hand-written index of correspondents using the existing information of the correspondent's name and the page number or letter number as a locator. However, this would not be entirely satisfactory, as a spot check of the completeness of the indexes to the letterbooks indicated that approximately 10% of the letters pasted into the letterbooks had no entries in the manuscript indexes. There is also the problem of inconsistency in the use of the name of an individual or the organization with which that individual is associated. On another project using the Carnegie letterbooks in order to determine whether or not the Museum had received free railroad transportation for the delivery of the paleontological material excavated in Wyoming and Utah between 1899 and 1923, it became clear that the manuscript index inconsistently used the name of individual railroad presidents, dispatchers and passenger agents as well

as various abbreviated forms of the corporate name of the railroads themselves, making it very difficult to locate all the letters sent to the Union Pacific Railroad. There are also problems with a direct transcription of the names included in the manuscript index, as the names are often abbreviated and occasionally unreadable, leading to inconsistencies or gaps in the created index. Another more labor-intensive approach would be to create a new index by examining each item of correspondence and capturing information on correspondent, the correspondent's organization, sender (since the letterbooks include letters sent by various assistants), date and letter number. Assuming that an authority file was created at the outset to regularize issues of variant names of individuals and organizations, a cumulative index would be an obvious benefit to searching the whole of the letterbook content.

Related Research Projects

In looking at a number of successful archival projects in natural history museum collections which have digitized correspondence or other collection records, it appears that while some have used various commercial database and OCR packages to help capture and provide access to existing printed, typescript or hand written content, it is more common that these have been augmented by substantial customized software and hardware applications. While many of these projects were intended to serve as models for future development, they have tended to be one-off, grant-funded research efforts. Projects typically address the issues of a particular resource or type of record, with little continuing work done at the conclusion of the funded project. While information on work flow, technical specifications and overall project management may exist in the project files, these projects are often dependent upon specific hardware or software as well as specific individuals, few of whom will remain with the institution beyond the term of the grant.

An examination of the 2007 special issue of the *International Journal on Document Analysis and Recognition* which deals with the analysis of historical documents reinforces these concerns of sustainability and generalizability. While many of the projects described present ingenious solutions to the myriad problems of integrating

legacy archival data into institutional information systems, they all seem to require specialized information science research environments in which to make them work. The editors of this theme issue identify five broad categories of challenges with the analysis of historical documents, beginning with digitization. The other challenges are dealing with the presence of artifacts of the page such as ink seepage and the subsequent opportunities for digital enhancement, understanding and utilizing the physical layout of the content as a means of analysis, recognizing the context in which the documents were created in order to extract significant elements of information, and word spotting, or the identification of shape-related features of significant keywords in the document and subsequent identification of other instances of that shape. The editors recognize the need to involve domain specialists, including archivists, in projects involving the analysis of historical documents, as it is important to “capture knowledge and semantic information from such experts as well as provide them with user-friendly means of specifying, interacting with, monitoring and validating the analysis process and its results.”⁷ The hybrid situation described in the paper dealing with the successful capture of museum specimen records which include both typescript and hand-written annotation combines a commercial OCR package with customized image analysis and text post-processing tools in order to generate database content is a tempting example of what can be done when sufficient resources are available for development.⁸

Diplodocus Archives

Two of the Carnegie Museum of Natural History’s correspondence collections had previously been digitized, and those digitized images converted to searchable text using optical character recognition (OCR) as part of the SmartWeb Exhibit project, funded by an Institute of Museum and Library Services grant in 2000-2002, which was designed to deliver information on demand to both onsite and online visitors. The Carnegie Museum of Natural History’s Diplodocus and Douglass Archives, a website hosted by Carnegie Mellon University, also in Pittsburgh, digitally captured correspondence, photographs and published popular and scientific articles documenting the Museum’s early paleontological discoveries, especially that of the Museum’s signature dinosaur, *Diplodocus carnegii*. Of particular interest to historians of paleontology is the

correspondence with the noted paleontological collector Earl Douglass, during the period 1894-1931, as his letters essentially serve as field notes. Not only do these letters provide extensive records of major discoveries, field conditions and collecting techniques, but they occasionally include drawings, such as that reproduced from the letter from Douglass to Holland on 6 November 1911. This drawing is the first illustration of the land on the border between Colorado and Utah that becomes the United States National Dinosaur Monument. For those interested in the creation and distribution of full-scale reproductions of the Carnegie Museum's specimen of *Diplodocus carnegii* to eleven of the world's museums, this correspondence collection includes correspondence and printed material dealing with the negotiations for and the installations of the eleven casts, beginning with the first presentation to the British Museum of Natural History in 1905 and ending with the gift of a replica to the Paleontological Museum in Munich in 1932. The only replica in South America was given to the Museo de la Plata in Buenos Aires and mounted under the direction of the Carnegie Museum's much-traveled director, W.J. Holland, in 1912.

Building on research developed by Carnegie Mellon University to digitize and provide access to the congressional papers of Pennsylvania Senator John Heinz, and known as HELIOS (Heinz Electronic Library Interactive Online System)⁹, the dinosaur archives project scanned two collections of correspondence, which included both typescript and handwritten material, during 2000-2002. The rationale for making the archival correspondence available directly to museum visitors was that increasingly limited labels describing the specimens on display made it difficult for visitors to gain any sense of the original scientific or cultural context of the object, as all they saw was its current placement within the artificial context of the Museum. One of the goals of the project was to use the correspondence in response to users' queries for more information on a particular exhibited specimen in the belief that the inclusion of "extensive correspondence between the museum's administrators, scientists and field collectors, the required field expense reports and the photographs [would] capture the museum's side of the story."¹⁰ However, the primary user access to the site in the system's browse mode mimicked the physical arrangement of the correspondence collection, *i.e.*, in folders, with

separate entries for each individual item within the folder. Groups of related material, such as sketches or newspaper clippings attached to an individual letters were maintained through the use of “bundles,” the term used to record the physical relatedness of the documents, and to allow higher resolution or the choice of individual file formats to better represent photographs, maps, field sketches, or other image-based material, but still retain the relationship of the parts to the whole document group. While OCR was attempted to provide access to the content of the correspondence, using an early version of TextBridge, it was not particularly successful. Extensive metadata about the individual items was supplied at the verification stage of the scanning, and transcriptions of the handwritten texts were included when available. Users could view the documents retrieved by their queries in either image or text view, the former being the scanned page image and the latter the display of the metadata and the OCR derived text.

Sitio Conte

A project which incorporated a wide variety of text and image material, including illustrated field notes, was the Sitio Conte website built by the University of Pennsylvania Archives. This project brought together the published and unpublished materials dealing with the University of Pennsylvania Museum’s 1940 expedition to Sitio Conte, Panama, providing extensive documentation of this Pre-Columbian cemetery dating to ca. AD 450-900. The online exhibition is a repurposing and extension of the popular exhibition, “River of Gold: Panamanian Treasures from Sitio Conte,” which began traveling to United States museums in 1996. This site allowed users to read the diary of J. Alden Mason, the director of the excavation, as well as the field notebooks, photographs and site plans prepared by Robert Merrill, the expedition’s surveyor, engineer and photographer. There was no keyword searching available to give direct subject access to specific passages in these texts. The archival material was supplemented by scanned images of corresponding published reports and contemporary newspaper clippings. Individual photographs as well as individual pages in the notebooks were included, with thumbnails available for preview and searching. Access was provided to an online database of all catalogued artifacts resulting from this expedition. However, this site

appears to have vanished from the University of Pennsylvania Museum's Archives site, illustrating yet another hazard of digital collections, that of impermanence.

American Museum of Natural History Congo Expedition, 1909-1915.

Released in 2002, this website brings together the field notes, diaries, photographs and drawings which documented the expedition to the Congo in 1909-1915 led by American Museum of Natural History mammalogist Herbert Lang, assisted by Columbia University undergraduate and budding ornithologist James Chapin. Supplemented with online access to the 160 scientific publications that incorporated these findings, sound recordings and video clips, this complex website presents not only a view of the Congo at the time of the expedition, but provides evidence of the continued impact of the expedition's collections on the zoological and anthropological research. An elaborate structure of hyperlinks connects searchable transcriptions of fieldnotes with the corresponding page images of the field notebooks and photographs of the corresponding specimen or artifact. Data from the fieldnotes and specimen catalogs are also searchable. In addition to the specimen photographs, there are 98 watercolor sketches by Chapin, over 2200 of the surviving 8000 photographs taken by Lang, and many images of the over 4000 anthropological objects cataloged into the collection. However, here again, while the content provided is exceptionally rich and the links relating specimens with their illustrations and documentation in field correspondence, collection records and published literature are extraordinarily valuable to the researcher, the information management system developed for this project has not been replicated elsewhere in the Museum and little documentation on the design and implementation process exists.

All three sites, each admirable in its own way, were one-off projects, and the technologies employed were not used elsewhere in the institutions to provide access to other collections. Lacking easily accessible technical specifications, their ability to influence future projects, particularly those under development in other museums, is limited.

Physical and intellectual aspects of the Carnegie Museum letterbook scanning project

In order to provide access to this material, the initial plan was to scan the portion of the Carnegie Museum letterbook collection that contains typescript material, provide access to the content by allowing users to search the text produced from the application of OCR software, and then display the corresponding page image in response to user's queries. There are many issues here, both technical and philosophical, not the least of which is the initial capture of the pages. Given that condition of the originals, which are on thin paper and bound in a variety of ways that makes the volumes difficult to open and use without damage, the original plan was to use an overhead or planetary scanner, the Minolta 5000, which the Carnegie Museum Library uses extensively to scan printed material from the Library's collection for interlibrary loan purposes. Initial experiments at scanning at 300 and 600 dpi produced files that included the background tone of the paper, which further complicated the success of the OCR. Some experiments were run on the sample scans by colleagues at the University of Pittsburgh Archives Service Center. Even by doing some limited manipulation of the resultant .tif file, the results were scarcely better than those resulting from the use of the Microsoft Document Imaging software that is part of the Microsoft Office XP suite. Ideally further experiments in OCR will be run on more robust OCR software such as the popular Omnipage and ABBYY FineReader to see if a more complete rendering of the text can be obtained.

Another decision that will have to be made is whether to leave the letterbook volumes intact or to disband them. Clearly, given the poor condition of the binding structures, even careful handling increases the risk of damage. However, if the pages are disbound, then considerable preliminary work will need to be done in order to clearly indicate their original order. The advantage of this careful examination of the individual items will result in a comprehensive statement of the completeness of the letterbooks, as some gaps in the pagination have already been noted by researchers. Simply using the existing numbering will not be sufficient, as some volumes have every page numbered and some volumes only have one number assigned for each multi-page letter.

Issues of authenticity with copies

Issues of authenticity, while not quite at the fever pitch of the debate today, were also present during the time of the popular use of the letterbooks. While there was less of a concern in telling the original from the copy than is presently the case with digitized documents, as the paper used for the copy was clearly different from that of the original, there was some concern with the enumeration of the characteristics of the ideal copy.

Rhodes and Streeter note that nineteenth century users of carbon copies objected to their use because “carbon copies were not necessarily exact duplicates of their originals. Only the original would have the sender’s signature on it. If corrections were made to the original after it was typed, they would have to be made on the copies as well, a tedious task that was sometimes neglected. Therefore, a carbon copy might not be an accurate representation of what had been sent to the recipient.”¹¹

Users of the copies also missed the visual attractions of the letterheads and the actual signature, as well as any annotations, changes or corrections on the letter itself. In the case of the Carnegie Museum letterbooks, it is not always clear just who the signer was, since the typescript closing of the letter merely gives the writer’s title, not his name. Curiously, a series of letters requesting gifts of representative economic botany material for a planned Gallery of Botany were marked as “signed “by the Custodian, Botanical Collections, but “approved” by the Director, perhaps indicating the relative responsibilities of the two. The identity of the individuals who would have signed the letters could probably be determined by those familiar with the Museum’s history, but the addition of those names as metadata in the record describing the individual letters would be advisable to assist more casual users if this content were to be made more widely available.

Another of the contemporary concerns with the creation of the copies of the letters, whether using one of the early press copy processes or using carbon paper, is an eerie precursor of the current discussions surrounding the capture and maintenance of electronic records, *i.e.* that the decision to make a copy had to be made at the time of the creation of the original. O’Toole notes that the technology of making the copy created an

artifact that differed in material and conventions of access and long-term storage, thus separating the logical and temporal sequence of correspondence files. “Filing and indexing the correspondence were also difficult: at a most basic level, it was necessary to file incoming and outgoing letters separately, since the former were loose sheets and the latter were pages of a bound copybook.”¹²

Conclusion

The value of correspondence files in natural history museums as supplements to the other more conscious forms of collection record is recognized, but not always utilized as the correspondence files are yet one more independently maintained body of records to be located, searched, and the information analyzed by the researcher. Even if the existence of these additional information sources is known, limited or non-existent access to specific content in the records may deter users unwilling to make a linear search through the entire collection. While the letterbook correspondence files, as a record group type, are typically arranged in chronological order, and typically include indexes of the names of correspondents built as part of the process of creating the files, users knowing the approximate time of an event, such as the announcement of a scientific discovery, or the name of an specific individual associated with a particular acquisition, may have better success in finding useful information. While linear search may still be required, at least there is a point identified in the continuum of the correspondence file from which to begin searching. Serendipitous digital discovery is also possible, provided references to the collection are harvested by the various web crawlers.

As with all recordkeeping systems, the principal challenges of identifying and then extracting useful information from correspondence files as a specific document type are the expected ones of identification of the specific access elements. The generally standardized form and structure of early twentieth century business correspondence allows for the identification of the correspondent and the date, elements most frequently used in associated indexes. Full-text search, providing that the OCR of the content has been successful and complete, increases the likelihood of discovery, although the hazards of variable terminologies, abbreviations and local jargon are well known. Mechanical

aids such as name indexes, archival finding guides, and item-level descriptions such as calendars of correspondence all help maintain the network of relationships that represents research in archival collections.

Perhaps the most compelling advantage of digitization is that these correspondence files will effectively be centralized – or their digital surrogates placed in virtual relationships with other similarly scattered collection records. The integration of this legacy data into the overall information infrastructure is a continuing obligation of cultural institutions as they migrate their paper-based record-keeping systems into the digital age.

WORKS CITED

Antonacopoulos, Apostolos and Andy C. Downton, “Editorial: Special Issue on the Analysis of Historical Documents.” *International Journal on Document Analysis and Recognition* 9(2-4), April 2007: 75-77.

Callery, Bernadette and Robert Thibadeau, “On Beyond Label Copy: Museum-Library Collaboration in the Development of a Smart Web Exhibit.” *Museums and the Web 2000*. URL: <http://archimuse.com/mw2000/papers/callery/callery.html>

Downton, Andy, Jingyu He and Simon Lucas. “User-configurable OCT Enhancement for Online Natural History Archives.” *International Journal on Document Analysis and Recognition*, 9(2-4), April 2007, 263-279.

Galloway, Edward A. and Gabrielle V. Michalek. “The Heinz Electronic Library Interactive Online System (HELIOS): Building a Digital Archive Using Imaging, OCR, and Natural Language Processing Technologies.” *The Public-Access Computer Systems Review* 6(4), 1995: 6-18. URL: <http://epress.lib.uh.edu/pr/v6/n4/gall6n4.html>

O’Toole, James M. “On the Idea of Uniqueness.” In *American Archival Studies*, edited by Randall Jimerson, 245-277. Chicago: Society of American Archivists, 2000.

Rhodes, Barbara J. and William Wells Streeter. *Before Photocopying: The Art & History of Mechanical Copying, 1780-1938: A Book in Two Parts*. New Castle, DE: Oak Knoll Books, 1999.

Wythe, Deborah. “The Museum Context.” In *Museum Archives: An Introduction*, edited by Deborah Wythe, 9-19. Chicago: Society of American Archivists, 2004.

WEBSITES REFERENCED

American Museum of Natural History Congo Expedition. URL: <http://diglib1.amnh.org/>

Sitio Conte, University of Pennsylvania Museum. Former URL:
<http://www.museum.upenn.edu/SitioConte/index.html>

SmartWeb: CMNH Diplodocus and Douglass Collection. Carnegie Mellon University Archives, Pittsburgh, Pennsylvania. URL: <http://diva.library.cmu.edu/CMNH/>

NOTES

¹ Carnegie Museum Letterbooks and Financial Ledgers, Carnegie Museum of Natural History Archives 2007-5, Letterbook 15A, 18 January 1907.

² Carnegie Museum Letterbooks and Financial Ledgers, Carnegie Museum of Natural History Archives 2007-5, Letterbook 15A, 24 January 1907.

³ Carnegie Museum Letterbooks and Financial Ledgers, Carnegie Museum of Natural History Archives 2007-5, Letterbook 15A, 11 January 1907.

⁴ Barbara J. Rhodes and William Wells Streeter, *Before Photocopying: The Art & History of Mechanical Copying, 1780-1938: A Book in Two Parts*. (New Castle, DE: Oak Knoll Books, 1999).

⁵ Deborah Wythe, "The Museum Context," in *Museum Archives: An Introduction*, ed. Deborah Wythe (Chicago: Society of American Archivists, 2004), 12.

⁶ Deborah Wythe, "The Museum Context," in *Museum Archives: An Introduction*, ed. Deborah Wythe (Chicago: Society of American Archivists, 2004), 14.

⁷ Apostolos Antonacopoulos and Andy C. Downton, "Editorial: Special issue on the analysis of historical documents." *International Journal on Document Analysis and Recognition*, 9(2-4), April 2007, 76.

⁸ Andy Downton, Jingyu He and Simon Lucas. "User-configurable OCR Enhancement for Online Natural History Archives." *International Journal on Document Analysis and Recognition*, 9(2-4), April 2007, 263-279.

⁹ Edward A. Galloway and Gabrielle V. Michalek, "The Heinz Electronic Library Interactive Online System (HELIOS): Building a Digital Archive Using Imaging, OCR, and Natural Language Processing Technologies." *The Public-Access Computer Systems Review* 6(4), 1995: 6-18.

¹⁰ Bernadette Callery and Robert Thibadeau, "On Beyond Label Copy: Museum-Library Collaboration in the Development of a Smart Web Exhibit." *Museums and the Web 2000*. URL: <http://archimuse.com/mw2000/papers/callery/callery.html>

¹¹ Barbara J. Rhodes and William Wells Streeter, *Before Photocopying: The Art & History of Mechanical Copying, 1780-1938: A Book in Two Parts*. (New Castle, DE: Oak Knoll Books, 1999), 128.

¹² James O'Toole, "On the idea of uniqueness," in *American Archival Studies*, ed. Randall Jimerson, (Chicago: Society of American Archivists, 2000), 261.