Historical Considerations in Biodiversity Informatics

Michael Khoo\(^1\), Gary Rosenberg\(^2\)
\(^1\)College of Computing and Informatics, Drexel University
\(^2\)Academy of Natural Sciences of Drexel University

**Abstract**

Historical collections of biological specimens are potentially rich sources of data for contemporary researchers. However, many technical issues have to be addressed in order to make these collections widely available. This paper reports on a qualitative study of historical and current data practices at the Academy of Natural Sciences, Philadelphia, which is seeking wider understanding of the historical dimensions of specimen metadata, in order to support migration to more global standards. A detailed case study of a single specimen shows how that specimen has been described in multiple ways and in multiple locations within the Academy, and the historically complex nature of the data and metadata contained in these descriptions.

**Keywords:** biodiversity informatics, communities of practice, darwincore, digitization, malacology, metadata, specimen labels

**Citation:** Editor will add citation with page numbers in proceedings and DOI.

**Copyright:** Copyright is held by the author(s).

**Acknowledgements:** The authors gratefully acknowledge support from the [redacted]

**Contact:** Editor will add e-mail address.

1 Introduction

Historical collections of biological specimens are potentially rich sources of data for contemporary researchers (e.g. Moritz, 2002; Beach & Ibarra, 2009; Heidorn, 2011). Once digitized, new metadata can be inferred from existing metadata, such as extracting geolocation metadata from the name of a collector and the date a specimen was collected. The Networked Integrated Biocollections Alliance’s (2010) *Strategic Plan for Establishing a Network Integrated Collections Alliance* envisages biodiversity collection digitization in the United States as supporting understanding of “the biodiversity dimensions and societal consequences of climate change, species invasions, natural disasters, the spread of disease vectors and agricultural pests, and other environmental issues.” A number of initiatives, such as iDigBio (http://www.idigbio.org/), are aimed at making biodiversity collections digitally accessible.

The digitization of biodiversity collections includes the imaging of specimens, the imaging and OCR of associated descriptive artifacts (labels, etc.), and the mapping of heterogeneous historical metadata to current metadata standards. Once digitized, individual specimen metadata can be added to departmental collections, then into institutional catalogs, and then crosswalked to global standards such as Darwin Core (TDWG, 2013) and databases such as the Global Biodiversity Information Facility (GBIF, 2014). These processes appear to be relatively straightforward if detail-oriented activities. In the reality of historical physical collections, however, a number of barriers are acknowledged, such as the heterogeneity of the original assets and their descriptions. Digitization involves addressing a wide range of specimen types and preservation techniques (dry specimens, alcohol-preserved specimens, microscope slides, etc.). It also involves capturing and transcribing a variety of descriptions (NIBA, 2010; Paul & Heidorn, 2013). Forms of description include labels next to or attached to dry specimens, labels on or in jars of alcohol-preserved specimens, numbers inked onto specimens, index cards in card catalogues, and entries in ledger books. These descriptions can vary widely. Old labels are often formatted in non-standard ways, with obscure handwriting, idiosyncratic formatting and abbreviations, and so on, all of which present barriers to extracting useful metadata. In addition, cataloging practices can vary between different departments within an institution, and between different institutions. In the case of older institutions, collections may have been founded on donations by early collectors, and their cataloging conventions may have set departmental standards which have subsequently endured.

Scaling digitization and description workflows to scales capable of processing and normalizing billions of specimens and records is a considerable challenge (Blagoderov & Smith, 2012; c.f Makris et al., 2012). The issues are not insurmountable, particularly with careful physical and digital curation, but at the same time there is a tension between the detailed work required to capture data in early descriptions, and the large number of historical specimens requiring digitization (Heidorn, 2011).
2 A multidisciplinary approach to biodiversity informatics

The research reported in this paper is part of a qualitative and ethnographic study of historical and current data practices at the Academy of Natural Sciences, Philadelphia (http://www.ansp.org/). One goal of the study is to gain a better understanding of data practices at the Academy, in order to support tools for wider networked data access for evolutionary biology, biodiversity, and other researchers. To gain this better understanding, the research is studying historical and current factors underlying data and metadata practices at the Academy. Among the research questions being addressed are:

- How are specimens and descriptions historically constituted and ordered?
- How can historical forms of metadata be mined in order to generate augmented metadata?

The theoretical component of the research is informed by Lave and Wenger’s theory of Communities of Practice (CoPs). CoPs are groups of people “who engage in a process of collective learning in a shared domain of human endeavor” (Wenger, 2013). Knowledge in CoPs resides with experienced community members, who share knowledge with novices when inducting them into the community. As this knowledge can become taken-for-granted by members and hard to articulate, it is often acquired over time through the medium of practice (Wenger, 1998; Lave & Wenger, 1991). CoP knowledge emerges over time within a duality of participation and reification, between participatory social practices, and the reified informational artifacts that both record and shape such practices: “[Reification] always rests on participation: what is said, represented, or otherwise brought into focus always assumes a history of participation as a context for its interpretation. In turn, participation always organizes itself around reification because it always involves artifacts, words and concepts that allow it to proceed.”

In this study, departments at the Academy of Natural Sciences of Drexel University are being studied as CoPs. Each department is being considered as a locale of historically constituted data practices. Current data practices and their historical antecedents are being analyzed, partly to support the mapping of this historical knowledge to wider information standards and infrastructures, such as Darwin Core and GBIF. One a priori assumption of the research is that as different departments within the Academy have different histories, they will also have developed different data practices, which will have to be mapped to common institutional metadata standards.

3 Data and metadata at the Academy

The 200 year-old Academy of Natural Sciences houses globally significant collections of 18 million biological specimens. The Academy has eight departments, each with multiple internal collections, only some of which have been digitized. Current data and metadata in the Academy includes descriptions of specimens (molluscs, birds, moths, fossils, etc.) in terms of taxonomy and date and place of collection, as well as descriptions of who collected, donated and identified the specimen, and when these actions occurred. While it would be useful for researchers to query an institution-wide database in terms such as: “I am interested in specimens of species S, collected in location L, between dates D1 and D2,” for a variety of reasons, such a search is currently often not possible. Data standards (e.g. for time and geolocation) vary between departments, and particularly with older records, data and metadata may be incomplete or missing, inconsistently formatted, or otherwise insufficient. These data issues make searches less efficient, and data mining in the Academy’s collections less effective. Ways are therefore being explored to enhance search and discovery processes. One option is to augment existing records with data and metadata from other sources. For instance, upper limits for dates of collection can be inferred from donation records and imprecise localities can be fleshed out from field notes. This approach faces its own challenges. The Academy’s departments each have their own individual catalogs and internal collections, and records can take multiple forms dating back to the nineteenth century, including labels on shelves, jars, and boxes, ledger books, and correspondence in archives. In order to usefully integrate these data, it is necessary to understand how they may be standardized and synthesized, and part of the research is therefore analyzing how historical and contemporary data have been and are created in each department.

4 Methodological approach

The research is following an action research approach. Action research studies and shapes changes in institutions through theory building, intervention and action-taking, with research outcomes mutually constituted over time in interactions between researchers and research subjects (Baskerville & Pries-Heje, 1999). The initial field site at the Academy is the Department of Malacology, where fieldwork is building a case study of historical and current data practices, which in turn will help to build data models,
and inform subsequent interviews and studies in other departments. The research data methods include interviews and ethnographic observations, as well as archival work, and the data include interview transcripts, photographs, field notes, and historical documents. These data are being analyzed empirically and inductively, following a grounded theoretical approach. From the point of view of ethnographic action research, this results in the production of (a) ongoing interim practical findings for the Department of Malacology, and (b) theoretical refinements for models of Communities of Practice. As an action research project, these findings emerge throughout the research cycle and are fed back into the field site on an ongoing basis. Some of these emergent findings are reported in the following sections.

5 Case study: Specimen 4295

Initial research is focusing on building detailed case studies of individual historical specimens. This section focuses on what might be called a biography of a single specimen from the Malacology Department’s collections. It illustrates the how a single specimen is described at different times in history, the changing characteristics of departmental data practices over time, and the possibilities for augmenting and enriching existing metadata. This particular specimen is number 4295 in the Department’s collections. It is a land snail, *Pleurodonte lucerna sublucerna*. *Lucerna sublucerna* is found in Jamaica, and one of the physical specimens held by the Department is also the type specimen for this particular species.¹

The current record and description of *Lucerna sublucerna* is available to outside researchers via the department’s web site (Figure 1). This is the record that researchers see initially if they query the department’s database. The Web descriptions are generated from a MySQL database, which in turn has selectively imported fields from a more comprehensive FileMaker Pro (FMP) database maintained in the Department itself (Figure 2). The FMP database contains additional fields, and has been developed offline within the Department to serve as a comprehensive aggregator for all existing metadata in the Department, and as a platform to support the development of wider Web access. The FMP record represents the most current description that the Department has of 4295. However, 4295 has not always been identified in this way, and an examination of other paper records in the department tells a more complex story.

Most of the paper records are collocated with the specimen itself. Both 4295, and a series of labels, are stored in a small card box in a tray in an archival cabinet (Figure 3). These labels represent a series of views of 4295 over the past 150 years. The oldest label (Figure 4) was created in the 1870s or 1880s, not by the Academy, but by the malacologist Albert Dod Brown (1841-1886), whose shell collection was donated to the Department some time in the 1880s. Compared with modern-day descriptions, Brown’s labels are relatively simple, and focus on unambiguous identification. This reflects scientific practices in the latter half of the nineteenth century, aimed at the classification of what were then often thought to be immutable species (Farber, 2000). Other descriptive elements, such as the date of collection, are not supplied, and the location of collection is simply ‘Jamaica.’ Brown assigned catalog numbers to his specimens, and his numbering scheme was subsequently adopted by the department for its own collection (the number assigned by Brown, 4295, is still present in the FMP database). While 4295 was cataloged and labeled by Brown, it was not however collected by him; two small initials added by Brown to the top left-hand corner of the label – “T. B.” – indicate that Brown acquired the shell from another nineteenth century malacologist, Thomas Bland (1809-1885), who collected in Jamaica in the 1860s. Brown’s identifications on this label are complex. Brown seems to have identified it originally as *H[elix] fuscolabris* Ad. (‘Ad.’ here refers to Charles Baker Adams, 1814-1853, another early collector and publisher of shells.) However, the *fuscolabris* is crossed out, and Brown provides two further possible identifications, *Helix* acuta ? var, and *Helix* lucerna ? var., indicating some uncertainty in the identification.

The second and third labels in the box (Figures 5 and 6), similar to each other, represent the first labels created by the Department. They are display labels, and include a blank patch of card where the specimen would have been glued for display, and handwritten annotations. At one point, these labels were one piece of card, but as specimen display and storage practices changed at the Academy, they were cut into pieces, to afford storage in the small specimen box. The bottom half of the second label

---

¹ “In taxonomy, the science of identifying, naming, and classifying species, the primary type specimen (or sometimes a series of specimens) serves as the scientific name-bearing representative for any animal or plant species. A secondary type specimen is a specimen of the type series other than a primary type. A primary type is the objective standard of reference for the identification and naming of species. Type specimens are important to scientists that study the classification of organisms and to all studies of comparative biology.” (National Museum of Natural History, 2008).
(Figure 5) omits Brown’s identification of *H. fuscolabris* from the previous label, and instead makes a tentative identification of *H. acuta* ? var. *lucern* or *H. lucerna* ? var. The top half of the label is in a different hand and provides a further new identification as *H. Acuta Lam var sublucerna*. This latter handwriting and identification is from Henry Pilsbry (1862-1957), who was appointed Conservator of the Conchology Section at the Academy in 1888. This label also retains the catalog number from the original label from Brown (4295).

*H. acuta* now seems like a positive identification, and this information is transferred to a newer display card (Figure 6). This card is on different cardstock, and the top half is now missing. At this point, any previous possible identifications now disappear from labels, along with any reference to Thomas Bland as the original collector. The data retained include the catalog number, Brown as the donating collection, and Jamaica as the collection location. Note also that, vertically on the right-hand outside edge of the label, it reads: EDP 26086, a reference to the Electronic Digitization Program. The EDP created a computerization record and allocated a unique serial number to a database. Early EDP work involved manual copying data from various labels onto data entry sheets, which had fields based on a cataloging scheme developed by the Smithsonian. The data sheets were then sent to the Smithsonian for actual data entry These data sheets are no longer available, having been discarded by the department.

The next description of 4295 is supplied not by a label, but by the acquisition ledgers for the Department, housed in the Department’s library in a separate room (Figure 7). The ledgers have ruled pages divided into a number of headed columns. In one ledger, 4295 is identified as *pleurodonte sublucerna P*, (where the ‘P’ stands for Pilsbry, who named it), and any previous identifications are now omitted. The ledger confirms the catalog number (4295), the location (Jamaica), the collector (‘T. B.’, or Thomas Bland), and the donor (Brown). A total of four specimens of *pleurodonte sublucerna P* are recorded in the ledger (something that was not obvious from the previous paper labels). Note that this ledger entry was created on “VII 27 1923” (July 27, 1923), approximately forty years after the shells were originally acquired by the Department (there is no recorded reason for this gap, but one possibility is that the shell was on loan to another institution in the intervening years).

No new labels were created until 1983, when the four specimens in the box were split out into a single specimen (a lectotype), and a lot of three specimens (paralectotypes). One shell (4295) was designated the lectotype and placed in its own box, with a new label (Figure 7a). This label includes the catalog number (4295), and notes: “3 paralectotypes split out, recataloged as ANSP 356836,” as well as the person who made the split (Det. A. Bogan) and the date of the split (12 May 1983). The three paralectotypes are kept in an adjacent box, which has a further label in the bottom (Figure 7b). This latter label, for the paralectotypes, now includes a new catalog number, 356836, and also points back to 4295 as the lot from which the paralectotypes were obtained. This label also includes an EDP (Electronic Data Processing) number, 26087, one higher than the EDP number in Figure 6 (EDP 26086). This suggests that these two sequential numbers were added after the paralectotypes were split out. The three paralectotypes themselves also have Department and Academy catalog numbers inked on them (Figure 8); these latter specimens also have red dots, which indicate ‘type specimens,’ although these dots are no longer added, as the ink is not archival.

The final label in the box is small and green, and has ‘4295’ written on it (Figure 10). This label refers to a digital imaging project.

6 Discussion

The brief and incomplete history of 4295 illustrates some of the historical and logistical issues associated with historical record digitization.

First, the labels for 4295 contain information that was relevant to the label creator at the time. For instance Brown, who had an extensive private collection, created a series of labels that both reflected and also shaped the (often amateur) descriptive practices of the mid-nineteenth century naturalists, aimed at placing individual types within relatively stable hierarchies of species and genera. At the same time, precise acquisition data (place, time) is not provided. Pilsbry’s labels continue the classificatory project, and also, in the display card format, reflect the growth and institutionalization of natural history in public museums. The early digitization of records reflect attempts to gather all existing data, as well as to provide a technology that can be used to manipulate these data. (It should be noted that, at least in the Department of Malacology, these digitization efforts have been the initiative of individuals over time.) The current Filemaker Pro database reflects concerns with developing a comprehensive database management system. Here, more modern concerns with detailed specimen provenance (such as date and locality of collection) come into play.
Second, there is some missing information. For instance, Brown acquired his collection from Bland, but we do not have any evidence of Bland’s labels (although it is possible that these might still exist somewhere). Again, some of the misidentifications on the Brown and Pilsbry labels did not make it into the digital catalog record, and these are potentially useful historical data, that could be triangulated with other nineteenth century publications. Again, as noted, the work sheets from the earlier digitization projects are also missing. All of these as well other sources could provide useful provenance metadata to augment existing descriptions and records.

So far in the research, several broad inter-related practical and theoretical themes have been identified. In practical terms, 4295 is described not just by one label, but by a sequence of labels and ledger entries over time. These descriptions evidence the contingent nature of the data that are of interest to researchers at the time. The Department has developed tools and workflows to capture this distributed metadata, and while much has been carried through to the electronic records, some has been dropped; for instance, it was realized in the course of this research that ’T. Bland’ needed to be added as ‘Collector’ back to the electronic record. Information such this is useful; archival work might be able to establish more precise details of Bland’s collecting activities, and this information could be used to ‘backfill’ catalog records where Bland is listed as the collector. In another example, the early identifications of 4295 that were crossed out have not been preserved; these earlier identifications could also be useful, as they could point to possible identifications reported under different names in the literature of the time. The case also confirms the complex issues involved in digitizing specimen labels, including handwriting recognition (and note also that some of the labels have names that have been crossed out). Finally, while it is not a specific component of this research, it is obvious that any technological attempt to automate the process of digitally imaging the specimen and its labels would face severe challenges. As noted in the introduction, this complexity has implications for data normalization and accessibility initiatives. Currently, initial research inquiries may be submitted to the department’s web site, but more detailed inquiries may necessitate contacting department staff, who have ‘local knowledge’ of such factors as the timeline of early pioneers and collectors, their contributions to the department, the ways in which they formatted their labels, their handwriting styles, the location of information such as ledgers and card indexes, cryptic abbreviations and annotations (such as “T. B.” for Thomas Bland), and how to navigate through archives of historic publications in the field. In these terms, the practice in the department appears to include not just these ‘facts,’ but how also to fit them together into field of knowledge that can be drawn on with facility and communicated to non-experts. In this sense they resemble Nardi’s description of librarians as keystone species in information ecologies who translate external inquiries into queries that can be mapped to the department’s increasingly rich data and information assets (Nardi & O’Day, 2000).

In theoretical terms, the research has led to a more nuanced understanding of how CoPs may emerge and evolve in a scientific discipline. So far, the outlines of at least three general historical CoPs have been identified in the history of the department, each of which has a distinctive participation-reification dynamic that generates different types of research and description practices. First, there is the earlier nineteenth century community of amateur natural historians; second, there is the emerging institutionalization, professionalization, and public nature of natural history in the second half of the nineteenth century; and third, there is adoption of database and digitization technologies in the later twentieth century. Each CoP is distinct in terms of practices and artifacts, with later CoPs also growing out of earlier CoPs. In theoretical terms, this offers evidence of how CoPs emerge over time, with new CoPs emerging from within the general milieu of existing CoPs (for instance, disciplines emerging within natural history, and specializations emerging within disciplines). It also offers some insights into the complexity of these processes, suggesting that CoPs are complicated synchronic and diachronic phenomena.

The implications of the study so far for wider data integration within the Academy will be investigated in further work. Further topics being investigated include further study in the department, comparative studies of descriptive practices in other departments, pilot archival work to see what can be gleaned about the Brown and Bland collections from documents at the Academy, and a general assessment of the time and workflow considerations associated with augmenting existing specimen metadata with various types of new provenance metadata.

7 Conclusion

Historical biodiversity collections could be important resources for researchers in a range of disciplines, if their holdings could be digitized and described. This case study has illustrated, with a case study, some of the complexities involved in digitizing biodiversity collections, including heterogeneous labels and diverse formatting. The paper has also suggested some new possibilities for metadata augmentation. In addition, the paper has introduced a CoP approach, to identify and describe some of the complex
knowledge that staff in the Department use to make sense of the collections, and the study yielded useful insights into the historical complexity of scientific CoPs. Overall, the study confirms some of the tensions between the work required to capture historical data, and the large number of specimens requiring digitization and description. Further work at the Academy will expand this study to other Departments, in order to develop comparative case studies.

References


Figure 1. Descriptions of specimen 4295, available on the Web.

Top:

Bottom:
Figure 2. Filemaker Pro data window.
Figure 3a. Specimen tray in an archival cabinet.

Figure 3a. Specimen boxes in a specimen tray.
Figure 4. Label, written in the 1870s (?), by A. D. Brown.
H. acuta Lam 4295

var sublucerna Pilsbry

---------------------------------
H. {acuta ? var. lucern
{lucern ? var. 4295
Jamaica.
A.D. Brown Colln.

Figure 5. Specimen card.
Figure 6.
Left-hand page:

<table>
<thead>
<tr>
<th>A.N.S. Number</th>
<th>Original number</th>
<th>NAME</th>
<th>LOCALITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>4295</td>
<td></td>
<td>pleurodonte sublucerna P</td>
<td>Jamaica</td>
</tr>
</tbody>
</table>

Right-hand page:

<table>
<thead>
<tr>
<th>Alcoholic</th>
<th>Dry</th>
<th>COLLECTOR</th>
<th>DONOR</th>
<th>Date of Presentation</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td></td>
<td>ditto (for T. B.)</td>
<td>A. D. Brown</td>
<td>VII 27 1923</td>
<td>Types for P. Acuta, ditto Sublucerna Pilsbry</td>
</tr>
</tbody>
</table>

Figure 7.
3 paralectotypes split out, recataloged as ANSP 356836
Det. A Bogan 12 May 1983

Figure 8. Type/paralectotype split
Figure 9. Labeled specimens

Figure 10. Digitization project label