Rapid Collections Surveying With Book Traces @ UVa

Kristin Jensen
University of Virginia Library, khj5c@eservices.virginia.edu

Carla H. Lee
University of Virginia Library

Kristin Jensen and Carla H. Lee, "Rapid Collections Surveying With Book Traces @ UVa" (2015). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284316266

Rapid Collections Surveying With Book Traces @ UVa

Kristin H. Jensen, Project Manager, University of Virginia Library

Carla H. Lee, Senior Director, Collections Access and Discovery, University of Virginia Library

Abstract

Many donated books in circulating collections have value as historical artifacts due to unique interventions by their former owners, such as marginalia, inscriptions, and insertions. These interventions can potentially offer a trove of evidence of how books have been consumed across time and what they meant to past cultures, but are generally undocumented and therefore undiscoverable through library catalogs. Moreover, as circulating copies, these books may be vulnerable to damage, loss, and withdrawal. Book Traces @ UVa is a two-year effort to survey pre-1923 books in the University of Virginia Library circulating collection for uniquely modified volumes and enhance our catalog to make them more readily discoverable. Because our target population is large—more than 92,000 volumes—we are developing techniques for rapidly sampling the collection through a randomized, statistically representative selection from each call number subclass. We believe our statistical samples will point the way for deeper exploration of subject areas where the books are especially rich in evidence of historical readership, and in this paper we present some of our preliminary findings as well as an overview of our survey and sampling techniques. We also suggest how the Book Traces experiment in considering non-rare books as historical artifacts can reshape libraries' approach to retention and off-site storage decisions.

Introduction

Book Traces @ UVa is a two-year research project in the University of Virginia Library, funded by a CLIR Hidden Collections grant. This project got started when an English professor named Andrew Stauffer noticed that many of the older books in our circulating collection contain traces left behind by readers of the past: marginal notes, gift inscriptions, sentimental objects tucked between the pages, and so forth. Professor Stauffer started a crowd-sourcing website called Book Traces to collect examples, informally, of uniquely modified volumes in academic libraries, focusing on circulating collections.[1] This website has gotten a lot of press and you may have heard about it.[2] Book Traces @ UVa is the next, more formalized step beyond the initial Book Traces crowdsourcing effort. Professor Stauffer teamed up with the UVA Library's director of preservation services, Kara McClurken, to launch a more rigorous and methodical survey of the UVA collection.[3] Later in my paper I am going to talk more about the methods that we are using for this survey, but I

[1] http://www.booktraces.org/

[2] Jennifer Howard, (2014, May 5), "Book lovers record traces of 19th-century readers," *The Chronicle of Higher Education*, Retrieved December 1, 2015, from http://chronicle.com/blogs/wiredcampus/book-lovers-record-traces-of-19th-century-readers/52415; Alexis C. Madrigal, (2014, May 7), "What Is a Book? Not Just a Bag of Words, but a Thing Held by Human Hands," *The Atlantic*, retrieved December 1, 2015, from http://www.theatlantic.com/technology/archive/2014/05/what-is-a-book/361876/; Allison Meier, (2014, May 8), "The Call to Action to Save Digitized Books from Oblivion," *Hyperallergic*, retrieved December 1, 2015, from http://hyperallergic.com/125215/the-call-to-action -to-save-digitized-books-from-oblivion/; Eve M. Kahn, (2014, December 18), "Of Magic Lanterns and Screen Gems," *New York Times*, Art & Design sec, Retrieved December 1, 2015, from http://www.nytimes.com/2014/12/19/arts/design/of-magic-lanterns-and-screen-gems.html. Additional press citations at http://www.booktraces.org/press-for-book-traces/

[3] Project website: http://booktraces.library.virginia.edu/

http://dx.doi.org/10.5703/1288284316266

want to start with an example of the kind of book we are finding. I hope this will give you a hint of the motivations behind the survey.

We recently found a mechanical engineering manual published in 1876.[4] There are two inscriptions on the title page. One is by the original owner, one R. B. S. Nicolson, who was an engineering student at UVA in the late 1870s. The second inscription shows us that the book was later donated to the UVa Library by his brother, John. The real story, however, is inside. Bound into the back of the manual are several pages of lined paper for taking notes, and written on one of them we found this message:

> New York City April 13th 1912.
>
> It seems a desecration almost for me to write in this book so exclusively associated with my brother—but I am led to look into it for the first time in many, many years this Saturday night, the anniversary of his birth. He was born that memorable day, fifty one years ago, on which the Civil War between the North and the South began—fifty one years ago!! How life is slipping by!
>
> This book is a relic of my brother's first ambitions—viz, to be a civil engineer—and of his course at the University of Virginia to this end. Instead of continuing to this goal, he went into our father's business in Savannah in 1880, coming however to an early end. He was drowned at Tybee Island Ga. July 10th 1881.
>
> John Nicolson

This book is more than just a container for intellectual content. It is a double memorial to the young man who was born the day the Civil War started and died tragically young, just twenty years later. John Nicolson made tribute to his brother's memory first when he wrote this personal reflection in the blank notepaper leaves of the book and again when he donated it to the University of Virginia. Moreover, this book is a record, not just of the Nicolson brothers' kinship, but also of their affective relationship with the University of Virginia. As an artifact, this engineering manual gives us a little piece of family history, Southern history, and University history all rolled into one.

With Book Traces @ UVa, the basic question we are asking is, what *else* is "hidden in plain sight" in the circulating collection? We want to identify unique volumes with preservation needs and see that they get appropriate treatment to preserve their artefactual value. We also want to identify uniquely modified volumes as such in the catalog, facilitating research on them as historic artifacts. We are doing all of this with an awareness that there are space pressures on libraries across the country, and that there is some concern among scholars of nineteenth-century studies that circulating materials from this era may be vulnerable to withdrawal due to low circulation or shared repository status. Part of what we are doing is considering what kind of historical evidence might be lost in the case of withdrawals, and we want to offer a rationale for retaining uniquely modified volumes as well as a model for surveying collections.

Survey Procedure: The Why and How of Statistical Sampling

Our project started in April of this year and runs through March of 2017. We are working primarily in Alderman Library, the main circulating library for humanities and social science research at the University of Virginia. We are also drawing some books from off-site storage, but our project does not go into special collections because we assume those books are already well preserved and

[4] John C. Trautwine, (1876), *The Civil Engineer's Pocket-book (*10th ed.), Philadelphia: Claxton. Catalog entry for the UVA Library copy in question: http://search.lib.virginia.edu/catalog/u916173. For more on the book and the Nicolson brothers' history, see Kristin Jensen and Maggie Whalen, (2015, October 23), "Book Find: A Brother's Memorial," *Book Traces @ UVa*, retrieved December 3, 2015, from http://booktraces.library.virginia.edu/book-find-a-brothers-memorial/

protected against loss and damage. We are surveying only books published before 1923, for a few reasons. First, titles in this population are out of copyright and likely to be available in Google Books or HathiTrust; we want to make a case that uniquely modified volumes in local holdings may be overlooked if users turn first to digital surrogates. Secondly, pre-1923 volumes are also generally low in circulation, making their unique features less likely to be discovered serendipitously, and also potentially making them candidates for withdrawal when libraries develop shared repositories. At the same time, the low circulation of these volumes makes it more likely that loose insertions may be found intact. Thirdly, setting a 1923 cutoff means that most of the books in our survey were printed during the industrial period from about 1820 through the early twentieth century when print culture exploded and the relatively low cost of books meant that people could afford to own more books, personalizing them and often treating them as sentimental possessions. Finally, the history of collection development at the University of Virginia gives us another reason to focus on books from the long nineteenth century. Books from this time period in our circulating collection are likely to have been acquired as donations following the decimation of the UVA Library collection by a fire in 1895. Thus, they are likely to show evidence of personal reading practices rather than institutional use.

Narrowing our shelf list to pre-1923 monographs still left us with over 92,000 books to examine in the Alderman Library alone. Surveying such a large collection presents some challenges. At the outset, we did not know for sure how many books we would be able to survey in 3,000 hours of student worker time, which is the amount budgeted in the grant. We also did not want to wait until the end of the two-year grant period to get a sense of what we had on hand.

The solution that we came up with in the early weeks of the project was to start with a statistical sample. We did a few pilot surveys to iron out workflow issues, and then we brought in a fourth year statistics major named Jackie Morrogh to help us create a statistical sampling scheme. (Unfortunately, Jackie could not come to Charleston with me as a co-presenter, but I will try to do justice to her work.) We divided the shelf lists by Library of Congress subclassification, and Jackie calculated how big a sample we would need to take from each range in order to determine the "hit rate" with 95% confidence and a 5% margin of error. When I say the "hit rate," I mean how many books we would find with unique modifications that meet a certain threshold of notability. Taking a sample from each Library of Congress subclassification reduced the total population initially being surveyed from a little over 92,000 to a little under 19,000. In other words, during the preliminary sample period we are only looking at one-fifth of the population of interest.

The statistical sampling approach gives us some choices. We could look at a smaller subset faster and draw our conclusions with less confidence, or we could look at a larger subset for better confidence but spend more time doing it. We chose a sample size that we felt would be feasible to complete and a confidence level that we believed would allow us to draw some reasonably reliable conclusions about the collection.

Once we had decided on a statistical approach, the technique for generating shelf lists was fairly straightforward. Our statistician determined the number of books we needed to sample from each subclass. She assigned a random number to each line of the shelf list, then sorted each subclass of the shelf list by those random numbers. From the top of the randomized list, she selected the number of books needed for our sample.

For subclassification ranges with very few holdings, we have to look at every book or nearly every book in order to survey them with a confidence level of 95% and a 5% margin of error, but for larger subclasses, a much smaller proportion gives us an adequate sample. For example, we have over 15,000 pre-1923 monographs in the Library of Congress PR subclass, English literature, but we only have to

look at 375 books to determine the "hit rate," the percent of PR books with unique modifications.

For another example, in the HQ range (covering the family, marriage, and women), we own 385 pre-1923 monographs and we are sampling 193 of them, about half. In call number subclasses with fewer than 140 books, the necessary sample size is so close to the population size, and so small, that we are actually surveying the entire population because it seems efficient to do so. By starting with the largest samples and working our way down the list to the smallest, we have been able to learn a lot about the largest subpopulations within our collection very quickly.

If we stratified the collection differently, such as by looking at all the A call numbers as one population, all the B call numbers as another population, and so forth, we would be able to take proportionally smaller samples and thus finish our entire survey even faster. The trade-off would be that what we learn about comparative hit rates would be accordingly large-grained. For example, if we had sampled all of the H call numbers as one population, we would not have learned that there is only a 5% rate of unique modifications among the HJ call numbers, representing books on finance, as compared with a 15% hit rate in the HB call numbers, representing books on economic theory and demography. We would still learn, however, that the hit rate is quite low in the A call numbers, representing general works, as compared with the H call numbers, representing the social sciences. And we would probably find a similar overall hit rate for the pre-1923 collection, around 12.5%.[5]

There are many benefits to starting Book Traces @ UVa with a statistical sample. Finishing the statistical sample has taken us only four months out of the two-year grant period, using about 500 student worker hours. While working on the statistical sample, we have also gathered data about our student workers' efficiency; knowing that they can survey about 38 books per hour allows us to plan our future work based on their efficiency as well as the hit rate data analyzed by our statistician. Even before finishing the sample, we had enough efficiency data to determine that we could finish surveying the entire Alderman Library population of pre-1923 imprints within the grant period, with time to spare. This was something we were not sure of at the outset. Now, knowing that we have time to survey other libraries within the UVA system, we can plan samples guided by our hit rate data from Alderman Library: we can choose to concentrate on subject areas that had the highest hit rates in the Alderman population in the hope that this will lead us straight to the richest seams of artefactual evidence.

Other Lessons Learned

I think doing the statistical sample has been the single most important choice we have made for learning a lot about our collection in a fairly small amount of time. That said, however, we have also benefited from putting some thought into designing an efficient process for carrying out the survey. Before hiring our student project assistants, I experimented with the surveying process, with an eye toward usability. Our process is designed to optimize a balance of speed and accuracy in recording data about the books in our open stacks. Our project assistants spend about 80% of their work time in the stacks, using a book truck as a mobile work station with a laptop, barcode scanner, and handheld "pen" scanner for text. Working from a list of pre-1923 monographs, the assistants pull each book from the shelf and give it a preliminary examination. If there is a bookplate, they use the pen scanner to record just the name on the bookplate; the bookplate census is a side project that will allow us to connect books donated by the same family. Books that have unique modifications meeting our criteria are set aside on the book trucks for further analysis; others are returned to the shelf. Near the end of each shift, the assistants do a second inspection of the books they pulled and record the

[5] At the time this paper was presented at the Charleston Library Conference, the survey had not yet been completed. It has since been completed and the results have been statistically analyzed. Our final overall hit rate for all pre-1923 monographs in the Alderman Library collection is 12.7%.

specific modifications, such as marginalia or insertions, using a Google form.

This two-step process promotes accuracy of description by allowing project assistants to concentrate on description separately from the task of finding and pulling books. It also accommodates one of the technical needs of the project: wifi is not fully reliable in the stacks area, so the assistants work offline using Excel spreadsheets for shelf lists in the stacks, then return to a central area to review the books they have selected and fill out the online descriptive form. Iterative design of the survey process has raised other technological considerations, too. For example, barcode scanners in raised cradles are more usable for this application than handheld scanners. For another example, laptop battery life is critical when students work in five-hour shifts and must roll their work stations through the stacks without plugging in.

I mentioned earlier that the statistical sample has helped us learn a lot about our collection in a short amount of time. Our findings are still preliminary: in fact, our project assistants will be wrapping up the last subclassification surveys in the next week or so, and the statistical analysis of our project data always lags two to three weeks behind the gathering of the data. However, we already have a lot of information about the relative hit rates in different subject areas. We expected to find a high hit rate in the P call numbers, covering languages and literature, and in fact we did.[6] We found a high hit rate in the B call numbers—representing philosophy, psychology, and religion—partly because we have a lot of books from two nineteenth-century philosophy professors who wrote in the margins a lot. It is also partly because we did some pilot work in the B call number ranges before starting the sampling scheme, and at that time our selection criteria were a little more inclusive.[7] We found a low hit rate in the A call numbers, which is not surprising: those represent "general works," which do not seem to call for a lot of personal engagement.[8]

To describe the types of unique modifications, or reader interventions, that we have been finding, we are using a modified set of terms from the RBMS Provenance Evidence Thesaurus[9] for research purposes, with the intention of eventually collapsing all the metadata into PET vocabulary for catalog description. The five most frequently found types of interventions are, in order: inscription, usually meaning the owner's inscription; verbal marginalia, which we are distinguishing from nonverbal marks; underscoring; nonverbal marginalia; verbal annotations, which we consider to be any sort of written marking that is not an inscription and does not interact as closely with the text as marginalia does; and finally gift inscriptions. Further down the list are things like insertions, authors' inscriptions, and doodles or artwork.

Another lesson we have learned is that it is considerably easier to survey books in bulk while they sit on open shelves in our main library. We have recently done a small pilot sample of books drawn from UVA's off-site storage facility, and the process was very labor intensive. When our project assistants are working in the open stacks, they spend roughly a minute and a half per book on our shelf list. By contrast, when we pulled books from off-site storage, the off-site staff spent roughly four to five minutes per book to retrieve each volume, route it to us, receive it back when we were done with it, and return it to the high-density shelving. Having discovered the labor requirements of surveying books in off-site storage, we are now making a point of surveying

[6] In the final statistical analysis (completed since this paper was presented at the Charleston Library Conference), 11 out of 16 sections in the P call numbers had hit rates above the library-wide average of 12.7%.

[7] In the final statistical analysis, 11 out of 15 sections in the B call numbers had hit rates above the library-wide average of 12.7%.

[8] In the final statistical analysis, 6 out of 9 sections in the A call numbers had hit rates lower than the library-wide average of 12.7%.

[9] http://rbms.info/vocabularies/provenance/alphabtical_list.htm

batches of books *before* they are scheduled to be moved to off-site storage.

The most important conclusion we have drawn is to confirm what Andrew Stauffer first discovered serendipitously. Many of the copies in our circulating collection have unique features that are not represented in digital surrogates created elsewhere. By considering each volume as a unique physical object, we can identify which ones have special interest as artifacts and as pieces of a distributed archive of the history of reading.