ARTICLE

Rarely Analyzed

The Relationship between Digital and Physical Rare Books Collections

Allison McCormack and Rachel Wittmann

ABSTRACT

The relationship between physical and digitized rare books can be complex and, at times, nebulous. When building a digital library, should showcasing a representative slice of the physical collection be the goal? Should stakeholders focus on preservation, high-use items, or other concerns? To explore these conundrums, a special collections librarian and a digital services librarian performed a comparative analysis of their library's physical and digital rare books collections. After exporting MARC metadata for the rare books from their ILS, the librarians examined the place of publication, publication date, and broad subject range of the collection. They used this data to create a variety of visualizations with the open-source digital humanities tool Tableau Public. Next, the authors downloaded the rare books metadata from the digital library and created illuminating data visualizations. Were the geographic, temporal, and subject scopes of the digital library similar to those of the physical rare books collection? If not, what accounted for the differences? The implications of these and other findings will be explored.

INTRODUCTION

As of August 2019, the Special Collections Division of the University of Utah J. Willard Marriott Library held over 256,000 printed works and archival collections. Approximately 22% of the collection, or just over 55,000 works, belongs to the Rare Books Department (https://lib.utah.edu/collections/rarebooks/), which contains not only books but serials, maps, manuscripts, ephemera, and other formats. The collection covers over 4,000 years of human history, with its earliest piece, a cuneiform tablet, dating to the mid-twenty-third century BCE; contains works from nearly 100 different countries; and represents a wide variety of topics, including the exploration and settlement of the American West and the history of the book. The Rare Books Department, a subset of Special Collections, specifically seeks to document the history of written human communication and actively collects historical items to enhance teaching and research at the University of Utah.

The Marriott Library has been adding digitized works from the Rare Books Department to its Digital Library (https://collections.lib.utah.edu/) for over 25 years. Approximately 780 works, or 1.42% of the rare books collection, has been digitized to date. However, no formal collection development plan was ever written, and rare books were selected for digitization by both curators and patrons. Unfortunately, the reason a particular item was digitized is not recorded in the system: it is unclear if age, research value, physical condition, a desire to bring forward underrepresented stories, or a combination of these and other factors influenced the decision to digitize a rare book. This piecemeal approach to digital library collection development, while not uncommon, made it difficult for library staff and patrons to determine the relationship between the digital and physical collections of rare books. It also presented challenges when library staff

Allison McCormack (allie.mccormack@utah.edu) is the Original Cataloger for Special Collections, University of Utah, University of Utah. **Rachel Wittmann** (rachel.wittmann@utah.edu) is the Digital Curation Librarian, University of Utah.

https://doi.org/10.6017/ital.v41i2.13415

attempted to communicate the scope and intent of the Digital Library to patrons, who assumed that the digitized items were representative of the overall collection. Given their expertise in library metadata, the authors decided to analyze both traditional library catalog records and digital library records for the rare books collection and explore whether the digital collection was proportionally representative of the physical collection or if it differed in geographic, temporal, or subject scope in a meaningful way. They then created a series of data visualizations to better communicate information about the library's rare books holdings.

LITERATURE REVIEW

While much has been written about methods and criteria for selecting special collections items to be digitized and the effects of digitization on collection accessibility, few authors have discussed the relationships between digital collections and the physical collections from which they were sourced. In their highly detailed treatise on selection strategies for digitization, Ooghe and Moreels identify representativity, a method that "aims for a final selection that provides a representative view of the original collections," as one of 25 selection criteria for digitization projects.[1] However, Alexandra Mills notes that "without a thorough understanding of the institution and collections, it is impossible to create truly representative collections."[2] Because many digitization initiatives are undertaken in response to user requests, preservation concerns, or the availability of project-based funding, it is likely that most libraries do not plan for their digital collections to be representative of their overall special collections holdings. As Peter Michel states, the digital collections at the University of Nevada, Las Vegas, were explicitly built with popular history and popular culture in mind and were never intended to be "surrogates of the collection."[3] Bradley Daigle of the University of Virginia explained that digitization could be undertaken to alleviate preservation concerns, respond to defined research needs, or to brand certain online content, but this approach could give the mistaken impression "that only the important materials are digitized."[4]

Despite the gaps in the literature, having an explicit collection development policy is still considered paramount; indeed, it is the very first principle listed in the National Information Standards Organization (NISO)'s framework for building "good" digital collections.[5] To investigate this type of documentation further, a Google search was employed using the search term "digital collection development policy site:edu". This yielded 10 publicly accessible digital collection development policies from academic libraries in the United States: [6]

- Amherst College Library (https://www.amherst.edu/library/services/digital/digitalcolldev)
- Emerson College Archives and Special Collections (https://www.emerson.edu/policies/digital-collections-development-policy)
- Colorado State University Libraries (https://lib.colostate.edu/digital-collection-development-policy/)
- Florida Atlantic University Digital Library (https://library.fau.edu/policy/digital-library-collection-development-policy)
- Georgetown University Library (https://www.library.georgetown.edu/digital-project-policy)
- Northern Illinois University Digital Library (https://digital.lib.niu.edu/policy/collection-development-policy)

- Oregon Health and Sciences University Digital Collections (https://www.ohsu.edu/library/ohsu-digital-collections-development-policy)
- University of North Texas University Libraries (https://library.unt.edu/policies/collection-development-digital-collections/)
- Wesleyan University Digital Library (https://digitalcollections.wesleyan.edu/about/what-we-collect)
- Williams College Special Collections (https://specialcollections.williams.edu/collection-development-policies/digital-collections/)

In reviewing the sample of 10 universities' digital collection development policies, homogenous content becomes apparent. Almost all of the policies include a mission statement, scope, and selection criteria for potential digital collection items. All policies include criteria that physical materials should meet in order to qualify for digitization. The most common criteria for digitization are materials that are rare or unique, high-use, fragile, important to institutional or regional history, and/or support campus curriculum or faculty research. In addition, the clearance to publish materials online is ubiquitous among the policies. Materials eligible for online display must either be in the public domain or intellectual property rights are held by the institution, and materials currently under copyright must receive permission from the copyright holder. A measured approach to digitization qualification has been employed by the University of North Texas (UNT) Libraries' Digital Collections and the Northern Illinois University Digital Library (NIUDL). UNT Libraries' Digital Collections policy lists levels of criteria that materials must meet in order to be digitized and included in the digital library; to qualify for digitization, all criteria on level one must be met while only one criterion from level two is needed. NIUDL includes a Priority Factor Rubric which includes criteria categories and corresponding numerical scale with a maximum point of 35, the higher value signifying an elevated priority. Six of the 10 policies include prioritizing materials that support diversity and inclusion missions on campus. Amherst College has leveraged their digital collection development policy to include content that would increase perspectives of underrepresented groups within the digital collections and traditionally underrepresented groups more broadly. NIUDL includes marginalized groups as a collection priority area in order to "deepen public understanding of the histories of people of color and other communities and populations whose work, experiences, and perspectives have been insufficiently recognized or unattended" and lists over 20 such groups. The collection candidate's relationship to other collections is outlined in four of the 10 policies. Georgetown University requires that "the materials form a coherent collection, fill gaps in existing collections, or complement existing collection strengths." Amherst College evaluates whether digitization would "enhance public awareness of Archives' Collection strengths." Another function of a digital collection development policy is to inform the public on the scope and provenance of contents in their digital library. The UNT Digital Collection Policy includes a section outlining the content contributors, including partners, which can be beneficial for large-scale digital libraries that host collections from multiple partners. UNT is also exemplary in defining collection curators and their responsibilities while underscoring the nature of this role, likely changing over time and not set to an individual. With no written digital collection development policy regarding special collections at the Marriott Library, the authors would first have to analyze both the physical and digital special collections before determining what factors may have influenced the digitization of these materials.

Libraries are gathering massive amounts of data, ranging from the metadata of their varied collections to patron usage statistics of both physical and digital collections. Interpretation of the

ever-growing accumulation of data can quickly become complex. By visualizing data, we are able to interpret large and often messy sets of data while processing multiple aspects of the data concurrently. For example, the Ohio State University (OSU) Libraries used Tableau Desktop to combine data from various departments in order to better manage and explore information.[7] Tableau was OSU's data visualization software of choice due to its ease of use and accessibility, and the program was also used to create dashboards that blend data from various sources for real-time visualizations.

BIBLIOGRAPHIC METADATA CLEANUP

To understand the Marriott Library's collections, one must first understand the relevant metadata, which for the Rare Books Department is in the Machine-Readable Cataloging (MARC) format. A popular criticism of MARC, commonly used in traditional library cataloging, is that the schema is highly regulated and, at times, redundant. However, for the purposes of this project, those qualities proved to be a boon. An older, uncorrected record in the Digital Library might list London as the place of publication for a particular book, but it was not immediately apparent if that referred to London, England; London, Ontario; or London, Ohio. However, a MARC record would not only list a book's city of publication in the 260 or 264 field but would also contain a two- or three-letter code in the 008 field that specified the country, US state, Canadian province or territory, or Australian state or territory in which it was published. For this reason, the authors decided to base their analysis on MARC record data from the physical collection instead of the Dublin Core metadata used in the Digital Library.

In order to tease out the relationships between our digital and physical collections, each of the approximately 55,000 rare books bibliographic records stored in Alma, the Marriott Library's cloud-based library services platform, would have to have a common set of data points that could be compared. For the purposes of this analysis, the authors chose to investigate the place of publication and the subject of each work. Despite the relative rigidity of MARC metadata, some of the Alma records lacked country of publication data in the 008 field. These records were not incorrect, but merely outdated: some had been copied directly from paper catalog cards when the library first transitioned to a computer-based cataloging system, while others were created using different metadata standards. Approximately 6,000 rare books either completely lacked a country code in the 008 field or had data that could possibly be enhanced by, for example, replacing a code for the United States with a code for a particular state.

Instead of editing all 6,000 records by hand, the cataloger wrote several metadata normalization rules in Alma to automatically correct the most obvious errors. Records that listed Chicago as the place of publication were assigned the MARC geographic code for Illinois, while those that were published in Lugduni Batavorum, the Latin designation for Leiden, were given the geographic code for the Netherlands. However, 3,000 records were unable to be enhanced in this manner, either because their place of publication was an ambiguous city name like Cambridge or because the place of publication was listed as unknown. The cataloger examined each record individually and was ultimately unable to assign a MARC geographic code to 1,682 records, most of which were Arabic manuscripts or advertising pamphlets that simply did not list a place of publication or creation. While these records would be excluded from the place of publication analysis, they could be mined for data on other topics. With the MARC records as complete as possible, the metadata was exported from Alma into an Excel spreadsheet and given to the metadata librarian for further manipulation.

METADATA TRANSFORMATION & VISUALIZATION CREATION

The next phase involved standardizing the raw metadata to create human readable data, rather than MARC codes, that are necessary to produce data visualizations. Once the physical rare books' bibliographic metadata was updated in Alma, it was then exported as a comma-separated values file. The raw data export produced a massive spreadsheet containing over 50,000 MARC records. These records included two- and three-letter location codes for the place of publication from the Library of Congress MARC Code List for Geographic Areas. Two-letter codes are used for most countries, while three-letter codes are used for states within the United States, provinces within Canada, and territories within the United Kingdom. While this additional level of location data was available for books from the United Kingdom and Canada, it was decided to review the collection at a country level for consistency and map display. Books from the United States, however, were analyzed on a state level, considering the research is germane to an American institution. Using a list correlating these codes to the location name provided by the Library of Congress (https://www.loc.gov/marc/countries/countries_code.html), a VLOOKUP formula was used in Microsoft Excel to add the location names to the MARC records. The VLOOKUP formula pulls in data from one table to another as long as the two tables have one data field in common. In this exercise, both tables of data contained the Library of Congress location codes, therefore the LC location codes were used to add the location names to the table containing the MARC metadata. Once the location names were added, there were some additional quality control steps required, as LC location names that included outdated country names posed issues to mapping the data to current country names and boundaries. For example, we combined the codes for East Germany and West Berlin for the one representing contemporary Germany. For countries that have since been dissolved and rezoned to multiple countries, e.g., the USSR and Czechoslovakia, these records were manually checked for city names and then added to the current country. Once this process was completed, the results showed the rare books were published in 97 countries and all 50 United States, as well as the District of Columbia.

Examining the subject content of the rare books physical collection was another aspect of analysis for this project. In contemplating this analysis, using the LC Subject Heading field was considered, however, faceting of LC Subject Headings and the structure of the exported data posed too many issues for a rather simple analysis. Instead, the Library of Congress call number was used to extract high-level LC classification information for each work by separating the first two letters of the call numbers included in the exported MARC metadata, which indicated LC class and subclass. To add the LC class and subclass names to these letters, a VLOOKUP formula was used again to match the letter codes to the list of LC classification categories. Once classification categories were added to the 55,000 records, works from all 21 LC master classes and 190 subclasses were represented in the rare books collection.

In addition to the physical rare books collection held at the Marriott Library, there is a selection of this collection that has been digitized and is accessible in the Marriott Digital Library. The Rare Books digital collection (https://collections.lib.utah.edu/search?facet_setname_s=uum_rbc) comprises 780 works, although this number includes unique records for individual volumes within a series and therefore is not a true comparison to MARC metadata records, which contain one record for a series. For example, the Silver Reef Miner, a newspaper "devoted to the mining interests of Southern Utah" published during the late nineteenth century, has 30 individual volumes in the Digital Library, but these are represented in just one MARC record. In order to compare the digital collection to the physical collection, the datasets would need to have

consistent data for comparison, namely place of publication and LC classification derived from call numbers. The digital collection metadata is in the Dublin Core schema, which does not include all of the metadata found in the MARC metadata, nor does it use the same format. While there is a Dublin Core spatial element used to capture geographic data on what the item is about, this does not always align neatly with the location of an item's publication. For example, *Reise in das innere Nord-America in den Jahren 1832 bis 1834* (2 volumes) is a book printed in Germany that documents an expedition to North America in 1832–1834 and includes illustrations of Native American people from the Swiss artist Karl Bodmer. For these volumes, the appropriate Dublin Core spatial data would include the specific regions the expedition traveled to in North America; in the MARC 26X field, however, it contains Koblenz, Germany, the city where the volumes were published. Call number data was included for many digitized works, but not in a consistent format. In order to use the same data to compare the physical rare books collection to the digital one, the digital collection metadata was updated with the improved/accurate call numbers found in the MARC metadata. Another improvement to the digital collection metadata was the addition of the Metadata Management System (MMS) ID unique numerical identifiers that aid in locating a record in the Alma system. When the rare books' descriptive metadata was originally converted to Dublin Core during the digitization process, some titles and call numbers were changed and became different from their physical counterparts. The inclusion of the MMS ID allows for a consistent identifier between the physical and digital collections.

When selecting data visualization software, being able to create a map of the places where books in the rare collection were published was a priority. Considering the goal of creating an easily replicable workflow for other libraries, the authors sought a freely accessible program that did not require advanced geospatial skills, unlike Esri's ArcGIS software. Tableau Software is a data visualization software package with both a public and desktop version. The Tableau Desktop version requires a subscription fee while Tableau Public is open access. For the purposes of this study, Tableau Public offered open access and mapping features that are enabled without any geospatial knowledge necessary.

ANALYSIS

Creating a variety of data visualizations allowed information about the Rare Books physical and digital collections to be more apparent than merely browsing entries in a spreadsheet. For example, there are numerous geographic disparities between the two collections of rare materials as shown in the American states in which works from the collections were published. While books from all 50 states are found in the physical collection (fig. 1), only 18 states are represented in the Digital Library (fig. 2), with New York being the state in which the highest number of books were published. As New York City has long been a major publishing center in the United States, the authors were not surprised by this. However, the subsequent states were quite different: California and Utah ranked second and third for the physical collection, while Massachusetts and Pennsylvania claimed those spots for the Digital Library. The authors believe several factors might influence this discrepancy. First, works can only be added to the Digital Library if they are no longer in copyright, and states with longer histories of European-American settlement are more likely to have published books that are now out of copyright. Furthermore, these older books are more likely to be in a fragile condition and therefore may have been digitized to decrease the amount of physical handling to which they are subjected.

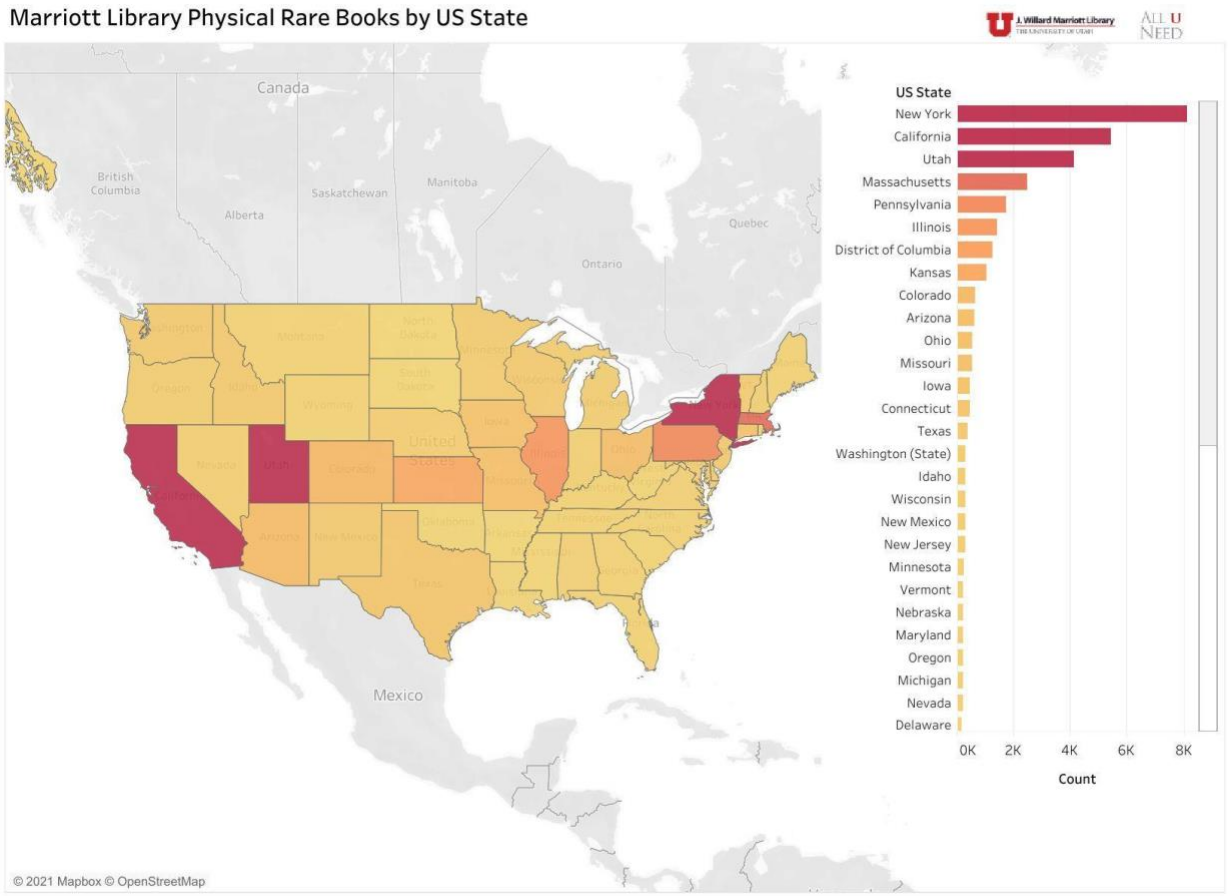

Figure 1. Marriott Library physical rare books by US state.

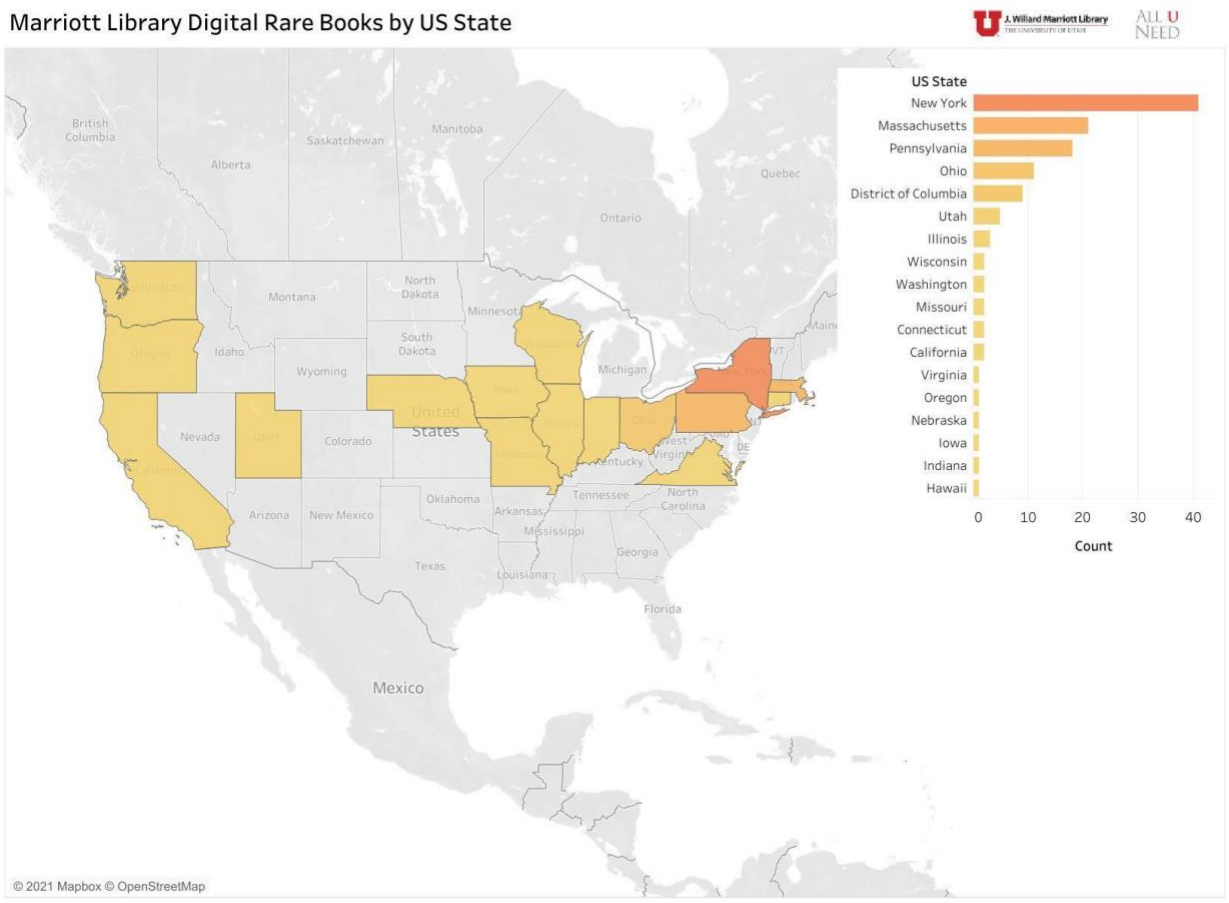

Figure 2. Marriott Library digital rare books by US state.

There are other discrepancies when comparing the country of publication between the physical (fig. 3) and digital collections (fig. 4). While 61% of the physical rare books were published in the United States, only 20% of the digitized works were published in this country. The authors expected to see Egypt rank highly in the physical collection, as many of the rare books were purchased by former University of Utah professor Dr. Aziz Atiya to support the Middle East Center for research he founded; similarly high in rank, Britain, Germany, France, and Italy were all major centers for the early printing and publishing trade in early modern Europe. However, there is strong geographic bias in the digital collection, as only North America, Western Europe, and one African country are represented online. Copyright may again play a factor, as the earliest books from non-Western countries in the collection often date to the twentieth century, but a Eurocentric or other bias cannot immediately be discounted. While the physical collection contains many more European imprints than from the Global South, it is much more diverse than the digital collection.

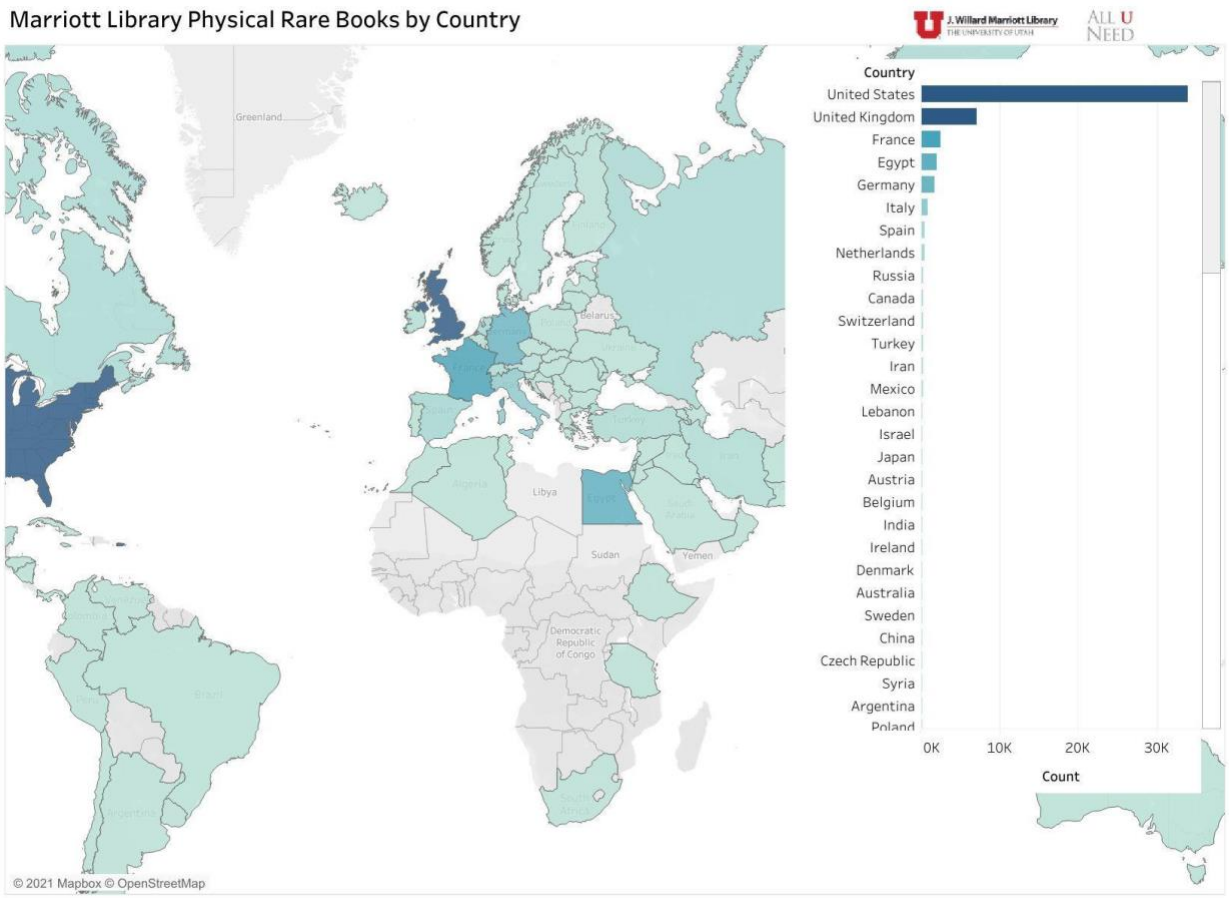

Figure 3. Marriott Library physical rare books by country.

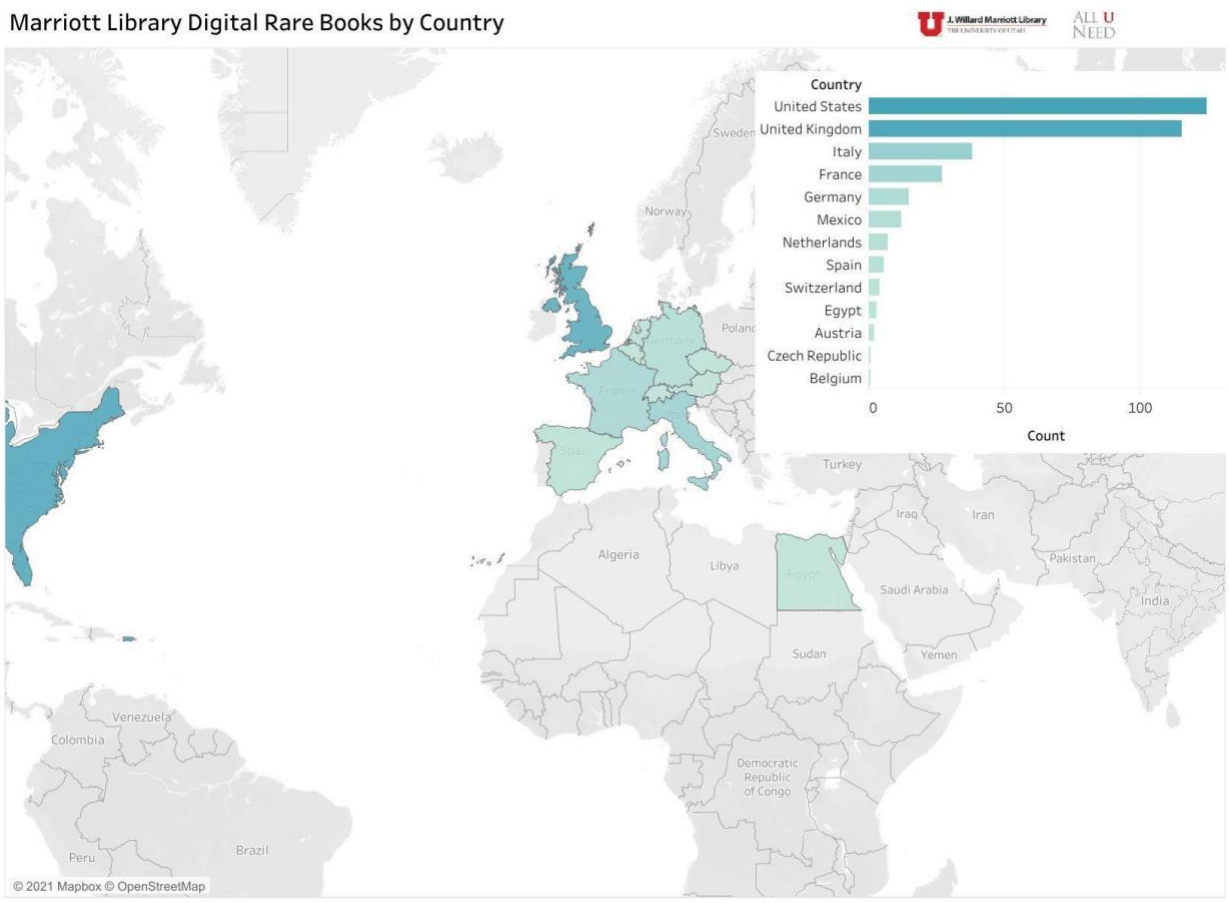

Figure 4. Marriott Library digital rare books by country.

The analysis of the subjects represented in the collection proved to be somewhat challenging to study. Due to the nature and structure of Library of Congress Subject Headings, which attempt to mirror natural language and may be composed of "strings" of phrases to represent complex topics, no Tableau Public visualization could be created that effectively grouped similar content areas together without looking quite fragmented. Instead, the authors based their analysis of subjects on Library of Congress classification numbers (i.e., call numbers) assigned to works, which, though not exact, can be understood as distillations of the subject of a work.[8]

Once again there were considerable differences between the physical and digital rare books collections (fig. 5). As in many generalized special collections, literature and history make up significant portions of the physical collection. However, works on bibliography, or the study of books and book history, comprise a notable percentage of the collection. Many of these are modern works on book history and special collections librarianship and therefore are unable to be digitized due to copyright law. Nearly 9% of the digital collection is on the sciences, though these works comprise only 3% of the physical collection. While this portion of the holdings may be relatively small, it contains many scientific high points such as Vesalius' *De Humani Corporis Fabrica*, early printings of ancient mathematical texts, and the journals of major scientific societies, which may have been digitized both for physical preservation as well as high interest on the part of students and faculty on campus.

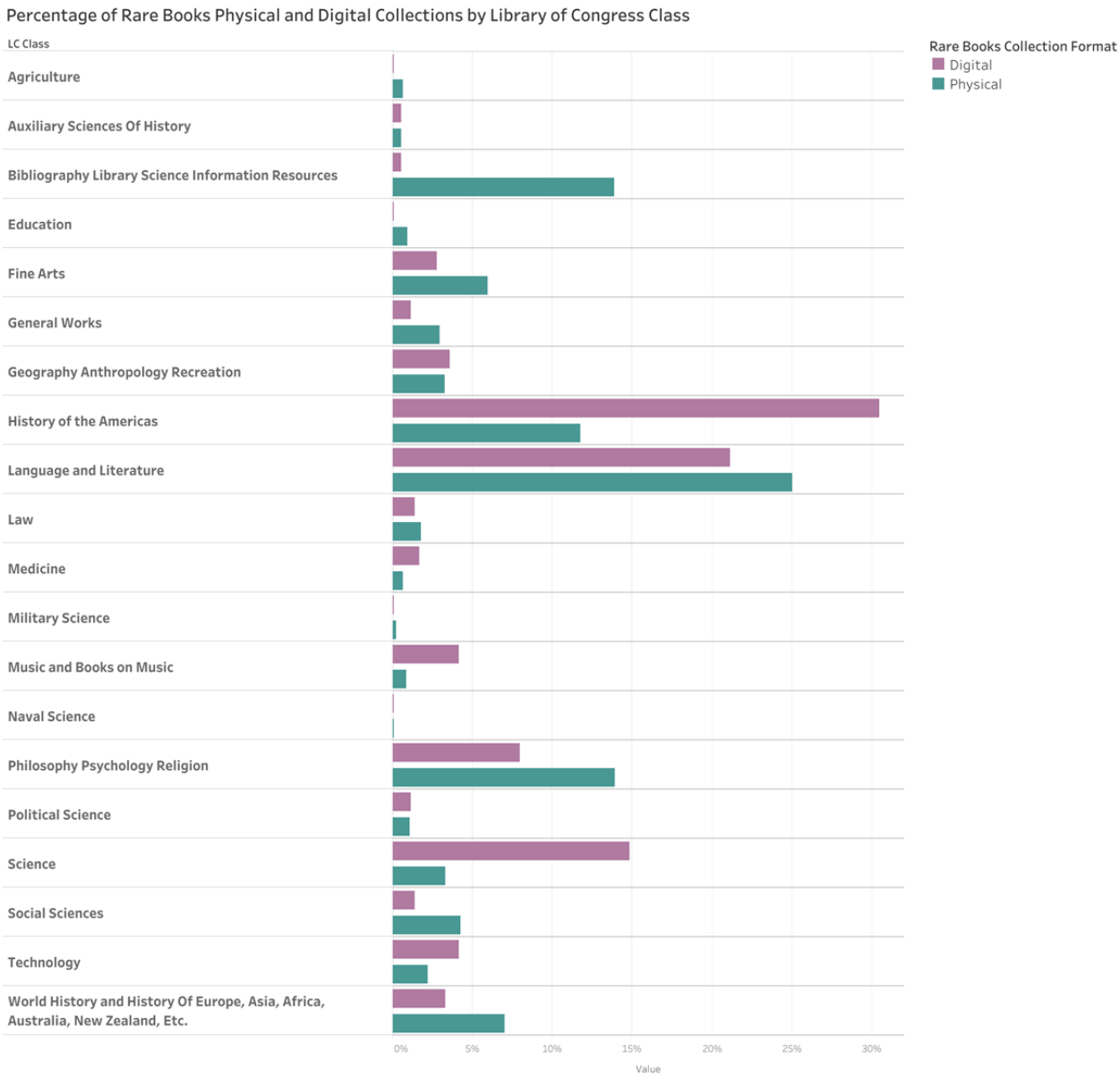

Figure 5. Percentage of rare books physical and digital collections by Library of Congress class.

NEXT STEPS

Now that the first phase of the project is complete, the authors would like to conduct additional analyses. First, they plan to compare the usage statistics of the digital rare books collection to the circulation statistics of the physical collection. This method of inquiry was not possible at the start of the project, as circulation information for the rare books was previously not tracked in the integrated library system. Now that rare books are checked out to patrons for use in the Special Collections Reading Room, this data can be quickly pulled from Alma. Once there is a year's worth of circulation data for the rare books unhindered by the changes necessitated by the coronavirus pandemic, the authors will compare the usage statistics of the digital collection for the same time period. Do patrons in the reading room look at similar materials to online patrons, or are their interests vastly different? Are some rare books used so frequently that they would benefit from the added physical security that digitization brings?

The authors also plan to pull annual usage statistics from the digitized rare books and share this with Special Collections Division leadership. Online patrons are still library patrons, and the division can use the viewing data to show the national and international reach of the collection. Relatedly, the authors will investigate the Digital Library usage data in more depth. Do patrons from Utah, the United States, and the world look at similar materials, or are there geographic divides among the online patrons? Do countries that are home to a majority of the University's international student body have higher viewership numbers?

Finally, the authors wish to convene a group of stakeholders to create a formal collection development plan for the rare books component of the Digital Library. Given the library's limited resources, it is imperative that digitization be done thoughtfully and systematically. There is a good rationale for creating a digital collection that is representative of the physical rare books collection as well as one that highlights certain collection areas. Both material fragility and the modern scholarly emphasis on highlighting the stories of people of color, women, and other underrepresented groups in library collections provide strong counterarguments to making digital libraries strictly representative of their physical counterparts. Since informal conversations with patrons of the Marriott Library revealed that they assumed the Digital Library was representative of the collection overall, it is imperative that this assumption be either confirmed or disclaimed in a publicly viewable statement.

In the case of the Rare Books Department, the authors are in favor of a focused, rather than representative, collection development policy. Firstly, many of the books in the collection are under copyright and therefore cannot be digitized, while other materials like reference sources for rare books librarians will be of limited interest to the general public. Furthermore, complex items such as artists' books are often poor candidates for digitization, as they may have movable components that cannot be captured accurately in a still photograph. As for what should be included online, the authors fully support equity, diversity, and inclusion efforts at the University of Utah and would like to see the Digital Library highlight materials from marginalized groups whenever possible. Usage statistics from the physical and digital collections, when they become available, should also inform the collection development policy to encourage traffic to the Digital Library. Whatever is ultimately decided, however, the clarity a written policy provides will help streamline decision-making and ultimately help both library staff and patrons understand and search within the Digital Library much more effectively.

ENDNOTES

[1] Bart Ooghe and Dries Moreels, "Analysing Selection for Digitisation: Current Practices and Common Incentives," *D-Lib Magazine* 15, no. 9/10 (2009), https://doi.org/10.1045/september2009-ooghe.

[2] Alexandra Mills, "User Impact on Selection, Digitization, and the Development of Digital Special Collections," *New Review of Academic Librarianship* 21, no. 2 (2015): 166. https://doi.org/10.1080/13614533.2015.1042117.

[3] Peter Michel, "Digitizing Special Collections: To Boldly Go Where We've Been Before," *Library Hi Tech* 23, no. 3 (2005): 382, https://doi.org/10.1108/07378830510621793.

[4] Bradley J. Daigle, "The Digital Transformation of Special Collections," *Journal of Library Administration* 52, no. 3–4 (2012): 253, https://doi.org/10.1080/01930826.2012.684504.

[5] NISO Framework Working Group, *A Framework of Guidance for Building Good Digital Collections* (2007), https://www.imls.gov/sites/default/files/publications/documents/framework3.pdf.

[6] The URLs in the following list were accurate as of March 2, 2022.

[7] Sarah Anne Murphy, "Data Visualization and Rapid Analytics: Applying Tableau Desktop to Support Library Decision-Making," *Journal of Web Librarianship* 7, no. 4 (2013): 465–76, https://doi.org/10.1080/19322909.2013.825148.

[8] Readers who do not work with MARC metadata may not be familiar with how Library of Congress call numbers are assigned. Created in 1891, the classification system is based on 21 classes designated by a single letter; subclasses add one or two letters to the initial class. Catalogers must choose which one of the classes to assign to a particular work. The subject headings may guide a cataloger towards a certain class, but there is not a 1:1 relationship between subject headings and call number classes.