Shakespeare and Company Project Data Sets

Joshua Kotin, Rebecca Sutton Koeser

Joshua Kotin, Rebecca Sutton Koeser, Princeton University
Peer-Reviewer: Rachel Buurma
Data Repository: https://doi.org/10.7910/DVN/QXPNLS
Article DOI: https://doi.org/10.22148/001c.32551

ABSTRACT

This article describes three data sets from the *Shakespeare and Company Project*. The data sets provide information about Shakespeare and Company, Sylvia Beach's bookshop and lending library in interwar Paris. The first data set focuses on the members of the lending library. The second, on the books that circulated in the lending library. The third, on the events—borrows, purchases, subscriptions, renewals, deposits, reimbursements—that connected members and books. Together, the three data sets promise to address and bridge concerns in modernist studies, the digital humanities, and the public humanities. Work on the data sets began in 2014. The first two versions of the data sets were released in 2020 and 2021, respectively. The current version, 1.2, was released in 2022. Over forty people have contributed to the data sets.

1. Overview

What if Sylvia Beach had had access to a computer at Shakespeare and Company, her bookshop and lending library in interwar Paris? What if she had used the computer to record the names and addresses of lending library members, along with the books they bought and borrowed? Gertrude Stein paying fifty francs for her first lending library membership on March 16, 1920. James Joyce moving to a new apartment in the seventh arrondissement on November 1, 1922. Ernest Hemingway borrowing Ivan Turgenev's *The Torrents of Spring* (1872, translation 1897) on October 27, 1925. Thousands of transactions. Thousands of addresses. Thousands of books. One computer.[1]

What if scholars and the general public had access to that computer today? What would we learn about Paris in the 1920s and 1930s; about the Lost Generation; about Stein, Joyce, and Hemingway? What books would we re-discover? Whose lives would we be able to explore? How would the computer change our understanding of modernism, and cultural and intellectual history more generally?

The *Shakespeare and Company Project* is that computer—or more accurately, that relational database and web application. The *Project* uses documents from the Sylvia Beach Papers at Princeton University Library and other libraries to give researchers and the general public access to the world of Shakespeare and Company. The *Project* focuses on the Shakespeare and Company lending library, which operated alongside the eponymous bookshop and publishing house. (Beach published three books under the Shakespeare and Company imprint—all related to Joyce. Most significantly, she published *Ulysses*, which she tenaciously kept in print through eleven printings between 1922 and 1930.)[2] Visitors to the *Project* website can browse and search the lending library's membership and holdings, explore maps of Paris, and track the borrowing histories of specific writers and the circulation of specific books. Want to know what Stein read in 1921? Want to know whether she read the same books as her neighbors in the 6th arrondissement—or as other Americans or women or forty-somethings? The *Project* can address these questions and many more.

The *Project* makes its data available in three data sets, which may be downloaded in CSV and JSON formats. The data sets allow for more ambitious, versatile, detailed, and efficient analysis. The first data set focuses on lending library members or "subscribers," and includes addresses, membership years, and demographic information. The second data set focuses on books and periodicals (and other formats), and includes borrow and purchase counts, circulation years, and bibliographic information. The third data set focuses on lending library events, and includes information about borrows, purchases, subscriptions, renewals, and other transactions. (These events connect members and books.) The data sets may be used individually or in combination: *Project* URIs serve as consistent identifiers across all three. The data sets are the foundation of the *Project*, a link between the world of Shakespeare and Company and our own.

This article provides an account of the *Shakespeare and Company Project* data sets—their sources, history, structure, limitations, and potential. The article has six sections. It begins with an introduction to the lending library and the *Project*'s three sets of archival sources. The article then provides an overview of the data collection and data structuring process before focusing on the individual data sets. The article concludes with a discussion of the limitations and potential of the data sets, as well as plans for future refinement and augmentation.

Work on the data sets began in 2014. The first version of the *Project* data sets was released in August 2020, the second in January 2021 (version 1.1), and the third in January 2022 (1.2). Over forty people—the *Project* team—have worked on the data sets, including thirty graduate and undergraduate students. (A full list of contributors is available on the credits page of the *Project* website.) This work was supported by grants and in-kind support from multiple departments, programs, and centers at Princeton University: The Center for Digital Humanities; the Office of the Dean for Research and the Innovation Fund for New Ideas in the Humanities; the Humanities Council and the David A. Gardner '69 Magic Grants; the Princeton-Mellon Initiative in Architecture, Urbanism, and the Humanities; the Bain-Swiggett Fund, Department of English; and the Office of the Dean of Faculty and the University Committee on Research in the Humanities and Social Sciences.

2. The Shakespeare and Company Lending Library

On November 17, 1919, Sylvia Beach opened Shakespeare and Company at 8 rue Dupuytren on the Left Bank in Paris.[3] From the beginning, it operated as a bookshop and a lending library. Beach modeled the business on La Maison des Amis des Livres, a French-language bookshop and lending library owned by Adrienne Monnier, Beach's life partner. English-language books were expensive and difficult to obtain in interwar Paris. Shakespeare and Company made a diverse range of English-language material available—canonical novels, avant-garde magazines, detective fiction, philosophy, poetry. In 1921, Shakespeare and Company moved to its famous location at 12 rue de l'Odéon, across the street from La Maison des Amis des Livres. The bookshop and lending library operated until 1941, when it closed

due to the Nazi occupation of France. Beach's Shakespeare and Company never reopened, but she loaned books privately from her apartment until her death in 1962.[4]

The lending library was a subscription business. Members paid a deposit and recurring fees to maintain access to the lending library's holdings. When Shakespeare and Company opened, members had eight subscription options: they could subscribe for one, three, six, or twelve months, and borrow either one or two books at a time. The least expensive option—one month, one book at a time—cost eight francs. Adding a second book raised the price to twelve francs. The most expensive option—one year, two books at a time—cost eighty francs. Members had to pay a deposit to join—initially seven francs if they chose to borrow one book at a time, fourteen for two. Members of La Maison des Amis des Livres received a twenty percent discount and did not have to pay a deposit, presumably because they had already paid one to Monnier. The deposit was reimbursed when the subscription ended.

Lending library books could be exchanged on the same day or borrowed for up to two weeks—one week for recent publications. Fines for overdue books were ten centimes per day. Borrowing recent periodicals cost an extra fifteen centimes per day. Subscription costs gradually increased over the next twenty years, as did the number of subscription options. Beach introduced various plans or subscription categories, including discounts for students and teachers, and for members willing to forego access to recent publications. Members living outside Paris had to pay ten percent extra and cover postage costs for mailing books. Members could purchase "day-by-day" memberships, or pay additional fees or "supplements" to increase their privileges.[5] At its height in January 1926, the lending library had 249 active members, who had access to thousands of books.[6] Members included American and English expatriates, including artists and writers, and businessmen and their families; French intellectuals; students; scholars; teachers; tourists.

3. Archival Sources

3.1. Collections

The data sets include information from archival collections at three libraries: the Sylvia Beach Papers at Princeton University Library Special Collections, the Joyce Collection at the University at Buffalo Libraries, and the Carlton Lake Collection at the Harry Ransom Center at the University of Texas at Austin.

In 1964, Princeton acquired Beach's papers and book collection.[7] Beach had connections to the town and university: her father, Sylvester, had been a minister at the First Presbyterian Church of Princeton, and she had worked at the university as a young adult. Her papers are extensive: 180 boxes, over seventy-eight linear feet. The papers include correspondence; lending library records and other business material (publisher receipts, tax returns, *Ulysses* business); manuscripts and typescripts from Joyce and others; drafts of Beach's memoir; photographs; and much else. Beach was an obsessive record keeper: she kept notebooks to track almost every aspect of her life, from routine chores (groceries, haircuts) to the distribution of *Ulysses*.

Beach's papers at Princeton are by far the most important source for the data sets and the *Shakespeare and Company Project* as a whole. But documents from the University at Buffalo and the Harry Ransom Center also play a role. In 1959, Beach sold much of her Joyce material to the University at Buffalo. In 1962, after her death, Buffalo acquired additional Joyce material from her estate. The James Joyce Collection, which includes additional acquisitions, comprises eighty-nine boxes—over twenty-two linear feet. In 1986, the Ransom Center acquired Maurice Saillet's Collection of Sylvia Beach and Shakespeare and Company—four boxes, almost two linear feet. The Saillet collection includes material related to Joyce, Shakespeare and Company, and Beach's personal life. The *Shakespeare and Company Project* only displays digital editions of material from Princeton on its website.

3.2. Lending Library Cards, Logbooks, Address Books

The data sets rely on information from three sets of sources: lending library cards, logbooks, and address books.

The lending library cards provide information about members and book circulation. (The *Project* uses "book" as a catch-all term: periodicals as well as articles, photographs, phonograph records, and prints circulated in the lending library.) The cards are 12 cm by 17 cm, and usually include the lending library member's name, address or addresses, membership dates, and membership duration and type— "1m1v," for example, stands for one month, one volume at a time. The cards also generally include the member's borrowing activity in three columns: the first column includes the date a book was borrowed; the second, the book's title; and the third, the date the book was returned. (Beach or one of her assistants would occasionally write "BB"—"brought back"—in the third column instead of a date.) The cards also often include notes about the fate of specific books and the activities of specific members. (See figure 1.) Beach kept separate records for book purchases at Shakespeare and Company, but occasionally recorded purchases and consignments on the cards. On Caresse Crosby's cards, for example, Beach lists five Black Sun Press books that Crosby consigned to Shakespeare and Company. (Crosby was the co-publisher of the press.) The purchases listed on the lending library cards do not offer a representative portrait of book sales at Shakespeare and Company.

The *Project* has identified cards or sets of cards for 654 lending library members. Some members have only one card. The most prolific member, Alice Killen, has twenty-seven: twenty-five of which are double sided. Killen's cards include information about 1,480 borrows. The earliest card in the data set is from November 17, 1919 and the latest is from June 28, 1962. Of the 654 cards or sets of cards, all but six are at Princeton. (Many Joyce family cards are at Buffalo.) The vast majority of the Princeton cards are in box forty-three of the Beach Papers: the *Project* displays digital editions of the entire contents of box on its website.[8]

Figure 1. Ernest Hemingway's Lending Library Card from 1925, Annotated

The logbooks provide information about daily revenue and expenses at Shakespeare and Company. The *Project* uses the logbooks for information about membership "events." The logbooks are organized by date, and generally include a member's name, subscription duration and category, fee, and deposit and reimbursement amounts. (See figure 2.) The logbooks also include information about supplemental fees. (Revenue usually appears on the verso page and expenses on the recto.) Beach or one of her assistants often indicate whether a membership was a new subscription or a renewal, although these indicators are not always reliable: renewals are occasionally marked as new subscriptions and vice versa. (New subscriptions usually include a deposit, but not always: members of La Maison des Amis des Livres were not required to pay a deposit, and Beach did not always require deposits

from friends or well-known writers.) The *Project* makes use of sixty-six separate logbooks, which include information about over 13,039 subscription events. Sixty-five logbooks are part of the Beach Papers; the other is held by the Harry Ransom Center and includes reimbursements that Adrienne Monnier paid lending library members after Shakespeare and Company closed. There are multiple versions of some logbooks—rough and fair copies—and gaps in logbook coverage, especially in the early and late 1930s.[9] The *Project* provides access to digital editions of all the logbooks at Princeton via a supplementary site hosted by Princeton University Library.

Figure 2. A Logbook Page from February 1922, Annotated

The address books provide information about lending library members: names, addresses, and membership dates and payments. (See figure 3.) The *Project* uses two address books from the Beach Papers. One address book includes information for 2,706 members who joined the lending library before 1935. The second features information for 321 members who joined after 1935. The first address book is a good example of Beach's meticulous, yet inconsistent record-keeping practices: it carefully lists members in alphabetical order and then stops midway through the letter "N."

Figure 3. The First Page of the Pre-1935 Address Book

3.3. Completeness, Complementarity, Consistency

As the above descriptions make clear, the *Project*'s archival sources are incomplete. The *Project* relies on cards for 654 members, but the logbooks and address books reveal that the Shakespeare and Company lending library had thousands of members. The *Project* relies on logbooks for daily lending library transactions, but there are gaps in logbook coverage. One of the two address books omits members with last names from the second half of the alphabet.

The sources are incomplete in other ways as well. The cards only list books by title. As a result, identifying books can be difficult. For example, the *Project* team is uncertain whether "Three Loves" refers to Max Brod's *Three Loves* (1929) or A. J. Cronin's *Three Loves* (1932). (All the borrowing events are from after 1932, suggesting that Cronin's *Three Loves* circulated in the lending library.) The cards and the address books occasionally include incomplete dates—a year and a month, or just a year or a month, or no date at all. Moreover, the cards and address books occasionally include incomplete addresses—hotels without street addresses, home addresses without street numbers. All three sources refer to married women by their husbands' names, and do not always or even often include first names. Finally, the handwriting of Beach and her assistants can be difficult to decipher. When possible, the *Project* team compensates for missing data using other sources in the Beach Papers and beyond: a notebook in the Beach Papers, for example, lists acquisitions to the lending library, and includes both titles and author names; the contemporary directories *Americans in France* supply first names of some lending library members.

Yet all three archival sources are also complementary. Sources substantiate one another, while compensating for gaps and absences. Most significantly, the logbooks and address books reveal information about members without extant cards. The address books and cards also supply membership information during gaps in logbook coverage; and the address books supply first names that are not in the logbooks or cards. Together, the three archival sources provide a near-complete portrait of the Shakespeare and Company lending library.

On rare occasions, sources are inconsistent. The most common discrepancy concerns the use of English and French personal titles or honorifics. The same member is often listed as both a "Miss" and a "Mlle." Occasionally, a member is listed as both a "Miss" and a "Mrs." (Beach and her assistants might have chosen titles based on assumptions about a member's age.) Another discrepancy concerns subscription and renewal dates on the lending library cards and the logbooks. Beach or her assistants would often back or forward date subscriptions and renewals on the lending library cards. For example, if a member's one-month subscription expired on March 1, 1936, and the member renewed on February 27, 1936, Beach or an assistant would start the new subscription on March 1, 1936 and include that date on the card. In such cases, however, the logbooks indicate the actual purchase date. The events data set includes membership start and end dates, and purchase dates—for more information about membership timelines refer to section 4.5 of this article.

When evaluating the *Project*'s archival sources, researchers should keep in mind that the sources reflect the work of a small and overwhelmed lending library staff, keeping records to facilitate the day-to-day operation of a business, not the work of future literary historians and data scientists. Why would Beach record the author of "Three Loves" when she knew exactly which book she had available to borrow? The archival sources also reflect the impact of time and the vicissitudes of interwar and wartime France. (Beach was interned for six months during World War II.) Additional factors affect the completeness of the sources as well. Beach would occasionally sell or give away archival materials. The Beach Papers themselves are imperfectly organized. For example, lending library cards have been discovered in miscellaneous folders in the Beach Papers.

4. History of the Data Sets

4.1. TEI: Structure

Understanding how the data sets were created is important for understanding their functionality and potential. The *Project*'s sources are not uniform. Standardizing and formalizing the information they contain was a complex, labor-intensive process. Working on the *Project*, the *Project* team confronted a series of challenges, and did

its best to identify solutions that respected Beach's work and the *Project*'s limited budget. From the beginning, the *Project* team recognized the value (and inevitability) of embracing a process of trial and error. The data sets are the result of that process.

The creation of the data sets began with a transcription and encoding project. Clifford Wulfman led a team of Princeton students as they transcribed and encoded the lending library cards, following a customized Text Encoding Initiative (TEI) schema. Soon a second team began to transcribe and encode information from the logbooks. This work led to three sets of TEI files. The first set included a complete transcription and encoding of the lending library cards. The second, a partial transcription and encoding of the logbooks, focusing exclusively on lending library information. The third, a personography of the people referenced on the cards and in the logbooks, using TEI tags to document names, birth and death dates, nationality, addresses, etc. Local identifiers were used across all three sets to link people. The *Project* team made the decision to partially transcribe and encode the logbooks for practical reasons: there wasn't sufficient funding or time to support full transcription and encoding.

4.2. TEI: Research

Work on the personography involved significant research. The *Project* team matched member names to Virtual International Authority File (VIAF) numbers, and identified longitude and latitude coordinates for the addresses. The *Project* team also corrected and standardized names and addresses, and began adding basic personal information about members. Finally, the *Project* team developed gender and nationality guidelines to support this research. Nick Budak's "Representing Gender in the Shakespeare and Company Project" (2019) provides an account of how and why the *Project* represents gender.[10]

During this stage of the *Project*, OpenRefine was used to group book titles in the lending library card files. (Beach and her assistants would often refer to the same book by different titles—"Portrait of the Artist," "A Portrait of the Artist," "A Portrait of the Artist as a Young Man.") This work resulted in the addition of title identifiers to the lending library card files.

4.3. TEI to Relational Database

In 2016, the *Project* team assessed the TEI and determined that a relational database would be better suited for connecting and analyzing information from the archival sources. In 2017, Jean Bauer designed a preliminary relational database structure. (See figure 4.) The structure added something that was previously implicit: "accounts" to connect members and events. This structure allowed the *Project* to accommodate cases where two members—for example, life partners, family members—shared an account.
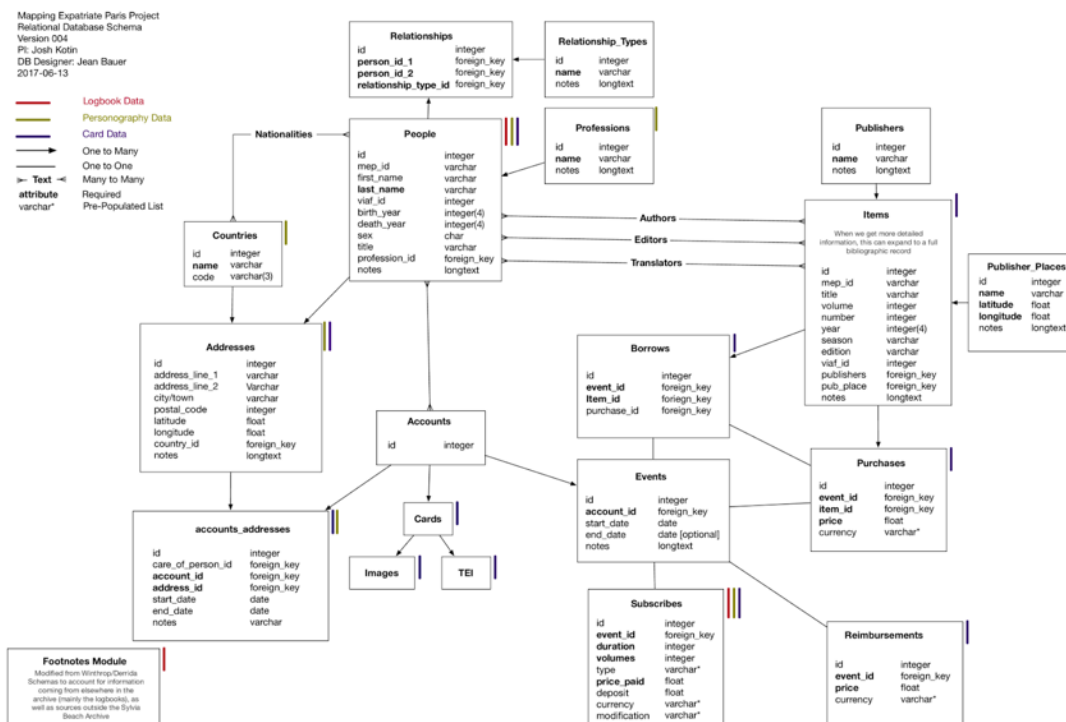

Figure 4. Initial Database Diagram, 2017

In 2017 and 2018, the database structure was iteratively implemented and refined using a series of scripts to extract data from the TEI documents. The database and scripts were implemented in the Python programming language, using the Django web framework. Data was progressively imported into the database, starting with the simplest data—the personography—and then the logbook and card data.[11] (See figure 5.) As each phase of import was completed, data work shifted from the TEI

to the administrative interface of the database. Eventually, the TEI data was documented as in-progress work toward the *Project* and archived.
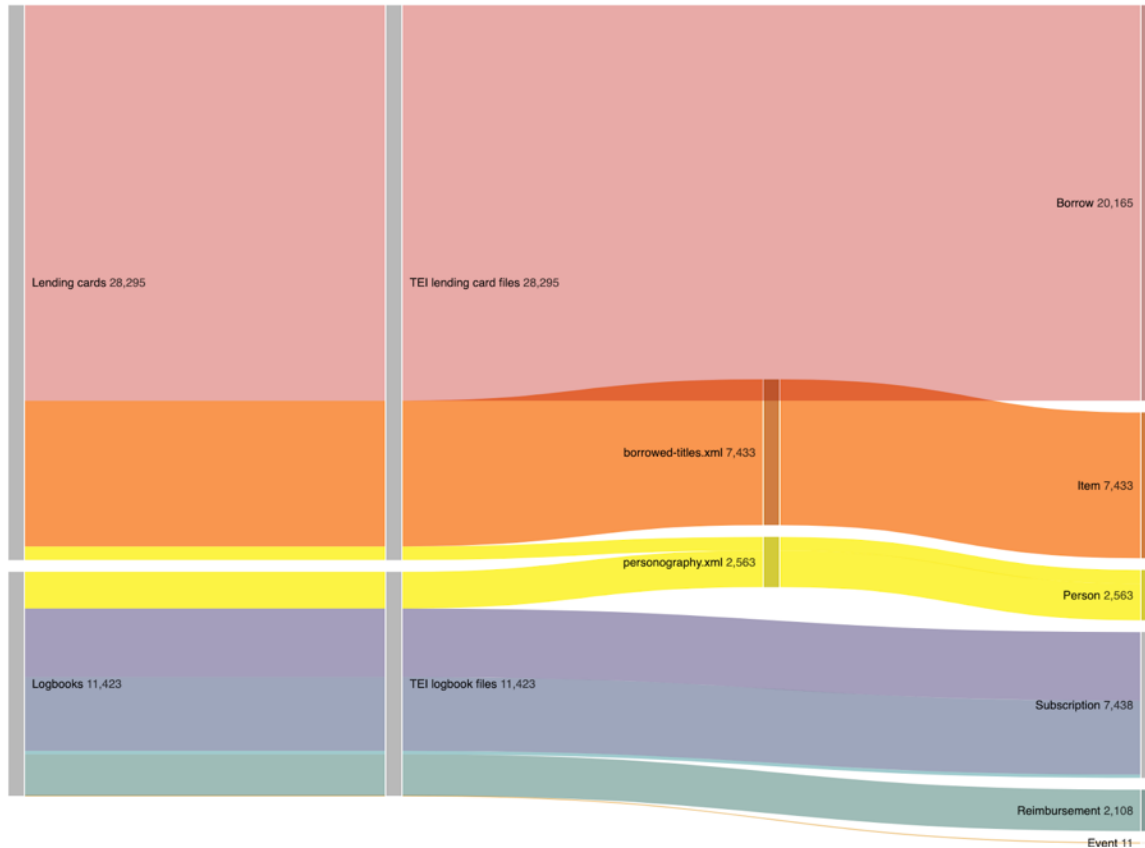

Figure 5. Flow of Information from Archival Souces to TEI Files to Relational Database

4.4. Relational Database: Structure

The database was designed around three modules, which would later define the structure of the data sets: people, books, and accounts. The people module stores information about members: name or names, gender, title, birth and death years, relationships, profession or professions, nationality or nationalities, and VIAF number. (See figure 6.) Gender, nationality, and profession are based on controlled lists. (The *Project* does not currently use the relationships or professions feature.) There is a Boolean field to differentiate "members" identified as organizations, rather than individuals. (The journal, *Nouvelle revue critique*, had a membership, for example.) The module can also store one or more informational URLs (such as links

to Wikipedia pages). The *Project* uses "slugs" to create URIs for people. The URIs are used for access on the *Project* website and as identifiers in the data sets. The database tracks past slugs for modified person records to resolve changed URIs. The person table also includes a datetime field that automatically updates whenever the main record is saved.
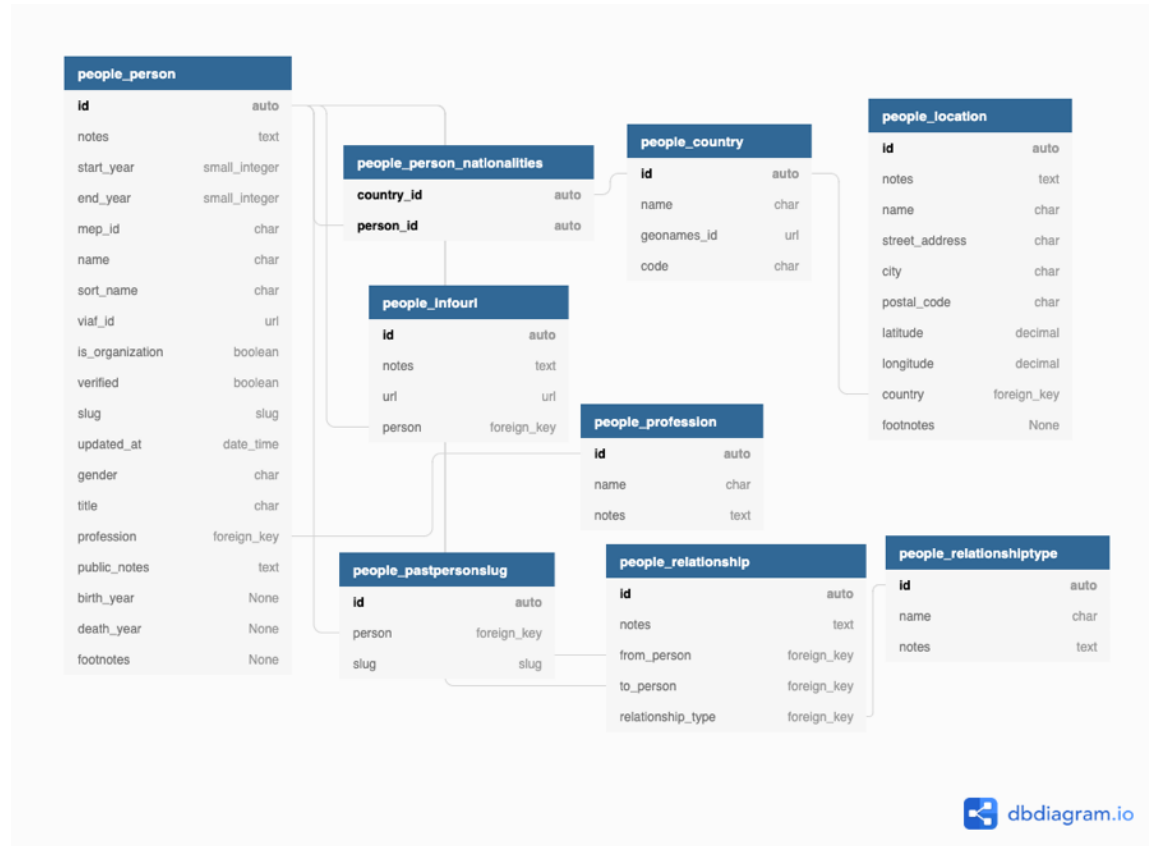

Figure 6. Database Diagram for People Module

The book module stores books as works. (See figure 7.) Works can have a format—book, periodical, photograph, etc.—and subjects and genres. Works can also have specific editions. The *Project* uses the editions feature to document volumes of multivolume works and issues of periodicals, not the specific editions of books that circulated in the lending library.[12] The module allows for a customizable list of creator types: author, editor, etc. Creators are also stored in the people module and associated with VIAF numbers, enabling the publication of a creators data set in the future. The books module can also store links to eBooks, which allow visitors to the
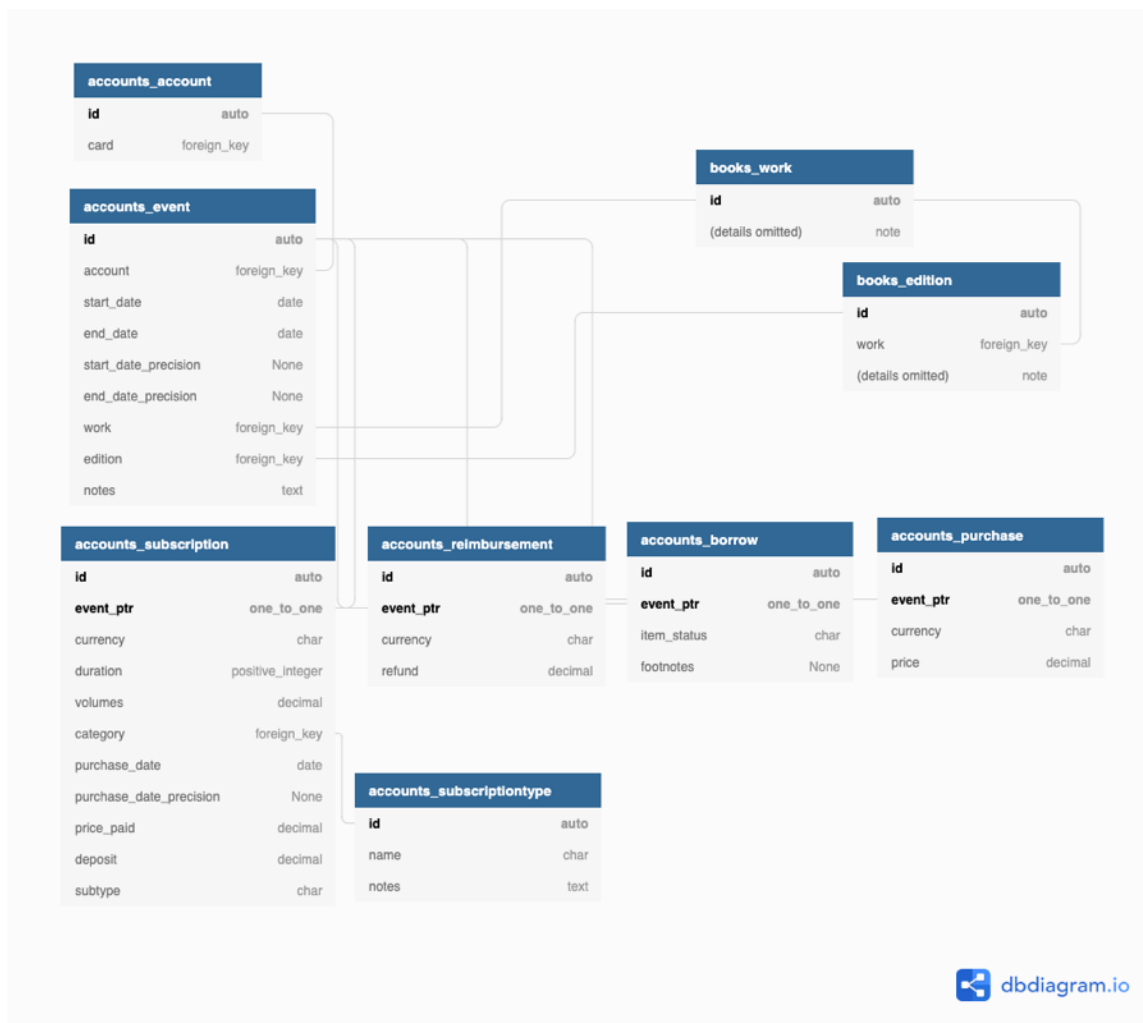
15

Project website to read the books that circulated in the lending library. The *Project* uses slugs for books in the same way it uses slugs for people.

Figure 7. Database Diagram for Books Module

The third module of the database—accounts—connects people and books, and stores information about library member events. (See figure 8.) Events are implemented using Django's multi-table inheritance, which means that there is a base table with common event fields: associated account, dates, etc. Event types—subscription, reimbursement, borrow, purchase—store information specific to that event in a table, and each record is linked to a base event with the common event fields. Subscriptions have a type or "category," which is used to track the different subscription plans, and a "subtype" used to distinguish new subscriptions, renewals, supplements, and separate payments.

Figure 8. Database Diagram for Accounts Module

All events support partially known dates. Date precision is stored in flags alongside the dates, and are used to calculate and display the date accurately based on the known portion.[13] (Initially, the *Project* only supported partially known dates for borrowing events. But the *Project* team eventually discovered the importance of supporting partially known dates for subscription events as well.) All events can be linked to a work and optionally to an edition. These links are stored on the base event because there are a few book-related events—gifts, requests, etc.—that do not fit the other event types. In some cases, these irregular event types are documented in a notes field.

People who are members of the lending library have associated accounts. The relationship between people and accounts was implemented as a many-to-many

relationship based on the original database design. In practice, however, no member should have more than one account. Accounts connect to addresses, which include both street addresses and latitude and longitude coordinates. Accounts use footnotes to link to bibliography records for the lending library cards. The bibliography records may have associated International Image Interoperability Framework (IIIF) manifests, if digital editions of the cards are available in that format and able to be displayed.

4.5. Relational Database: Research

After the migration to a relational database, the *Project* team worked to augment and refine the data about members, books, and events in the established fields. The *Project* team also developed a set of conventions for identifying members and books, and representing gaps, errors, and ambiguities in the archival sources.

Work on the lending library members developed on two fronts. First, the *Project* team attempted to identify and "merge" members with multiple accounts. In the logbooks, members appear each time they join the lending library, renew their memberships, and receive a reimbursement. After the migration, these members appeared multiple times in the database, associated with multiple accounts. To decide whether members with the same name were, in fact, the same person, the *Project* team developed a set of "merging guidelines." Most basically: the team merged accounts only if they belonged to members with the same name, and the event in one account clearly fit an event in the other account. "Fit" describes contiguous subscriptions and renewals, and subscriptions with deposits that match reimbursements. Consider the case of "Miss Wilson." The name appears in the logbooks five times, associated with subscription events on November 9, 1923; April 10, 1926; August 20, 1927; September 7, 1927; and September 14, 1927. In the database, Miss Wilson appeared five times as well, associated with five separate accounts. The *Project* team inferred that at least two of these events belonged to the same Miss Wilson: on August 20, 1927, she joined the lending library for one month, paying a 100 francs deposit, and on September 14, 1927, she received a 100 francs reimbursement. Accordingly, the team merged the two accounts. The team, however, wasn't certain about how the other events were connected, so kept them separate. Consequently, there are now four Miss Wilsons in the database, instead of

five. (The *Project* would rather leave members unmerged than merge them incorrectly.) The 1.2 data sets include 5,235 members. That number surely counts some members multiple times. (For example, the four remaining Miss Wilsons may in fact be the same member.) The *Project* invites researchers to help identify members and suggest merges.

Second, the *Project* team developed naming guidelines to complement the gender and nationality guidelines. Our aim was to provide the full name of each lending library member as well as the name in the archival sources. Accordingly, the *Project* represents Katherine Anne Porter as "Katherine Anne Porter" and "Mrs. Katherine Anne Pressly," and Claude Cahun as "Claude Cahun" and "Mlle Lucie Schwob." (See figure 9.) The *Project* also lists variant (and often incorrect) spellings of names that appear in the sources. Lincoln Steffans, for example, is also listed as "Lincoln Stephan"; and Marie Boscq is also listed as "Marie Bosq."

Figure 9. Database Page for Katherine Anne Porter

Work on the books was even more complicated. The lending library cards list books by title only. After the migration, the database was populated by thousands of titles without author or publication information. Starting in 2018, the *Project* team worked to match these titles to author names and dates of first publication. (The team used dates of first publication because of the challenge of identifying the specific editions that circulated in the lending library.) For books in English translation, the team used the date of the book's first publication in English. For books written before the

advent of the printing press—Plato's *Republic*, Augustine's *Confessions*—the team left the publication date blank.

The *Project* team has identified ninety-two percent of the 6,020 items that circulated in the lending library, according to the extant lending library cards. The remaining eight percent are label "uncertain." These uncertain books could not be adequately identified for one of two reasons. Either the title was "generic"—that is, it could refer to multiple books: for example, "Three Loves" and "Katherine Mansfield." (Beach and her assistants would often refer to biographies, and collected and selected editions by their subject's or author's names.) Or the title cannot be matched to any book: for example, "12:30 from Heaven," which Darsie Rutherford Gillie borrowed in 1937. Books labelled "uncertain" in the book data set connect to an explanatory note in the notes field. For "12:30 from Heaven," the note suggests that the book might be Freeman Wills Crofts's *The 12:30 from Croydon* (1934).

The *Project* team had to do other work with the books. For instance, the team occasionally had to "demerge" and "merge" books. Due to the work with OpenRefine during the TEI stage, some book titles were incorrectly matched in the database and some still needed to be matched. For example, Gertrude Stein's *Three Lives* (1909) was incorrectly conflated with "Three Loves." To demerge a book, the team created a new book record and re-associated events from one record to the other. To merge books, the team re-associated events and deleted the duplicate record. A notes field in the database tracks this work, but it does not appear in the data sets. (See figure 10.)

Figure 10. Database Page for Gertrude Stein's Three Lives

As the *Project* team manually corrected book records, Koeser implemented and ran a script to match the books with OCLC entries using the WorldCat Search API. If a match was found to all the search criteria, she used it to update the record with an OCLC Work URI, sample edition URI, format type, genres and subjects. In spite of this effort, however, the OCLC data proved insufficiently reliable. (OCLC has discontinued using Work URIs.) Thus, the *Project* only includes information collected about format in the data sets, and much of this information had to be corrected manually.

Finally, the *Project* team had to identify specific volumes and periodical issues, and, when possible, add links to full text versions of books. To identify volumes and periodical issues, the team used information on the lending library cards. For full text versions, the team linked to versions available on the Internet Archive. In rare cases, the team linked to other digital archives or specific digital editions. For example, a link to the Blue Mountain Project provides a complete run of the journal *Broom* (1921–1924).

Work on the events focused on the membership timelines of individual members. As discussed in section 3.3, a member would occasionally renew a membership before it was set to expire. Initially, the *Project* conflated purchase and start dates, resulting in truncated membership timelines. To address this problem, the *Project* team revised the database to differentiate between the two, and applied a data migration

to adjust start dates for renewals purchased before the previous membership period expired. Consider the case of Frances and Morrill Cody, who were members between 1924 and 1926. (The Codys were married and shared an account.) The Codys joined the lending library for six months on July 7, 1924, and then renewed for another six months on November 11, 1924. In the database, the renewal has a purchase date of November 11, 1924; a start date of January 7, 1925; and an expiration date of July 7, 1925. (See figure 11.) Members would also renew *after* their previous membership expired. The *Project* team does not backdate the start dates of these renewals unless there is clear evidence that Beach did so. As a result, some membership timelines might be slightly too long.

Figure 11. Purchase and Start Date of a Renewal by Frances and Morrill Cody

Work refining and augmenting the data about members, books, and events is ongoing. Between the first and third versions of the data sets, the *Project* team added information from the address books—addresses, first names, events. The team also improved and corrected hundreds of member and book records. Much of this work was aided by visitors to the *Project* website. The credits page acknowledges their assistance.

5. Data Sets

5.1. Members

The members data set includes information about 5,235 lending library members in the following fields: member URI, member name, sort name, title, gender, individual or organization, has lending library card, birth year, death year, membership years,

VIAF URL, Wikipedia URL, nationalities, addresses, postal codes, arrondissements, longitude and latitude coordinates, notes, last updated.

Member names appear in two fields. "Member name" includes the member's full name and (if different) the name in the archival sources. The "sort name" is the member's full name, with the last name first to facilitate alphabetization. The "title" field includes titles from the archival sources. At an early stage of the *Project*, the *Project* team hypothesized that titles might be a clue to nationality: French titles (Mlle, Mme, M.) for French members, and English titles (Miss, Mrs., Mr.) for non-French members. This hypothesis proved incorrect: the language of the titles—and even the presence of titles—is largely arbitrary. Due to the way names from the database appear on the *Project* website, the *Project* team does not include titles for members who were also "creators." If the team were to include titles for these member-creators, the author of *The Sun Also Rises* (1926) would appear on the *Project* website as Mr. Ernest Hemingway, instead of Ernest Hemingway. Titles appear in the name and sort name fields of members without known first names or first initials.

Of the 5,235 members, 564 have extant lending library cards. The information on the cards facilitates identification. In version 1.2 of the data sets, the *Project* team has been able to link 520 members to VIAF and 358 members to Wikipedia. (Thirteen members have a Wikipedia page, but no VIAF number.) (See figure 12.) Significant obstacles to identifying members include partial names, missing addresses, and, most obviously, the fact that most members were not public figures. (Shakespeare and Company was not only—or even primarily—a community of writers and artists.) The *Project* team has not yet attempted to identify members using public records.
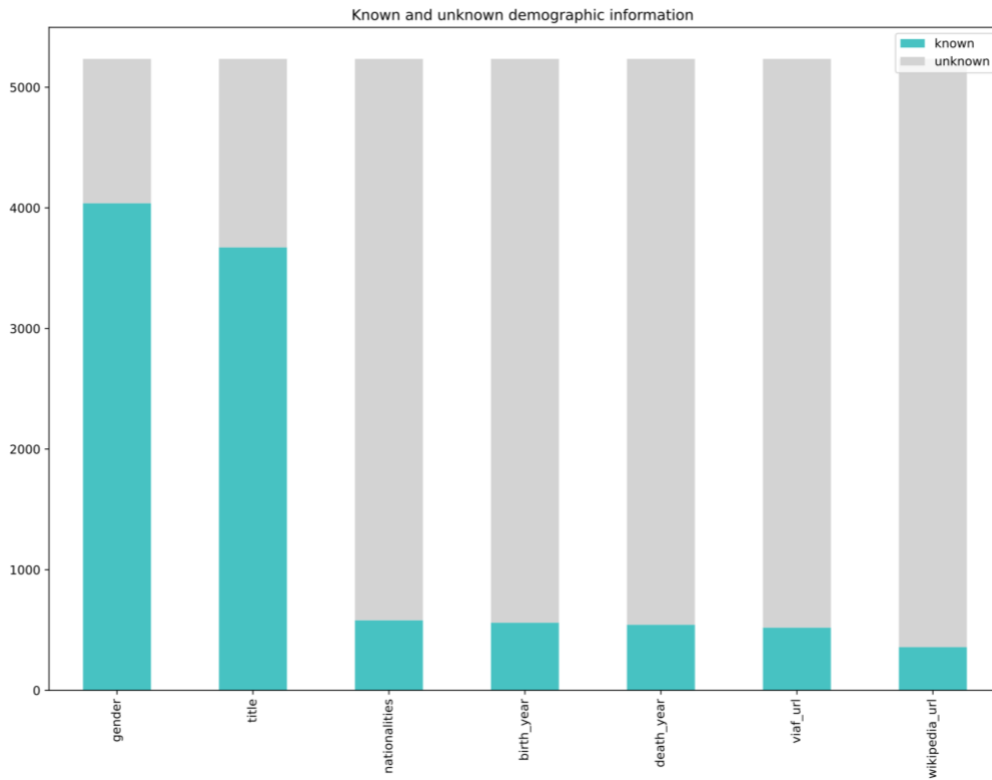
Figure 12. Members with Known Values for Demographic and Informational Fields

To facilitate analysis without the events data set, the members data set includes the years a member was active in the lending library. There are fifty-eight members with no membership years in the data set. The archival sources for these members do not include membership years, and the *Project* team has been unable to accurately date their activities. Consider the case of Louis Aragon. He has one lending library card, which indicates that he borrowed two books by Joyce on October 11. The card does not include a year. The card is likely from 1921—Aragon first read installments of *Ulysses* earlier that year—but the *Project* team has not yet been able to confirm the date. 1920 is also a possibility. Or consider the case of Julia Wright, Richard Wright's daughter. As a child, in the late 1940s and early 1950s, she visited Beach's apartment with her father. A single card lists four books in her own handwriting and includes the date, June 27. The *Project* team contacted Wright, but she could not remember the exact year. (Wright is the only living member of the lending library.) Other members without years derive from the address books, which do not always include membership dates.

The data set includes member addresses. These addresses are all from *Project* sources. The *Project* team corrected—but did not supplement—addresses using secondary sources. Some addresses do not have latitude and longitude coordinates. Many of these are hotels: Beach included the hotel name, but not a street address, and the *Project* team has been unable to pinpoint a specific location.

The data set includes a notes field, which displays information about sources and shared accounts. This information is displayed on the *Project* website as well. The "updated" field includes a timestamp, which automatically updates when a member record is modified. The timestamp does not update, however, when an event associated with a member's account is modified.

Between the 1.0 and 1.2 versions of the members data set, the *Project* team reduced the total number of members 5,726 to 5,235 by merging records belonging to the same members and removing mistaken records. The team also added 1,055 addresses from the sources. The database maintains a record of old identifiers after member records are merged, but this information is not included in the members data set.[14] The status of a member from the previous data set who is no longer available in the current version can always be determined by performing a HTTP GET request on the member URI. If the response is a 301 Moved Permanently, then the member record was merged with another member, and the new location is the new member URI. If a 404 response is returned, then that member record was removed.

5.2. *Books*

The books data set includes information about 6,020 books in the following fields: book URI, title, author, editor, translator, illustrator, introduction, photographer, year of publication, format, uncertain, eBook URL, volumes/issues, notes, event count, borrow count, purchase count, circulation years, last updated.

Ninety-two percent of the 6,020 items that circulated in the lending library were books. Seven percent of the items do not have a format listed—these items have not been adequately identified. The majority of the remaining items were periodicals. Two photographs, one article, and one print also circulated. The *Project* team

identifies multivolume works published in the same year as one item—for example, Thomas Mann's two volume novel, *Buddenbrooks* (1901, translated 1924). Periodicals also appear as one item. Multivolume works published over multiple years appear as separate items—for example, the first twelve volumes of Dorothy Richardson's thirteen volume, *Pilgrimage* (1915–1967).

The books data set includes a Boolean field indicating whether an item can be adequately identified—475 items, eight percent of the total, are marked as uncertain. (Section 4.5 of this essay describes the difficulty of identifying items. Some items marked as uncertain have identifiable formats—hence the discrepancy between the eight percent of items marked as uncertain and the seven percent without a listed format.) A "notes" field in the data set provides likely identifications. These notes are also visible on the *Project* website.

The data set also includes 3,155 links to digital editions of the books that circulated in the lending library. These links are available on the *Project* website and are intended to encourage visitors to read the books that circulated in the lending library. The links—most of which are to the Internet Archive—may have limited value as data for computational analysis.

To facilitate analysis without requiring use of the events data set, the books data set includes summary information about the events associated with each title: the total number of events, total borrows and purchases, and a list of years that a book circulated. The number of borrows can be used to reveal the most popular books and other formats. Joyce's *A Portrait of the Artist as a Young Man* (1916) was the most frequently borrowed book with fifty-six total borrows. Kotin and Koeser's "Shakespeare and Company: Top Ten Lists" (2020) gives an overview of the most popular books and periodicals.[15] As discussed in section 3.2 of this essay, researchers are strongly encouraged to focus on borrowing events (as opposed to total events) when analysing book data.

There are thirty-five books without circulation years in the data set. Each book only has one event, and those events do not have known years. Like the members data set, the books data set includes a timestamp which is automatically updated whenever a book record is modified.

Between versions 1.0 and 1.2 the books data set, the *Project* added nine books and corrected numerous book records. Changes were often minor: adding the name of the author of an introduction, correcting a date of publication. The books data set only includes books (and other items) listed on extant lending library cards. The Beach Papers, however, contain catalogs of lending library holdings, which include items not listed on the cards. The *Project* may eventually use these catalogs to create a more comprehensive database of lending library holdings.

5.3. Events

The events data set includes information about 35,523 lending library events in the following fields: event type, start date, end date, member URI, member name, sort name, subscription price, deposit amount, duration, duration in days, volume limit, category, purchase date, reimbursement amount, book borrow status, borrow duration in days, book purchase price, currency, item URI, title, volume, author, year of publication, notes, source type, source citation, source manifest, source image. The events data set includes high level information about members and books to facilitate analysis without requiring use of the other data sets.

Every event has a type—either a subscription event type (subscription, renewal, supplement, reimbursement, separate payment) or borrowing event type (borrow, crossed out, generic, gift, periodical subscription, purchase, request.) Ninety-four percent of all borrowing events are, in fact, borrows. The remaining six percent involve other kinds of events, including events that are crossed out on the lending library cards. (Occasionally, a member would decide to borrow a different book at the last minute and Beach or one of her assistants would cross out the title listed on the card.) Researchers are encouraged to examine the digital editions of the lending library cards to learn about the range of "generic" events. "Separate payment" refers to an uncommon event: a member visits Shakespeare and Company to complete a payment initiated by an earlier subscription event. (For example, a member pays for a subscription on one day and the subscription deposit on another.) Supplements are usually payments to increase borrowing limits—for example from one volume at a time to two. The *Project* lists the number of extra volumes paid for by a supplement in the volume field. Occasionally, supplements extend subscriptions.

Every event has a start and end date, although one or both may be missing, and dates that are present may be only partially known. Unknown dates are indicated by the empty string in the start or end date field. Single-day events, such as purchases and reimbursements, are represented by the same date in the start and end date. Other events, such as subscriptions and borrowing events, have duration, and use different dates to indicate the duration of the event when known. Event dates are in ISO 8601 format. The events data set includes 383 events with no date; 159 events with year only; 494 events that only include year and month but no day; and 55 events with a day and month but no year expressed as --MM-DD. In the 1.2 data sets, 290 reimbursements have partial or no dates, and no reimbursement amount. Information about the vast majority of these reimbursements come from the address books, which often simply indicate whether a member received a reimbursement or not, and do not include a date.

Analyzing the number and distribution of events illuminates both the operations of Shakespeare and Company and the completeness of the Beach Papers. The Beach Papers include a greater proportion of lending library cards from the 1930s than the 1920s. As a result, the number of overall events in the data set reaches its peak in 1939. (See figure 13.) The small number of events after 1941 reflects the fact that Beach continued to loan books after Shakespeare and Company closed.
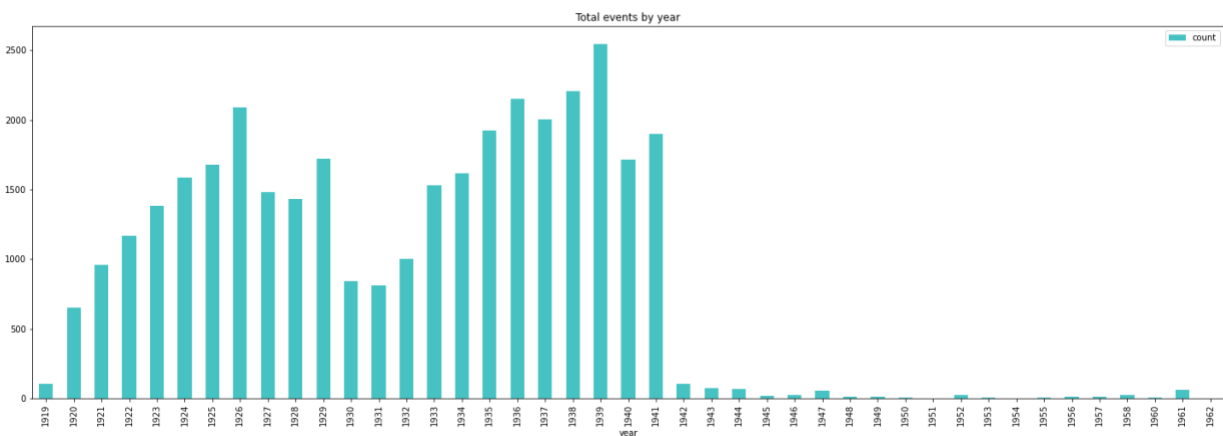

Figure 13. Total Events by Year, 1919–1962

28

Focusing on subscription events alone, however, reveals that Shakespeare and Company had the greatest number of active members—249—in 1926, at the height of the roaring 1920s. (See figure 14.) This discrepancy is due to the incompleteness of the *Project* sources. The small number of members in 1942 reflects still active subscriptions that were purchased in 1941.
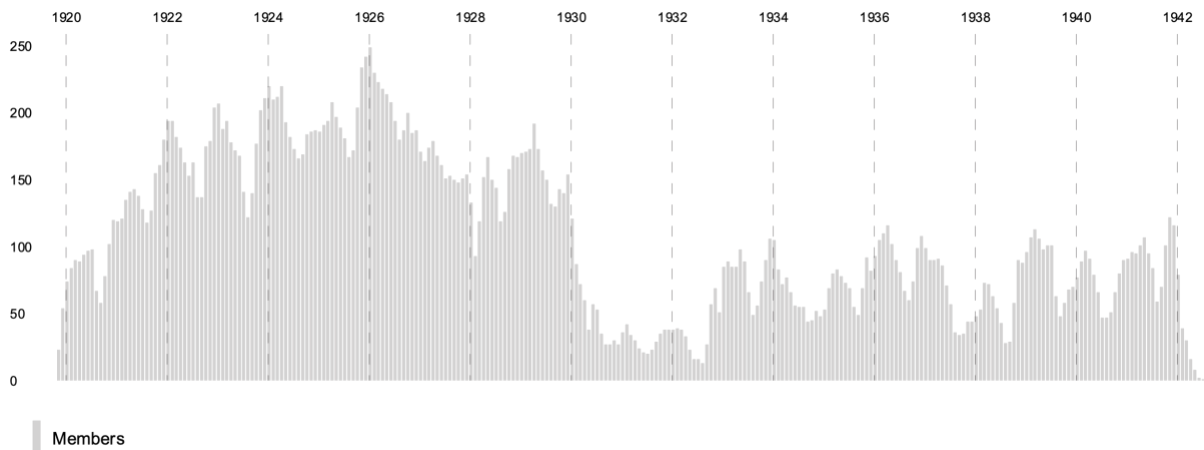
Figure 14. Total Membership by Month and Year, 1919–1942

Every event is associated with one or two library members—two in the case of shared accounts. (The *Project* includes 49 shared accounts.) Member information is provided in three forms: a URI, for connecting the data set with the members data set and for accessing the member page on the *Project* website; a display name; and a sort name.

Subscriptions, renewals, and supplements have a "subscription price" and a "subscription deposit." Subscription reimbursements have a "reimbursement refund amount." Book purchases and periodical subscriptions have a "purchase price." The currency field indicates the currency of all payments. The vast majority of transactions—99.95%—were in francs, with only a handful of transactions in British pounds or US dollars. The *Project* includes the currency field to avoid introducing errors via conversion. Missing amounts in these fields reflect missing amounts in the *Project*'s sources.

Subscriptions and borrows include duration information when it can be calculated. Duration is calculated as a relative delta between the start date and the end date, including the start but not the end date in the total—or half each of the start and end dates. Durations are only provided when both start and end dates are known or if both start and end date have a known month and day but no year. In the latter case, dates are assumed to be in the same or subsequent years. The data set also includes a human-readable duration which is used on the *Project* website and in the administrative interface. Researchers should use the day fields provided for calculations. For completeness, however, the following explains the readable duration: any duration less than twenty-eight days is expressed as weeks (for multiples of seven) or days; for longer durations, timedelta between start and end dates is expressed as a combination of years, months, and days.[16]

The events data set also includes information about the "volume limit" of subscriptions (that is, the number of volumes included in the subscription); the subscription category (or plan); and subscription purchase date, since purchase dates and start dates may not be identical. Beach would occasionally allow members to borrow more than their permitted number of volumes at one time. In the 1.2 datasets, 1,035 subscriptions and renewals do not indicate a volume limit. Plans range from free memberships to "day-by-day" memberships to discounts for students, teachers, and members of La Maison des Amis des Livres (abbreviated as AdL).[17]

A "borrow status" field indicates how each borrowing event ended, if known. The field documents information provided on the lending library cards that might not otherwise be apparent: if a borrowing event has no known end date, a book might still have been returned, bought (thirteen events), or missing (four events). All borrowing events with a known end date are also marked as returned. Of the 21,059 borrowing events in the data set, 613 (three percent) have an unknown status.

As of the 1.2 release of the data sets, all events include source information; these fields may be plural, reflecting the fact that some events are documented in multiple sources. (See figure 15.) Source information includes the type of source (lending library cards, logbooks, address books); a citation for the particular item or, in the case of the logbooks, group of items; and when possible, IIIF manifest and image URLs, which provide a more specific location and access to the digital edition.
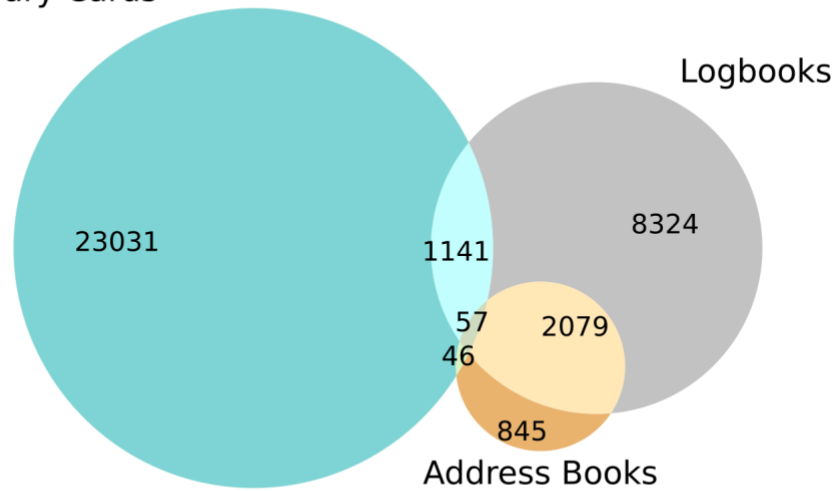
Figure 15. Events by Source Type

The sources from the Beach Papers have been digitized by Princeton University Library and made available using IIIF Presentation and IIIF Image APIs. The source references for events from lending library cards held at Princeton have been linked to the corresponding digital editions. On the *Project* website, source image references display events with images of the relevant lending library card. The data set includes the IIIF resources as a reference and to encourage researchers to analyze the images alongside the data. The images will not likely be useful for handwriting recognition, since considerable work has been done deciphering, sanitizing, and standardizing the references on the cards.

6. Limitations, Potential, and Future Plans

The main limitation of the data sets is the incompleteness of the *Project*'s sources. A secondary limitation is the *Project*'s scope. The *Project* focuses on the Shakespeare and Company lending library, and thus do not provide a full portrait of Beach's bookshop or publishing activities. The financial information in the data sets does not provide an account of the overall business of Shakespeare and Company.

Yet these limitations are minor in light of the *Project*'s potential. The fact that the *Project* sources exist at all is a marvel. The American Library Association (ALA) has long advised libraries to protect patron privacy by "purg[ing] circulation […] records when they are no longer needed for library operations."[18] Shakespeare and Company was not affiliated with the ALA, and Sylvia Beach did not take this advice. She preserved individual member records, as well as more general circulation and membership information. Her decision creates an array of research and outreach opportunities at the intersection of modernist studies, digital humanities, and public humanities.

The *Shakespeare and Company Project* data sets promise to address major questions in modernist studies. The data sets illuminate the reading practices of individual writers and artists, as well as the connections between and among them. For example, researchers can discover what Stein read, and whether Stein and Joyce (or Stein, Joyce, and Hemingway) read anything in common. In this way, the data sets promise to support biographical research and critical studies of influence and referentiality. The data sets also illuminate the lives of individual books. For example, researchers can track the circulation of *Ulysses* or Virginia Woolf's *Mrs. Dalloway* (1925) or Havelock Ellis's *Studies in the Psychology of Sex* (1897–1928). Researchers can also use the data sets to track the circulation of particular books around Paris—from address to address, neighborhood to neighborhood. The high price of *Ulysses*—150 francs in 1922, about $200 today—made purchasing a copy difficult for many potential readers. The data sets reveal a new readership for the novel.

The data sets also promise to address more general questions in modernist studies. How did the Lost Generation move around Paris between 1919 and 1941? How did Anglophone and Francophone modernists interact? What was the relationship between so-called "high-brow" and "middle-brow" culture? Did modernist classics circulate with bestselling novels—and if so, how? Did members, for example, read Woolf's *The Years* (1937) and Margaret Mitchell's *Gone with the Wind* (1936) concurrently or sequentially? What were the most popular books and periodicals at the lending library? Does the answer challenge how scholars and the general public understand the modernist canon?

The data sets also allow scholars to address questions in literary theory. Does borrowing duration reveal how certain books—and genres—were read? Does a book's circulation years reveal its "staying power"? Is there a connection between proximity and taste, and between class and taste? Researchers could begin to address these questions about taste by examining whether lending library members in the same arrondissement were more likely to borrow the same books than members in different arrondissements. Did men and women read different books? Did men and women have different canons? What about different age groups and nationalities? Can different networks of lending library members illuminate the nature of artistic communities or the power of personal recommendations? Can the reading practices of different communities contribute to the history and theory of reading? Can the circulation of the most popular (and unpopular) books at the lending library contribute to theories of innovation and influence? The questions are endless!

The data sets also create opportunities to test new methods for working with partial information. Archives are almost always incomplete. The complementarity of the *Project* sources, however, provides a clear picture of what the Beach Papers are missing. Time series forecasting and backcasting could be used to compensate for gaps. Analyzing borrowing behavior against subscription limits, for example, might make it possible to calculate the lending library's "borrowing capacity," which might, in turn, help estimate the scale of undocumented borrowing activity. Indeed, a combination of network community detection and feature analysis of full-text works might make it possible to model what members without extant cards read. Such speculative approaches to archival gaps and "silences" are an important area of research.

Finally, the data sets promise to connect communities: scholars of modernism and data scientists, and these overlapping communities with a wider non-academic public. The Lost Generation holds a prominent position in the popular imagination. The exploits of Hemingway and F. Scott Fitzgerald, the lesbian communities on the Left Bank, the salons and cafes of Montparnasse—all serve as material for popular books and movies. The data sets promise to fuel this work, revealing new stories and new histories. A single address can inspire a television mini-series. Consider, for example, Reid Hall at 4 rue de Chevreuse on the Left Bank—the data sets reveal that

twenty-three different members of Shakespeare and Company lived there. This fact, once recognized, is unsurprising: Reid Hall was a center for American women studying in Paris. Yet the fact is still compelling: twenty-three lives, twenty-three stories—all connected by Shakespeare and Company, which was a twenty-minute walk away through the Luxembourg Gardens. The *Project* generates thousands of such stories.

In the future, the *Project* team plans to release address and creator exports to facilitate further research. The creator export will allow for demographic analysis based on creators of the books that circulated in the lending library. (All the creators are associated with VIAF numbers, which can be used to identify birth and death dates, gender, and nationality.) In the long term, the *Project* is considering integrating information from other sources in the Beach Papers—for example, book inventories and information about purchases of *Ulysses.* But those decisions depend on how the current data sets are used and the feedback we receive from the *Project*'s overlapping communities. In 2022, Kotin and Koeser plan to edit a special feature, "The World of Shakespeare and Company," in collaboration with the *Journal of Cultural Analytics* and *Modernism/modernity*, with contributions from scholars from these overlapping communities.

Notes

[1] Would Beach have used a computer if she had had access to one? Not likely. In her memoir, *Shakespeare and Company* (1959), Beach describes how she managed the lending library: "My lending library was run on what Adrienne [Monnier] called, though I never knew why, 'le plan américain.' It would have horrified an American librarian, with her catalogues and card indexes and mechanical appliances. It was quite suitable for a library such as mine. There was no catalogue—I preferred to let people find out for themselves how much was lacking; no card index—so unless you could remember, as Adrienne, with her wonderful memory, was able to do, to whom all your books were lent, you had to look through all the members' cards to find out what had become of a volume." See Sylvia Beach, *Shakespeare and Company* (Lincoln: University of Nebraska Press, 1991), 21.

[2] For more information about Shakespeare and Company as a publisher, see Joshua Kotin, "Shakespeare and Company: Publisher," in *Publishing Modernist Fiction and Poetry*, edited by Lise Jaillant (Edinburgh: Edinburgh UP, 2019), 109–134.

[3] To learn more about Beach and Shakespeare and Company, see Noel Riley Fitch, *Sylvia Beach and the Lost Generation: A History of Literary Paris in the Twenties and Thirties* (New York: Norton, 1983); and Sylvia Beach, *The Letters of Sylvia Beach*, edited by Keri Walsh (New York: Columbia University Press, 2011).

[4] The current English-language bookshop in Paris named Shakespeare and Company was opened in 1951 by George Whitman. The bookshop was originally named Le Mistral. Whitman changed the name to Shakespeare and Company in 1964 to honor Beach.

[5] For a more detailed account of how the lending library worked, see Joshua Kotin, "Becoming a Member of the Shakespeare and Company Lending Library," *Shakespeare and Company Project* (Princeton: Center for Digital Humanities, 2020).

[6] All the numbers in this article are from Joshua Kotin and Rebecca Sutton Koeser et al, *Shakespeare and Company Project Dataset: Lending Library Members, Books, Events*, version 1.2 (Princeton: Dataspace, Princeton University), doi.org/10.34770/dtqa-2981.

[7] For more information about the acquisition of the Beach Papers, see Howard C. Rice, Jr., "The Sylvia Beach Collection," *The Princeton University Library Chronicle* 26.1 (1964): 7–13.

[8] Box forty-three in the Beach Papers includes notes and other ephemera in addition to the cards. The *Project* has connected all the material in box forty-three to specific members. The one exception is a small notebook detailing sales of Havelock Ellis's seven-volume *Studies in the Psychology of Sex* (1897–1928). The *Project* links to a digital edition of the notebook, but does not associate it with a lending library member or include it in the data sets.

[9] For an account of the gaps in the logbook coverage, see Joshua Kotin and Rebecca Sutton Koeser, "The Shakespeare and Company Lending Library Cards in Context," *The Shakespeare and Company Project* (Princeton: Center for Digital Humanities, 2020).

[10] See Nick Budak, "Representing Gender in the Shakespeare and Company Project," *Shakespeare and Company Project* (Princeton: Center for Digital Humanities, 2019).

[11] For more information about the transition from TEI to a relational database, see Rebecca Sutton Koeser et al., "Documents to Data: The Evolution of Approaches to a Library Archive," Poster, ACH2019 (July 24, 2019), doi.org/10.5281/zenodo.3277320.

[12] The *Project* only includes information about the specific edition (or printings) of one book: *Ulysses*. In general, the lending library cards do not include information about specific editions. Both American and British editions of books circulated in the lending library.

[13] For implementation details, see Rebecca Sutton Koeser, Nick Budak, Gissoo Doroudian, Xinyi Li, Benjamin Hicks, and Kevin McElwee, "Princeton-CDH/mep-django: V1.4," Zenodo, January 22, 2021, doi:10.5281/zenodo.4458099.

[14] Versions of the members data set after 1.0 include a supplemental file with information about members merged, renamed, and deleted since previous versions.

[15] See Joshua Kotin and Rebecca Sutton Koeser, "Shakespeare and Company: Top Ten Lists," *Shakespeare and Company Project* (Princeton: Center for Digital Humanities, 2020),.

[16] Calculations are made using the relativedelta module of the python-dateutil library.

[17] For more information about various plans or subscription categories, see "What Are the Lending Library Plans?" *Shakespeare and Company Project* (Princeton: Center for Digital Humanities, 2020).

[18] "Library Privacy Checklist 1: Overview," American Library Association (January 31, 2017).