Journal of Cultural Analytics

Koeser, Rebecca Sutton and Rustow, Marina. 2026. Princeton Geniza Project Datasets. *Journal of Cultural Analytics*, 11(3): pp. 1–35. DOI: https://doi.org/10.22148/jca.1084

Princeton Geniza Project Datasets

Rebecca Sutton Koeser, Center for Digital Humanities, Princeton University, rebecca.s.koeser@princeton.edu
Marina Rustow, Near Eastern Studies/History, Princeton University, mrustow@princeton.edu

The Princeton Geniza Project (PGP) is a database devoted to the documentary sources of the Cairo Geniza. Beginning in 2020, as the database approached its thirty-fifth anniversary, the Princeton Geniza Lab (PGL) and the Center for Digital Humanities (CDH) formed a partnership to launch a major overhaul of the database. The overhaul connected previously siloed data, and it modernized the data formats, front-end and back-end workflows, and interfaces. Now, for the first time, the PGP has published datasets from this long-running project in the hope of bringing the materials to a wider community of digital and computational humanists. This essay describes the project's history, explains the structure of the data, and provides context to explain the data's nuances, unevenness, and gaps. Our overview of the data includes the relationship between fragments and documents; the cataloging metadata linked to each document (types, descriptions and tags); the documents' structured languages and scripts; the format of dates and the various calendars; and the bibliographic entries, including transcriptions and translations. Finally, we share our analysis of how the data have changed over time. This essay is intended both as an invitation to encourage interest in this fascinating, rich dataset and as a reference for anyone working with the PGP and its data.

Journal of Cultural Analytics is a peer-reviewed open access journal published by the Center for Digital Humanities. © 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.
OPEN ACCESS

Introduction

Figure 1: Letter from a North African trader in a Red Sea port to his brother, 1103 CE demonstrates a typical fragmentary, handwritten *geniza* document with the text laid out at multiple angles. (Cambridge University Library, T-S 12.392. Princeton Geniza Project, PGPID 9690.)

"Then we boarded a boat that contained not a single nail of iron, may God protect us with his shield" (**Figure 1**, "Letter: T-S 12.392"). Thus wrote a North African Jewish trader in 1103 to his brother back home. He had made it as far as Dahlak, an island in the southern Red Sea, but most of his voyage still lay before him: like many northern African traders of his generation, he hoped to make it across the Indian Ocean to the Malabar coast, where he would buy pepper, iron, and other commodities. Having made it up the Nile, across Egypt's eastern desert and as far south as Dahlak, the new obstacle

now before him was a type of craft he'd never seen before, one whose planks were tied together with ropes made of coconut coir. Little did he know, this was a perfectly safe way to travel. Boats of this kind had been utilizing the regularity of the monsoon winds to ferry humans across the Indian Ocean since antiquity.

The Cairo Geniza

The richness and detail of this trader's story are known to us thanks to an extraordinary cache of historical sources: the Cairo Geniza, one of the densest and most coherent corpora for premodern history. A *geniza* is a repository for discarded texts in Hebrew script. Jewish custom dictates avoiding casual destruction of Hebrew writings lest they contain the name of God, and one *geniza* in particular, in a medieval synagogue in Cairo, preserved hundreds of thousands of pages that are now dispersed in dozens of institutional collections.

The texts are mostly, though not all, in Hebrew script. By current estimates, ten percent are documents, including letters, legal deeds, lists, accounts, and other ephemera of everyday life, and these are mainly in Judaeo-Arabic, the term modern specialists use to describe Arabic written in Hebrew characters. *Geniza* fragments have not yet been completely cataloged, let alone published or translated. Adding to the complexity of working with them is the fact that many of the texts are fragmentary, and they are now housed in seventy libraries worldwide, so that parts of the same page can be separated by an ocean.

Historical documents from the Cairo Geniza—the "documentary *geniza*" for short—are especially precious because they convey unique information often not preserved in other sources. They are also unusually candid, written without an eye to posterity because their owners never intended them to be deciphered by dogged scholars hundreds of years in the future. And due to a fortunate confluence of circumstances, when the *geniza* came into existence in the 1020s, Egypt was the economic and cultural powerhouse of the Mediterranean and the endpoint of trade networks across the Indian Ocean. The people reflected in the documents were remarkably mobile geographically, like our North African with the transportation concerns. The world the documents represent centered on Egypt, but it stretched from Spain to Sumatra.

The Princeton Geniza Project

The sheer volume of new information *geniza* documents offer is uncommon in premodern history. But precisely this abundance—and the fact that the texts are often incomplete and linguistically recondite—makes special demands of scholars. That's where digitization comes in.

Early in the personal computing era, in 1986, a Princeton-based team began digitizing transcriptions of *geniza* documents to make them searchable. This was the birth of the Princeton Geniza Project (PGP), a database devoted to the documentary sources of the Cairo Geniza.[1] In 2020, as the PGP approached its thirty-fifth anniversary, the authors launched a major overhaul of the database. Now, as it reaches its fortieth, we are publishing its vastly expanded data to bring them to a wider community of digital and computational humanists and make them more tractable.

Although the PGP is a long-running project, this is the first time its data have been published comprehensively as a dataset, or separately from the interface for accessing the documents. This essay therefore describes the project's history, explains the structure of the data, and provides some context for their nuances and unevenness. We intend it as a frame of reference for anyone who wants to work with them.

PGPv4

Between 2020 and 2022, the Princeton Geniza Lab (PGL) and the Center for Digital Humanities at Princeton (CDH) formed a research partnership to update the Princeton Geniza Project database. Over the previous decades, the data had become siloed and unwieldy. Our goals were to streamline the database and allow researchers to interact directly with it rather than through the mediation of software engineers. Especially urgent was implementing a solution for managing document transcriptions, a pain point and bottleneck for years. Transcriptions are the gold standard for researchers because a stable, expert-reviewed, machine-readable version of texts makes it easier for historians to extract information from the manuscripts. We therefore modernized the database's infrastructure by building in best practices for crediting researchers' contributions and designed an interface to allow working with different kinds of evidence in tandem.

Over more than five years—the first two as a CDH–PGL partnership and years 2–5 working with Performant Software Solutions—we restructured and combined the existing project data, migrated them to a new relational database, built a transcription interface to display transcriptions and images side-by-side, developed a way to link transcriptions to images, and built modules to track the relationships among people, places, and documents.[2]

---

[1] Readers may be familiar with the "Scribes of the Cairo Geniza" Zooniverse project, which launched in 2017 (Esten). The two projects share some overlapping content and staff but are otherwise completely distinct.

[2] Project charters for the first and second year of the collaboration between PGL and CDH document the significance, scope, goals, and project team members, and they include high-level roadmaps for the planned technical implementation for each phase (Budak et al.; Rustow et al.).

In 2022, as we neared the launch of the first iteration of the new database, we decided to label it version 4 (PGPv4) in acknowledgment of its long history and the decades of scholarship and engineering that had gone into it since the 1980s.

The larger aim of our efforts was to accelerate research on the documentary *geniza*. In this, we succeeded. In May 2021, when we launched the new admin interface, the database contained around 17,000 document entries; over the next two years, the team doubled the number of document entries, dramatically enriching the quality of information on these sources. The ease of adding new materials and metadata to the database allowed the team to accelerate their pace of research. But acceleration wasn't the only benefit. The new database allows team members to share their research with each other and with the public in real time. It has also invited a new pool of researchers to interact with the material, including undergraduates and even some high school students. Any good engineering tool helps not only to improve data interfaces, but also to streamline workflows and clarify them to new users.

Although PGPv4 is backed by a relational database, the PGP team also planned to export data in a standard format because PGP researchers were already accustomed to working with the data in a shared spreadsheet and often made their own datasets as they gathered evidence. Publishing the data expands that capacity for data-driven research, providing different modes and scales of access to allow for what Ryan Cordell calls "zoomable, scalable, or macroscopic reading" and analysis (Cordell).

Before we describe the data, we offer a history of the project and its data.

The History of the PGP and Its Data

The PGP datasets now available are the result of a long history of work by many researchers and software engineers. The history of the PGP *is* the research and technical work that led to these datasets, and the current data are still informed and impacted by their origins and many transformations.

Origins of the Project

The year 1985 was a turning point in the study of the documentary *geniza*. The founder of the field, S.D. Goitein (1900–85)—an extraordinarily committed and assiduously organized researcher who opened up the material for the first time at scale—died in February 1985. His PhD students weren't sure the field could survive without his superhuman energy and philological expertise.

Goitein had developed what he called a "lab" of research notes, transcriptions, and translations, most of which remained unpublished. These included 29,000 index cards containing notes on the people, places, names, professions, objects, and obscure

vocabulary in the documents, as well as more than 2,000 of his document transcriptions and translations. To get a sense of what this means in the life of an ordinary researcher, consider that it can take anywhere from half a day to several weeks to transcribe and translate a document. The texts don't always use standard grammatical forms or vocabulary that appears in dictionaries with relevant meanings, and the fragments are often faded and torn. Then there is the problem of cross-referencing words and phrases without digital aids. Goitein's index cards, transcriptions and photocopies were his analog database, and his students wanted to preserve them. A fortunate confluence of events intervened to allow them to do so.

Goitein had retired from the University of Pennsylvania in 1969 and moved to the Institute for Advanced Study in Princeton in 1971. Mark R. Cohen and A.L. Udovitch, who had studied with Goitein, were both then teaching in the Near Eastern Studies Department at Princeton University and were in close touch with their mentor. In 1984, IBM's Advanced Education Project donated a fleet of computing equipment to Princeton and eighteen other United States research universities in order to encourage their faculty members to develop classroom instruction materials utilizing personal computers. This was the dawn of the personal computing era; IBM wanted to accelerate it by embedding computers in university research. It is unclear whether the corporation imagined humanists taking advantage of computational infrastructures.

Goitein bequeathed his research materials—his lab—to the Jewish National and University Library (now the National Library of Israel) in Jerusalem. Before sending the material from Goitein's home in Princeton to Jerusalem, Cohen and Udovitch photocopied and cataloged it, with the indispensable help of two doctoral students, Paula Sanders and Amy Singer. As they worked through the materials and indexed them, they came up with the idea of leveraging the IBM grant to turn Goitein's unpublished transcriptions into a full-text retrieval system. This was the birth of the Princeton Geniza Project, and of the Princeton Geniza Lab that came to house it.

Using IBM software, six computer terminals, and a laser printer, Cohen, Udovitch, their doctoral students, and several typists and developers created an electronic database of *geniza* document transcriptions. They digitized most of Goitein's published and unpublished transcriptions and then added several hundred transcriptions by his former doctoral students. They also checked each transcription against the original manuscripts (or photostats and microfilm images of them), resulting in extraordinarily high-quality, reliable data. This was the first version of the Princeton Geniza Project database, which our current team retroactively dubbed PGPv1, and it contained around 1,800 transcriptions.

Making the transcriptions searchable required the PGP team to solve a number of engineering problems. Latin script was the default option for coding, indexing, and display, and technology did not yet support search in right-to-left, non-Latin scripts such as Hebrew. The markup language specialist Michael Sperberg-McQueen solved the problem by adapting existing DOS word-processing software to handle Hebrew script.

Searching the corpus was also challenging because of its size: 10 MB, laughably small by today's standards. But it was difficult for early PCs to index the data. Between 1994 and 1997, a solution arrived with an educational software engineer named Peter Batke, who took an existing program for MS-DOS called WordCruncher 4.5 and adapted it to the needs of right-to-left text. Batke possessed a combination of specializations that at the time was, in Batke's own words, "quite esoteric" ("Background and History of the Project"). He had devised search solutions for other textual corpora and also co-designed the Duke Language Toolkit, a solution for computing in non-Roman alphabets. An advantage of WordCruncher, which was designed at Brigham Young University for indexing, searching, and retrieving content from large textual corpora such as the collected works of Shakespeare, the Bible, and the Book of Mormon, was that it could index texts relatively quickly, using small random access memory regions. It could also produce concordances that helped researchers find terms they didn't yet know to search for, a pronounced advantage for *geniza* documents due to their surprisingly variable orthography.

But then technology continued to develop apace. Batke had solved the indexing and searching problem, but he had done so in the soon-to-be-outmoded DOS environment. Then there was the problem of how to access the database without coming physically to Princeton. Cohen remarked that at an earlier stage of the project, as the team backed up its work on floppy disks, "we … bided our time, waiting for a way to manipulate the database and disseminate it to others" (Cohen 40). As of 1996, the team was still planning to distribute its data and software on CD-ROMs. The growth of the internet made this plan obsolete.

PGPv2

The advent of the internet led to the development of PGPv2, designed by Rafael Alvarado, Batke's successor as a humanities computing consultant at Princeton. Alvarado made the content of PGPv1 available online for the first time by migrating it to a web application he called TextGarden, which stored the documents in Extensible Markup Language (XML) and took advantage of a flexible relational model by indexing them as a network. TextGarden built on the text encoding and interface affordances

of dynamic HTML, representing Hebrew and Arabic script using Unicode. Thus began the first major PGP migration from local terminals and floppy disks to a globally available website.[3]

Migrations can be both liberating and treacherous for data. This one came with a vast improvement and some minor losses. The improvement was the Unicode text encoding system, which allowed multiple writing systems to be displayed at the same time. Before Unicode, web pages were able to display text in only one non-Roman writing system at a time.[4] This was a problem for the sizable subset of *geniza* documents written in both Hebrew and Arabic script. These included letters in particular because when Jewish traders wrote to their colleagues in Judaeo-Arabic, they would often write the addresses in Arabic script for mail carriers who didn't read Hebrew. Before Unicode, there was no way to display transcriptions of bi-alphabetic documents accurately. The team devised a workaround—transliterating the Arabic addresses into Hebrew script—but this misrepresented the original and also departed from best practices in the field. After Unicode, the texts could be displayed as written. (As of 2026, our team still occasionally finds addresses in Hebrew characters only to discover in comparison with the manuscript that they should be in Arabic characters; these are lingering artifacts of the pre-Unicode workaround.)

As for the minor losses with the migration to PGPv2, they included some improper conversion of characters and some corruption of library shelfmarks (the equivalent of call numbers for manuscripts). These were relatively easy to fix once researchers identified them. "The process of conversion was daunting," Cohen later remarked, and "was accomplished with a certain amount of imperfection"; but he and his team managed to eliminate most of that imperfection over the subsequent years (Cohen 43).

PGPv2 precipitated a revolution in the field of *geniza* studies. Before PGPv2, researchers accessed Cairo Geniza documents in one of four ways: via transcriptions published in books and journal articles, via leads from footnotes, in Goitein's unpublished files (in Jerusalem or Princeton), or by paging through manuscripts in holding institutions. It was all needle-in-a-haystack work. PGPv2 made it possible to work from one's home or library and slice through a subset of the documentary *geniza* corpus by searching it with keywords. The Cairo Geniza was at the time believed to have preserved around 10,000 documentary texts, of which PGPv2 included transcriptions of 2,250—enough to fuel dozens of doctoral dissertations.

---

[3] Our thanks to Rafael Alvarado for contributing to this write-up of his work on TextGarden.

[4] Unicode Standard Version 1.0, Volume 2 was printed in 1992. The work leading up to it included substantial efforts to support a large number of Arabic ligatures ("Chronology of Unicode Version 1.0"). Unicode and font implementations still make assumptions about languages that don't necessarily hold for the languages used in PGP materials, for instance the need to use Arabic vowels with Hebrew consonants.

PGPv3

The database continued to expand thanks, in part, to a new funding stream. In the late 1990s, the Friedberg Genizah Project (FGP) in Jerusalem began working to catalog all Cairo Geniza fragments and create digital images of them, including the subset of documentary texts on which the PGP focused. The FGP's work established for the first time the total number of Cairo Geniza fragments—400,000—but no one knew how many of those were documentary texts because they hadn't all been cataloged. Around 2004, the FGP began supporting the PGP team's work identifying and transcribing documentary fragments. That led to the addition of another 2,100 fragments, bringing the total number of document transcriptions in the database to 4,350.

In 2004, a new engineer joined the PGP team: Ben Johnston, an Educational Technology Consultant at Princeton's McGraw Center for Teaching and Learning. Johnston's contributions to the PGP began with updating the transcriptions by lightly encoding them using the Text Encoding Initiative (TEI) technical standard. He also began storing them in Bitbucket, Atlassian's Git-based source-code repository.

But this was also the end of an era. Udovitch retired from Princeton in 2008, and Cohen in 2013. In 2015, Eve Krakowski and Marina Rustow replaced them as faculty in the Near Eastern Studies Department, and Rustow became the PGP's director. They had been using the PGP for years and were eager to continue its work and expand it. The first order of business, Krakowski and Rustow decided, was to find out, once and for all, how many documentary *geniza* texts existed. Revisiting the standard estimate of 10,000 eventually led to a shift in the PGP's purpose and back-end workflows.

From the project's origins until 2015, the goal had been to make transcribed *geniza* documents searchable for specialists. The project had, then, always been geared toward specialists who knew how to read the documents and merely needed a way to search the texts. The PGP had never focused on creating English-language descriptions of each document, let alone full translations, because the specialists toward whom it was geared could do that on their own. Because producing transcriptions was a slow and painstaking process that required many hours of expert labor, the number of documents in the database had hovered in the low thousands for three decades. As long as the field believed the total number of documentary *geniza* texts to be around 10,000, the PGP's having digitized transcriptions of nearly half of them was a major achievement. But no one had revisited that estimate in many years. It was due for reexamination.

Since PGPv1 and PGPv2 had covered only documents for which transcriptions existed, the focus of the technology behind them was text search and retrieval. PGPv3 introduced a new set of affordances, including displaying transcriptions and English-language descriptions of each document, as well as a robust set of typologies or genres to help organize the documents. Rustow and Krakowski worked with those affordances

to refocus the team's work on generating as many English-language descriptions of documents as possible, even of documents for which transcriptions didn't yet exist. In effect, they turned the PGP from a text-retrieval system into a descriptive database of the documentary *geniza*, with the goal of making it a useful tool not only for specialists but for any researcher interested in the historical material it contained.

The result was that in just two or three years, the footprint of the database increased from 4,350 transcriptions to around 17,000 descriptions (**Figure 2**). That rapid increase, in turn, emboldened Krakowski and Rustow to revisit their estimate of the total size of the documentary corpus. They now reckoned it to be around 20,000 to 30,000 documents (still, as it turned out, too low). They therefore refocused the team's work on identifying new documents and writing descriptions that nonspecialists could access.

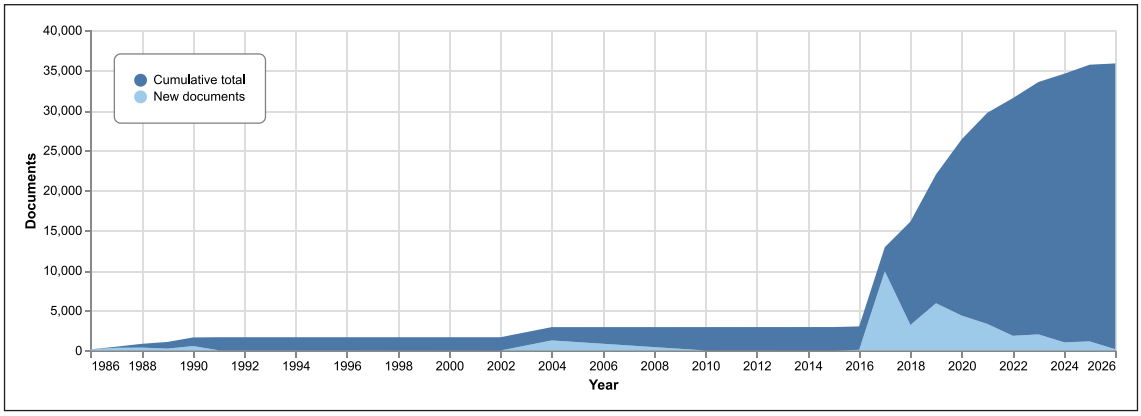

Figure 2: Charting documents added to PGP by year shows a visible spike between 2016 and 2018 due to the shift from transcription to description.

Beyond the new estimate, there were other implications of shifting focus from transcriptions to descriptions. Writing descriptions allowed researchers the latitude to add documents to the database even if they couldn't decipher every word. Adding difficult documents to the database would also, the team hoped, bring them to the attention of researchers and encourage them to come up with their own transcriptions of them.

Also, documents can become easier to decipher in batches. By reading dozens of sales contracts, for instance, you can clarify the textual structure and patterns they use, and once you've read fifty of them, you can anticipate what even a fragmentary or faded contract should be saying in the bits you can't decipher. By capturing as many documents as possible and grouping them into types, Krakowski and Rustow hoped not simply to arrive at a more realistic picture of the documentary *geniza* and its scope but to accelerate the transcription of documents.

To support the rapid expansion of the database, Johnston rebuilt the system to ingest descriptive metadata from a Google spreadsheet. Before, any change to the database went through the project directors and two or three trusted researchers, but the directors now gave roughly one dozen researchers access to the spreadsheet so they could add or edit descriptions—an improvement in the flexibility and collaborative capacity of the PGP.

That collaborative capacity extended not only to individual researchers but also to collaboration with teams at other institutions, most notably the Cambridge University Library (CUL), which houses both half the contents of the *geniza* and a dedicated Genizah Research Unit (GRU). Since the early 1970s, the GRU had employed PhD- and postdoctoral-level scholars to write tens of thousands of descriptions of *geniza* fragments, around 7,000 of which were documentary texts. In 2017, the Cambridge team graciously gave Rustow their document descriptions. The PGP team ingested them wholesale, giving credit to CUL, checking and expanding them over the subsequent years. The team also received metadata from the FGP database, which resulted in more than 2,000 new descriptions.[5] PGPv3 thus resulted in a massive and relatively painless expansion of the database, bringing the total number of documents well beyond the historical estimate of 10,000.

But then the expansion of the PGP's focus and data pushed its infrastructure to the breaking point. The Google Sheets metadata spreadsheet had made adding document descriptions astonishingly easy. But the number of entries grew so large that the spreadsheet became unwieldy to load on many machines. Rustow, Krakowski, and Johnston recognized that they would eventually need to rebuild the system, and they began to take stock of the problems they needed to solve.

In addition to the cumbersomeness of the metadata spreadsheet, PGPv3 presented another workflow challenge. It was still cumbersome to make even minor edits to the transcriptions. In the old days, when researchers found a typo or came up with a better reading of a text, they had sent emails to the PGP director, who in turn passed the edits on to Johnston, who meanwhile had taught himself to read the Hebrew and Arabic alphabets and inserted the changes himself. Johnston had simplified this cumbersome process by giving Rustow, Krakowski, and two of their research assistants access to the Bitbucket repository so they could edit the texts directly. But Bitbucket didn't handle right-to-left text well, and the process was cumbersome enough to discourage editorial changes, or even frequent correction of erroneous texts.

[5] These FGP descriptions were crowdsourced, so instead of ingesting them wholesale, the team hired Amir Ashur to review and revise them, a task he accomplished from 2022 to 2023.

Rustow and Krakowski began consulting with the Center for Digital Humanities in 2017. It was only then that they became aware of the long-term need not only for a different infrastructure but also for better project documentation, stability, and sustainability. Their contemplation of what the PGP should look like in a decade (or four) was the immediate motivation for the PGL-CDH research partnership that began in July 2020.

PGPv4

The lead researchers on PGPv4 were Rebecca Sutton Koeser, who led the software design and implementation team, and Marina Rustow, who led the research team. They worked together with another software engineer (Nick Budak, replaced in 2021 by Ben Silverman of Performant Software Solutions), a UX designer (Gissoo Douroudian, replaced in 2022 by Chelsea Giordan of Performant Software Solutions), and PhD students and postdocs in *geniza* studies and related fields who served as project managers (Stephanie Luescher [2020–21], Rachel Richman [2020–present], Zohar Berman [2021–22], Ksenia Ryzhova [2022–present], Amel Bensalim [2023–present], and Pratima Gopalakrishnan [2023–present]). (In what follows, "we" refers to this expanded team.) The expansion of the team created two needs that the PGP had never before had to face: holding regular meetings and creating documentation. Thanks to the CDH's experience with similar partnerships, the new team workflows fell into place relatively quickly over the second half of 2020. The collaborative effort required the engineers on the team to learn the basics of *geniza* studies and the *geniza* researchers to learn about data modeling and software engineering workflows.

As we designed PGPv4, the first and most important decision we made was to migrate the data from a flat model to a relational database. Knowing how we wanted to structure it required three months of weekly conversations on data modeling. At its core, we decided, there was a many-to-many relationship among fragments and documents.

A document, such as a letter, contract, or list, is a discrete unit of text. A fragment is the physical piece of paper or parchment on which the document is written. Fragments have their own shelfmarks (identifiers used by the holding institution to reference the material object).[6] Often fragments contain more than one document. Paper and parchment were not exactly scarce, but they weren't cheap either, and medieval scribes were nothing if not frugal. Writers often reused materials, flipping over a letter they

[6] In some cases, libraries gave a single shelfmark to unrelated fragments, typically when they were small and needed to be encapsulated within the same "multifragment" page of large binders (Katz). Multifragment shelfmarks can be found at Cambridge for very tiny fragments that so far haven't been cataloged in the PGP and at JTS for larger fragments, many of which are documents in the PGP.

had received to scribble down a financial account, or drafting a legal document on the back of another legal document, or writing a letter around the text of an official communiqué that a state bureaucrat had discarded. We wanted to be able to display each of the documents on a single fragment discretely, but without losing its relationship to the other document(s) on the same piece of paper or parchment.

The converse situation was nearly as common. Just as each fragment could contain more than one document, some documents were composed of more than one fragment. Many documents had been torn up when they were discarded, or else dismembered in the jumble of the synagogue *geniza* chamber. Sometimes fragments containing parts of the same document ended up in different libraries, or in the same library under different shelfmarks. When specialists reunite these pieces of what used to be a single page, they call it a "join." Finding one is a cause for celebration—perhaps not as momentous as discovering a new subatomic particle, but worthy of a champagne toast. Like completing a puzzle, you're staving off just a little bit of entropy in the world. We wanted to display joins as unitary texts linked to multiple fragments.

Some joins yield not a single text but many, for instance when scribes wrote on a used writing support that then got torn apart. Thus, a document can occupy more than one fragment, and a fragment can contain more than one document. That many-to-many relationship was the basis of our decision to build a relational database.

Doing so required us to walk away from a longstanding informational hierarchy that the field had taken for granted. Many of us had tacitly assumed that there is a base unit of analysis in the *geniza*, whether the fragment or the document; one always seemed to be the star of the show. This implicit hierarchy had been baked into previous printed scholarship and digital tools and structured the display of information in different ways. Since libraries are the custodians of physical objects, they tended to privilege the fragment. The FGP database, because it was born as an imaging project, also favored the fragment and displayed images of fragments. But scholarly editions tended to focus on textual units such as documents, since their task was to present and analyze coherent textual units. Since the PGP had been developed by historians, it tended to favor the document. We were also aware that doing so entailed some loss of the relationship to the physical object in its totality. For example, if you had a legal document on one face of a fragment and a business letter on the other, previous versions of the PGP displayed them as two different entries. The connection between the two was evident only to researchers tracking shelfmarks. We wanted to make that relationship clearer. The fact that the author of the second document had somehow gotten their hands on the first is relevant information to a historian.[7]

[7] For example, the circulation and reuse of materials provides a window into the operations of the Fatimid caliphate (Rustow, *The Lost Archive*).

We managed to sidestep this longstanding hierarchy and acknowledge the importance of both the fragment and the document, depending on the context, by developing a relational data model. The fragment is what you request when you go to a library or search for a digital image, and the document is what you transcribe when you create a text edition. PGP needed to account for both, and for the relationship between them.[8]

The many-to-many relationship between documents and fragments (**Figure 3**) is but one example of the choices governing our underlying data structure. It is a representative example in that it demonstrates the necessity of understanding something about *geniza* documents and the history of research on them in order to work with the data. In the next part of the essay, we shift from the history of the project and the challenges that drove its software development to the nature and structure of the current PGP datasets.

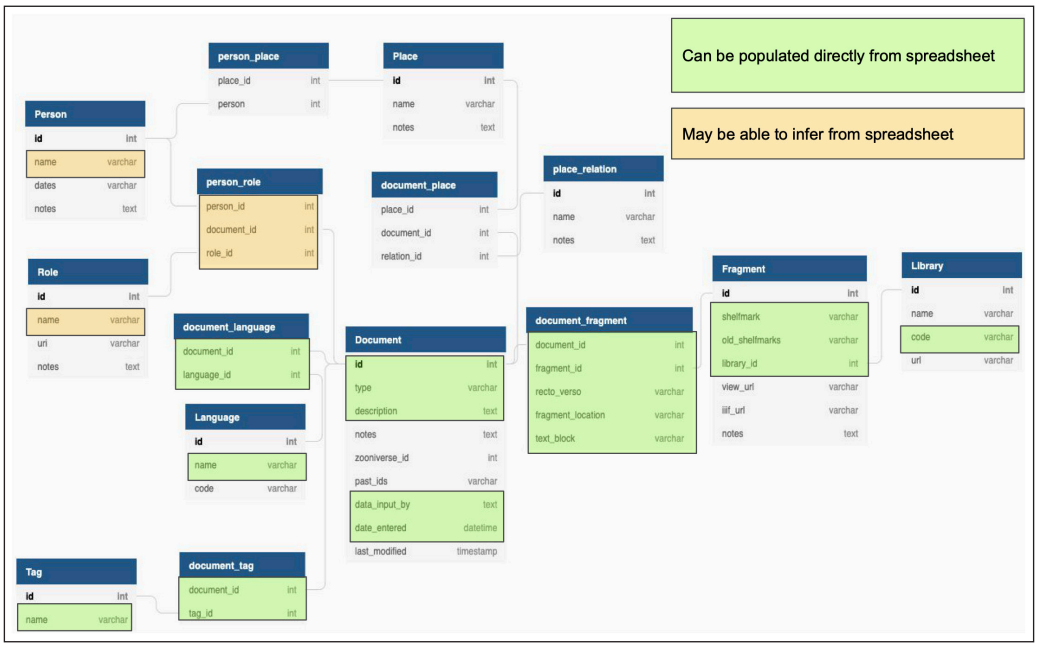

Figure 3: An annotated version of a preliminary database diagram from project planning documents from October 2020 shows the proposed data migration from the PGPv3 metadata spreadsheet.

The Nature and Structure of the Datasets

The process of rebuilding the PGP began in July 2020 and is now complete—a five-and-a-half-year process from start to finish. Nonetheless, the rewards came long before the work wrapped up. In May 2021, we launched the new admin interface. A good research infrastructure should encourage and inspire research, and ours quickly did just that, accelerating the expansion of the data.

[8] It also serves as a powerful opportunity to correct cognitive biases in a field that had long been dependent on printed editions of manuscript texts (Rustow, "Fragmentology").

The PGP datasets that the rest of this essay describes are exported from the PGPv4 web application as multiple flat, tabular data files.[9] The application is implemented in Python with the Django framework and is backed by a relational database, PostgreSQL. Familiarity with the structure of the underlying database and choices made in modeling the data should make it easier to understand and work with and recombine the exported data.

Fragments and Documents

In the PGPv4 web interface, each fragment is correlated with a shelfmark, a holding institution, and a collection within that institution. On the public site, fragments are identified by the shelfmarks assigned to them by their holding institutions. Shelfmarks rarely change. When they do, it is a decision of the holding institution, and the PGPv4 interface displays both the new and the old shelfmark.

Documents are correlated with other kinds of structured information, such as descriptions, tags, and the languages in which the documents are written. Documents have unique numeric identifiers called PGPIDs. While shelfmarks rarely change, PGPIDs sometimes do; for instance, when researchers find duplicate document entries and merge them into a single PGPID. The PGPv4 web application tracks both historic shelfmarks and old PGPIDs to find records when identifiers have changed, but that information is not currently included in the data exports.

Documents written on the same fragment are linked under "related documents." This relational data structure makes it easier to find and access documents that are written on the same fragment.

In SQL, a many-to-many relationship requires an additional database table in order to track pairs of related objects. A minimal implementation of this relationship would be a list of document and fragment id pairs, where each identifier occurs once for each related object. A document written on a single fragment would have one entry, whereas a document written across two fragments would have two entries with the same document id and two different fragment ids. PGPv4 stores additional information about the relationship between a document and a fragment in this connecting table (TextBlock in **Figure 4**). These extra fields enable researchers describing documents to explicitly specify the order of fragments or to indicate an uncertain relationship.

The majority of PGP documents—34,237, or more than 95%—are written on a single fragment. They are not joins so they are simple from a data-modeling perspective. There are 1,618 documents that are joins, most of them between two fragments, and

[9] Numbers and charts are based on the 1.1 version of the datasets (Rustow et al., *Princeton Geniza Project datasets*).

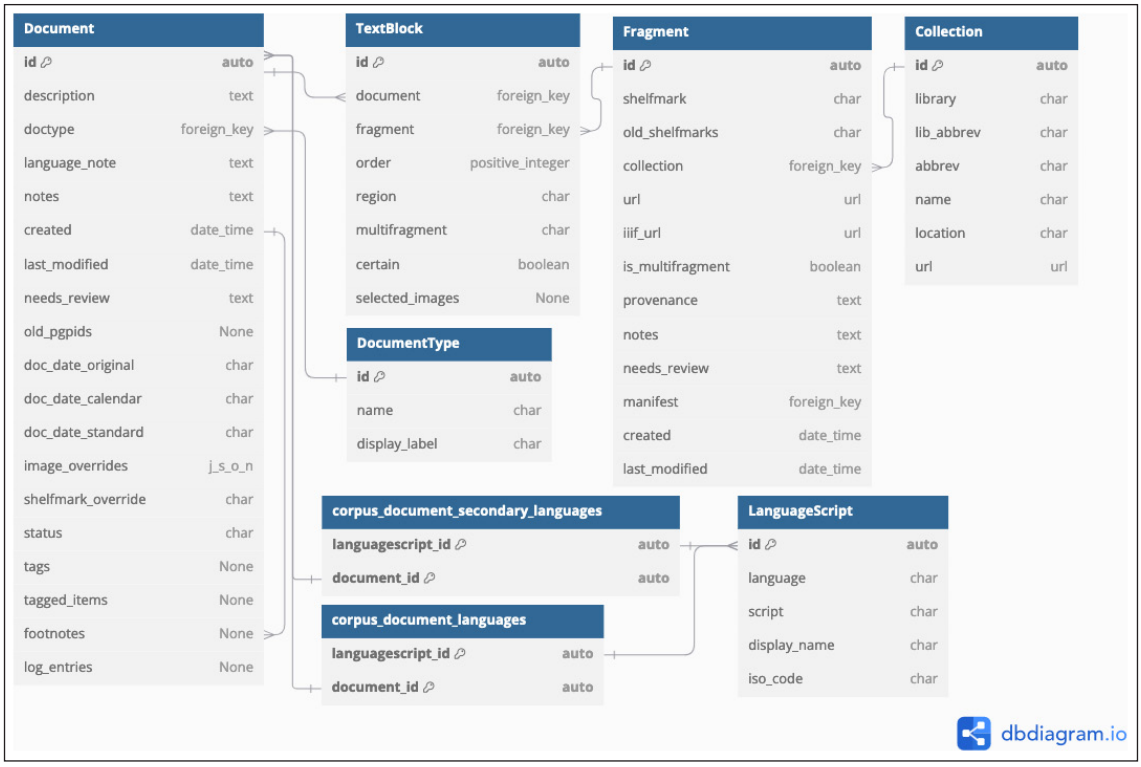

Figure 4: This database diagram shows the many-to-many relationship between documents and fragments and other associated information in the PGPv4 database (generated from models in PGP Django codebase v4.19; translation text fields are omitted for simplicity).

some among more than two fragments (**Table 1**). A small set of outlier documents are correlated with an unusually large number of fragments: 32, 34, and 64 respectively. These are multipage documents, and they are all notebooks containing the archives of a legal court.

Fragments	Documents	%
1	34,237	95.49%
2	1,224	3.41%
3	197	0.55%
4	91	0.25%
5	46	0.13%
6	27	0.08%
7	8	0.02%
8–64	25	0.07%

Table 1: An overwhelming majority of PGP documents are written on a single fragment.

If we flip the question and ask how many fragments contain more than one document, again we find that the majority, nearly 90%, contain a single document (**Figure 5**).

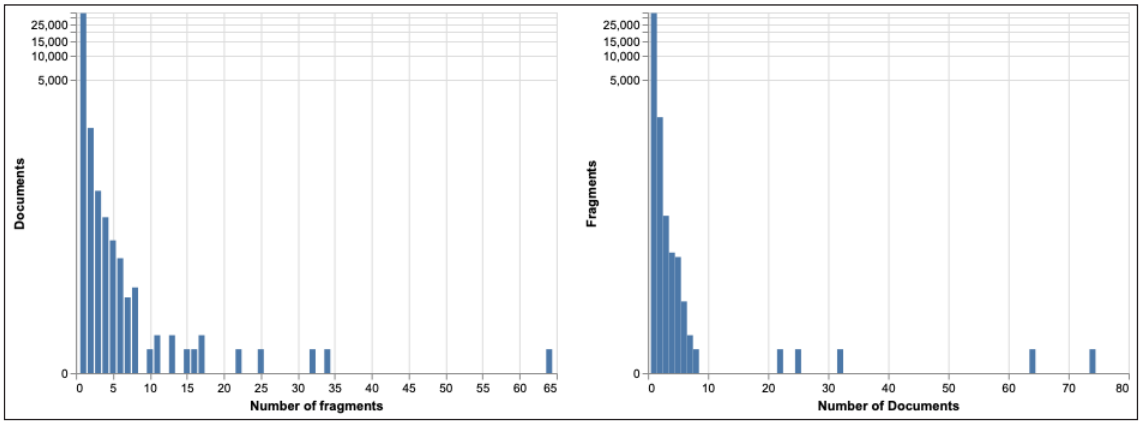

Figure 5: Whether analyzing documents by number of fragments (left) or fragments by number of documents (right), the vast majority are single documents on single fragments.

The more complicated cases provide an opportunity to draw insights beyond the textual and visual contents of individual documents. The co-occurrence of different texts on the same fragment enables scholars to trace new histories of paper and parchment circulation and reuse. For instance, two petitions to the Fatimid ruler Sitt al-Mulk ("State Document: Bodl. MS Heb. B 18/23"; "State Document: ENA 3974.3") were glued together by a Jewish scribe in need of a long vertical scroll on which to copy a liturgical text. By making the relationship between documents and fragments more explicit and easier to navigate, PGPv4 makes it easier to work across related documents and reused fragments.

Types, Descriptions, and Tags

Each document entry in the database includes three main metadata components: types, descriptions, and tags (**Figure 6**). Every document has a description, even if it is short. Nearly all documents have a type.

Types are ways of categorizing documents according to their textual structure. Each document is categorized under a single type. PGP uses an intentionally simplified set of categories that Krakowski and Rustow developed over several years of discussion and workshopping with other colleagues. The simplicity of this typology makes it easier to enter new documents into the database and to group like with like.[10]

[10] We omit here the recently added type, Inscription, which in the 1.1 dataset has only been applied to five documents.

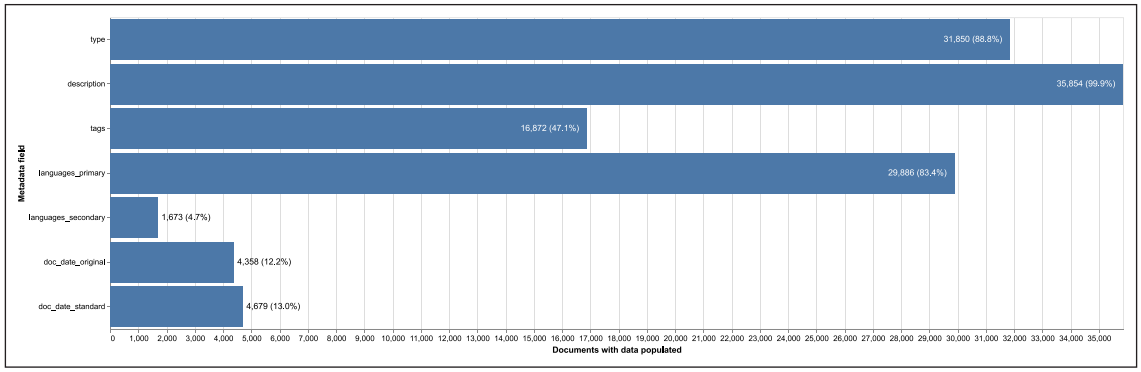

Figure 6: Document metadata fields are all optional, but type, description, and primary language are typically available.

1. **Credit instrument or private receipt**: letters of credit, commercial receipts, records of debt, records of payment (excluded: financial transactions mentioned in legal documents)
2. **Legal document**: bills of sale, testimonies, quittances drawn up by courts (as distinct from private receipts), marriage contracts, divorce deeds, summaries of court cases
3. **Legal query or responsum**: special types of legal texts either describing a case in order to solicit a legal opinion or rendering a nonbinding legal opinion on it
4. **Letter**: correspondence between business associates or family members; official administrative letters from leaders of Jewish communities
5. **List or table**: commercial accounts, grocery lists, genealogical lists, charity distribution tables, inventory lists, jottings in list or tabular form
6. **Literary text**: not documents but included in the PGP on a case-by-case basis if they capture significant historical information relevant to our documentary texts, for instance poems dedicated to known individuals; historical chronicles
7. **Paraliterary text**: quasi-documentary but highly stereotyped texts such as amulets, calendars, magic spells, prescriptions, and recipes
8. **State document**: official communiqué to or from government officials, including decrees, petitions, internal memoranda, and tax receipts
9. **Unknown type**: illegible, unidentified, or uncategorized texts that require further study

This nine-part typology addresses the need to reduce friction at two key stages of working with the data: entering a document for the first time and then transcribing it. It reduces friction at the data-entry stage by avoiding complex decision trees. Although there are many more than nine templates that the scribes of the documents were following (or devising on the fly), when entering a document for the first time,

picking through a massive list of document types would place an undue burden on the researcher. These nine types cover nearly all the existing cases.

The typology reduces friction at the transcription stage by making it easier to find similar documents for reference, allowing researchers to compare documents easily to facilitate transcription. For a researcher faced with the task of transcribing a document that is faded, torn, or full of holes, looking at similar documents helps fill in the blanks. Even though scribes used formulaic phrases the way experienced chefs use ingredients rather than slavishly following recipes, the limited typology helps bring similar documents together.

There are currently 35,855 documents in PGP. The most common types are letters (11,281) and legal documents (8,096). The current dataset includes 4,005 documents of unknown type, which usually means they have not yet been cataloged and described, or they don't fall into existing categories (**Figure 7**).

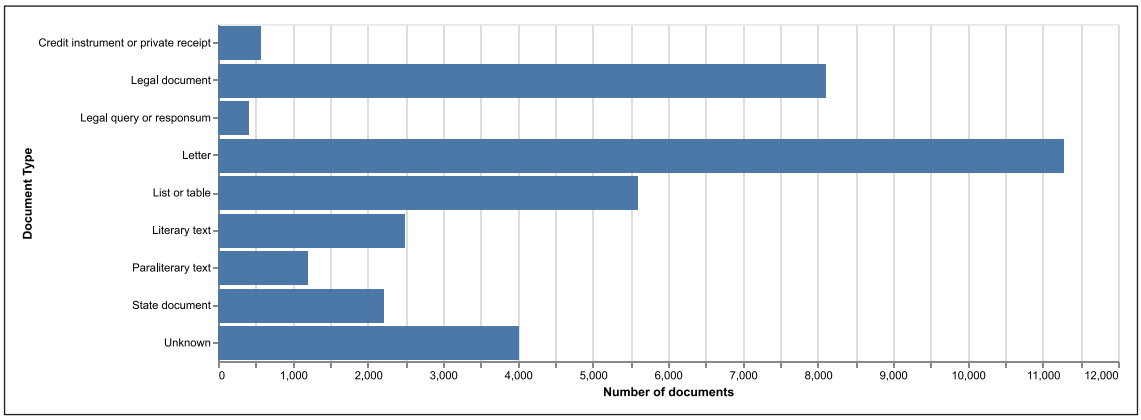

Figure 7: A chart of documents by type shows the distribution of documents across this controlled taxonomy. The most common types are letters and legal documents.

All documents have descriptions. These are English-language summaries of the main contents of a document. (Some documents also have Hebrew-language descriptions culled from existing Hebrew-language scholarship.) Descriptions don't follow a strict template and can range from terse to expansive. Most are brief, with an average of 271 characters and 45 words (technically, "tokens"). They are generally structured as an inverted pyramid with the main information—author, recipient, locations, dates—followed by the details. For instance:

> Letter from Nahray b. Nissim in Alexandria to his uncle Abū l-Khayr Mūsā b. Barhūn al-Tāhartī in Fustat. Dating: Friday, [probably 25] Nisan [4810] AM = 20 April 1050 CE. April 1051 CE (Gil's dating) is less likely but also possible. Nahray reports among other things that he had forgotten to bring his capitation tax receipt "for the year 441" on a business trip. ("Letter: ENA 2805.14")

Descriptions are, by their nature, one place where the data's unevenness emerges. Sometimes descriptions are short due to a lack of existing research on a document; in other cases, they are short because the documents were entered into the database in the 1990s, before the team started including descriptions for nonspecialists. Some of the longest descriptions are of documents that the PGP team identified recently: they wanted to put as much information about them as possible into the database for future researchers. Others are lengthy for the opposite reason: they've been in the database for a long time or have received copious attention from scholars and accumulated a long trail of research and publications.

Tags are equally uneven. There is no restriction on the number of tags that can be applied to a document (although the interface has an auto-complete function to help researchers select preexisting tags instead of multiplying them needlessly). Existing tags do not reflect a concerted effort by the team to tag all our documents; rather, they represent the team's current and past research interests.

For example, one research team member, Alan Elbaum, wrote his master's thesis on how people in the world of the *geniza* described their physical ailments. His research project resulted in a large number of tags related to illness. Since it is impossible to anticipate all the possible questions that future historians will bring to *geniza* documents, our goal is not to produce tags for some theoretical future researcher, but to flag themes of interest to us now. Tags don't, then, represent the current state of documentary *geniza* research. They are not library subject headings but rather PGP researchers' post-it notes.

The datasets contain 2,695 unique tags, of which 1,064 are used only once. Some of the most commonly used ones are #account (755 documents, flagging a future attempt to decipher traders' accounts); #communal (751 documents, reflecting the organizational levels of the Jewish community); #illness and #illness letter 969–1517 (694 and 651 documents, respectively).

Languages and Scripts

Geniza materials are written in a variety of languages and scripts, including Judaeo-Arabic (a range of Arabic dialects and registers written in Hebrew characters), Hebrew, Aramaic, and Ladino (a range of medieval Romance dialects and registers written in Hebrew characters).

To simplify the data structure, we modeled languages and scripts as a single entity. For instance, Judaeo-Arabic is a single entity, not Arabic language and Hebrew script. **Table 2** shows the languages that appear on the largest number of documents. (Documents often use more than one language, so these totals may include the same documents more than once.)

Language/Script	Documents
Judaeo-Arabic	16,137
Arabic	9,970
Hebrew	7,847
Aramaic	1,761
Greek/Coptic Numerals	976
Ladino	374
Unidentified language and script	39
Unidentified (Hebrew script)	23
Unidentified (Latin script)	8

Table 2: *Geniza* documents are most frequently written in Hebrew and Arabic scripts and languages, but there are some documents in unidentified languages and scripts (totals based on primary and secondary languages).

One special case on the table is Greek/Coptic Numerals—not a language but a writing system specifically for numbers. Many scribes who knew neither Greek nor Coptic wrote alphanumerals derived from Coptic (and, in turn, from Greek). Since they were part of what was effectively a mercantile notation system, we thought they were worth tracking separately. Greek/Coptic Numerals is therefore a language/script category distinct from either Greek or Coptic.

The data also includes a few unknown languages and scripts, which may be of interest to scholars who want to tackle a challenge. **Table 2** also includes tallies for documents with unidentified or partially unidentified languages and scripts. These designations indicate that a researcher has examined the document and was either unable to read or identify the script or could identify the script but not the language it was used to write.

Because the authors of *geniza* documents often employed a mix of languages, our data model allows documents to have multiple languages. Of the 29,886 documents with primary languages, 6,736 (22.5%) have multiple languages. The most common combinations are Judaeo-Arabic with Hebrew and Judaeo-Arabic with Arabic (**Figure 8**).

We also distinguish between primary language, a designation that indicates that most of the document is written in it, and secondary language, for languages used only incidentally. Common examples of secondary languages are Arabic addresses on Judaeo-Arabic letters or Greek/Coptic numerals in Judaeo-Arabic letters and accounts. The distinction between a primary and a secondary language is not always clear-cut: the research team hasn't developed a quantitative threshold, which would have slowed data-entry. What one researcher labels as a secondary language may be primary to another. For instance, some may argue that the incidental use of Hebrew and Aramaic

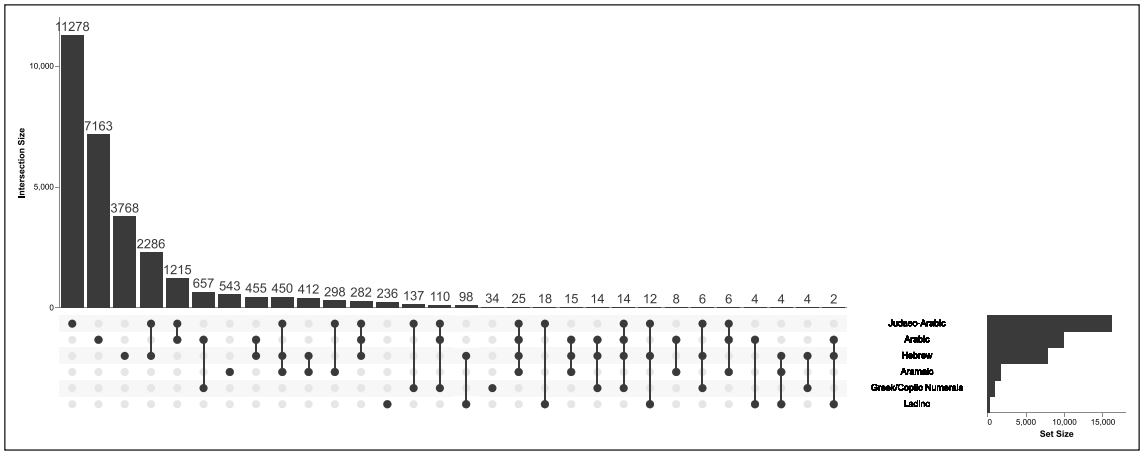

Figure 8: An UpSet plot of documents by languages shows the distribution and combination of languages across the collection (limited to languages that occur at least 200 times). The upper bar chart shows the number of documents for different groups of languages and language combinations; the bar chart at right shows the size of each language category across all groups. The matrix provides a visual indicator of the categories at left and the combinations represented in the upper bar chart. For more on UpSet plots and the challenges of visualizing overlapping combinations with more than four sets, see "Visualizing the Collections" (Koeser).

is inherently part of Judaeo-Arabic and should not be marked as secondary languages at all. In practice, the way researchers structure secondary languages depends on their language proficiency and attention to detail, and even seasoned *geniza* researchers may miss some code-switching because they are so accustomed to it. These categories may be reviewed and refined in the future, perhaps with the help of Handwritten Text Recognition (HTR) or other machine learning–based philological tools.

Dates and Calendars

Geniza documents date from the ninth century to the early twentieth, but they are not evenly distributed. Most documents come from the eleventh, twelfth, and thirteenth centuries, with significant later clusters from the sixteenth and nineteenth that have received less attention than they deserve (**Figure 9**).

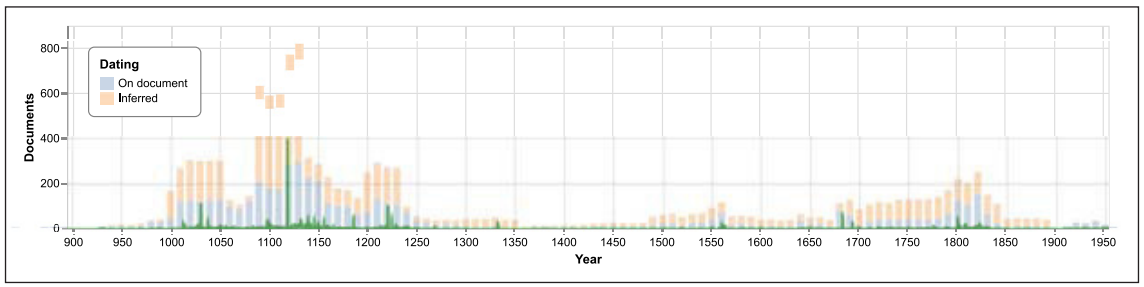

Figure 9: This chart shows the distribution of documents by date. The green line graphs the totals for documents by year, based on a midpoint date for date ranges. The stacked bar chart indicates totals for dated documents by decade with any overlap for date ranges.

Situating documents chronologically is essential for historians. It allows them to interpret and contextualize them more accurately. But fragmentary documents can be challenging to date. Rustow has written elsewhere about the challenges and different types of uncertainty involved in dating *geniza* documents (Rustow "Dating Problems?"). Even texts that contain explicit dates may be incomplete. Letters, for example, often include the month and day but omit the year because it would have been obvious to both sender and recipient.

When a document explicitly notes the date, it is called "date on document" and is considered authoritative. This is not always the date on which the document was written—it can also be a date that happens to be mentioned—but it is usually close enough to serve as a reference point. Documents rarely have more than one date. When they do, we use whichever one is latest. When the document doesn't note the date explicitly, it can sometimes be inferred based on the people, events, or types of coins mentioned (many coins were minted and circulated in limited periods; see Dudley and Elbaum). Experienced *geniza* scholars can infer the dates of documents based on handwriting, either because an individual's hand is well-known or because a style of handwriting occurred only in a certain period. Examples include the scribe Ḥalfon b. Menashshe, a Jewish legal clerk in Fustat active in the first half of the twelfth century, with a distinctive handwriting that allows for confident document dating (Elbaum and Ryzhova); and a highly cursive Iberian style of writing that became common in Egypt starting in the late fifteenth century.

The PGP currently has 5,961 dated documents (16.6% of all documents). Of these, 4,679 have an explicit date, and 1,282 have been dated inferentially. Since dates are usually not precisely known, the range of possible dating can be quite large (**Table 3**; see **Figure 9**).

Dating type	Documents	Mean	Max
On document	4,679	2	102
Inferred	1,282	46	701

Table 3: Documents may have a date on document or an inferred dating, and dates on documents are more common and tend to be more precise. This table shows the average and maximum date precision in years for the two kinds of dates.

Most of the surviving dates on documents use non-Gregorian calendars. These include Jewish (*anno mundi*), Seleucid, Islamic (*hijrī*), and the Egyptian fiscal (*kharājī*) calendars (**Figure 10**). Months include the Jewish, Coptic, and Islamic months, and sometimes a single date will mix and match calendars.[11]

[11] For instance, a document dated "23 Ḥeshvan (Shawwāl) 521 AH"; the description notes that "it is unusual but not unheard of to combine Hebrew months with the Hijrī calendar" ("Legal Document: T-S 6J3.5" 1127).

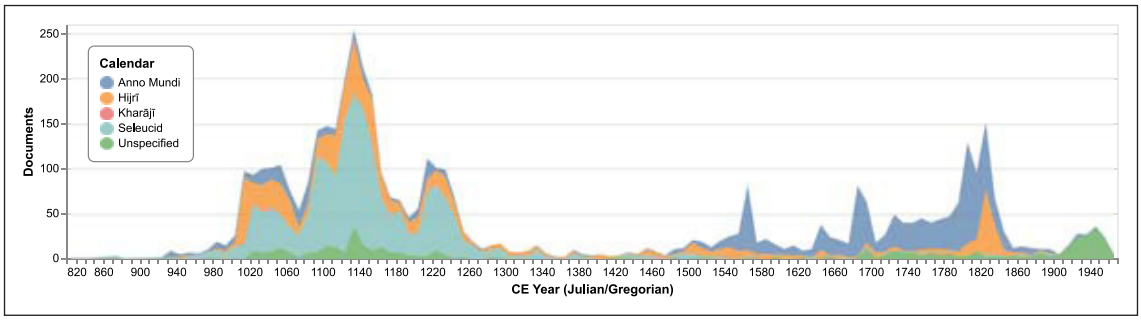

Figure 10: Documents are dated according to multiple calendars and then converted to Common Era dates for use and analysis together.

Starting with PGPv4.5 (June 2022), PGPv4 automatically parses and converts Seleucid, *hijrī*, and *anno mundi* dates to Common Era dates (Julian before October 4 1583, Gregorian after).[12] Converting to a common calendar enables documents to be sorted, filtered, and compared in a unified chronology as well as to be understood by non-expert audiences.[13]

Like the PGPv4 web interface, the datasets provide both the original calendar and the converted dates because they are useful for different kinds of analysis. In the datasets, the "date on document" consists of three fields: original date as written, original calendar, and a standardized date. Inferred dates are similarly composed of multiple fields: a display field with textual or human-readable information, a standardized date field, a rationale documenting the basis for the inferred date, and an optional notes field to provide more information about the dating. In both cases, standardized date fields use Extended Date/Time Format (EDTF) for a single date or date range.

Bibliographic Entries

One of the driving principles of PGP was that any encounter of a human mind with a *geniza* document is worth preserving: the material is so challenging that it helps to see prior attempts to understand it, even if they were inconclusive. This same principle seems to have been one of Goitein's motivations for retaining his research notes in such an organized and systematic fashion.[14] Like other digital humanities projects, PGP is both "a work of scholarship and an instrument of scholarship": it publishes and builds on existing research while also enabling new research (Kotin and Koeser, "The World of Shakespeare and Company" 2).

---

[12] Calendar conversion is implemented using the Python convertdate library, with additional logic to handle the Seleucid calendar. We do not yet support conversion of kharājī dates, since there is still too little evidence about how they coordinated with the hijrī calendar.

[13] Before built-in support for automatic conversion, researchers converted dates manually using online tools such as Theodore Beers, *Calendar Converter for Near East Historians*, https://www.muqawwim.com.

[14] See Goitein's, especially section 4, which in retrospect reads uncannily like a blueprint for PGPv4.

All the more important, then, for the database to distinguish clearly between existing and new research. The PGP builds on a substantial collection of published and unpublished materials, most notably but not solely Goitein's transcriptions, translations, and index cards. (Goitein produced so many thousands of these that we had to split them into pseudo-volumes in our bibliography.) Other scholarship in the database includes books, book chapters, articles, dissertations, and blog posts, as well as unpublished transcriptions by Goitein's students, their students, and their students' students, down to current project researchers (**Table 4**).

Type	Sources	Footnotes
Unpublished	287	15,289
Article	250	672
Book	97	7,208
Book Section	56	224
Dissertation	21	1,021
Blog	1	1

Table 4: PGP draws content from a variety of scholarship sources and includes a substantial number of unpublished transcriptions.

PGPv4 tracks these research outputs—and unpublished products of scholarly research—by linking documents to sources with what we call footnotes, a solution adapted from previous projects developed at CDH (Koeser, *Derrida-Django v1.3*; Koeser et al., mep-django).[15] Footnotes are implemented with a Django "Generic Foreign Key," which uses a combination of content type (e.g., document, person, place) and object identifier (e.g., document PGPID, or person or place database id) to link to any record in the database. This builds on work from the Shakespeare and Company Project, which used a similar approach to link book-borrowing data to the archival sources from which the information was drawn (Kotin and Koeser, "Shakespeare and Company Project Data Sets" 18). For PGPv4, we augmented the footnotes with a document relation field that indicates what information the source provides for this specific document—currently, transcription ("Edition" or "Digital Edition"), translation and/or discussion. In the PGP datasets, the sources data provide a select bibliography with citation information and the number of footnotes associated with that source.

[15] The CDH footnote module was originally inspired by Jean Bauer's approach to database footnotes, which were used to allow for conflicting evidence, indicating "whether or not that source supports the information in the record" (Bauer).

Transcriptions and Translations

Transcriptions have long been the gold standard in documentary *geniza* research. Because the texts are so difficult to read and contain unique, historically valuable information, every professional effort to read one is worth preserving. Translations are equally valuable, resolving ambiguities in the original and making sense of unclear syntax, sometimes in different ways. And, of course, translations make the material available to a wider pool of researchers.

The PGPv4 database stores and manages transcriptions and translations in the W3C annotation format with minimal HTML formatting. Annotations are linked to images when they are available. The dataset exports include a plain-text version of the transcription and translation content in the content field of the footnotes data. The document relations, "Digital Edition" and "Digital Translation," indicate that a record has a transcription or translation available. Footnotes that note "Edition" or "Translation" indicate that there is a scholarly source containing a transcription or translation of a document, but the content is not yet available in PGPv4. The language of the translation can be determined based on the language of the associated source, typically English or Hebrew.

Researchers should be mindful that transcriptions and translations include editorial symbols such as square brackets, parentheses, curly brackets, ellipses, and double brackets. These are customary in philological transcriptions of historical texts. Over the past century, a set of best practices has evolved for scholarly transcription known as the Leiden conventions. (For instance, square brackets enclose an editor's suggested reconstruction of missing text, as when the manuscript is torn, faded, or abraded.)[16] In 2021, the PGP team developed a more streamlined version of the Leiden conventions for new transcriptions.[17] (Some of the older printed transcriptions included in PGP data follow a slightly different version of the Leiden conventions and not all have yet been brought into line with new standards.) Computational work on transcription content will likely require removing these indicators as a preprocessing step for most use cases. That said, indications of missing or inferred text may also provide an interesting object of study.

Images of Fragments

We decided that PGPv4 should include digital images alongside text whenever possible. It is significantly easier to make sense of *geniza* documents when you can access the

[16] See "Leiden Conventions," *Wikipedia*, https://en.wikipedia.org/wiki/Leiden_Conventions.

[17] See Geniza Lab, "Transcription Conventions in Princeton Geniza Project Database," https://docs.google.com/document/d/e/2PACX-1vR8kR4zZdnDZXjoLrYrABZn58PRyzrKfEiixQzE9vAzNfzI4Enxs0jU9KO5rTdiH1ZMTPwfqm31mFuX/pub.

fragments themselves or digital images of them. The visual information contained in a fragment can enable an experienced *geniza* scholar to determine the genre of the text at a glance and whether it was an official version of a document or the notes of a scribe preparing to write one. It also allows you to see whether the document is fragmentary and where the lacunae are. High-resolution images also enable scholars to correct the transcriptions that previous generations had made from photostats, microfilms, or the original fragments.

We included images by leveraging International Image Interoperability Framework (IIIF) APIs. IIIF provides interoperable mechanisms and associated tooling to publish, use, and reference images and groups of images. It also makes it possible to work with aggregated content managed and preserved by different institutions. Libraries and museums around the world are increasingly adopting IIIF to share digitized content. Fortunately, the three largest *geniza* collections in the world use IIIF to publish their materials. These include tens of thousands of fragments from the Cambridge University Library, which houses around 200,000 fragments; the 43,000 fragments at the Jewish Theological Seminary (JTS) in New York, which are currently hosted by the Digital Princeton University Library (DPUL); and the 11,000 fragments at the University of Manchester. We hope the remaining collections will be made publicly available online via IIIF over the coming years through the National Library of Israel.

In some cases, reliance on IIIF required waiting for institutions to migrate older digitized images into modern systems, for example, the *geniza* materials at the University of Pennsylvania, which were imported into PGPv4 in 2023. In other cases, IIIF image publication required additional organizational and technical creativity.

For instance, JTS holds 10.75% of the *geniza*, but it is a small institution without the technical infrastructure to host IIIF images. In this case, by building on the Princeton Geniza Lab's previous agreements to share JTS materials for research, we developed an agreement enabling the DPUL to host and publish JTS images via IIIF. In another case, the Bodleian Library at the University of Oxford, with more than 12,000 fragments, publishes digitized images and metadata online with permissive licenses, which enabled us to convert their XML metadata into IIIF manifests and lean on PUL infrastructure to serve the images as IIIF. In yet another case, the University of Manchester images are available as IIIF, but each image is published separately with references to their "obverse" image. Since IIIF allows for remixing, the PGPv4 team wrote scripts to combine these images into the more common structure of one manifest per fragment. These manifests preserve information about original ownership and rights, and images that were already published as IIIF use a partOf relationship to link back to the original manifest.

Whenever images are available in PGPv4, the links are included in the fragments data. Records may include the `url` and the `iiif_url`. The first provides a human-viewable format, and the second is a machine-readable version (IIIF Presentation JSON manifest). Manifests published on princetongenizalab.github.io are the locally customized or remixed versions described in the previous paragraph.

Using IIIF has another advantage. It enables PGPv4 to display the images of multiple fragments that make up a join, or single document, even when those fragments are held by different institutions. The admin interface allows data curators to suppress, reorder, and rotate images in order to display only those relevant to a particular document, in the correct sequence and orientation, for example, omitting the verso if only the recto is needed. PGP datasets do not yet include this document-centric arrangement of images. We recommend toggling between the datasets and the PGPv4 web interface to view images in context. In the meantime, images are accessible through the fragments associated with a document.

People and Places

The PGP datasets include information about people and places associated with the documentary *geniza* corpus. The current version of the dataset includes 1,802 people and 486 places, and our data will continue to grow. Information about people and places will fuel the analysis of trade and social networks. The information also holds the potential to facilitate the dating of undated documents, such as when a person with a known lifespan appears in an undated document.

People and places can be linked to documents in the database and also to one another. But the datasets do not currently include those relationships, so they are not yet computationally rich in the way that the rest of the datasets are. The current exports include the number of related documents. We hope that future versions of the datasets will include more details on the rich information connecting people, places, and documents.

Dataset Versioning and Frequency of Change

PGPv4 is the site of active, ongoing scholarship as team members describe, categorize, transcribe, and publish their research. The team also occasionally makes larger, collaborative efforts for targeted work, for instance, when introducing a new document type and updating document records with that new label, such as when the team recently split out "Legal query or responsum" from the larger "Legal document" category. Likewise, the team regularly prepares batch imports of transcription content from published scholarship. Eventually, they will also ingest a batch import of machine-generated transcriptions from the HTR4PGP project.

Both documents and fragments data files include timestamps for the date the record was initially entered and when it was last modified in the database. As analysis shows, hundreds of documents are modified every week (**Figure 11**). A look at the historic totals of major entities in the PGP datasets shows a steady growth across the data, with occasional jumps enabled by changes in workflow or technology, such as support for transcription editing in PGPv4.9 (October 2022) or a significant push to work on or import new content (**Figure 12**).

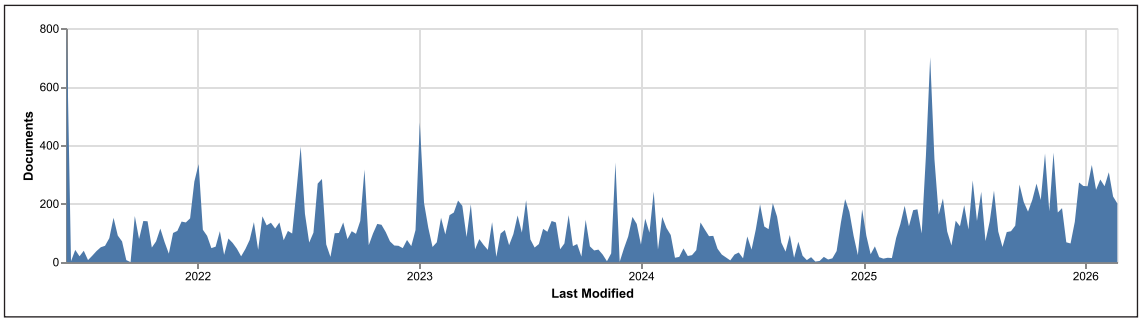

Figure 11: This chart of documents by last modification date in the database shows regular activity with spikes of more intense work.

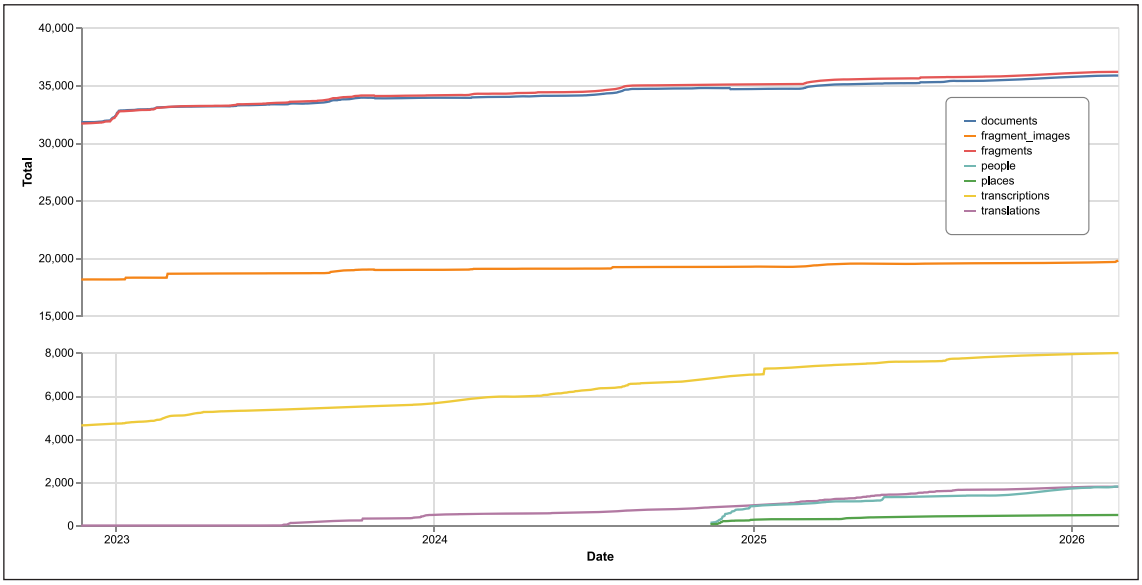

Figure 12: A chart of historic totals of various aspects of PGP data (documents, fragments, images, transcriptions, translations, and later people and places) shows steady growth at different scales.

Because the database is of extraordinarily ancient vintage—forty years is an eternity in digital humanities terms—the datasets include records entered as early as 1986. This longue durée metadata has some potential in itself for analysis.

New versions of the PGP datasets will be published quarterly in order to provide reasonably recent, stable snapshots that can be used for citation and reproducible research as well as to provide larger checkpoints for tracking changes in the long-lived data of a large-scale research project.[18]

Ongoing and Future Work

The Cairo Geniza has long had a fascinating, puzzle-box like quality for scholars. Specialists still routinely make new discoveries, some important, some whimsical. Alan Elbaum recently discovered an unknown document written in the handwriting of Moses Maimonides (1138–1204), a philosopher and physician who spent the last forty years of his life in Fustat and is perhaps the most famous Jewish person of the Middle Ages (Ashur and Elbaum; "List or Table: T-S AS 202.396"). Elbaum also discovered a Fatimid government report from 1109 CE about battles with Crusaders during the siege of Tripoli. The fragment provides an astonishing new perspective on a well-known historical event (Elbaum et al.; Rustow et al., "Fragment of the Month"). Koeser came across a delightful drawing of a Nile boat (**Figure 13**) while implementing the random sort feature for PGPv4.2 in March 2022. As it turns out, Goitein had known

Figure 13: Recto of T-S K5.82 includes a drawing of a Nile boat ("Paraliterary Text: T-S K5.82").

[18] Version 1.0 was published in July 2025 and version 1.1 in February 2026. In the future, the process will be more streamlined to support regular publication.

of the fragment, but no one had yet described it in the PGP ("Paraliterary Text: T-S K5.82"). Most recently, Elbaum has been working with Gideon Bohak and Oded Zinger to decipher encrypted letters (Schmierer-Lee), an effort the transcriptions collected in these datasets may be able to accelerate. The development of PGPv4 and the publication of these datasets bear the potential to open up the materials and their many remaining discoveries to work by scholars with new areas of expertise.

The Princeton Geniza Lab is continuing to expand and improve the PGP data, including filling out data on people and places and adding new transcriptions and translations. The project has drawn engagement from students eager to work with humanities data and to contribute to active research projects. Undergraduate projects have included network graphs and visualizations, automatic image rotation, Handwritten Text Recognition (HTR), and data cleaning and refinement to help with cataloging efforts.

An ongoing partnership between Rustow and Daniel Stoekl Ben Ezra ("Handwritten Text Recognition"; Stoekl Ben Ezra et al., *MiDRASH Geniza 01 HTR Model*) has been leveraging the eScriptorium platform to train and fine-tune HTR models to segment and transcribe PGP materials written in Hebrew script. The project uses the existing PGP scholarly transcriptions as a starting point for training data. The transcriptions these models help generate will eventually be incorporated into the PGPv4 database and datasets, expanding the textual content available for researchers to analyze.[19]

As the amount of full-text content increases, so, too, will the data's research potential. The need will also become more acute to leverage scalable computational methods, including natural language processing (NLP), machine learning, and multimodal language models. One current effort, led by the Princeton Geniza Lab's research software engineer, Mohamed Abdellatif, is adapting and refining an existing solution for transliterating Judaeo-Arabic to Arabic (Weisberg Mitelman et al.) in order to apply it to PGP transcriptions of Judaeo-Arabic texts, which make them more broadly accessible to the Arabic-reading public (Abdellatif et al.). This effort will also make the content more computationally accessible: transliteration will allow researchers to work with the material using existing Arabic language models and other NLP tools.

Other projects have also been inspired by PGPv4. Koeser is now leading a collaboration to develop the Python library *undate* for reasoning with uncertain and partially known dates with mixed precision and calendars (Koeser et al. "Undate"; Koeser et al., *Undate Python Library*). This library incorporates and generalizes PGPv4's

[19] Some of these efforts have been folded into recent, more ambitious work to generate automatic transcriptions of the full Cairo Geniza under the auspices of the MidRASH project (Stoekl Ben Ezra et al. "MiDRASH Automatic Transcriptions").

solutions for calendar conversion and ambiguous dates, with improved parsing and temporal logic, and may eventually be reintegrated into PGPv4 to improve date handling. It has the potential to make it easier to work with the dates in the PGP datasets and others like them.

Other possibilities include mining specific subsets of PGP materials for information, such as reconstructing epistolary networks from letters, or figuring out what percentage of real estate in medieval Cairo was owned by women (Richman). The availability of images in combination with descriptions and transcriptions offers possibilities for multimodal analysis, such as clustering documents visually to identify similar scribal characteristics or individual hands, or identifying other unsuspected commonalities across the materials. PGP images bear further analysis, and torn fragments could be digitally restored either within the PGPv4 or based on future versions of the dataset that provide access to document-centric arrangements of images. Analysis of PGP data in combination with other datasets and sources, such as the historic coins identified by Dudley and Elbaum or the historical places in the World Historical Gazetteer, offers yet more opportunities for new discoveries.

Finally, as the product of a long-running project, the PGP dataset also offers possibilities for meta-analysis of digital research projects. How have the affordances of changing technologies affected data entry? Tagging and description by researchers with particular interests have resulted in some unevenness of the data, and are footprints detectable across the archive that surface some details but miss others. How will the integration of machine-generated transcription transform PGP research?

By publishing and writing about the PGP datasets, we hope to open up the documentary *geniza* puzzle box for a new set of scholars and new modes of inquiry. We look forward to seeing what new discoveries will be made by cultural analysts and digital and computational humanities researchers, and to the fruitful interchange between scholars with different domain expertise and training.

Acknowledgments

The authors would like to thank Mohamed Abdellatif, Amel Bensalim, Rachel Richman, Ksenia Ryzhova, Laure Thompson, and Jeri Wieringa for their helpful comments and suggestions on previous drafts of this essay, and Ksenia Ryzhova for her logistical support during the final stages of completion.

Competing Interests

The authors have no competing interests to declare.

Works Cited

Abdellatif, Mohamed, Joel U. Bretheim, and Marina Rustow. "Machine Transliteration of Long Text with Error Detection and Correction." *Journal of Data Mining & Digital Humanities* NLP4DH (I. Historical and Linguistic Approaches), 7 Mar. 2025. https://doi.org/10.46298/jdmdh.15020.

Ashur, Amir, and Alan Elbaum. "New Maimonidean Documents." *From the Battlefield of Books: Essays Celebrating 50 Years of the Taylor-Schechter Genizah Research Unit*, edited by Nick Posegay, Magdalen M. Connolly, and Ben Outhwaite, Brill, 2024, pp. 10–20. https://doi.org/10.1163/9789004712331_003.

"Background and History of the Project." *Princeton Geniza Project*. https://genizaprojects.princeton.edu/pgpsearch/about.html. Accessed 7 Jan. 2025.

Bauer, Jean. "Tales from the Port: Part 2—Migrating the Database." *Jean Bauer*, 4 Oct. 2012. https://jeanbauer.com/packets/2012/10/tales-from-the-port-part-2-migrating-the-database/.

Budak, Nick, Marina Rustow, Rebecca Sutton Koeser, Gissoo Doroudian, Natalia Ermolaev, Stephanie Luescher, and Rachel Richman. *CDH Project Charter—Princeton Geniza Project 2020–2021*, Center for Digital Humanities, 2020. https://doi.org/10.5281/zenodo.4081714.

"Chronology of Unicode Version 1.0." *Unicode*, Unicode Consortium. https://www.unicode.org/history/versionone.html. Accessed 10 Jan. 2025.

Cohen, Mark R. "The Princeton University Geniza Project: Using the Internet for Jewish and Islamic Research." *Language, Culture, Computation: Computing of the Humanities, Law, and Narratives; Essays Dedicated to Yaacov Choueka on the Occasion of His 75th Birthday, Part II*, edited by Nachum Dershowitz and Ephraim Nissan, Springer, 2014, pp. 38–46. https://doi.org/10.1007/978-3-642-45324-3_3.

Cordell, Ryan. "What Makes Computational Evidence Significant for Literary-Historical Argument?" *Ryan C. Cordell*, 27 Jul. 2017. https://ryancordell.org/research/dh/computational-evidence-for-literary-historical-argument/.

Dudley, Matthew, and Alan Elbaum. "Coins of the Cairo Geniza: A Paleographical Glossary for Dating Textual Fragments." 2023. https://geniza.github.io/paleographicalglossary/.

Elbaum, Alan, and Ksenia Ryzhova. "Ḥalfon b. Menashshe Ha-Levi," *Princeton Geniza Project*, n.d. https://geniza.princeton.edu/en/people/halfon-b-menashshe/.

Elbaum, Alan, Paul M. Cobb, and Marina Rustow. "'The Franks, May God Forsake Them': A Fatimid Government Document from 1109 about Skirmishes in the Levant." *Crusades*, vol. 24, 2025. https://doi.org/10.1080/14765276.2025.2605631.

Esten, Emily. “Scribes: Issue Introduction.” *Startwords*, no. 2, Dec. 2021. https://startwords.cdh.princeton.edu/issues/2/.

“Handwritten Text Recognition.” *Princeton Geniza Lab*, n.d. https://genizalab.princeton.edu/projects/handwritten-text-recognition.

Katz, Menachem. “Reuniting Minute Fragments.” *Genizah Fragments: The Newsletter of Cambridge University’s Taylor-Schechter Genizah Research Unit*, no. 73, Apr. 2017, p. 3.

Koeser, Rebecca Sutton. “Visualizing the Collections.” *Princeton Prosody Archive*, Jan. 2020. https://doi.org/10.17613/VBF16-3D987.

Koeser, Rebecca Sutton, Benjamin Hicks, Kevin Glover, Kevin McElwee, Nick Budak, Xinyi Li, and Jean Bauer. *Derrida-Django v1.3*. *Zenodo*, 25 Oct. 2021. https://doi.org/10.5281/zenodo.5602038.

Koeser, Rebecca Sutton, Cole Crawford, Julia Damerow, Malte Vogl, and Robert Casties. *Undate Python Library*, *Zenodo*, 3 Sep. 2026. https://doi.org/10.5281/zenodo.18270524.

Koeser, Rebecca Sutton, Julia Damerow, Robert Casties, and Cole Crawford. “Undate: Humanistic Dates for Computation.” *Computational Humanities Research* 1, 2025, p. e5. https://doi.org/10.1017/chr.2025.10006.

Koeser, Rebecca Sutton, Nick Budak, Gissoo Doroudian, Kevin McElwee, Benjamin Hicks, and Xinyi Li. *Princeton-CDH/Mep-Django: v.1.5.4*, *Zenodo*, 3 Sep. 2025. https://doi.org/10.5281/zenodo.3834178.

Kotin, Joshua, and Rebecca Sutton Koeser. “Shakespeare and Company Project Data Sets.” *Journal of Cultural Analytics* 7, no. 1, 2022. https://doi.org/10.22148/001c.32551.

Kotin, Joshua, and Rebecca Sutton Koeser. “The World of Shakespeare and Company: An Introduction.” *Journal of Cultural Analytics*, vol. 9, no. 2, 2024. https://doi.org/10.22148/001c.116905.

“Legal Document: T-S 6J3.5.” 1127. *Princeton Geniza Project*. https://geniza.princeton.edu/en/documents/3637/.

“Letter: ENA 2805.14.” 1050. *Princeton Geniza Project*. https://geniza.princeton.edu/en/documents/695/.

“Letter: T-S 12.392.” 1103. *Princeton Geniza Project*. https://geniza.princeton.edu/en/documents/9690/.

“List or Table: T-S AS 202.396.” n.d. *Princeton Geniza Project*. https://geniza.princeton.edu/en/documents/22676/.

“Paraliterary Text: T-S K5.82.” n.d. *Princeton Geniza Project*. https://geniza.princeton.edu/en/documents/8483/.

Richman, Rachel. *Women’s Labor and Property in Cairo Genizah Documents*. Princeton University, 2026. PhD dissertation.

Rustow, Marina. *The Lost Archive: Traces of a Caliphate in a Cairo Synagogue*. Princeton UP, 2020.

Rustow, Marina. “Dating Problems? Ask the Princeton Geniza Project Team.” *The Center for Digital Humanities at Princeton*, 18 Nov. 2020. https://cdh.princeton.edu/blog/2020/11/18/dating-problems-ask-princeton-geniza-project-team/.

Rustow, Marina. "Fragmentology: A Geniza Manifesto." *Digital Philology: A Journal of Medieval Cultures*, vol. 13, no. 2, 2024, pp. 247–76. https://doi.org/10.1353/dph.2024.a940698.

Rustow, Marina, Alan Elbaum, and Paul M. Cobb. "Fragment of the Month: January 2026: A Fatimid Interoffice Memo about the Franks, May God Forsake Them." *Cambridge University Library*, Jan. 2026. https://www.lib.cam.ac.uk/collections/departments/taylor-schechter-genizah-research-unit/fragment-month/fotm-2026/fragment.

Rustow, Marina, Rebecca Sutton Koeser, Nicholas Budak, Gissoo Doroudian, and Rachel Richman. *CDH Project Charter—Princeton Geniza Project Year 2, 2021–2022*. *Zenodo*, 11 Apr. 2022. https://doi.org/10.5281/zenodo.7841822.

Rustow, Marina, Rebecca Sutton Koeser, Rachel Richman, Ksenia Ryzhova, Amel Bensalim, and Mohamed Abdellatif. *Princeton Geniza Project datasets*. *Zenodo*, 20 Feb. 2026. https://doi.org/10.5281/zenodo.18716581.

Schmierer-Lee, Melonie. "Q&A Wednesday: Secret Codes and Erotic Prose, with Alan Elbaum." *Genizah Fragments*, 22 Jun. 2022. https://genizahfragments.lib.cam.ac.uk/2022/06/22/qa-wednesday-secret-codes-and-erotic-prose-with-alan-elbaum/.

"State Document: Bodl. MS Heb. B 18/23." 1024. *Princeton Geniza Project*. https://geniza.princeton.edu/en/documents/39034/.

"State Document: ENA 3974.3." 1021. *Princeton Geniza Project*. https://geniza.princeton.edu/en/documents/19304/.

Stoekl Ben Ezra, Daniel, Marina Rustow, Benjamin Kiessling, Luigi Bambaci, Jessica Parker, Gavin McDowell, Nachum Dershowitz, et al. *MiDRASH Geniza 01 HTR Model*. *Zenodo*, 22 Feb. 2026. https://doi.org/10.5281/zenodo.18732245.

Stoekl Ben Ezra, Daniel, Luigi Bambaci, Benjamin Kiessling, Hayim Lapin, Nurit Ezer, Elena Lolli, Marina Rustow, et al. *MiDRASH Automatic Transcriptions of the Cairo Geniza Fragments*. *Zenodo*, 27 Nov. 2025. https://doi.org/10.5281/zenodo.17734473.

Weisberg Mitelman, Daniel, Nachum Dershowitz, and Kfir Bar. "Code-Switching and Back-Transliteration Using a Bilingual Model." *Findings of the Association for Computational Linguistics: EACL 2024*, edited by Yvette Graham and Matthew Purver, 17–22 Mar. 2024, pp. 1501–11, Association for Computational Linguistics. https://aclanthology.org/2024.findings-eacl.102/.