Manuscript Studies

A Journal of the Schoenberg Institute for Manuscript Studies

VOLUME 5, NUMBER 2

(Fall 2020)

Manuscript Studies (ISSN 2381-5329) is published semiannually
by the University of Pennsylvania Press

The Schoenberg Institute
for Manuscript Studies
University of Pennsylvania Libraries

MANUSCRIPT STUDIES

VOLUME 5, NUMBER 2
(Fall 2020)

ISSN 2381-5329

Published by the University of Pennsylvania Press,
3905 Spruce Street, Philadelphia, PA 19104.

Printed in the U.S.A. on acid-free paper.

Manuscript Studies brings together scholarship from around the world and across disciplines related to the study of premodern manuscript books and documents, with a special emphasis on the role of digital technologies in advancing manuscript research. Articles for submission should be prepared according to the *Chicago Manual of Style*, 16th edition, and follow the style guidelines found at http://mss.pennpress.org.

2020 SUBSCRIPTION INFORMATION:
Single issues: $30
Print and online subscriptions: Individuals: $40; Institutions: $94; Full-time Students: $30
International subscribers, please add $19 per year for shipping.
Online-only subscriptions: Individuals: $32; Institutions: $82

Please direct all subscription orders, inquiries, requests for single issues, address changes, and other business communications to Penn Press Journals, 3905 Spruce Street, Philadelphia, PA 19104. Phone: 215-573-1295. Fax: 215-746-3636. Email: journals@pobox.upenn.edu. Prepayment is required. Orders may be charged to MasterCard, Visa, and American Express credit cards. Checks and money orders should be made payable to "University of Pennsylvania Press" and sent to the address printed directly above.

One-year subscriptions are valid January 1 through December 31. Subscriptions received after October 31 in any year become effective the following January 1. Subscribers joining midyear receive immediately copies of all issues of Manuscript Studies already in print for that year.

Postmaster: send address changes to Penn Press Journals, 3905 Spruce Street, Philadelphia, PA 19104.

Visit *Manuscript Studies* on the web at mss.pennpress.org.

Scribes, Scholars, and Scripts: Reviewing Data from *Scribes of the Cairo Geniza*

Emily Esten
University of Pennsylvania Libraries

Scribes of the Cairo Geniza (scribesofthecairogeniza.org) is an international partnership led by the University of Pennsylvania Libraries and the Zooniverse (zooniverse.org), the world's largest platform for online crowdsourced research. The citizen science project invites members of the public to help classify and transcribe fragments of medieval and premodern manuscripts from the Cairo Geniza. This corpus of more than 350,000 documents, the majority of which date from the tenth to thirteenth centuries, was kept in a storeroom (or "geniza") of the Ben Ezra synagogue in Fustat (Old Cairo) until the late nineteenth century.[1] Geniza fragments serve as a time capsule of Mediterranean Jewish history in a period when 90 percent of Jews lived in the Islamic world. With the crowdsourced classification and transcription data produced through this project, Scribes of the Cairo Geniza has the potential to rewrite the

1 An overview and introduction to this project is available at Laura Newman Eckstein, "Of Scribes and Scripts: Citizen Science and the Cairo Geniza," *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies* 3, no. 1 (2018): 208–14, https://doi.org/10.1353/mns.2018.0008.

history of the premodern Middle East, Mediterranean and Indian Ocean trade, and the Jewish diaspora.

Crowdsourcing and citizen science projects like *Scribes of the Cairo Geniza* engage members of the public (referred to here as "citizen scientists") with professional researchers to complete a research project through small tasks, with an aim toward converting public efforts into measurable data and discoveries. Citizen science projects have been most effective in cases when the tools provided are appropriate for the crowd and tasks at hand, project teams support the production of high-quality data, and data are useful and accessible by the research communities for whom they are created.[2] *Scribes of the Cairo Geniza* is unusual not only as a non-English digital humanities project (though this number is steadily growing in the field at large), but as a multilingual undertaking in the crowdsourcing community. The project interface is available in Arabic, English, and Hebrew in order to engage and support the participation of non-English-speaking communities. Over 9,500 citizen scientists from twenty-seven different countries had participated in the project as of 31 January 2020.

The project utilizes digitized images from Geniza collections at the University of Pennsylvania Libraries, the Library of the Jewish Theological Seminary, the Taylor-Schechter Genizah Research Unit at Cambridge University Libraries, the University of Manchester Library, the Bodleian Libraries at the University of Oxford, Columbia University Libraries, and the National Library of Israel. The current number of viewable pages is 68,434 (an estimated 20 percent of the total Geniza). These pages vary in content from the religious works expected in a typical geniza to economic, political, social, and communal documents as well as personal correspondence. The publicly available metadata vary greatly, depending on the size, scope, and interest of the image partner's collections. According to the Friedberg Genizah Project, a privately funded effort to put all Geniza

2 Samantha Blickhan, Coleman Krawczyk, Daniel Hanson, et al., "Individual vs. Collaborative Methods of Crowdsourced Transcription," *Journal of Data Mining and Digital Humanities* (2019): 1–33, https://hal.archives-ouvertes.fr/hal-02280013v2.

fragments online, 30 percent of all Geniza shelfmarks have an existing catalog record; less than 15 percent, however, have transcriptions.[3]

Our project goals are the following:

- Provide our community of citizen scientists opportunities to view and decipher Cairo Geniza fragments
- Contribute to the classification of fragments by script type and content
- Produce transcriptions of the material as open data that will help in the work of historians, linguists, and other scholars of this material.[4]

Results of the Sorting Phase

As discussed previously, in order to transcribe subjects (the Zooniverse term for individual pages or fragments), we first needed more information about the subjects themselves. In the first phase of the project, which the project team refers to as the "initial sorting phase," we asked citizen scientists to sort Cairo Geniza fragments based on script type, style, and other visual characteristics.

The initial sorting phase began on 8 August 2017 and was completed on 8 February 2019. In this phase, citizen scientists sorted 40,109 subjects from the University of Pennsylvania, the Library of the Jewish Theological Seminary, and the Taylor-Schechter Genizah Research Unit at Cambridge University. Over seven thousand users participated in this phase of the project.[5]

For each subject, a volunteer answered the following questions:[6]

---

3 "Friedberg Genizah Project," Friedberg Genizah Project, 2013, https://fgp.genizah.org.

4 "About," *Zooniverse.org*, https://www.scribesofthecairogeniza.org/about.

5 This data is publicly available at Emily Esten, "Dataset: Scribes of the Cairo Geniza, Sorting Phase, August 2017–February 2019," Scholarly Commons (Judaica Digital Humanities, Kislak Center), 31 January 2020, https://repository.upenn.edu/cairogeniza/1/.

6 A full explanation of these questions and the reason for their selection can be found in Eckstein, "Of Scribes and Scripts," 208–14.

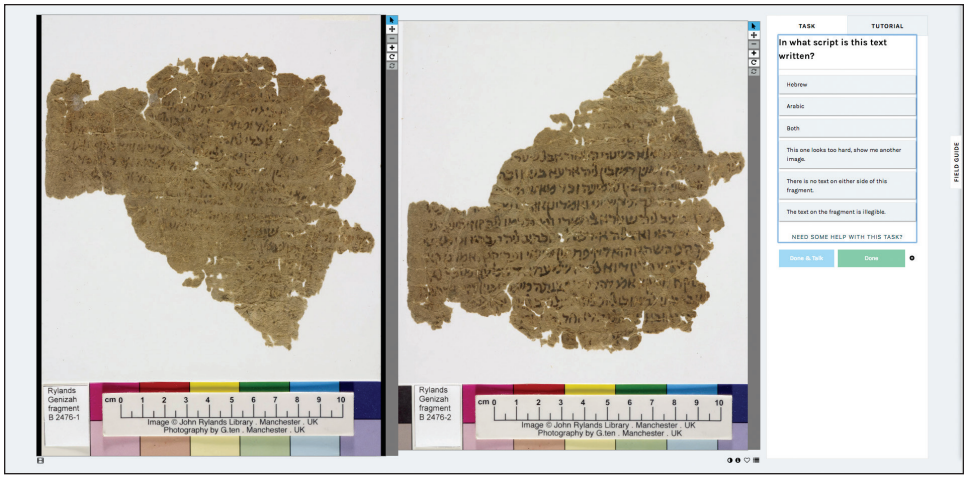

FIGURE 1. A screenshot of the Scribes of the Cairo Geniza sorting interface in English, starting on the first question.

- In what script is this text written?
- Is the script written in a formal or informal style?
- Do you see evidence of binding, justified margins, or top-corner page wear?[7]
- Identify (using the point tool) visual characteristics: colons (Hebrew), diacritics (Hebrew), diagonal and/or perpendicular text in the margin (Hebrew and Arabic), or seals (Arabic).[8]

Following these questions, a volunteer may select "done & talk" to provide additional information they may have, ask researchers questions, or tag a subject.[9]

---

7 This question was added in June 2019 to assess if a subject was once part of a book.

8 Explanations for the inclusion of these specific characteristics can be found in Eckstein, "Of Scribes and Scripts," 210–11. Identification of a horizontal line above a word in a Hebrew script subject was removed in June 2019.

9 The figures below come directly from the report of the phase 1 data. This data is publicly available at Emily Esten, "Reviewing Sorting Phase Data," Scholarly Commons (Judaica Digital Humanities, Kislak Center), 31 January 2020, https://repository.upenn.edu/cairogeniza/2/.

- 17,031 subjects (42.4 percent) are classified as Easy Hebrew.
- 460 subjects (1 percent) are classified as Easy Arabic.
- 18,516 subjects (46 percent) are classified as Challenging Hebrew.
- 1,397 subjects (3 percent) are classified as Challenging Arabic.

This data set is the first comprehensive list of Arabic fragments in the Cairo Geniza, with specific counts from the given collections. Most notably, about 7 percent of the Geniza fragments were found to be out of scope for transcription, whether due to difficulty, blank pages, or subjects beyond repair. This is extremely helpful for the project to know which fragments may require scholarly attention or are not of interest at all.

Regarding the visual characteristics, the following numbers mean at least one volunteer identified the fragment as having the specific characteristic:

- 9,108 subjects (22 percent) were classified as having evidence of diagonal and/or perpendicular text in the margin.
- 416 subjects (1 percent) were classified as having evidence of seals.
- 3,734 subjects (9 percent) were classified as having evidence of a horizontal line above a word.
- 11,978 subjects (30 percent) were classified as having evidence of a colon in the text.
- 4,398 subjects (11 percent) were classified as having evidence of a dot, vowel, or diacritic.
- 6,457 subjects (16 percent) were classified as having evidence of justified margins.
- 5,958 (14.8 percent) subjects were classified as having evidence of binding.
- 3,707 subjects (9 percent) were classified as having evidence of top corner page wear.

Of particular note, for 65 percent of subjects, citizen scientists were in unanimous agreement of the script type (Hebrew, Arabic, or both). Of the remaining subjects, a majority (26 percent) of citizen scientists chose between two responses, and citizen scientists disagreed on 7 percent of subjects whether

the script type was Hebrew or Arabic. Considering the range of expertise of the citizen scientists involved in the project, this signifies that the collective wisdom of the crowd was largely in agreement.

As noted in all crowdsourcing projects, this does not mean these subjects are definitely Hebrew or Arabic, are formal or informal, or have these visual characteristics. However, the data, feedback, and consensus from citizen scientists give the project team confidence that the crowdsourcing model continues to be an effective solution for sorting texts.

As the research team reviews this sorting data, they plan to improve upon the field guide and sorting workflow for effective and accurate identification with future collections. The sorting workflow remains open for classification: new subjects are routinely added as image partners join the project, and these new subjects are appropriately sorted into the transcription workflows on an ongoing basis.

Launch and Initial Results of the Transcription Phase

Starting with the 17,031 Easy Hebrew and 460 Easy Arabic subjects sorted by citizen scientists, the second phase of the project launched on 6 March 2019. In this phase, which the project team refers to as the "transcription phase," we are asking citizen scientists to transcribe Cairo Geniza fragments that have been sorted as Easy Hebrew and Easy Arabic script. The transcription workflows have been broken into levels of difficulty so that citizen scientists can participate based on their level of confidence in the task.

Zooniverse developers created the custom front-end interface shown in figure 2.[10] Upon selecting a transcription workflow (either Easy Hebrew or Easy Arabic), citizen scientists are presented with a fragment for transcription. To the left (the opposite in Hebrew and Arabic interfaces) is a toolbar for viewing the image by panning, zooming in and out, rotating, or inverting

10 The code for the custom front end is available at "Scribes of the Cairo Geniza," Github (Zooniverse), 6 January 2020, https://github.com/zooniverse/scribes-of-the-cairo-geniza.

FIGURE 2. A screenshot of the custom Scribes of the Cairo Geniza transcription interface in English, including the onscreen clickable keyboard.

colors. To the right (the opposite in Hebrew and Arabic interfaces) is a Subject Info toolbar, containing shelfmark and attribution information about the subject at hand.

Citizen scientists select "add transcription" from the toolbar, and then place a dot at the start and end of a chosen line of text. A transcription box automatically appears after the second dot is placed. This project took technical inspiration from *Ancient Lives,* a 2011 Zooniverse project that invited the public to transcribe ancient Greek fragments using a clickable onscreen keyboard. For Hebrew transcriptions, citizen scientists have the option to choose between modern characters and images of twenty different script types to aid in transcription. Using paleographical classifications developed by Judith Olszowy-Schlanger at the École Pratiques des Hautes Études, Paris, Laura Newman Eckstein at Penn Libraries developed keyboards that use images from Geniza fragments in place of each character. The volunteer is then able to compare the unidentified characters they are transcribing with images of identified character types from Geniza fragments.

Using a keyboard that matches the style of the fragment allows for a user who does not necessarily know Hebrew to transcribe solely by matching letter

shapes. For Arabic transcriptions, an adapted version of the modern North African keyboard, combined with a 1920s Egyptian typewriter, showcases each letter's different forms (isolated, initial, medial, and final).[11] For both Hebrew and Arabic transcription workflows, users also have the option to use their physical keyboards for transcription rather than the onscreen one.

This project uses an independent workflow transcription approach, as piloted in Zooniverse's *Shakespeare's World*.[12] In this workflow, citizen scientists transcribe fragments independently of one another, and the results are aggregated to make a single, "best" version. The aggregated result of each line is also given a confidence score, which shows how much (or how little) agreement existed in each line for each subject. When a confidence score of 3 is reached or the line has been transcribed by at least seven people, the line is grayed out and retired from transcription. A page is retired when all lines on a page have been grayed out.[13]

Citizen scientists also are provided additional content within the interface for assistance. The following options are provided as text modifiers for transcription to indicate special occurrences within the text:

- *Insertion*: for text that has been added in
- *Deletion*: for text that has been crossed out
- *Damaged*: for text that is obscured or destroyed due to physical damage
- *Drawing*: to mark drawings that interrupt lines of text
- *Grid*: to indicate tabular text
- *Divine Name*: to use in place of the Divine Name, for citizen scientists who, on the basis of religious, moral, or ethical beliefs, oppose transcribing the Divine Name as written on the page.

---

11 "About," *Zooniverse.org,* https://www.scribesofthecairogeniza.org/about,.

12 "Shakespeare's World," *Zooniverse.org,* 8 December 2015, https://www.shakespearesworld.org.

13 More information about the process of aggregating transcriptions can be found in Coleman Krawczyk, "Aggregating Annotations in the Anti-Slavery Manuscripts Project," *Boston Public Library Blogs*, 26 November 2018, https://www.bpl.org/blogs/post/aggregating-annotations-in-the-anti-slavery-manuscripts-project.

As of 31 January 2020, nearly eight hundred citizen scientists had attempted to transcribe fragments. Eighty-nine Hebrew fragments and fifty-three Arabic fragments have been retired, meaning a volunteer noted that all lines of that fragment have been completely transcribed. This accounts for 10 percent of all Easy Arabic fragments: an estimated time to completion for the Easy Arabic workflow is less than one year. These subjects are currently under review by the content specialists on the research team for quality control.

When subjects are retired from the sorting workflow, they transfer into the appropriate transcription workflows on an ongoing basis. While there is no announced date for the launch of the challenging workflows, the research team plans to take under consideration volunteer feedback from the first year of the transcription phase in order to best assist citizen scientists in the challenging transcription process.

As noted in the *Ancient Lives* project, the data pipeline for these crowdsourced transcriptions has massive potential to redefine how scholars of Judaic and Near Eastern studies can interact with the Geniza through digital humanities tools.[14] The transcription results of the project are still limited but have already been applied by partner research teams to further Geniza research. Collaborators at the Princeton Geniza Lab at Princeton University, headed by Marina Rustow, will make use of the consensus transcriptions toward their goal of a technological infrastructure that links images, transcriptions, translations, and previous research materials in mapping the entire documentary Geniza corpus. By dividing the arduous process of mapping the Geniza into discrete parts—identification, description, and translation—we can extend the accessible corpus quickly. This assists in furthering communication among scholars and volunteer colleagues, and opens the products of research.

Collaborators at the e-Lijah Lab at the University of Haifa are using initial transcription data as part of their ongoing work to apply handwritten text recognition to medieval Hebrew texts. During a 2019 hackathon at the

14 A. C. Williams, John F. Wallin, Haoyu Yu, et al., "A Computational Pipeline for Crowdsourced Transcriptions of Ancient Greek Papyrus Fragments," 2014 IEEE International Conference on Big Data, Washington, DC, 2014, 100–105. DOI: 10.1109/BigData.2014.7004460.

University of Haifa, participants led by Vered Raziel-Krezmer developed a prototype for automatic identification and cataloging of Geniza fragments. This prototype uses consensus transcriptions of literary Geniza fragments to quickly align with matches in Sefaria, an open-source library of Jewish texts and their interconnections, in Hebrew and in translation. Comparing transcriptions of Geniza materials to existing Jewish texts, scholars may highlight changes in tradition over the millennium or identify older materials (as occurred in the "Scribes of the Seder" initiative).[15] Lastly, the project team hopes that all the data produced through *Scribes of the Cairo Geniza* will be the basis for a future database of the project.

Engaging Citizen Scientists

The classification and transcription data produced through *Scribes of the Cairo Geniza* is certainly worthy of note, and provides a strong model for deciphering a large corpus of difficult texts through public digital scholarship. This is further recognized through the project's public engagement efforts, which have brought compelling, contemporary scholarship to a burgeoning community of citizen scholars.

For the project team's purposes, the easiest form of engagement occurs in the Talk boards. As identified in previous Zooniverse projects like *Shakespeare's World,* the use of online forums for tagging, linking, and discussion play a crucial role in the knowledge production process.[16] While there is still a clear distinction between volunteers and researchers, the Talk forums allow for "more open and collaborative form of knowledge production," letting the public ask and take charge of questions and inspiring new avenues

15 The "Scribes of the Seder" initiative in March 2018 had volunteers tag portions of various haggadot to invite people to look critically at the texts used for Passover. The patchwork haggadah is available on Sefaria at https://www.sefaria.org/sheets/105137?lang=bi. More on the initiative can be viewed at https://medium.com/@judaicadh/sederscribes-1866981146e6.

16 Frauke Rohden, Christopher Kullenberg, Niclas Hagen, and Dick Kasperowski, "Tagging, Pinging and Linking—User Roles in Virtual Citizen Science Forums," *Citizen Science: Theory and Practice* 4, no. 1 (2019): 19. DOI: http://doi.org/10.5334/cstp.181.

for conversation.[17] Over 16 percent of fragments were tagged on the Talk boards by citizen scientists during the sorting phase. The majority of tags (over 600 unique tags in the project) were user-generated and grouped into the following categories:[18]

- *Project*: These tags refer to comments about the project (#interface, #notgenizah, #mismatched) or are subjective tags (#weird, #unusual, #beautiful).
- *Language/Script*: These tags refer to the scripts (#arabic_script, #hebrew_script, #latin) or languages (#coptic, #english, #italian, #ladino) on the fragment.
- *Condition:* These tags refer to the condition of the subject (#microfragment, #faded, #damaged, and #reuse are in this category).
- *Feature*: These tags refer to specific markings (#aleph, #charakteres, #strikethrough), visual characteristics (#colons) or distinctive features (#binding, #diagonal_text, #marginalia) of a fragment.
- *Genre*: If it is not a religious text, these tags help identify what type of fragment it might be. These might be themes/terms referenced in the text (#agriculture, #magical, #literary) or types of fragments (#titlepage, #reed-trial, #legal_document).
- *Judaica*: These tags make up the bulk of the Cairo Geniza—they vary from historical persons to biblical references, holidays to specific genres of literary texts.

Similarly, the comments on the Talk boards reveal personal engagement with the project and its work. Regular conversations, shared on the project's blog as the "Talking the Talk" series, demonstrate engagement between scholars and citizen scientists working together to teach each other, and generating conversations about Jewish history and culture, linguistics, and the work of libraries in the preservation of materials. Citizen scientists have

17 Rohden et al., "Tagging, Pinging and Linking."
18 Esten, "Reviewing Sorting Phase Data."

favorited and created collections of interesting fragments for personal use, and even downloaded images onto their personal computers for sharing.

Citizen science projects also offer the opportunity to engage volunteers in analyzing the data and materials produced through their efforts. Several engaged citizen scientists have taken the initiative to begin their own research into Geniza materials, including a crowdsourced collection of decorative pictorial marginalia within the Geniza, glossaries of various spellings of the Divine Name, and identification of material of fragments (paper or parchment).[19] Following the questions from user-generated research has helped spark renewed interest in these materials. As one librarian tweeted after promoting a volunteer discovery, "Why to collaborate, reason #5235: You might have curatorial responsibility for a collection, might even have digital images for a manuscript in that collection, and yet you might not know about a gorgeous illumination in a manuscript in same."[20] Celebrating volunteers' knowledge, research, and expertise as researchers in their own right has helped the project grow into a space of learning, communication, and collective ownership over the crowdsourced data.

In the 2019–2020 survey of participants in the project, 84 percent of respondents identified their interest in participation as "Contributing—I like to contribute to historians' work."[21] As one respondent explained, "Like most people, I will never handle an ancient piece of writing, only see pictures or film of it. But digitally manipulating these documents and helping scholars eventually interpret them makes me feel a very real connection to the source material, the people who wrote it, and the time they lived in."[22] These fragments provide compelling and meaningful insights

---

19 Emily Esten, "Cultivating Community with the Cairo Geniza," *Medium*, 7 November 2019, https://medium.com/@judaicadh/mcn2019-cultivating-community-with-the-cairo-geniza-eaf5182c28cd.

20 Michelle Chesner, Twitter post, 22 August 2019, 12:32 p.m., https://twitter.com/hchesner/status/1164576132869623810?s=20.

21 Emily Esten, "Who Are the #GenizaScribes? 2019–2020 Community Survey Report," Scholarly Commons (Judaica Digital Humanities, Kislak Center), 4 February 2020, https://repository.upenn.edu/cairogeniza/3/.

22 Esten, "Who Are the #GenizaScribes?"

into the medieval world, and an opportunity for citizens to pursue project-level research with researcher support.

The Cairo Geniza, once stored for burial in a synagogue in Fustat, now lives online as documentary evidence of a community's collection of its history and culture. As *Scribes of the Cairo Geniza* enters its third year, the project continues to bring together researchers, institutions, and citizen scientists around the globe to create a new, virtual community around these fragments, recording their own experiences along the way.