

Roni Shweka, Yaacov Choueka, Lior Wolf, Nachum Dershowitz, Masha Zeldin

Automatic extraction of catalog data from Genizah fragments' images

The Friedberg Genizah Project, Israel; Tel Aviv University, Israel

The Cairo Genizah is a collection of about 250,000 manuscript-fragments of mainly Jewish texts discovered in the late 19th century. Most of the fragments were written between the 10th and the 14th centuries. Today, the fragments are spread out in libraries and private collections worldwide. The Friedberg Genizah Project ([www.genizah.org](http://www.genizah.org)) is in the midst of a multi-year process of digitally photographing all of the extant fragments. As of March 2011, the virtual library of the project holds over 250,000 digital images, and 200,000 additional images are expected to be integrated over the next few years. Unfortunately, this huge collection is far from being entirely cataloged, despite the ongoing effort to document and catalog all extant fragments. Moreover, the existing catalogs differ greatly in the amount and type of the data they present. Many of them record briefly the content of the fragment, without any information regarding its physical attributes.

We present a system for collecting all catalog data that can be extracted automatically from the fragment's image, mainly: the exact dimensions of the fragment; number of columns; number of lines; size of the margins; the fragment's physical status (torn vertically, horizontally, missing corners); and several additional features. Our system differentiates between bifolios and single pages, and in the first case collects the above data for each page separately. Besides the above attributes, which are expected to be found in every modern catalog, the system extracts some finer data that may be relevant to paleographic studies, such as density of lines (line height, inter-line space) and density of characters (number of characters in a fixed unit of length).

In addition to the detailed physical description of a single fragment, the huge database generated by the system serves for supporting identification of "join" candidates in the Cairo Genizah. A *join* is a set of manuscript-fragments that originate from the same original codex, but are scattered today under different shelfmarks, possibly in several different libraries. In a previous work, we described a system for the automatic identification of joins by ascertaining the degree of handwriting similarity between pairs of fragments. By querying the database

and applying some basic rules for a good match, taking into account the completeness or incompleteness of the fragments, we can significantly improve on the quality of the results obtained by just analyzing the handwriting similarity.

Another aspect introduced in this paper is the proper conditions for taking digital images of manuscripts that are necessary for achieving this kind of results. We argue that, today, the function of such digital imaging is not only conservation and accessibility, but these images should be considered as potential inputs to image-processing algorithms and processes, and the computer should be therefore taken into account as one of the “clients” of the images. Hence, appropriate conditions should be considered in advance when digitizing manuscripts. Among these conditions we mention the following:

- **Choosing the optimal background for foreground-background separation.** The background color should contrast, not only with the color of the fragment material, (vellum or paper), which is some hue of light brown, but also with the color of the ink, usually dark brown or black. Otherwise, text will be erroneously recognized as part of the background, and characters will be interpreted as holes in the fragment. The common practice in some libraries to digitize manuscripts on white, brown or black background should be considered therefore as an imperfect one, because these colors do not contrast well with the manuscript and the ink colors. Our study shows that the best contrast for these colors is provided by a blue background. Indeed, when we started digitizing the huge Genizah collection at the Cambridge University Library, we used blue as the standard background color for all images and the same practice was followed in the digitization of the British Library Genizah collection. Note that since with such contrasting colors the computer can very effectively differentiate between the fragment and its background, it is possible to automatically change the color background to any color desired by the user.
- **Avoiding the use of clips, weight bags, notes, etc.** Every significant element in the image should be easily identified and recognized by the computer, and the best segmentation is achieved by color separation. On the other hand, when there is a need for use of extra elements with no significance to appear, such as elements to hold the fragment or keep it flat, we recommend that they be of the same color as that

of the background. Notes (such as shelfmark numbers) should be of a fixed size and shape, with some apparent icon on them, so as to enable the software to identify them easily.

- **Use of a ruler in the image.** Placing a ruler in the image enables the software to automatically determine the exact dpi of the image, and thus assess the various measures in some recognized unit, such as *cms* or *inches*. This practice is crucial especially when different images are taken with different lenses or when the camera is not fixed in the same position throughout the entire process. The ruler should be distinctive from the fragment; hence a wooden brown ruler or a see-through plastic one will not make a good choice.

Unfortunately, when such aspects are neglected, the application of computerized methods as described above and harvesting their results become unnecessarily difficult, and the quality of obtained results is adversely affected.

#### **Related literature:**

Lerner, HG & Jerchow, S 2006, 'The Penn/Cambridge Genizah fragment project: issues in description, access, and reunification', *Cataloging & Classification Quarterly*, vol. 42, no. 1, pp. 21–39.

Reif, SC 2000, *A Jewish archive from Old Cairo: the history of Cambridge University's Genizah collection*, Curzon Press, Richmond.

Stinson, T 2009, 'Codicological Descriptions in the Digital Age', in M Rehbein, P Sahle & T Schaßan (eds.), *Codicology and Palaeography in the Digital Age - Kodikologie und Paläographie im Digitalen Zeitalter*, Schriftenreihe des Instituts für Dokumentologie und Editorik, vol. 2, BoD, Norderstedt, pp. 35–51.

TEI Consortium 2011, 'Manuscript Description', TEI P5: Guidelines for electronic Text Encoding and Interchange, Version 1.9.1, Viewed 15 March 2011, <<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html>>.

The Friedberg Genizah Project, <<http://www.genizah.org>>.

Wolf, L, Littman, R, German, T, Mayer, N, Dershowitz, N, Shweka, R & Choueka, Y 2011, 'Automatically identifying join candidates in the Cairo Genizah', *International Journal*

*of Computer Vision* (forthcoming. Online publication:

<<http://www.springerlink.com/content/p227026r1124xj30/fulltext.pdf>>).