Genizah Research Enters the Computer Era

Yaacov Choueka

The Fifteenth World Congress of Jewish Studies
Jerusalem, August 2009
A special session: "The Friedberg Genizah Project- Objectives and Accomplishments".

Discovered in 1896 in the attic of a synagogue in the old quarter of Cairo, the Genizah is a large collection of discarded codices, scrolls, and documents. These were written mainly in the 10th to 15th centuries, and mainly in Hebrew and Arabic (usually in Hebrew characters). The documents and fragments are now dispersed in over fifty libraries and collections around the world.

The philanthropically-funded Friedberg Genizah Project is in the midst of a multi-year process of digitally photographing (in full color, at 600dpi) most of the extant manuscripts. The entire collections of the Jewish Theological Seminary in New York, the Alliance Israelite Universelle in Paris, the recently rediscovered collection in Geneva, and many other collections from Strasbourg, Vienna, Budapest and elsewhere, have already been digitized. They comprise about 90,000 images (recto and verso of each fragment).

All the images are being made available to researchers online at www.genizah.org . This site provides a very convenient web interface, with zoomable images, bibliographic and catalogue data, transcriptions and translations, search facilities, and discussion forums.

In this talk, three recent achievements of the project will be announced:

1. A few months ago, FGP signed an agreement with Cambridge University Library, for a joint three-year project, funded by FGP, during the course of which Cambridge will digitize their entire Genizah collection. This will result in some 400,000 additional images, to be

delivered at the rate of 10,000 per month (starting soon), in what is probably one of the largest ever digitization efforts attempted in the world of manuscripts.

2. FGP has begun applying computerized image processing to the digital photographs. This effort -- directed by Dr. Roni Shweka and realized by Rotem Littman -- includes separating the document from its background, segmenting the image into written areas and blank space, straightening the image, and automatically inferring the dimensions of the fragment, the written area, and the individual lines.

3. Because of the unique circumstances of the Cairo Genizah, the leaves of most of the original documents were recovered unbound and are to be found today dispersed among different libraries. Over the past century, scholars have expended a great deal of time and effort on identifying such pages and rejoining them. Despite the few thousands of such joins that have been found by researchers, very much more remains to be done. In this regard, FGP has embarked on an ambitious project -- in collaboration with Professor Nachum Dershowitz and Dr. Lior Wolf of Tel Aviv University and their students -- to use modern machine-learning techniques in a bold computer-aided effort to identify new joins. A highlight of the talk will be a sampling of the hundreds of new joins discovered by the computer in this fashion, based solely on image similarity, including several joins of noteworthy interest.