Where is my Other Half?

- Ben-Shalom,Adiel
  Friedberg Genizah Project
- Choueka,Yaacov
  Friedberg Genizah Project
- Dershowitz,Nachum
  Tel Aviv University
  nachum@cs.tau.ac.il
- Shweka,Roni
  Friedberg Genizah Project
- Wolf,Lior
  Tel Aviv University

Summary

We describe the design and implementation of a plan to integrate a manuscript matching scheme into a coherent and efficient system that can help scholars find the best candidates for potential joins for any given fragment.

1. Introduction

One of the most challenging issues in the analysis of large collections of historical manuscripts or handwritten fragments is the "joining" of fragments, either piecing together torn parts of a mutilated folio – which may have disintegrated because of wear and tear over the ages – or reconnecting two or more folios originating from the same original manuscript, but which have since been separated and dispersed to diverse locations – perhaps on account of trading activities between institutions. A striking case in point is the Cairo Genizah, discovered towards the end of the nineteenth century in the attic of an old synagogue in Cairo, which contains more than 320,000 fragments (deriving from tens of thousands of individual documents, almost all in Hebrew characters – though not necessarily in the Hebrew language) spanning a thousand years of writing and copying. Various parts of the Genizah finds are currently located in more than sixty university collections and public libraries spread out on different continents all over the world. Until just recently, a Genizah researcher holding half a page in hand and seeking its other half, did not have any resources with which to achieve that goal beyond erudition, a few catalogs, a gifted memory, and a fair measure of luck.

In recent work *[1]*, we described an automated scheme for image analysis and processing that culminates in enabling the computer to compare two images and compute a similarity score based solely on the individual handwriting style. A series of benchmarks and tests convinced us of the reliability and utility of this metric. Moreover, many hundreds of "joins" of interest to humanities scholars have already been identified *[2]*. Here, we describe the design and implementation of a follow-up plan devised to integrate the matching scheme into a coherent and efficient system that can help scholars find the best candidates for potential joins for any given fragment. A system with somewhat similar goals for processing and joining fragments of frescoes is described in *[3]*.

2. Joins and Jigsaws

The following steps have been implemented:

A – The basic idea is to match each of the Genizah fragments with one another so as to obtain a similarity score for each pair of fragments. We used a combination of local descriptors (SIFT) and learning techniques (OSS *[4]*, SVM, and others). Out of an estimated total number of 320,000 fragments, about 230,000 fragments were available to us, represented by 450,000 digital images, with two images per fragment (recto and verso). For every fragment, a numerical signature vector was computed, encapsulating aspects of its writing style. With a specially designed software component that measures the readability of every fragment, we eliminated from this scenario most fragments with poor legibility, those that most likely would not contribute true joins but rather would deteriorate the effectiveness of the system. These included blank or almost-blank pages, illegible or very dark texts, minute fragments, etc. After eliminating these problematic items, we were left with a total of 158,000 fragments to be compared with one another. That gave a total of 12.4 billion *pairs* that needed to be measured for similarity, a huge number indeed. Some twenty different similarity scores were computed and stored for each pair. These were generated by using four different algorithms to represent the handwriting style of each document and by using different similarity measures between documents. The different similarity scores can be "stacked" together to achieve higher accuracy.

B – Twenty CPU's from the Computing Lab of the Blavatnik School of Computer Science at Tel Aviv University ran together continuously for 37 days (the equivalent of some 18,000 computing hours), and the task was accomplished. This computer run is probably one of the most intensive ever implemented in a digital humanities context, in terms of computing resources. Four terabytes of output were generated in the process.

C – An efficient and compressed database was built to preserve these results in a structure that is easy to manipulate within a reasonable on-line response time. For each fragment, the top 300 similar fragments were precomputed.

D – A simple program, *Propose Joins,* was then integrated in the operational software of the Genizah website, available at http://www.jewishmanuscripts.org. Any user can input an image number, and the system will respond immediately by giving a list of the best 100 candidates that might qualify as joins for the given fragment, sorted from most similar to least, accompanied by the actual images of these candidates. See Fig. 1. A user can then mark some images as worthy of further investigation as potential joins, passing them over to a second program, called "Jigsaw Puzzle", described below.

It is our assumption – backed up by Genizah researchers' expressed attitudes – that a competent user would not mind spending an hour or so examining these images, even if he or she does not end up finding any join in the set, since this is the only way to systematically look for such a join were there indeed any. Our experience shows, in fact, that if there is a join in the Genizah world for the given fragment, it will be found – almost always – in this set.

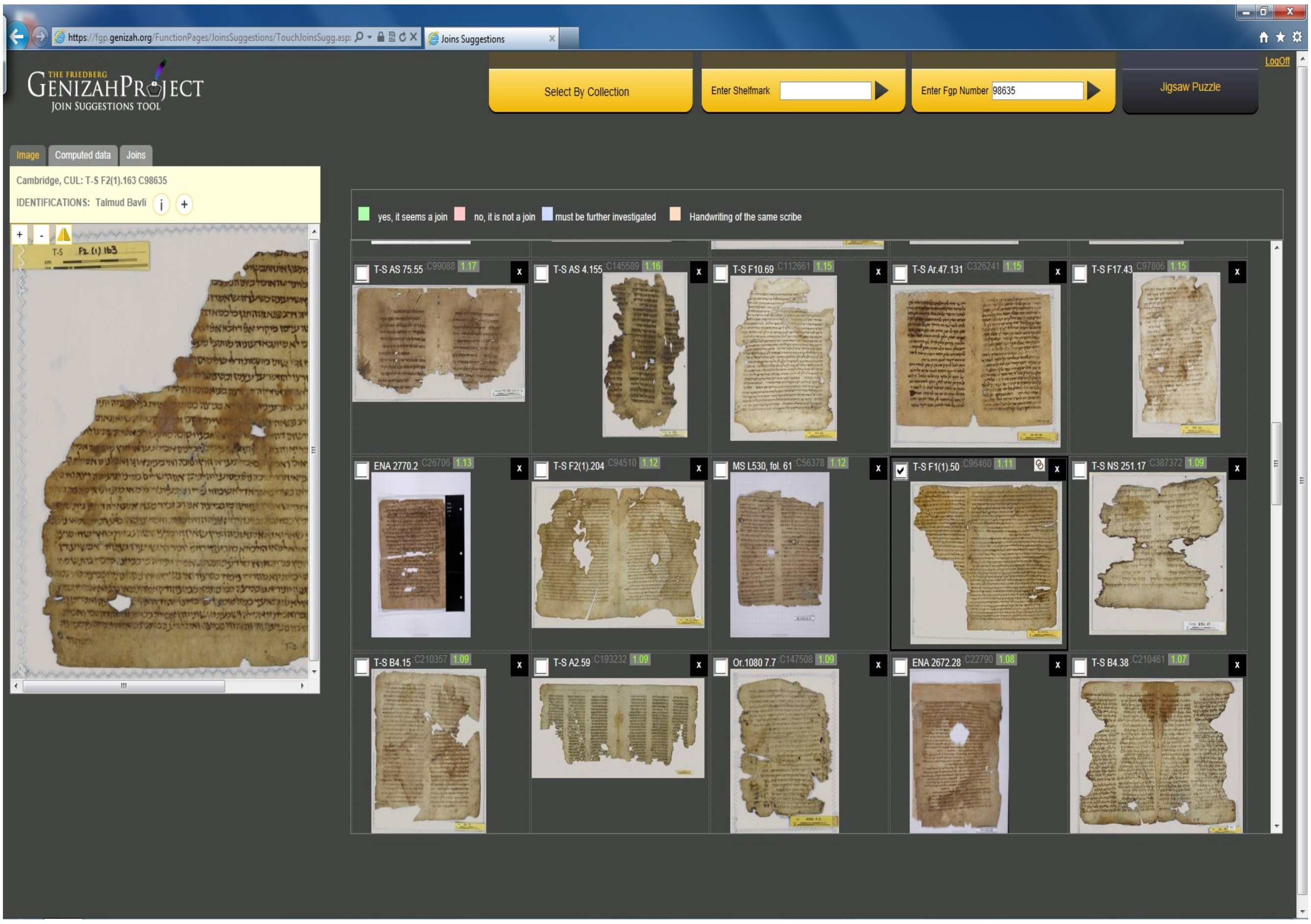
E – The *Jigsaw Puzzle* program displays the additional images designated by the user together with the original image on the screen, each image already restricted to the fragment's physical contour, and, using the mouse and a few tabs, a user can magnify or reduce each of them, move any image, rotate it by any angle in any direction, "flip" the image over to display the verso (say) of the fragment instead of its recto, calibrate the images at their original proportions in order to check if the geometric features and the running text of the various pieces indeed fit neatly into a join. See Figs. 2-4. If a join is found, it may be incorporated in the website for all users to be made aware of it, with the identifying scholar's name inscribed as the join's composer.

F – To make the system even more user-friendly and intuitive to researchers, even if they are not completely at ease with computers, a large 42" touch-screen was installed in the Genazim lab, as a prototype, with an attached PC on which the website and the software were installed, all completely transparent to the user. See Fig. 5. Using a virtual keyboard, the user approaches the system by inputting an image number, receives back the images of the 100 best potential candidates, marks some of the relevant ones, passing them over to Jigsaw, where they can be easily manipulated – moved, rotated, flipped, calibrated – by just touching the screen with one's fingers, much like what one is used to doing nowadays with smartphones, and as naturally as one might arrange a jigsaw puzzle spread out on a table. See Fig. 6.

Discussion

We expect the overall scheme, with all the steps detailed above, to be of relevance in many other similar contexts, although, admittedly, the Genizah case is rather unusual in its scope and complexity. The join-matching tool is quite sophisticated and is constantly undergoing improvement. The jigsaw tool is relatively simple but has already proved very appealing to scholars. We are currently applying the join tool, more or less as-is, to other corpora, including the Dead Sea scrolls and papyri *[5]* and Tibetan manuscripts and xylographs. Other potential applications include the 2,000,000 images of 70,000 pre-1900 Taiwanese deeds and court papers from the Taiwan History Digital Library *[6]* and Yad Vashem’s now publicly available Holocaust archives (http://www.yadvashem.org/yv/en/resources/index.asp). Furthermore, we hope to make the jigsaw tool more widely available. In addition, we have begun to use machine-learning tools based on the same signature vectors to help answer palaeographical questions for such corpora *[7][8][9]*.

We are still left with the major problem of trying to reconstruct the original state of the entire Genizah collection, that is, to find, once and for all, the entire network of true joins in this collection, with a reasonable level of completeness and precision, and to do this efficiently and in a relatively short time. A crucial step in achieving this goal is to find effective methods for recognizing and eliminating large quantities of false joins and non-joins, even if that may be at the cost of losing a few correct ones. This is currently a topic of intense investigation, involving elements of graph theory, clustering techniques, data mining and related methodologies.



https://fgp.genizah.org/FunctionPages/JoinsSuggestions/TouchJoinsSugg.asp

Enter Fgp Number 98635

Cambridge, CUL: T-S F2(1).163 C98635

IDENTIFICATIONS: Talmud Bavli

Fig. 1: After the user gave the system the i.d. number of the fragment being studied, the system responds by displaying that image on the left side, and 100 suggested joins, sorted by decreasing order of similarity, on the right.

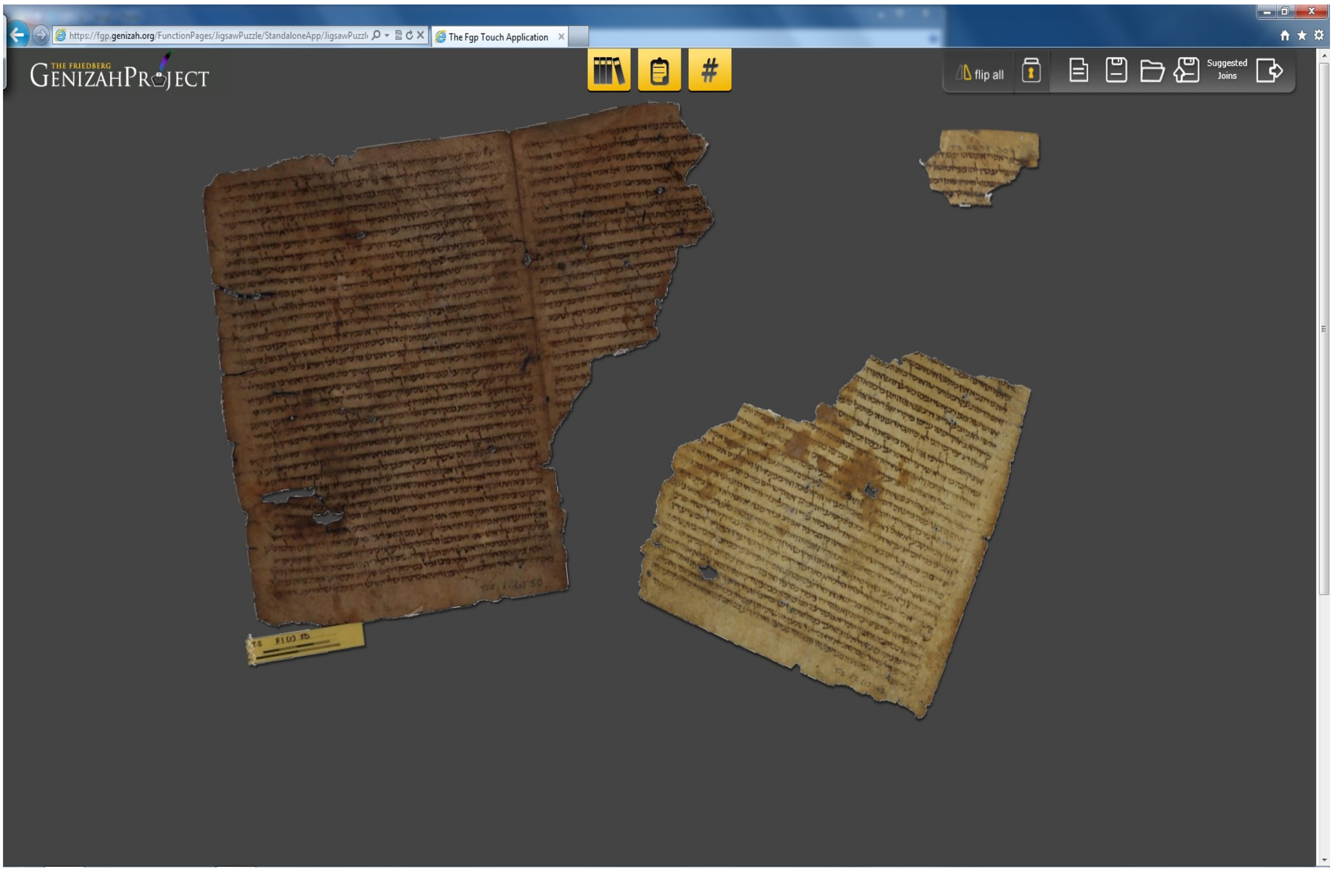

https://fgp.genizah.org/FunctionPages/JigsawPuzzle/StandaloneApp/JigsawPuzzl
The Fgp Touch Application
THE FRIEDBERG GENIZAH PROJECT
flip all
Suggested Joins

Fig. 2: The given fragment and two of the suggested joins.

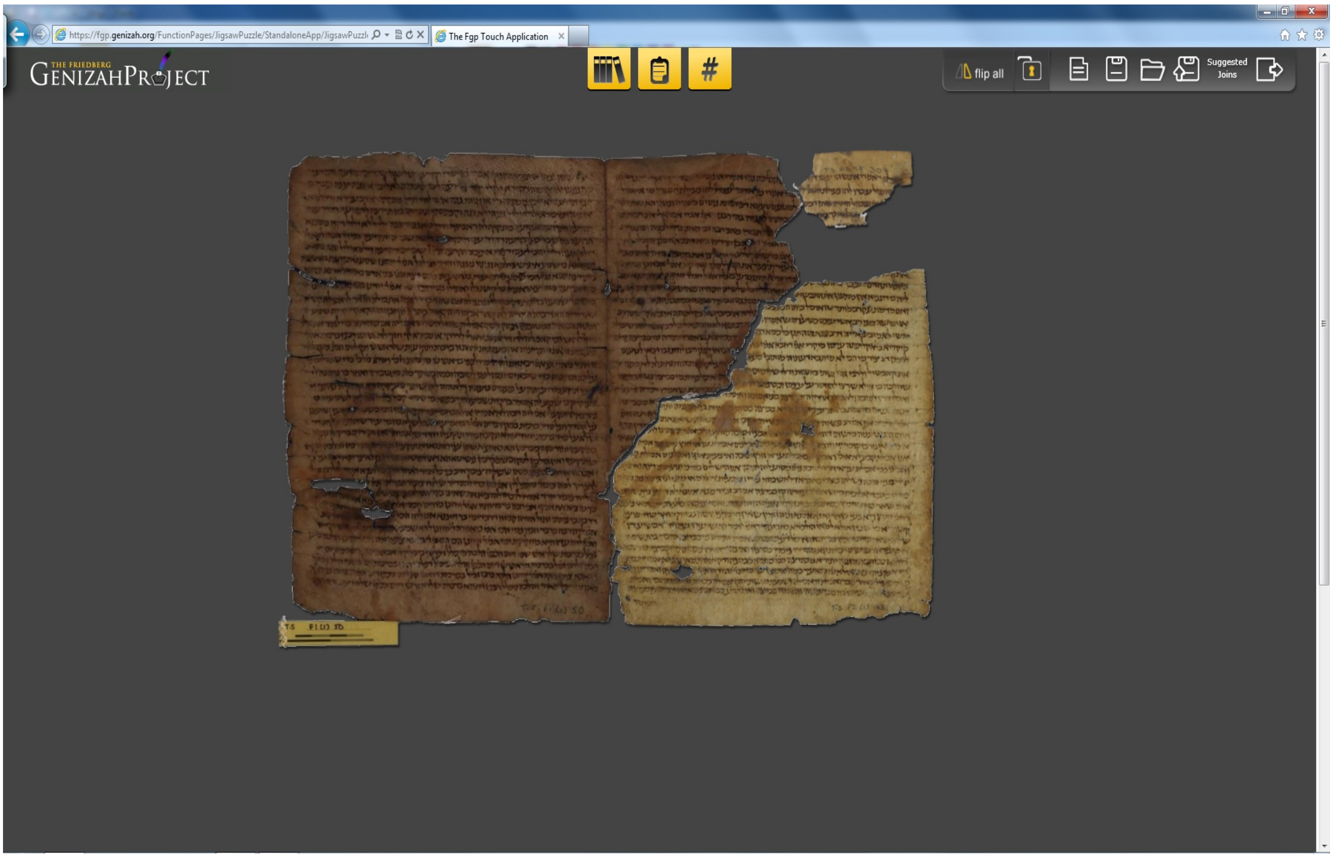

https://fgp.genizah.org/FunctionPages/JigsawPuzzle/StandaloneApp/JigsawPuzzl
The Fgp Touch Application
THE FRIEDBERG GENIZAH PROJECT
flip all
Suggested Joins

Fig. 3: Two fragments have been “glued together” by the Jigsaw program; the third is on its way.

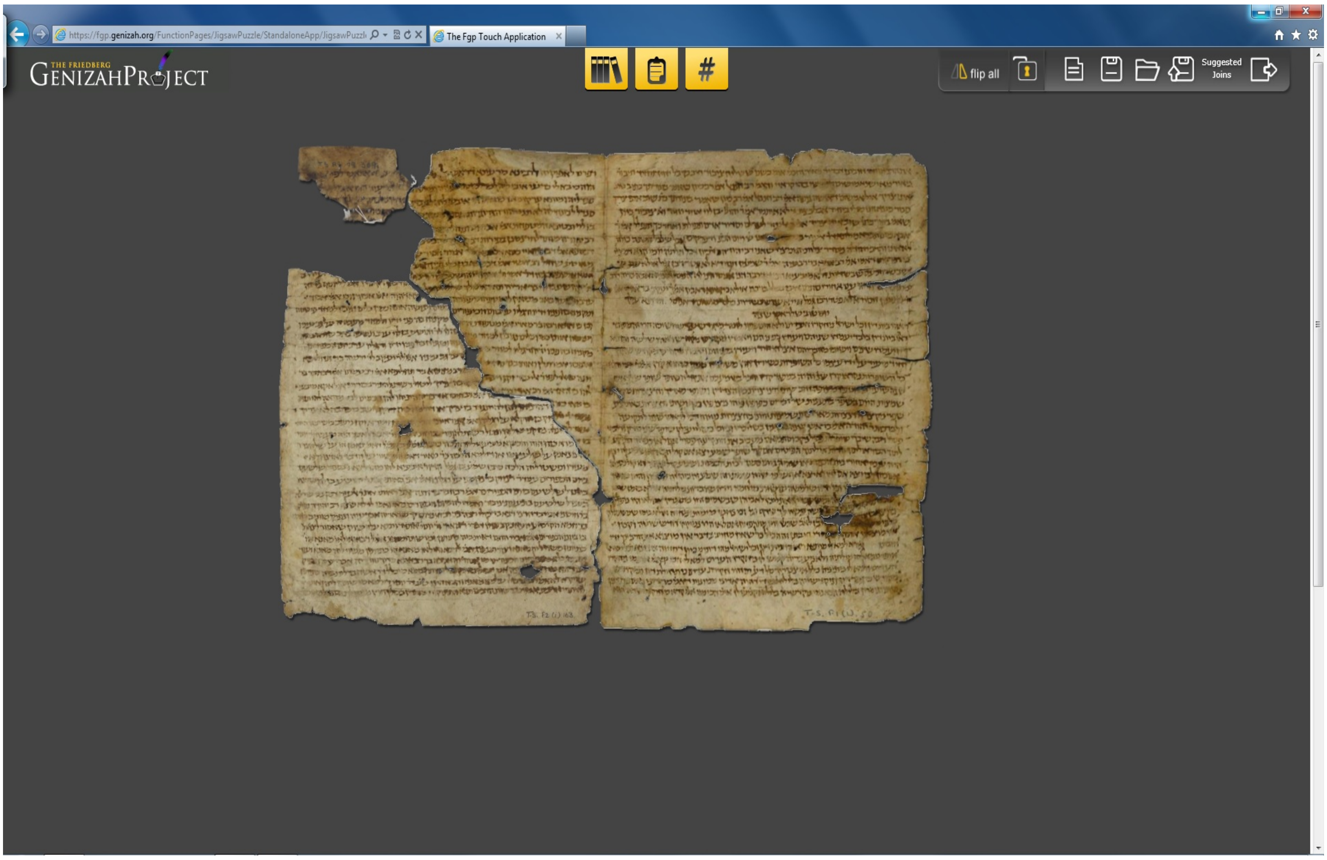

https://fgp.genizah.org/FunctionPages/JigsawPuzzle/StandaloneApp/JigsawPuzzl
The Fgp Touch Application
THE FRIEDBERG
GENIZAH PROJECT
flip all
Suggested Joins

Fig. 4: The final join (verso).

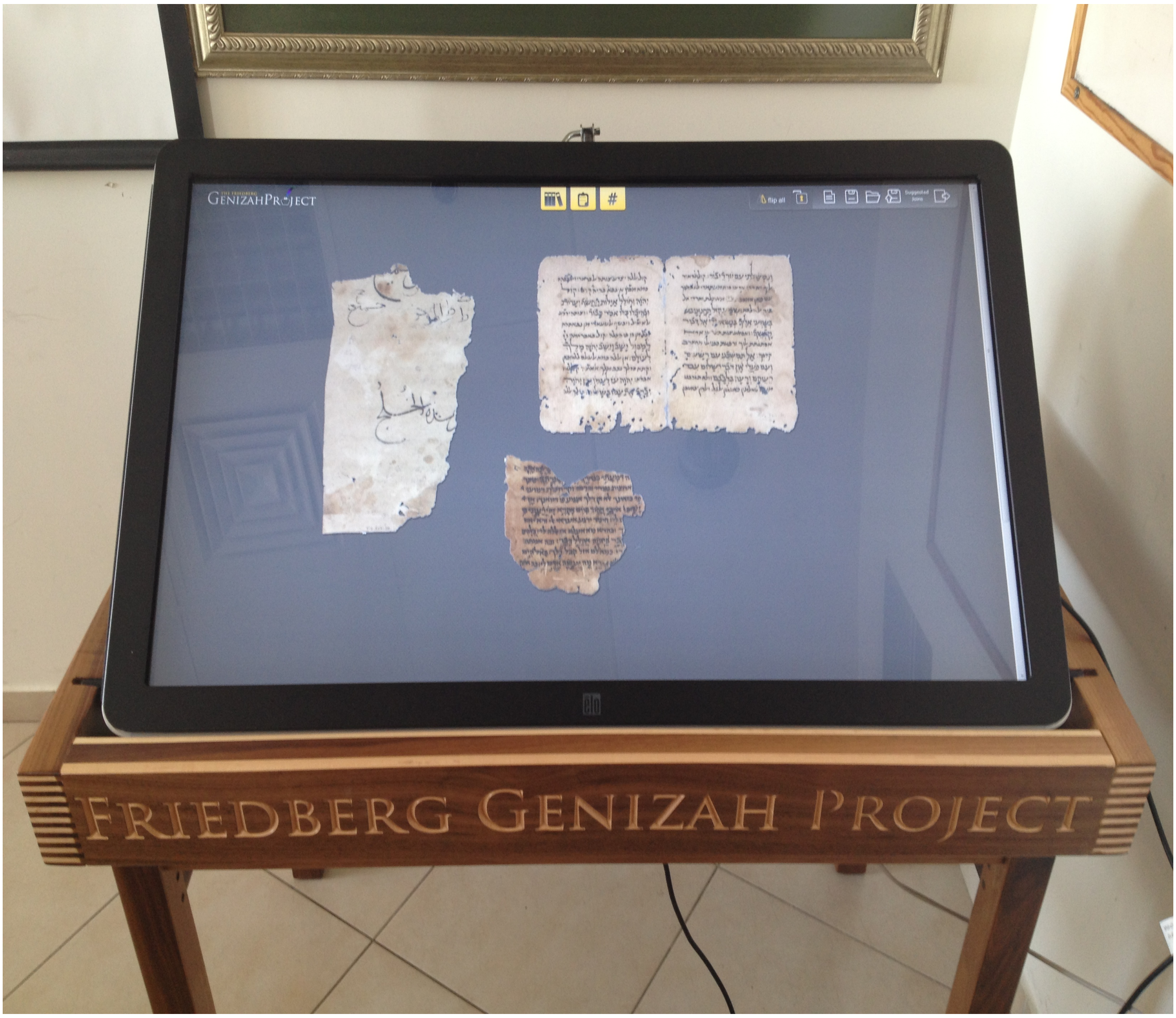


Fig. 5: The large touch screen mounted on a specially designed chassis, showing the fragment at hand together with 2 suggested joins.

Fig. 6: One of the authors (Y.C.) using the Jigsaw program to manipulate the suggested join.

References

1. Wolf, L., Littman, R., Mayer, N., German, T., Dershowitz, N., Shweka, R. and Choueka, Y. (2011). Identifying Join Candidates in the Cairo Genizah. International Journal of Computer Vision, 94(1): 118-135.
2. Shweka, R., Choueka, Y., Wolf, L. and Dershowitz, N. (2011). "Veqarev otam ehad el ehad": Zihuy ktav yad vetseruf qit'ei hagnizah beemtsa'ut mahshev (Identifying Handwriting and Joining Genizah Fragments by Computer), Ginzei Kedem, vol. 7, pp. 171-207. (In Hebrew.)
3. Brown, B. J., Toler-Franklin, C., Nehab, D., Burns, M., Dobkin, D. P., Vlachopoulos, A., Doumas, C., Rusinkiewicz, S. and Weyrich, T. (2008). A System for High-Volume Acquisition and Matching of Fresco Fragments: Reassembling Theran Wall Paintings. Proceedings SIGGRAPH 2008, ACM Trans. Graph., 27 (3).
4. Wolf, L., Hassner, T. and Taigman, Y. (2009). The One-Shot Similarity Kernel. IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan, pp. 897-902, Sept. 2009.
5. Bearman, G. (2008). Imaging the Dead Sea Scrolls for conservation purposes. SPIE Newsroom, December 29, 2008.
6. Hsiang, J., Chen, S.-P. and Tu, H. C. (2009). On Building a Full-Text Digital Library of Land Deeds of Taiwan. Proceedings of Digital Humanities 2009, College Park, MD, June 2009, pp. 85-90.
7. Ben-Shalom, I. (2013). Automatic Paleographic Grouping by Script Styles and Scribal Identity in Large Medieval Collections. M.Sc. thesis, Tel Aviv University, Nov. 2013.
8. Dershowitz, N. and Wolf, L. (2013). Automatic Scribal Analysis of Tibetan Writings. Abstracts of the 13th Seminar of the International Association of Tibetan Studies, Ulaanbaatar, Mongolia, July 2013.
9. Wolf, L., Dershowitz, N., Potikha, L., German, T., Shweka, R. and Choueka, Y. (2011). Automatic Paleographic Exploration of Genizah Manuscripts. In: Kodikologie und Paläographie im Digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age 2, Fischer, F., Fritze, C. and Vogeler, G., eds., Schriften des Instituts für Dokumentologie und Editorik, vol. 3, Norderstedt: Books on Demand, Germany, pp. 157-179.