1,000-year Manuscripts meet Computer Technology

The Cairo Genizah Case

Yaacov Choueka

A – The Genizah Collection

The Cairo Genizah is a remarkable set of fragmented Jewish manuscripts discovered towards the end of the nineteenth century in the loft of an old synagogue from the medieval part of Cairo.

The collection, the way it was discovered and its contents later dispersed all around the world, its importance, and the very great and profound impact it had on Jewish Studies and on the study of medieval communities of the Mediterranean basin in general, is already well described in the other papers of this bulletin. We shall therefore content ourselves with detailing some of its basic features that affect the methods and technologies that had to be developed in order to computerize this collection and its research context, i.e. to build the Computerized Genizah Research World.

While some of the ideas presented here are very specific and closely geared to the Genizah collection, others – as will be clear below - may be applied to various very large collections of historical manuscripts that represent treasures of cultural heritage in its truest meaning.

***

The Genizah contains – as we finally know today – about 300,000 "fragments", a fragment being (mostly) a handwritten piece of writing material (mostly paper, but sometimes parchment, vellum, papyrus, etc), so, in fact, a "page" or rather a "folio", but more usually a torn, mutilated, stained, sometimes minute (no more than a few square inches) part of an original folio or bifolio, itself sometimes part of an entire manuscript. These fragments, almost all in Hebrew characters, were put by their owners into that doorless room in which they were found, so as not to disrespectfully throw something with the name of God in it in the garbage. Gradually, however, people got used to throw there not only old prayer books, bibles, Talmud pages or rabbinical writings, but anything written in Hebrew characters. The room was supposed to be emptied now and then and its contents respectfully buried with a special ceremony in a Jewish cemetery. By some miraculous chance, however, the room was never emptied, and thus generations upon generations, for almost a thousand years, deposited there not only religious writings but

also letters, contracts, marriage or divorce papers, bills and business transactions, recipes, medical prescriptions, magical amulets, Jewish courts' archives, and more. Arabic being the principal language of most Mediterranean communities (Jewish ones included) at these times, many such Jewish writings – including philosophy and ethics books, even Jewish law codes - were commonly written in a special Jewish dialect of Arabic (Judeo-Arabic), but (for good reasons) with Hebrew characters, and thus reached us today and were salvaged for posterity. Although containing occasionally some fragments in Aramaic, Ladino, Persian, Yiddish, and more, the bulk of the Genizah is in Hebrew or Judeo-Arabic, almost always in Hebrew characters.

B – The Friedberg Genizah Project

Extensively researched for a hundred years, the study of this collection revolutionized not only Jewish Studies, but also our knowledge about the history, economy and social life of Jewish communities in the Mediterranean basin, and their interaction with the Islamic communities there.

In order to further advance the study of the Genizah in various ways, Dr. Dov Friedberg from Toronto initiated almost ten years ago a vast not-for-profit project – the Friedberg Genizah Project (FGP) - funded by him. Many Genizah research teams were created, both outside Israel, such as in Cambridge, Princeton and Manchester, and in Israel itself, at the Hebrew University, Tel Aviv University, Hanegev University, Ben-Zvi Institute, etc, whose aim was to extensively study, identify, describe, catalog, and transcribe as much as possible of the Genizah fragments, by looking either at the original manuscripts – wherever they reside - or at their microfilm substitutes, created in the late sixties by the Institute for Microfilmed Hebrew Manuscripts (now part of the National Library in Jerusalem). These efforts resulted in a flurry of Genizah-related activities, such as the compilation of a large set of data on some of the fragments, as well as the production of a large number of research publications (papers and books), a scholarly yearly journal devoted entirely to Genizah research, special university courses and seminars, many M.A. and PhD theses, and the like.

More information on the Genizah, on FGP, and on the scholars and scientists that run it can be found at www.genizah.org .

C –The Computerization Mission

In January 2006, a new unit of FGP, Genazim, located in Jerusalem, was created, whose mission was to computerize the entire Genizah Research World, including the research results derived from decades of international scholarly activity including that from the FGP research teams work.

How does one go to develop this mission of Computerizing the Research World of a large collection of historical manuscripts?

It was clear from the very beginning that one of the ultimate goals of Genazim was to develop and maintain a website - freely open to researchers as well as to the general public - which will contain, organize, manipulate, process and display as much as possible of the Genizah resources, using a rich and sophisticated set of software tools for searching, querying, filtering, sorting and displaying that data.

This website was indeed developed in the space of four years, and is now fully operational, with more than one million pertinent data items available there. More than 1,200 researchers are already registered to this website, and routinely access and use it daily for their research needs.

Previous to that, however, we had to set up the major principles that would guide us in this work. Following is a few of them:

a - In the twentieth century, the Genizah was researched as part of the vast world of Hebrew manuscripts, and it was studied in specific domains by experts in that domain. A Bible researcher would go over tens of thousands of fragments (as much as he could) discarding everything except for biblical material. The same line would be followed by a Talmudic scholar, for the Talmudic fragments, and the same by a linguist for linguistic material, etc. Thus most of the Genizah would be looked upon again and again, probably tens of times but only a tiny part would be analyzed by the various researchers.

Our approach was that the Genizah collection is an integral corpus in its own, and that we are going to open its gates to all researchers in all domains and for all fragments in all collections. You never know what great surprise a fragment can represent until you actually identify and analyze it.

b – As befits an exact-science discipline, and as typical of computers' operations, our work will have to be precise, comprehensive and up-to-date in every aspect. Thus, for example, we shall try to trace down every single fragment of this 350,000-extensive set.

c – Every data item that we integrate in our databases and ultimately display in the website, will have to be appended by the source of this information: a book, a catalog, a

paper, a scholar, etc. Although we are in constant dialog with Genizah researchers, we don't allow ourselves the liberty of identifying or transcribing a single fragment.

d – We should be not – and are not – arbitrators between scholars. As happens more often than not, scholars may differ on the identification of a fragment, its contents or its author; we display all the differing – many times contradictory – opinions in our site with their sources, and let the user choose.

e – The project is by its very nature open-ended, in the sense that for years and years, as research progresses, new data will become available and will have to be integrated and displayed in the website. Thus the software system once reaching its aims - should be stabilized and amenable to the adding of information directly from the researchers to its databases, with a minimal interference of a small group of programmers solely dedicated to the necessary system maintenance over the years.

f – The *shelfmark* of a fragment is the number given to it by the library in which it resides, in exactly the same way as a book is "shelf-marked". The shelfmark helps the librarian retrieve the fragment when needed, but, more importantly, it is the unique "identity number" by which it is internationally recognized, mentioned or discussed in the literature. The world of unique shelfmarks is expect6ed therefore to be a well-defined rigorous and exact world, still, for reasons we have no place to detail, it can rather be termed as a "loosely controlled chaos". In any case, we decided to take the shelfmark of a fragment as the central axis to which every single datum of information on that fragment will be attached. Our system does not recognize and can not deal with a fragment to which no shelfmark has been assigned by its owner.

D – Two critical tasks

Faithful to the policy detailed above, we took upon ourselves at the very beginning of Genazim activities two important and critical tasks that we thought should form the backbone of our development.

1 – Inventories

The first task is to compile a complete, computerized, precise, and up-to-date inventory of the formal shelfmarks of all Genizah fragments in all Genizah collections all over the world, large or small, "important" or not, public or private. No cataloging data or identifications of any sort were included in that task; we needed only the shelfmark and the number of fragments that were included by the librarians in that shelfmark. We wanted to have this list compiled as much as possible not from outdated catalogs or lists,

but by having the compiler looking actually at every fragment and recording the corresponding data, including for example cases in which there was an envelope with a shelfmark but the pertinent fragment was missing, being on loan or in conservation, or just disappeared. In certain cases, as in Cambridge University Library, the Jewish Theological Seminary of New York, or the Alliance Israelite Universelle in Paris, the inventory was compiled by the library staff itself in cooperation with Genazim; in others, such as in Bibliotheque Nationale et Universitaire in Strasbourg, British Library in London or Oxford University Library we sent our Genizah experts who accomplished the task in cooperation with the library staff.

Currently, the inventories residing at the Genizah servers in Jerusalem comprise more than 250,000 shelfmarks and account, we believe, for more than 99.9% of all shelf-marked Genizah fragments. To achieve this, and assure the comprehensiveness of the process we had, among others, to compile for the first time a complete list of the more than 75 Genizah collections in the world, tracking down even "collections" that contain just one fragment.

As became clear later, this effort not only allowed us, from that moment on, to attach all available and future data on any fragment to its shelfmark, but also prompted Genizah researchers to use the formal shelfmarks as defined by the relevant libraries and s appearing in our inventories, in their publications, thus encouraging a trend of standardization much needed in that context. Moreover, a fragment would commonly change its shelfmark either because it changed owners or even because of a restructuring of the library shelves and naming procedures in that library. We made a large effort to collect all the older or alternate shelfmarks, record them and attach them to the currently valid one, so as to correctly attach data that may be appended by an older shelfmark to the newer one.

2. Digital Images

Up to the early seventies of the 20th century, the only way for a scholar to study a Genizah fragment would have been to travel to the city and the library where it resides, and, assuming he has the proper credentials, to spend his time at the library (while it is open) looking at that fragment (usually with a magnifying glass). During the seventies of the 20th century the Institute of Microfilmed Hebrew Manuscripts in Jerusalem was well on its way for microfilming the totality of all Hebrew manuscripts in the world, among them the Genizah ones. Life become easier for scholars in Jerusalem (or Israel), but still the interested researcher was restricted by the opening hours of the library, by the

availability of microfilms, and especially by the rather poor quality of microfilms' readers and by their rigidity.

Thus the decision was taken at the onset of Genazim activities to produce or acquire full-color high-quality (600 dpi resolution) digital images of all Genizah fragments, and to make them available through the Internet to any interested user, who can then look at them and study them any time and from anywhere. This move necessitated intense negotiations with the libraries in which such collections reside, and the signing of suitable legal agreements to protect their copyrights. In many cases, such as JTS (New York), AIU (Paris), Geneva, Strasbourg, Vienna, and others, FGP sent its own expert-photographers who accomplished the complex task (that has to take into account among others the exact naming of the produced files), using special techniques in quite record times. In other cases, such as in CUL (Cambridge) British Library (London), and others, the digitization was accomplished by the digitization laboratory of the library itself, in cooperation with Genazim and with FGP financial support.

We insisted on always digitizing both sides of every fragment, large, small or tiny, even when they seemed to be blank or un-readable. We also recorded missing fragments by taking the image of the corresponding envelope (or even a simple page) with a "Missing" caption on it.

The digitization of the Cambridge collection, expected to produce about 350,000 images, and that of the British Library, are well on their ways, and are planned to be finished, the former by 2012 and the later by the end of this year.

All in all, our website contains now about 200,000 digital images, representing the complete digitization of many collections. We assume that by 2012 the site will contain more than half-a-million images, covering about 95% of all Genizah fragments; this could probably be considered as one of the largest digitization efforts of historical manuscripts ever attempted.

E - The Data

Designing a computerized system for the research world of a very large collection of historical manuscripts that has been under intensive study for more than a hundred years should be supported by two axes: data and software.

What kind of Genizah data should be collected, stored and processed to be finally displayed in the website?

The following eights types of data (no more and no less) were found to be the appropriate ones to play this role:

1. Shelfmarks: about 247,000 shelfmarks, representing the comprehensive inventory of 66 collections, and covering 320,000 fragments, form now the skeleton of the system.
2. Digital images: till now we have completely digitized 31 collections, two more are being digitized, and some 200,000 images are available at the website.
3. Bibliographical references, i.e. detailed references to any publication that discusses or even mentions a specific Genizah shelfmark. A complete set of references for all publications in any language, from the Genizah discovery times till 2004, for CUL Genizah fragments, compiled by the CUL Genizah Research Unit, is integrated in our databases. Moreover, all references to non-Cambridge shelfmarks in Hebrew publications of 1966 – 2004, as well as many references in other languages to such fragments is also in the website. In total, 226,000 such references are recorded in the databases.
4. Cataloging records, that would specify, in a (mostly) coded form, any available information pertinent to a given fragment, found in scientific catalogs, books, and scientific publication in general, both on its physical aspects: outer and inner (text-blocks) dimensions, number of lines, writing material, margins, corners, holes and tears, etc, and on its contents aspects: domain (Bible and Biblical commentaries, Talmud and Talmudic Commentaries, Philosophy and Ethics, letters and historical documents, medicine, magic etc, - there are more than 30 such domains), title of work, author, scribe, date of copying, etc. About 70 such fields are included in the cataloging record, and 202,000 such records participate in the website, some rather lean with just a couple of fields marked, others quite complete.
5. Scans of all Genizah-related entries in any catalog of Jewish manuscripts whether published or only printed, handwritten, or electronic. Almost no library - and certainly no researcher - can afford to have all the 45 relevant catalogs easily at hand, so that giving the researcher the ability to see, with just a click, clear scans of all the original entries related to a certain shelfmark, is indeed a major research help. More than 70,000 such scanned entries are currently available in the website.
6. Transcriptions: Because of the sometimes difficult calligraphy, and no less because of the physical status of the fragment, recording the text of a fragment

into computer is almost always a difficult task done exclusively by researchers. About 13,700 transcriptions are currently available.

7. Translations: as noted above, many of the Genizah fragments are in Judeo-Arabic, so a translation to Hebrew may be useful. About 2900 such translations (including a few translations into English) are in the website.
8. "Joins": One of the most critical issues in Genizah research is that of discovering "joins", i.e. different fragments - folios or parts of folios - originating from the same manuscript or the same folio, that have been relocated (through the unavoidable tearing and wearing of the originals during so many centuries and the random acquisition and trade of manuscripts) in different libraries, one fragment being found, say, in Cambridge and the other in Vienna. During a hundred years of research, about 4,000 of such "joins" were discovered through the mere erudition, memory and intelligence of Genizah scholars, and all are recorded in the databases.

F – The Website

Obviously a robust software system is needed to absorb these data, integrate. process and manipulate them in order to make them useful to the researchers.

The software system is composed mainly of 3 parts:

- the databases, in which the data is checked, organized, connected and updated;
- the input module, which allows the research teams, and later all accredited users, to directly input the data in the databases;
- the website, which is in fact the only interface between the researchers and the data, and through which the whole Genizah research world is supposed to be transparent and available to the user.

We shall focus here on the website, which can be accessed through www.genizah.org by clicking on the "login" button (a free and simple registration is needed). We shall content ourselves here with a general outline of the website and its various functions, since anyone can access it, and directly manipulate its various functions.

There are essentially two ways of querying the website:

A user can select a fragment using a drop-down menu, by selecting the city where the collection resides (this is a common procedure for the Genizah), then choosing the sub-collection, the volume, etc, and, finally the specific shelfmark (each of the 75

collections has its own structure, and the menu is specifically adapted to each such structure). Alternatively, if he knows the exact shelfmark, he can directly type it. He can then chooses between six different functions that display all the available data pertinent to this shelfmark: a high-quality image of (both sides) of that fragment; a scanned image of any entry in any Genizah catalog that is related to this shelfmark; the set of all bibliographical references to any publication that mentions this shelfmark; content-identifications of the fragment; full cataloging record; any transcriptions of this fragment, when available, and its translation, and finally all joins (if any) in which this fragment has a role.

The browsing between these functions is totally dynamic; the user doesn't have to retype or choose the fragment's shelfmark again.

On the other hand, a user can submit a query to the system, in order to get the list of all shelfmarks that satisfy a given set of conditions. Six different types of such a search are available.

- by the cataloging data: The user can specify any set of Boolean conditions ("and", "or", "not") on the values of 15 relevant fields of the cataloging record, asking, for example, for the shelfmarks of all biblical fragments from Exodus that have cantillation signs, originate from the 12^{th} to 14^{th} centuries, contain at least 5 lines, form a join with another given fragment, and have a "colophon" (date of writing) at their end;
- by the bibliographical references data, e.g. all fragments for which there is at least (or at most, or exactly) n (including zero) references, from a set of specified journals, or a set of specified authors, in some specified years, etc;
- by the 'identifications" data, e.g.. all fragments from a certain domain, for which there is at least n identifications, attributed to given source or a Boolean combinations of sources, etc;
- by Genizah catalogs, i.e. all fragments for which there is an entry in a given catalog, or in a specified combination of catalogs;
- by transcriptions and translations, i.e. all fragments for which there is a transcription by a given source;
- finally, by a full-text search of the transcriptions (or translations, or the Genizah catalogs' text), e.g. all fragments that contain a given word or combination of words, or a specified exact phrase, etc.

The way a query is submitted is simple and user-friendly, and the way the results are displayed conforms to the user expectations, giving the maximal amount of relevant information in the minimal amount of structured space. The reader is invited to convince himself about that by actually browsing in the website.

Finally, a few additional modules are available to help users conduct their research efficiently.

- a "Quickview" function allows the user to browse very quickly through (low quality) consecutive images of a given collection's shelfmarks (typically tens of images in a few minutes), so as to focus on the fragments that interest him;
- an individual workspace is available to every user where he can store (manipulate, add or delete) in his own designed structure a small set of fragments which he is currently researching, and which is kept for him from session to session;
- a public forum is available where users can exchange information, discuss issues about given shelfmarks, add ort correct information, etc. Any user can also build a "restricted" forum for him and his small set of colleagues, for internal discussions;
- "notes" can be added by any user on any shelfmark, to be displayed for all users;

and more.

G - Research achievements

In this last section, we would like to describe briefly some remarkable results in the area of computer-assisted analysis of high-quality digital images of historical handwritten manuscripts, which were achieved by a cooperation of our team with researchers from the Computer Science Departments of Tel Aviv University and of the Hebrew University.

The starting point was to try and find out what physical attributes of a Genizah fragment can be automatically deduced by the computer through a fine analysis of its digital image.

We were able to develop a few software modules that, through such an analysis, can recognize and follow the exact contour of the textual part of the fragment, thus separating it from its background and from some artifacts that may be in the image (such as rulers or color charts), can count the number of lines in that fragment, and can measure the fragment's outer and inner dimensions, the average written-line

width and length, the average inter-line width, the density of letters-per-line, the existence of margins and their average dimensions, and more.

It was thus proved that this type of data, considered essential in the study of manuscripts and that can be found in many catalogs of manuscripts' collections, which has been marked until now manually by scholars, with a notable waste of precious research time, can now be extracted automatically from the fragment's digital image with much more accuracy and efficiency.

A crucial further step however was achieved when we were successful in developing a complex program that can analyze the handwriting of two fragments' images. and assert the probability that they were both written by the same scribe. This is not done however through the analysis of the individual handwritten letters and their shapes, but rather through a global comparison scheme, vaguely similar to the way with which two portraits are compared by the computer and found to be of the same person. This way we were able indeed to discover hundreds of hitherto unknown joins.

in many respects an astonishing result whose consequences may completely change the research platform of Genizah studies in the very near future.

We are in the process of implementing these findings on the set of images currently in our databases, a set which will reach, in a couple of years, the mark of half-a-million images. The data derived from the computerized analysis of these images will be ultimately integrated in our databases and displayed in the Genizah website. Using these techniques we hope that we'll be able, ultimately, to prepare a complete "Genizah catalog of Joins", and by this, *to reconstruct the Genizah original manuscripts.*