Computerizing the Cairo Genizah: Aims, Methodologies and Achievements

Yaacov Choueka*

Since its discovery towards the end of the nineteenth century, the Cairo Genizah has been the subject of intense research and analysis. Hundreds of books and monographs and thousands of papers and other publications, of which this journal, devoted entirely to Genizah research, is one relatively recent example, attest to the fruits of this endeavor. Nevertheless, it was only a few years ago that an ambitious, far-reaching plan for the computerization of the Genizah research world was initiated. This plan has been in a continuous process of design, development, and implementation since then. Now, six years later, with almost all of the research and development elements already in place; with an operational website, www.genizah.org, being accessed hundreds of times daily by a large group of registered scholars, researchers, academics, rabbis, and laymen curious about their Jewish heritage; and with the new technologies currently available slowly re-shaping the profile of Genizah studies and holding the promise of completely re-defining its horizon, it is time to tell the story of this computerization project, its aims, its methodologies and its achievements. That is the purpose of this paper.

A. How it all started

In the course of the Fourteenth International Congress of Jewish Studies, held in Jerusalem in May 2005, a meeting attended by the Friedberg Genizah Project sponsor representative (R. Rubelow), the Project's directors (Prof. M.

* Currently head of "Genazim," the computerization unit of the Friedberg Genizah Project.

Ben-Sasson and Prof. H. Ben-Shammai), and this paper's author (henceforth Choueka) — then a freshly retired *Professor Emeritus* of Computer Science from Bar-Ilan University — took place. The subject of the meeting was computers and the Genizah; i.e. how computer technologies could successfully be applied to the world of Genizah research, after a few previous attempts to achieve this goal had failed. The Project's directors suggested — in fact, recommended — that Choueka take this challenge upon himself. At that time, neither the recommending nor the recommended parties had the faintest idea of what *computerizing the Genizah* might really mean. That was, indeed, *the* challenge. In any event, Choueka, rising to this challenge, took upon himself the study of the issues involved, and towards the end of 2005 presented a report outlining his vision for such an endeavor and a rough plan for implementing it in four-to-six years. The plan was endorsed by all parties involved, and in January 2006 a Jerusalem-based computerization unit named "Genazim" was created, as part of the Friedberg Genizah Project, to carry out this project. Offices were rented, staff (computer programmers and consultants) recruited, hardware bought and installed, and an exciting, intense and demanding journey effectively started.

B. The Cairo Genizah: a reminder

The Cairo Genizah, that huge collection of fragmented manuscripts discovered in the loft of an old synagogue in Cairo; the way in which it was discovered and its contents dispersed around the world; its importance and the very great and profound impact it has had on Jewish Studies and on the study of the medieval communities of the Mediterranean Basin in general, are all probably well known to the readers of this journal. I shall therefore content myself with detailing some of the basic features of the Cairo Genizah collection that impact the difficulties, the methods and the technologies related to its computerization, i.e. to the building of the computerized Genizah research world.

First, one must note the staggering extent of the Genizah collection. As can

finally be asserted today, this collection contains about 320,000 "fragments," a fragment being sometimes a page, but more commonly a torn, mutilated, stained, often minute (no more than a few square inches) fraction of an original folio or bifolio — itself, again, part of an entire manuscript.

Second, since almost all of the Genizah documents, with very few exceptions, are written in Hebrew characters, the computerization scheme was restricted to only this set of characters.

Third, although occasionally containing some fragments in Ladino, Persian, Yiddish, and additional languages, the bulk of the Genizah material is in Hebrew, Judeo-Arabic and Aramaic, and this again had the potential to affect some of the decisions in the definition of the project.

Fourth, although the bulk of the Genizah fragments (about 60%) was transferred, through the initiative and personal efforts of Salomon Schechter and with the financing of Sir Charles Taylor, to the Cambridge University Library (henceforth Cambridge), the remainder of it was dispersed between some 70 public libraries and private collections all over the world, with different pages from the same codex, even different fragments from the same folio, often being deposited in libraries not only in different cities, but even on different continents. Because of its dispersal in so many libraries, cities and continents, Genizah research has been hindered by many serious difficulties. Time-consuming travels were necessary in order to study the original manuscripts and, when such travel was impractical, researchers had to content themselves with studying their oftentimes low quality microfilm substitutes, created in the late 1960s by the Institute for Microfilmed Hebrew Manuscripts, now part of the National Library of Israel in Jerusalem, where they would have to cope with curtailed operating hours, a small number of microfilm readers, etc. Many fragments, being stained or obscured by age, were difficult or even impossible to decipher, at least with a reasonable degree of confidence, the only tool available to the expert to remedy this situation being, in fact, a standard magnifying glass.

Finally, the analysis of a torn folio is unsatisfactory unless the other parts of that folio (or other folios originating from the same manuscript) are found and

"joined" to this fragment, an activity which had to rely almost exclusively on the expert's memory, or, with some luck, on information randomly mentioned in catalogs or research papers. Such information naturally accounted for a very small percentage of the Genizah material.

C. The Friedberg Genizah Project

In order to further promote the study of the Cairo Genizah and rejuvenate interest in this field of study, a vast international non-profit humanities venture, the Friedberg Genizah Project, was established in 1999 by Dr. Albert (Dov) Friedberg of Toronto, Canada. A number of 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, Ben Gurion University, the Ben-Zvi Institute, and more. The aim of these research teams was to extensively study, identify, describe, catalog, and transcribe as many as possible of the Genizah fragments, by looking either at the original manuscripts — wherever they resided — or at their microfilm substitutes, housed primarily in Jerusalem. Each team was focused on a particular domain of Genizah material and was headed by a specialist in that domain. Thus, teams were created for Judeo-Arabic Biblical exegesis, for Talmudic commentaries, for philosophy and ethical works, for responsa material, for documentary material, for magic and magic-related fragments, for language-related material, etc. These efforts resulted in a flurry of Genizah-related activities, such as the compilation of a large amount of data on the fragments studied, the production of a large number of research publications (papers and books), the publication of a yearly scholarly journal devoted entirely to Genizah research (*Ginzei Qedem*), special university courses and seminars, many M.A. and Ph.D. theses, and the like.

About six years later, the stage was set for a new vision: that of implementing advanced computer technologies in the world of Genizah research, not only in order to make Genizah research more easy and efficient, but in order to

(hopefully) develop new ways for implementing automated procedures in that research which would radically change its horizons.

D. Methodological issues

A project with the scope and complexity of the computerization project described here must naturally be guided by a set of methodological principles which clearly define its contour, both in terms of the vision to be achieved and in terms of the practical limitations of the development efforts. A list of such principles follows; most of them were adopted at the starting point of the project, others were defined and adopted slightly later, in the course of taking the first steps towards its implementation.

1. 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 those domains. A Bible researcher would go over tens of thousands of fragments (as many as he practically could), discarding everything except for biblical material. The same line would be followed by a Talmudic scholar for the Talmudic fragments, the same by a linguist searching for linguistic material, etc. Thus, most of the Genizah would be looked at again and again, many times over, but only a tiny part of it would be thoroughly analyzed by the various researchers and their findings recorded in appropriate publications.

The approach of the computerization Project is that the Genizah collection is an integral corpus of its own, and that its gates should be opened to all researchers in all domains and for *all* the fragments found in *all* of the Genizah collections dispersed around the world. Many times we received advice not to process fragments from this or that domain because it "was not important." We ignored all of this advice; one never knows what great surprises a fragment can present until one actually identifies and analyzes it, and one can't know which domain, neglected today, will be the fashionable research topic of tomorrow. We

believe in building the infrastructure of the computerization system to satisfy not only today's needs, but tomorrow's as well.

2. As befits an exact-science effort, and as typical of computers operations, the work of the computerization project must be precise, comprehensive and up-to-date in every aspect. Thus, for example, the project should attempt to trace down every single fragment of this 320,000-piece set, large or tiny, clean or corrupted, even if it appears to be yet another copy of an already known text.

3. Although primarily intended to aid experts in Genizah research, the system should be open to any interested researcher or layman; in fact, one of the project's aims is to, so to speak, "popularize" the Genizah world, transforming it from an esoteric topic that interests a few tens of dedicated scholars in the world, to one that can be consulted and appreciated by thousands of occasional users. Who wouldn't be excited to read an 11th century version of *kaddish* or of *birkat hamazon*, or to find a manuscript with a different version of a problematic *sugyah* he happens to be learning in *daf yomi*?

4. "Genazim" should be an independent group, unaffiliated with any university, institute or library anywhere, so as not to be biased by such an institution's inclinations or interests. A Steering Committee, composed of the most renowned Genizah scholars, with varied areas of expertise and from many institutions, was established to advise "Genazim" on questions of principle in the course of the project's development.

5. Genazim is a computerization group, and even though some of the staff is well-versed in Genizah studies, it is — in principle — not involved in Genizah research of any kind. Of the hundreds of thousands of Genizah data items accumulated in its databases, not one was contributed by Genazim staff. Similarly, every data item integrated in "Genazim's" databases and ultimately displayed in the website is appended by the source of this information: a book, a catalog, a paper, a scholar, etc. Although we are in constant dialogue with Genizah researchers, we don't allow ourselves the liberty of identifying, cataloging, or transcribing a single fragment.

6. We should not be — and are not — arbitrators between scholars. As

happens more often than not, scholars differ on the identification of a fragment, its contents, its author or some other property attributed to it. We display all of the differing — oftentimes contradictory — opinions on our website, indicating their various sources, and let the user choose. As a consequence, only the author of an item may later correct it; any corrections received from other experts will be added as supplementary data.

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

8. The *shelfmark* of a fragment is the name (number) given to it by the library in which it resides, in exactly the same way as is done in every library for any book in its possession. The shelfmark helps the librarian retrieve the fragment when needed, but, more importantly in the Genizah context, it is the unique "identity number" by which it is internationally recognized, mentioned or discussed in the research literature. The world of unique shelfmarks would therefore be expected to be a well-defined, rigorous, fixed and rigid world; however, it can rather be described as "loosely controlled chaos." While many librarians give a unique shelfmark to every fragment, others may give a shelfmark to a group of (many times unrelated) fragments, with no standard system available to name the individual fragments within that shelfmark; librarians may decide to reorganize their libraries and shelves and change shelfmarks accordingly; collections are bought or sold and change ownership and shelfmarks, etc. Still, it was decided to take the shelfmark of a fragment as the central axis to which every single datum of information on that fragment would be attached. Our system does not recognize, and cannot deal with, a fragment to which no shelfmark has been assigned by its owner. Faithful to the policy detailed above, we took upon ourselves, at the very beginning of

Genazim's activity, two important and critical tasks which we believed should form the backbone of our development efforts: the creation of inventories and of digital images.

E. Inventories

The first task undertaken by "Genazim" was the compilation of a comprehensive, precise and up-to-date computerized inventory of the formal shelfmarks of all the Genizah fragments in all of the Genizah collections around the world, whether large or small, "important" or not, public or private. No cataloging data or identification of any sort was included in that task; at this stage we documented only the shelfmark and the number of fragments included by the librarian in that shelfmark. We made every effort to compile this list, not from outdated catalogs or handlists, but by having compilers actually look at every fragment in a given collection and record the corresponding data — including, for example, cases in which an envelope with a certain shelfmark existed but the pertinent fragment (assumed to be inside) was missing, having been loaned out, undergoing conservation, or having simply disappeared. In certain cases, such as those of the 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 the Bibliotheque Nationale et Universitaire in Strasbourg or the British Library in London, "Genazim" sent its Genizah experts, who accomplished the task in cooperation with library staff.

Currently, the inventories residing at the Genazim servers in Jerusalem and displayed on the Genizah website contain about 247,000 shelfmarks and account, we believe, for the totality of Genizah shelfmarks — or, to be on the safe side, for more than 99.99% of all shelf-marked Genizah fragments.

To achieve this, and to assure the comprehensiveness of the process, we had, among other things, to compile, for the first time, an exhaustive list of all the public and private Genizah collections in the world, tracking down even

"collections" that contain a single fragment (such as The Goldsmith Museum of Chizuk Emunah Congregation in Baltimore or The Temple Israel of Hollywood in Los Angeles).

As later became evident, this effort not only allowed us to attach, from that moment on, all available and future data on any fragment to its shelfmark, but also prompted Genizah researchers to make use of the precise and accurate shelfmarks (as formally defined by the relevant libraries) appearing in our inventories in their publications, thus encouraging a much needed trend of standardization in that context. Moreover, since fragments may often change shelfmarks, for the reasons detailed above, we made a sustained effort to collect all of the older or alternate shelfmarks of a given collection, to record them and to attach them to the current one, so as to correctly attach data that may have been appended to an older shelfmark to the newer one.

The Computerized Inventory List of all Genizah shelfmarks, sorted by collections, resides now in the "Genazim" servers in Jerusalem and is displayed on the Genizah website.

F. Digital Images

In the early years of the 21st century, the only alternative available for a researcher desiring to study a specific Genizah fragment, other than traveling and examining it wherever it resided, was to use the corresponding microfilm, available principally at the Institute for Microfilmed Hebrew Manuscripts in Jerusalem, with all the inconveniences and shortcomings typical to this solution.

The decision was therefore made, at the onset of Genazim activity, to produce full-color high-quality digital images of all Genizah fragments and to make them available through the Internet to any interested user. This would enable users to manipulate the images and study any Genizah fragment at any time and from anywhere.

This decision necessitated intense negotiations with representatives of every library that housed a Genizah collection, convincing them of the importance (and

practicality) of digitizing their collection and of cooperating with the Friedberg Genizah Project in this task, persuading them to allow us to display a copy on our website, and negotiating and signing suitable legal agreements to protect their copyrights. In many cases, such as in the Jewish Theological Seminary in New York, the Alliance Israelite Universelle in Paris, and the libraries of Geneva, Strasbourg, Vienna, and others, the Friedberg Genizah Project sent its own expert photographers to accomplish the complex digitization task according to the rigorous standards and parameters set by "Genazim," using novel "running belt" techniques to accomplish the task in record times. In other cases, such as in Cambridge, the British Library and others, the digitization was accomplished by the digitization laboratory of that library, in cooperation with "Genazim" and with its financial support.

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

Every image was allocated a unique "Genazim" number that (unlike the shelfmarks) is fixed and will never change. We encourage researchers to mention this number in their publications (in addition, of course, to the shelfmark), and this recommendation is slowly being implemented.

Currently (May 2012) the Genizah website contains more than 400,000 digital images of Genizah fragments. With the digitization of the Cambridge and the British Library Genizah collections well on their way and expected to be completed by the end of the summer of 2012, we expect this number to rise to 600,000 images, representing probably more than 98% of the Cairo Genizah manuscripts (the exceptions being the collections of Oxford and St. Petersburg, and a couple of very small private ones). This digitization effort of the Genizah collection is probably one of the largest digitization efforts of historical manuscripts collections ever attempted, for any culture or language and by any institution.

G. The Data Axis

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

To begin with data, what kind of Genizah data should be collected, stored and processed, to be finally displayed in the website? After analyzing the Genizah research activities, eight categories of data were found to be the appropriate ones to be collected and processed. I list them below, together with statistics on the amount of such data that has been collected and is currently included in the system.

The first two data items have already been presented above, including in their quantitative aspects: 1. *the shelfmarks* and 2. *the images*. We turn now to the other six:

3. *Bibliographical references*: It was decided to include, in our website, detailed references to any publication that discusses or even mentions any specific Genizah shelfmark, anywhere, at any time and in any language. A complete set of references for all publications in any language that mention the shelfmark of a Cambridge Genizah fragment, from the discovery of the Genizah until 2008, compiled by the Cambridge Genizah Research Unit, was integrated into our databases courtesy of the Cambridge University Library. Moreover, all references to non-Cambridge shelfmarks in Hebrew publications until 2004, and an almost complete set of references to non-Cambridge fragments in publications in non-Hebrew languages, are being compiled by the Friedberg Genizah Project bibliography teams and are also available on the website. In total, almost 200,000 such references are recorded in the databases.

4. *Cataloging data*: To every Genizah shelfmark we append (when available) a cataloging record that specifies, in (mostly) coded form, all available information on that shelfmark. Such data can be related either to the fragment's

physical aspects: outer and inner (text-block) dimensions, number of lines, writing material, margins, corners, holes and tears, etc., or to its "content" aspects: domain (such as Bible and Biblical commentaries, Talmud and Talmudic commentaries, philosophy and ethics, documentary material, medicine, magic etc.; there are about 30 such domains), title of work, author, language, script, scribe, date of copying, etc. About 70 such fields are included in the cataloging record.

About 270,000 such records are currently available on the website; a few of them rather "lean," with just a couple of fields marked, others more complete.

5. *Scans*: To every shelfmark we append scans of all entries that appear in any Genizah-related catalog, whether published or printed, electronic or even just handwritten, that relate to this fragment. Besides the data extracted from an entry in such a catalog and included in the Cataloging Record mentioned above, a chance is given to the user to actually see an image of that entry as it appears in the catalog. With 34 Genizah or Genizah-related catalogs available, very few libraries — and certainly no researcher — can afford to have all these catalogs easily available, 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 tool in itself. More than 70,000 such scanned entries are currently available on the website.

6. *Transcriptions*: Because of the sometimes difficult calligraphy and the physical state of many of the fragments, deciphering the text of a fragment is almost always a difficult task, done mostly by researchers. We therefore made an effort to attach to the image of a given fragment, whenever possible and available, its transcription. About 15,000 such transcriptions have been collected (or transcribed, when needed, to computer-readable form) and integrated in the Genizah databases, and are currently displayed on the website.

7. *Translations*: As noted above, a large part of the Genizah fragments are in Judeo-Arabic, and many of these have been translated to Hebrew. A few fragments have also been translated (either from Judeo-Arabic or from Hebrew) to English. About 3,000 such translations are included in the website.

8. *"Joins"*: One of the most critical issues in Genizah research is that of discovering "joins," i.e. different fragments — parts of folios or folios — originating from the same folio or from the same manuscript that have been dispersed (due to the unavoidable wearing and tearing of the originals over many centuries and to the random acquisition and trade of manuscripts) in different libraries; one fragment being found, say, in Paris, and the other in Vienna. During a hundred years of research, about 4,000 or 5,000 of such "joins" were discovered through the erudition, memory and intelligence of Genizah scholars, assisted sometimes by available catalogs. Of these, about 3,000 are recorded in the system databases and clearly noted in the website.

No additional type of data was found to be important or useful enough to be included in the system, when taking into account the "costs" of its implementation. Thus, for example, we are not adding to the system copies of the (tens of thousands) of Genizah-related papers (and certainly not books) in any format.

The data currently integrated in the website were collected by the systematic inspection of three main sources: the output of the Friedberg Genizah Project research teams, Genizah-related catalogs, and Genizah-related books. From time to time sporadic additions or handlists were received from Genizah researchers.

Finally, even though it would certainly have been very useful and important to systematically examine all the tens of thousands of papers that discuss Genizah material in order to extract from them relevant data (especially identifications and cataloging data) to be added to the databases, such a project is understandably too formidable, in terms of the necessary manpower, resources and time, to be implementable. The only viable solution to this problem, i.e. the integration on the website of all the published data on Genizah fragments, to the extent possible, can only be achieved through the voluntary contribution of the Genizah-interested public, in a Wikipedia style. To this end, we have recently added a new module to the system, which allows

accredited users to easily add such data to the website, which will display it already on the following day (see H 3.6 below).

H. The Software Axis

Obviously, a robust software system is needed to absorb the data described above and to integrate, process and manipulate it in order to make it useful to researchers. The software system we built is composed of 3 major parts:

- the input module, which allows the research teams, and later all accredited users, to directly input data into the databases;
- the databases, in which the data is stored, reviewed, organized, inter-linked and updated;
- the website, which is in fact the only interface between the researcher and the data, and through which the entire Genizah research world is intended to be transparent and available to the user.

I 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). I shall content myself here with a general outline of the website and its various functions, since anyone desirous of doing so can access it and directly use and manipulate its various functions.

There are essentially two ways of querying the website:

1. Searching for data on a specific shelfmark

Users can select a particular 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 collections has its own structure, and the menu is specifically adapted to each such structure). Alternatively, if one knows the exact shelfmark, one can directly type it. One can then choose between the following six different functions, which display all the available data pertinent to

this shelfmark: a) high-quality images of (both sides) of all fragments included in this shelfmark; b) a scanned image of any entry in any Genizah catalog that is related to this shelfmark; c) full bibliographical references to any publication that mentions this shelfmark; d) identification of the fragment (i.e. a short "running title" that describes and identifies the fragment); e) full cataloging records; f) transcriptions of the fragment; g) translations; and finally h) "joins" in which this fragment participates.

One final note on the images:

The images are displayed through the ViewOne viewer (enhanced by functions developed by our computer scientists for this project), which allows for 4 different types of (repeated) magnification and 4 types of fitting the image to the screen. The viewer also allows the user to flip the image, reverse the display from "black-on-white" to "white-on-black", mirror the text (in cases where it was inscribed in mirrored writing), measure the distance between any two points on the image and the angle of any two lines drawn on the image, adjust the contrast, brightness and luminescence of the image, rotate it by different angles (to allow, for example, for the easy reading of margins written vertically or diagonally), store the resulting image for further processing in coming sessions, and more.

Browsing between these functions is completely dynamic; the user doesn't need to retype or re-choose the fragment's shelfmark again and again when switching from one function to another.

2. Queries

Alternatively, a user can submit a query to the system and receive a list of all shelfmarks that satisfy a given set of conditions. The criteria can be a mix of all the data attached to the shelfmark. Following are some examples. One could search for:

- shelfmarks of all biblical fragments from Exodus that have cantillation signs, originate from the 12th to 14th centuries, contain at least 5 lines, and form a join with another given fragment;

- shelfmarks 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. (as an interesting example: the set of all shelfmarks from the Manchester collection that were never mentioned in any publication);
- shelfmarks from the "magic" domain, for which there are at least N different identifications, and for which there is an entry in catalog A.

3. Additional tools

Besides the two options described above, many additional modules are available on the Genizah website to help users conduct their Genizah research efficiently.

1. **QuickView**: a standard digital image, having a high resolution, may take some time to display. The QuickView function allows users to browse very quickly (typically tens of images in a few minutes) through (low quality) consecutive images of a given collection's shelfmarks, so as to focus on the fragments that interest them;
2. **Full-text**:a full-text search is also available, and can be applied to the transcription/translationtexts, the Genizah catalogs' text, the "running title" or the free text section of the cataloging record, etc. Likewise, a list of all the different words in the transcriptions' corpus, with their frequencies in that corpus, is available, and can be sorted alphabetically or by (increasing or decreasing) frequencies, for browsing by the user;
3. **Workspace**: an individual workspace is available to every user, where they can store and manipulate in their privately designed structure a small set of images which they are currently researching, and which is stored for them from session to session;
4. **Forum**: a public forum where users can exchange information, discuss issues about given shelfmarks, add or correct data, etc., is available to all users. Any user can also build a "restricted" forum for internal discussions between himself and his restricted set of colleagues;

5. **Notes**: Short notes can be written by any user, to be appended to a specific shelfmark and displayed to all users;
6. **Input**: A special module ("FOLUS" — Friedberg Online Users' Input) allows accredited users to add information (identifications, cataloging data, joins, transcriptions, etc.) to the system, which will be integrated in the databases and displayed on the website (with their names as the source) the very next day.
7. **Jigsaw**: When trying to test a hypothesis about the possibility of joining 4 or 5 fragments, say, into one folio, a researcher may invoke the function "Jigsaw," giving it the numbers of these fragments' images. The images will then be displayed on his (preferably large) screen, where he can rotate or move any of them in an effort to fit them physically together, as in a real puzzle. If satisfied, he can then store the final image on the website.
8. **Website instances**; A user can open, on his (again, preferably large) screen, several (up to 4) instances of the website, processing each of them independently, thus looking simultaneously at a fragment's image on the first instance, at its catalog record on the second, at its transcription on the third, etc.

While some of the ideas presented here are specific and closely geared to the Genizah collection, others may be applied to various large collections of historical manuscripts which represent treasures of cultural heritage in its truest meaning (the Dead Sea Scroll collection comes to mind as a case in point). In any case, it seems that no even remotely similar website, with its huge set of images and its rich research and manipulation options, has ever been developed before for any collection of handwritten historical manuscripts.

I. Research Achievements in Digital Image Analysis

In this last section, I would like to briefly describe some remarkable results in the area of computer-assisted analysis of high-quality digital images of historical handwritten manuscripts, which were achieved by Genazim AI group

researchers (Choueka, Dr. Roni Shweka and Adiel ben-Shalom) in cooperation with researchers from the Computer Science Department of Tel Aviv University (Profs. Lior Wolf and Nachum Dershowitz and their assistants). These results, never achieved or applied before to any collection of handwritten manuscripts, were presented at various international conferences and published in many journals and books (for a list of references see the "Conferences and Papers" page on the Genizah website).

1. Why digitize a manuscript?

Traditionally, the claim is put forward that there are basically two important reasons for digitization: conservation and accessibility. Digitizing a manuscript produces the closest possible surrogate to the original (some would even say a better one), in case the original is destroyed either by a force of nature (fire, earth-quake, inundation) or just through malpractice. Making this image available on the internet, on the other hand, provides access to such a surrogate for any interested user, anywhere, anytime, and thus usually saves him the time and trouble of traveling and, more generally, makes him independent of any institution's operational procedures.

We claim, however, that a digital image is necessary also because it is the only format a computer can "understand" and analyze. The idea is that we should look at the computer as a "potential user," in fact even treat him as a "privileged" one. If we make the effort of digitizing the manuscript with the parameters best suited to computer use, it will reward us a hundred times over by supplying us with data, information and suggestions that may save us a lot of tedious labor, and maybe also point us in interesting new directions in our research tasks.

2. How to digitize a manuscript?

Following are some of the main conditions we found it important to apply when digitizing collections of historical manuscripts (these parameters were

presented in[1] below and have already been adopted by Cambridge and by the British Library):

1. The resolution (dots per inch) should be set at approximately 600 dpi. Less than that will not give a satisfactorily detailed image, and much more than that will make it too "heavy" (in terms of size in megabytes) and difficult to manipulate, and will make it impractical to display online on the internet.
2. In order to allow the computer to delineate the fragment and extract it from its background, the background should have a color which is maximally contrastive to that of the average writing material and of the ink typically used in such a hand-written fragment. We found, experimentally, what this average is for Genizah material, and accordingly what the contrastive color should be. It was found to be a certain kind of blue that can be precisely defined in technical color schemes standards.

It is important to add that if a particular library thinks that this background color is not suitable for its users, the background can be automatically changed to any background color desired, since, if the background color was chosen as specified, the computer can successfully recognize the background areas almost to the pixel, and therefore can color the background with any desired color.

3. The use of external artifacts such as clips, weights, etc., should be avoided as much as possible. If absolutely necessary, these artifacts should be colored in the blue hue mentioned above so as to allow the computer to easily recognize them as parts of the background.
4. It is necessary to insert a ruler in the image, alongside the fragment but of course without covering any part of it, so as to allow for the calibration of the image.

Assuming these parameters are followed, let us now focus on two of the

1 Roni Shweka, Yaacov Choueka, Lior Wolf, Nachum Dershowitz, and Masha Zeldin. "Automatic Extraction of Catalog Data from Genizah Fragments' Images" *Digital Humanities* (DH 2011), Palo Alto, CA., June 2011.

remarkable results that were achieved by the advanced computerized analysis of a manuscript's digital image. The starting point was to try and discover what physical attributes of a Genizah fragment could be automatically deduced by the computer through a fine analysis of its digital image.

3. What physical attributes of a fragment can be automatically identified by the computer?

We were able to develop a few software modules that, through a computerized analysis, can allow the system to:

- recognize and follow the exact contour of the textual part of the image, thus separating it from its background;
- measure the fragment's inner and outer dimensions;
- count its number of lines;
- compute the average written-line width and length, the average inter-line width, the average "text density" (the number of letters in a specified measure unit);
- compute 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 partially found in catalogs of manuscript collections, which until now had been marked 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.

We are in the process of implementing these findings on the set of images currently in our databases and, ultimately, on the complete set of Genizah images. The data derived by this process will be integrated into our databases and displayed on the Genizah website.

4. Suggesting joins

A crucial further step was achieved when we succeeded in developing a complex program capable of analyzing the handwriting in the images of two different fragments and asserting the probability that both were written by the same

scribe (and so, perhaps, originate from the same manuscript). This is not done through the analysis of the individual handwritten letters and their shapes, but rather through a global comparison scheme, vaguely similar to the way in which two portraits can be compared by computer and found to be of the same person. In this way we were able to discover, in a rather short time, hundreds of hitherto unknown joins. A paper recently published by Roni Shweka in this journal[2]describes the "joins" component of the project and lists more than a hundred new joins in various Genizah domains which were discovered in this way.

It should be added that even when a similarity between the handwriting of two images does not point to a "strong" join, meaning that the two fragments originate from the same manuscript, it may still point to a "weak join," meaning that both fragments were written by the same scribe, an important fact that can be — and has been — used many times to draw far-reaching conclusions on authorship, identifications and other parameters related to such manuscripts.

5. The Grand Vision

Our grand vision now is to use these techniques in order to compare every Genizah fragment's image to the image of every other one, so as to find all possible strong or weak joins between all Genizah fragments. Our problem in achieving this in the coming few years is two-fold: the mind-boggling number of comparisons required (hundreds of billions!) and the need to filter the true joins suggested by the system from the "false positive" ones, i.e. cases in which the system was fooled by the images into assigning them a high probability of being joins while they are not really so.

We are now trying various approaches that may help solve this complex problem, including the very recent "Citizen Science" one, in which the general

2 Roni Shweka, Yaacov Choueka, Lior Wolf, and Nachum Dershowitz. "'Veqarev otam ehad el ehad': Zihuy ktav yad vetseruf qit'ei hagnizah beemtsa'ut mahshev (Identifying Handwriting and Joining Genizah Fragments by Computer)." *Ginzei Qedem*, vol. 7 (2011), pp. 173–209 [Hebrew].

public, which most probably doesn't have any expertise at all in this area, is asked to contribute of its time and common sense to propose conjectures in special cases in which the human brain can do much better that the computer, and thus help alleviate the problem.

If and when these problems are solved, we shall be ready to attempt *the reconstruction of the original Genizah library*, a crucial step which may completely change the horizons of Genizah research in the near future.