The Penn/Cambridge Genizah Fragment Project: Issues in Description, Access, and Reunification

Heidi G. Lerner
Seth Jerchower

ABSTRACT. The University of Pennsylvania Library and the Taylor-Schechter Genizah Research Unit at Cambridge University Library in England have embarked on a project to digitize their joint holdings of manuscript fragments from the Cairo Genizah. One goal of this collaboration is to develop and implement an online catalog and image database for the University of Pennsylvania's collection of Genizah fragments, which will provide the foundation for a global electronic repository and catalog of the entire Cairo Genizah. The project staffs have developed preliminary guidelines for standardized descriptive metadata. The authors discuss the

Heidi G. Lerner is Hebraica/Judaica Cataloger, Stanford University Libraries (E-mail: lerner@stanford.edu).

Seth Jerchower is Public Services Librarian, Center for Advanced Judaic Studies, University of Pennsylvania (E-mail: sethj@pobox.upenn.edu).

The authors would like to thank Michael Ryan (Director, Annenberg Rare Book & Manuscript Library, University of Pennsylvania), Arthur Kiron (Curator of Judaica Collections, University of Pennsylvania Library), Greg Bear (Manager, Schoenberg Center for Electronic Text & Image, University of Pennsylvania Library), Stefan Reif (Director, Taylor-Schechter Genizah Research Unit, Cambridge University Library) for their vision and work in creating the University of Pennsylvania/Cambridge Genizah Project; Ezra Chwat (Institute of Microfilmed Hebrew Manuscripts) for expertly describing the Genizah fragments; Zachary Baker (Reinhard Curator of Judaica and Hebraica, Stanford University Libraries) for reading this paper and providing valuable editorial comments and suggestions.

Cataloging & Classification Quarterly, Vol. 42(1) 2006
Available online at http://www.haworthpress.com/web/CCQ

doi:10.1300/J104v42n01_04

issues and difficulties specific to cataloging these fragments, how an online catalog can facilitate this ambitious task, and why MARC tagging was adopted for this purpose.

KEYWORDS. MARC, manuscript cataloging, descriptive metadata, Unicode, digital repositories, Cairo Genizah

BACKGROUND

The Penn/Cambridge Genizah Fragment Project has been established as a model to reunite scattered manuscript fragments from the Cairo Genizah via the World Wide Web and other information technologies. The University of Pennsylvania Library and the Taylor-Schechter Genizah Research Unit of the Cambridge University Library initiated this project and created a web-based catalog and image database (http://sceti.library.upenn.edu/genizah).[1] In the following pages, the authors introduce the contents of the Cairo Genizah and a look at some of the earlier types of catalogs used to describe the fragments, describe how and why Machine-Readable Cataloging (MARC) tagging was adopted and interpreted for this project, and give a brief overview of the imaging technology.

INTRODUCTION TO THE CAIRO GENIZAH

A *genizah*[2] (plural *genizot*) is a storeroom or repository for old, used and damaged books, Torah scrolls, and other documents containing the name of God, whose destruction Jewish tradition proscribes. The tradition of setting aside volumes containing sacred Hebrew texts rather than destroying or disposing of them is an ancient one, found in practically every Jewish community. Yet very few *genizot* have survived, since their contents are typically buried. The *genizah* of the Ben Ezra synagogue in Fustat,[3] Egypt (a Byzantine outpost, whose founding in 643 C. E. predates that of Islamic Cairo) is unique for a number of reasons:

1. It survived because the majority of the fragments were never removed for burial. Worn-out volumes and leaves were deposited in a second floor chamber located behind the women's gallery (in

later periods this entrance was closed off and the chamber was accessible only through an exterior passageway). Remarkably, they survived fires and acts of vandalism.

2. The quantity of materials, estimated at 220,000+ fragments.[4]
3. The time span that its contents cover. We know that the Ben Ezra Synagogue, or *Kanīsat al-Yerūshalmīyin* (or *al-Shāmiyīn*), was ordered destroyed by the Shiite caliphate in the early part of eleventh century, and rebuilt about 1040 (one Muslim source states that the Coptic Patriarch was forced to sell the church of St. Michael to the Jewish community in 882[5]). The fragments extend from the eighth or ninth century (and even earlier, as the palimpsests are counted and examined) up through the nineteenth century, with large concentrations of materials dating from the tenth through the fifteenth centuries.
4. The importance of the Fustat community. The Ben Ezra Synagogue was the center of the Egyptian and indeed Mediterranean Jewish world during the Fatamid period (969-1171), and home to the Egyptian *nagid*[6] (in Islamic countries, the head of the Jewish community). Prior to this, the *Kanīsat al-Yerūshalmīyin* was the seat of the Palestinian Jewish community, one of two Rabbanite communities (the other being Babylonian), which coexisted with a Karaite community.[7]
5. Throughout much of the nineteenth century, various collectors gained limited access into the Genizah. The community gave some items as gifts; others made their way into the marketplace. In 1896, Solomon Schechter, then Reader of Rabbinic Literature at Cambridge University, became aware of the Genizah's potential importance for Jewish studies. With the intellectual and pecuniary support of Charles Taylor, Master of St. John's College at Cambridge, the balance of the fragments–today estimated at more than 140,000–was acquired from the Cairo Jewish community and brought to Cambridge.

PROBLEMS INHERENT TO GENIZAH STUDIES

In its present state, the Cairo Genizah presents, depending on one's proclivity, a cataloger's paradise or a cataloger's nightmare. Cambridge, with more than 140,000, holds the largest collection of fragments in the world.[8] Of the original estimate of 220,000, where are the remaining 80,000 to 100,000 fragments?

England

Cambridge, University Library– > 140,000; Westminster College– ± 2,000.
Manchester, John Rylands University Library– ± 10,000.
Oxford, Bodleian Library– ± 5,000.
London, British Library– ± 5,000.
Birmingham, Selly Oak Colleges, Mingana and Mittwoch Collections– ± 40.

United States

New York, Jewish Theological Seminary of America– ± 30,000.[9]
Philadelphia, University of Pennsylvania's Center for Advanced Judaic Studies– > 500; University of Pennsylvania Museum of Archaeology and Anthropology–28 (not all Cairo Genizah proper).
Cincinnati, Hebrew Union College-Jewish Institute of Religion– ± 250.
Washington, D.C., Smithsonian (various)–114.

France

Paris, Alliance israélite universelle– ± 4,000; Jack Mosseri Collection– ± 4,000.
Strasbourg, Bibliothèque nationale et universitaire– ± 1000.
Austria: Vienna, Österreichische Nationalbibliothek, Rainer Collection– ± 150.
Hungary: Budapest, Academy of Sciences– ± 650.
Russia: St. Petersburg, National Library of Russia: Antonin Collection– ± 1200; Firkovich Collection–several thousand.
Ukraine: Kiev, Academy of Sciences, Abraham Harkavy Collection– several dozen.
Israel: Jewish National and University Library– ± 300.[10]

The Cairo Genizah in its "original" state was not a collection as much as a completely disorganized and unattended mass of discarded materials, subject to perusing and plunder. At present, its contents are better described as "scattered" than "distributed." Individual leaves from any one particular manuscript, and fragments of individual leaves are dispersed among different institutions. From the time of the first divisions of fragments into personal and institutional holdings, collections have been sold; institutions have come and gone; two world wars have been

fought; and maps, states, governments, and ideologies have changed. To say that these events have complicated any inventorial assessment would be an understatement.

CATALOG TYPOLOGY

Over the past century and a quarter, many catalogs of the Cairo Genizah fragments have been produced. One type of catalog is organized around a local collection. As early as 1886, Adolf Neubauer's *Catalogue of the Hebrew Manuscripts in the Bodleian Library and in the College Libraries of Oxford*[11] [reissued in 1994 with addenda and corrigenda by Malachi Beit-Arye] included entries on the Bodleian's collection of Cairoene fragments. Another, perhaps the most exasperating, although engaging example is the one published in 1921 for the Elkan Nathan Adler collection[12] (now housed at the Library of the Jewish Theological Seminary of America). Catalogs of great value were produced for some of the smaller collections, such as that of the Smithsonian (*Fragments from the Cairo Genizah in the Freer Collection*).[13] The Freer catalog treats its collection as an integral and publishable entity, and contains extensive descriptions, photographic facsimiles, transcriptions, and translations of its 52 fragments. Benzion Halper's *Descriptive Catalogue of Genizah Fragments in Philadelphia*,[14] provides neither facsimiles nor transcriptions. However, it is organized topically. The 487 fragments it describes are now housed at the University of Pennsylvania's Center for Advanced Judaic Studies Library. As such, it provides an important base for the Penn/Cambridge project.

Another type of catalog is organized solely by a single topic or genre. Some are oriented on specific genres within local collections, such as Lewis-Gibson on the Syriac palimpsest fragments at Cambridge.[15] Neil Danzig's 1997 catalog of Rabbinic fragments at the Library of the Jewish Theological Seminary of America,[16] while oriented upon the holdings of one local collection, exhaustively provides cross-matches and concordances. The best results transcend borders. As early as 1901, the *Facsimiles of the Fragments Hitherto Recovered of the Book of Ecclesiasticus in Hebrew*, edited by Solomon Schechter,[17] was published containing 60 photographic facsimiles of all extant Ben Sira Ecclesiasticus fragments. This edition, although sparse on physical description, gathers in one volume fragments from the Taylor-Schechter, E. N. Adler, British Museum, Lewis-Gibson, Bodleian, Consistoire israélite de Paris, and Gaster collections, identifies and collates what

had originally been four separate codices, and provides (for its time) exhaustive bibliographies.[18] The late Michael Klein's catalogs of Palestinian Targumic fragments assess fragments within the context of the *Targumim*[19] (translations of the Bible into Aramaic), reconstruction of the original codices, and their distribution throughout the current collections.

Another noteworthy endeavor is the CD-ROM distributed by the Saul Lieberman Institute Database of Talmudic Versions.[20] This CD-ROM includes information regarding all Talmudic Genizah materials, is subject to periodical updates, and represents the first major multi-tiered and searchable electronic catalog of the Genizah. It is not restricted to Genizah fragments, but includes all manuscripts and early printed versions of the Talmud.

CATALOGING ISSUES[21]

To date, the above-mentioned catalogs have shared the disadvantage that plagues printed catalogs: the contained data (or "descriptive metadata," as it were) are static. The advantage first presented by an online catalog is its dynamic relation to the data. Data can be virtually input and distributed, as well as updated as needed. An online catalog also provides and facilitates exponentially greater search capabilities. This type of catalog is best adapted for the handling of Cairo Genizah materials. Data may be entered locally and stored centrally, enabling an ideal level of information exchange and one that was previously unattainable.

However, in addition to the concept of static vs. dynamic data there is another concept relative to the typology of data, i.e., of information. Scholars and special collections librarians often refer to collections of many now defunct Jewish libraries. These former collections, such as the David Sassoon Collection whose holdings were broken up over time and redistributed, have well documented printed catalogs (e.g., *Ohel Dawid*[22]). Although, these collections are no longer intact, their catalogs still offer a valid point of reference. This is not the case for the Cairo Genizah. As long as it existed and functioned as a *genizah*, it could be considered, only in the most generous of descriptions, a repository. During its existence as an entity under one roof, no efforts to catalog its contents were ever made. We have already discussed the random and international scattering of individual codices, leaves, and pieces of leaves. By necessity, each entry produced by the Penn/Cambridge collaboration includes its respective Halper reference. A Halper refer-

ence contains *inter alia*, important provenance information, e.g., Cyrus Adler, Amram, Sulzberger, etc.

Conceptually, Cairo Genizah fragments contain two types of information: intrinsic and extrinsic.[23] The Genizah's unusual distribution renders necessary an expansion of the definition "extrinsic data." Any information regarding cross-matches, whether intra- or extra-institutional, textual, or codicological, is data extrinsic to the fragment itself.

DESCRIBING THE GENIZAH FRAGMENTS

The initial phase of the project focused on the digitization of the fragments and the creation of online catalog records for the individual fragments held by the Center for Advanced Judaic Studies (CAJS) at the University of Pennsylvania. The aim was to create a searchable Web-based image database, which allows scholars to locate and identify individual fragments by title, author, institution, language, physical characteristics, subject, or bibliographic history. A template was developed that provides the descriptive elements used for individual Genizah fragments. These elements were defined and then mapped to the corresponding MARC 21 tags.

Metadata has become widely discussed in the library, scholarly, computing, and publishing communities. Information professionals in particular are excited about its potential to improve access to electronic materials. Any institution that begins a project utilizing metadata to describe its resources should very carefully develop its strategy to address the technical, organizational, and human challenges involved in such a project. Careful collaboration and planning between individuals and among institutions is ideal. Many digitization projects are discovering that it is expedient to integrate metadata into existing library systems and take advantage of well-defined standards of organizing information.

MARC originated in the 1960s as a means of exchanging library catalog records. It is made up of a data structure and encoding procedure that implements national and international standards. Today *MARC 21 Format for Bibliographic Data*[24] is the encoding format most commonly used by libraries in North America. Europe and the international cataloging community are rapidly adopting the MARC 21 standard for the creation and processing of bibliographic data. Significantly for this project, the MARC 21 format provides an expedient means for integrating descriptive metadata of the fragments into existing library systems.

Up until the present, the manuscript community has not embraced the MARC standard for its cataloging purposes. In codicology, the traditional methods for locating manuscripts have been printed catalogs. With the advent of such SGML- and XML-based electronic technologies and projects such as the Text Encoding Initiative (TEI),[25] the Digital Scriptorium,[26] or the European project known as Manuscript Access through Standards for Electronic Record (MASTER),[27] we are seeing more manuscript metadata on the Internet. These projects adhere to encoding standards, already accepted and used by humanities scholars, which allow for uniform searches within and across databases. Their adherents feel that the MARC standard does not adequately support manuscript description, and that SGML and XML can be used by tools beyond those found in library and archive communities.

The University of Pennsylvania Library and Taylor-Schechter Genizah Research Unit agreed that in addition to creating the website, a key component of the project is to integrate bibliographic records for the digitized images into Penn's local MARC-based catalog (Franklin) that runs on Voyager.

PENN/CAMBRIDGE GENIZAH PROJECT AND THE USE OF MARC 21

As much as possible, the project aims to adopt MARC 21 encoding procedures for the cataloging of these fragments, and to provide cataloging that is reasonably compatible with *Anglo-American Cataloguing Rules (AACR2)*,[28] and *Descriptive Cataloging of Ancient, Medieval, Renaissance, and Early Modern Manuscripts*.[29] The resultant records provide bibliographic control over the fragments, which, owing to their unique linguistic, religious, intellectual, historical and literary value, require precise and detailed identification. The appropriate MARC 21 tags also allow linking from the online catalog record to the digitized fragment. The *Library of Congress Subject Headings* are used to provide controlled subject access. Personal, corporate, and title headings provide a unique challenge. For the most part, the records use the authorized headings that already exist in the *LC/NACO Name Authority File* (NAF) or headings that have been created according to AACR2 guidelines.

Again though, the complexities of cataloging these items have to be emphasized. Most of these fragments are incomplete documents, with their mates scattered among many different institutions and collections,

or even in different volumes within the same collection. They very often lack a title or colophon. Individual volumes may include multiple and even unrelated texts.

A main entry (100, 110, or 130) field is provided when applicable, in as many cases as possible. The fragments cover a variety of types of material including literary fragments, liturgical works, biblical and rabbinic texts and their related commentaries, and other philosophical, scientific, and linguistic writings. Also included in the collections are a number of legal documents, communal and commercial records, educational documents, and private letters. For those works in which there is an identifiable author, a personal name heading goes into the 100 field. Ideally, the heading matches the form established in the LC/NACO NAF. If no heading exists, the author's name should be formatted in accordance with current AACR2 cataloging and ALA/LC romanization standards. Alternatively, a fragment emanating from an administrative or communal body, institution or synagogue will have its issuing body recorded in a 110 field. A fragment containing liturgical, biblical, or rabbinic texts should have a uniform title with an indication, for example, of its part, version, language, and translator.

For a fragment that contains a formal title, the title proper [245 field] will reflect the exact wording, which appears on it, or is extracted from one of the appropriate printed descriptions of the collection. If the fragment does not have an identifiable title, the title statement is provided by the cataloger based on an existing description of the fragment or from direct examination of it. If the manuscript contains several unique items bound together, a constructed title is provided that represents the themes found in the group of items.

Many of the fragments are works that are, or include translations, or are parts of a larger work. When an author or corporate body is included as a main entry in a record for a fragment, a uniform title is provided in the 240 field.

Alternative titles as might be extracted from various existing printed catalogs such as B. Halper's *Descriptive Catalog of Genizah Fragments in Philadelphia* can be recorded in a 246 field.

The 260 field contains information concerning the place and date that a manuscript was copied. The collation, i.e., foliation, and unique physical characteristics such as the physical state of the fragment are noted in the 300 field. Since this field records the physical description of an item, the "subfield b" [Other Physical Details] can be used to describe and index the information on the condition of the fragment. Tagging for this

information is not specified elsewhere in *MARC 21 Format for Bibliographic Data*.

Codicological and paleographical features are crucial in providing the most precise description and identification of the fragments. One challenge currently facing the development of the template is posed by the limitations of *MARC 21 Format for Bibliographic Data* in encoding these unique characteristics. In addition, the encoding of these features is what ultimately needs to be done to enable the most precise location and identification of the fragments. The 340 field normally contains the physical description for an item that has special conservation or storage needs. This field is adapted to record information such as the material of the fragment, the dimensions of the fragment (given in centimeters), the medium of writing and how it was used to inscribe the text, the layout (number of columns, blank sides), and binding.

The 500 fields provide detailed descriptions for those fragments which contain more than one work, or whose author and title are unidentifiable. Very often a detailed overview of the contents of the fragments needs to be provided with precise listings of the various passages. Passages from anonymous texts are often quoted.

The contents of the Cairo Genizah include almost anything written in Hebrew script, i.e., Hebrew, Judeo-Arabic, Ladino, Judeo-Greek, Jewish Aramaic, Judeo-Persian, and Yiddish. There are also fragments that are written in non-Hebraic scripts and in non-Hebraic languages such as Arabic in Arabic script, Coptic, Ethiopic, Syriac, and even Chinese. Many of these fragments are/or include translations. Information on the language/dialect of the item and details such as the type of script, vocalization and other linguistic and calligraphic details are currently recorded in the 546 field with the corresponding MARC language codes appearing in the 041. It is hoped that a more exact method for indexing these paleographic elements can be developed.

Another key element in identifying the Genizah fragments is a detailed bibliographic history. The 510 field has been adapted to inform scholars where a particular fragment has been listed or described, such as in a catalog or bibliography. Producers, i.e., the contributors to the database; and end-users, i.e., scholars and researchers need to know such things as copyright and reproduction information. The 540 field displays the terms governing the use or reproduction of the described materials.

Provenance/acquisition information providing details on where the original of the digitized fragment is held, as well as former ownership, is recorded in the 561 field.

The 580 field provides information that can link part of one incomplete manuscript to its mate(s) in another collection or collections. This field also links individual collections (as subsets) to the superset of the Cairo Genizah.

The 581 field has been adopted to provide scholars with information on published descriptions of the item, when or where it has been cited or published, or to lead researchers to articles or monographs that are based on research that emanates from the collection.

Topical identification of the fragments is crucial. This can include identification of periods in which the original text was written (such as *tannaitc*, *gaonic*), etc.; bodies and genres of literary texts such as *midrash*, *piyyutim*,[30] legal responsa, and philosophical tracts; subject matter of the documents; identification of persons cited or involved in personal or commercial transactions; religious, rabbinic, biblical and liturgical works and their related commentaries. This information is to be recorded in the relevant 600, 610, 630, 650, or 651 fields.

The 700 fields provide an adequate means of giving access to the many additional titles, people, and institutions that may be identified with a fragment. The 700, 710, 730, and 740 fields are used to record added entries for people or corporate bodies that are partially responsible for the document, as well as texts that co-exist, are related, or are included alongside the work that is being described. These can include an additional author, translator, commentator, witness, owner, editor, or signatory; court, synagogue, or school. Fragments frequently contain multiple literary entities such as *piyyutim* and identifying each individual work is imperative for researchers and scholars. These are recorded by title (or opening refrain) in the 740 or 730 fields. Often fragments will contain text that covers more than one book of the Bible or tractates of Talmud, and these additional books and chapters will be recorded in the 730 field as added uniform title entries. The 787 field is used to provide a link to other fragments whose relationship has been described in the 580 field.

One of the major goals of this project is to utilize these online records as the basis for search and retrieval of the digitized documents themselves. The 856 field records the electronic location of the digitized fragment. Ultimately, this field can also provide an electronic link to digitized images of related fragments that may be found in other collections.

In a collection as large and diverse linguistically and bibliographically as the Cairo Genizah, it is obvious that multiple forms of personal and institutional names will be present. These forms differ both within

the documents themselves as well as the way they are cited in descriptive catalogs. Name and title headings are submitted to the LC/NACO NAF via the NACO Hebrew Funnel (see Figure 1) as part of the initial project.

CREATING THE RECORDS AND THE IMAGES

The initial phase of the project is nearly complete. The electronic versions of the University of Pennsylvania's Cairo Genizah fragments and corresponding descriptions are now housed at the University Library's Schoenberg Center for Electronic Text & Image (SCETI).

A manuscripts scholar may not be an expert in the *MARC 21 Format for Bibliographic Data* and other standard cataloging tools. Conversely, cataloging librarians who are familiar with the conventional resources

FIGURE 1. An example of an authority record contributed by the Penn/Cambridge Genizah Fragment Project to the LC/NACO NAF.

RLG's RLIN21(TM) -- FIND Record ID NAFR200330971

ID:NAFR200330971

VST:d 2003-10-13 **ST:**p **MS:**c **EL:**n

001 nr2003030971

003 DLC

005 20031011162018.0

008 030922n?·acannaabn··········?a·aaa·····c

010 __ ‡a nr2003030971 ‡z nr2003033038

040 __ ‡a PU-CJS ‡b eng ‡c PU-CJS ‡d PU-CJS

100 0_ ‡a Shemaryah ben Aharon, ‡c ha-Kohen, ‡d 12th/13th cent.

400 0_ ‡a Shemariah ben Aaron, ‡c ha-Kohen, ‡d 12th/13th cent.

400 0_ ‡a Shemaryah, ‡c ha-Kohen, ‡d 12th/13th cent.

400 0_ ‡a Shemarya ben Aharon, ‡c ha-Kohen, ‡d 12th/13th cent.

670 __ ‡a Collected liturgical poems in honor of the bridgeroom, 12th century-13th century? ‡b (name not given)

670 __ ‡a Ency. Judaica, c1972: ‡b v. 13, col. 595-596 (Shemariah b. Aaron ha-Kohen, Babylonia, 12th/13th cent.)

670 __ ‡a Catalogue of the Hebrew manuscripts in the Bodleian Library and in the college libraries of Oxford, 1886-1906. Manuscript, Heb. f. 58 (Cowley 2853, 9): ‡b v. 2, col. 1331 (Shemaryah ha-Kohen [acrostic])

670 __ ‡a Institute of Microfilm Hebrew Manuscripts, online catalog, Sept. 23, 2003 ‡b (hdg.: Shemaryah ben Aharon ha-Kohen)

670 __ ‡a Śefune shirah, c1967: ‡b p. 144 (Shemaryah ha-Kohen [also in acrostic]) p. 146, etc. (Shemarya ha-Kohen ben Aharon)

and guides used in cataloging may not have sufficient scholarly background to provide the complexity of descriptive information required for manuscript cataloging. Within the scope of this project, experienced codicologists and librarians have demonstrated that they can work together to provide the best and most accurate information to describe and index these manuscript fragments. A highly skilled Hebraic and Judaic manuscripts specialist provided detailed bibliographic descriptions of the individual fragments.

The entries were originally created as Unicode text documents,[31] in which each line corresponded to a specific MARC 21 bibliographic data field. The array of field, indicator, and subfield codes were presented to the cataloger as a predefined set or template. The cataloger was instructed to denote indicators with a respective numeral (or underscore "_" in the case of a blank indicator) and subfield code delimiters with the dummy symbol "|" followed by an appropriate alphanumeric. It was decided that quality control regarding field code, delimiter, and subfield code values, as well as content, would be analyzed, and if necessary, corrected by proofing editors once the tagged data was submitted as Microsoft Word files. Any plain text editor could be used; at the time, Word 2000 was used for its Unicode compatibility in handling romanization symbols, and Hebrew and Arabic characters. Individual records were not saved as single files. Rather, the cataloger batched multiple records in a single file, in which two blank lines separated individual records.

The batches of records were combined to form a single batch, and saved as Unicode text. This file was then given to the proofing editors. MarcEdit, a free MarcMaker, MarcBreaker, and editing utility,[32] was employed to convert this file to individual MARC compliant records. While many corrections in formatting and the addition of leader and directory control fields were required to convert and break the file into individual MARC records, nearly all such modifications were expedited through global "find and replace" commands.[33] The individual MarcBreaker records were then uploaded into the University of Pennsylvania Library's Voyager integrated library system (see Figure 2). A Hebrew language monographic cataloger was hired to ensure that the bibliographic descriptions adhered to the local and national cataloging standards that were adopted and adapted for this project (correct MARC 21 tagging, appropriate use of and creation of subject headings and access points, proper formatting of bibliographic data). The corrected MARC 21 bibliographic records were exported from Penn's Voyager database (Penn's current version of Voyager only supports the

FIGURE 2. Example of a Voyager record for a Genizah fragment (Halper 109). The 787 fields represent links to other fragments of the same manuscript, housed at the Taylor-Schechter Research Institute (Cambridge University Library).

MARC-8/ALA character set[34]) and saved as 16-bit text files. All Hebrew and Arabic script was manually "pasted" back from the earlier batch file, and then the records were uploaded into the SCETI website which resides on a UTF-8 compliant version of an Oracle database (see Figure 3). At the SCETI website, they were converted into Dublin Core for public display[35] using a MARC to Dublin Core crosswalk model developed by Cornell University Library[36] for public searching.

Digitized facsimiles of all the leaves associated with each fragment were made with images of both the recto and the verso scanned. These 24-bit RGB full-color facsimiles have been digitized at 600 dpi and archived as TIFF images on the SCETI network server. For web delivery

FIGURE 3. A portion of a 500 note (the same bibliographic record as Figure 2, after it was loaded into the SCETI database [Dublin Core element "Description"]): the Hebrew script appears in the SCETI generated bibliographic record.

DESCRIPTION

Content. General Notes

Catchword on the bottom of verso; article endings marked with a punctus; three line addendum in margin of recto. Short comments, closer to the sporadic style of Rashi, than the flowing style of the Gaonic glosses. This is not Rashi's gloss, nor Judah ben Nathan's. No sources are mentioned by name. This hand is known to have copied other Talmud glosses that may have been from pre-crusade Worms or Mayence, that were later to be obscured by Rashi.The recto contains comments on the end of 22b to the end of the Mishnah on 23b. The verso contains comments on 24b שיא תשא לע ועטנה to the beginning of 25a

purposes, digital surrogates are converted from the original TIFF images using LizardTechs Multi-Resolution Seamless Image Database[37] (MrSID) format (see Figure 4). The MrSID images are stored on the SCETI network server, and loaded onto clients as JPEG images for quick downloading, while retaining the high resolution details of the TIFF. Four different image sizes are available, as are functions for magnification, rotation, and comparison of multiple fragments. All the metadata from the bibliographic records is available directly from the web site. In addition to being available through the SCETI site, the bibliographic records of the University of Pennsylvania fragments will be accessible and searchable through the University of Pennsylvania Library's Franklin OPAC. Access to the respective SCETI page is made possible through persistent URLs (PURLs) contained in 856 fields in the holding records. The PURLs are maintained by SCETI.

Although not within the current scope of the project, transcriptions of the fragments could potentially be incorporated. The Princeton Geniza Project, based at the Dept. of Near Eastern Studies at Princeton University[38] is engaged in placing S. D. Goitein's transcriptions online.[39] Links could be made available from within the individual records to fully searchable, online transcriptions of the fragments in both plain and

FIGURE 4. Center for Advanced Judaic Studies Cairo Genizah fragment Halper 109 [Early talmudic gloss on Yevamot 22b-23b] (see Figure 2 for Bibliographic description).

marked-up text format. Traditionally, because of technical considerations, manuscript texts of this type have been transcribed or transliterated into Latin characters, sometimes with the addition of diacritics. However, with the implementation of Unicode, the potential to encode non-Roman language characteristics in texts is greatly improved. Expanded uses of Unicode are also being explored to enable the display of the diverse scripts and character sets of the Genizah authority records for headings generated by the database.

CONCLUSION

This has been a brief overview of the elements that are to be used in describing the digitized fragments. Of course other types of information or metadata are required to "define" or encapsulate a digital collection. Producers and end-users need to have recorded such things as digitization information, hardware and software requirements, and so on. The team from the University of Pennsylvania Library and Tay-

lor-Schechter Genizah Research Institute includes librarians, software and application developers, and scholars. As project requirements and needs are examined and discussed; reviewing, testing, and refining of results will continue throughout.

The project leaders at the University of Pennsylvania Library and at the Taylor-Schechter Genizah Research Institute hope that the success of this project will encourage other institutions and repositories to join our efforts in digitizing and providing dynamic cataloging for the Genizah fragments. This ultimately is our most economic means of re-unifying the scattered Genizah materials "under a virtual roof."

Received: September, 2004
Revised: June, 2005
Accepted: June, 2005

NOTES

1. Penn/Cambridge Genizah Fragment website see: http://sceti.library.upenn.edu/genizah [accessed Jan. 12, 2005].
2. "Genizah." *Encyclopedia Judaica*, c1972.
3. See under "Cairo." *Encyclopedia Judaica*, c1972.
4. Benjamin Richler, *Hebrew Manuscripts: a Treasured Legacy* (Cleveland: Ofeq Institute, 1990), p. 116-118.
5. Stefan C. Reif, *A Jewish Archive From Old Cairo* (Richmond, Surrey: Curzon, 2000), p. 4.
6. "Nagid." *Encyclopedia Judaica*, c1972.
7. Two Jewish sects co-existed in the Fustat community: the Karaites rejected the rabbinic tradition and laws followed by the Rabbanite Jews. See "Karaism." Encyclopedia Judaica, c1972.
8. Richler, p. 116.
9. From the JTSA website see: http://www.jtsa.edu/library/about/specialcoll.shtml#manu [accessed Jan. 12, 2005].
10. Richler, p. 63-65.
11. Adolf Neubauer [ed.], *Catalogue of the Hebrew Manuscripts in the Bodleian Library and in the College Libraries of Oxford* (Oxford: Clarendon, 1886-1906).
12. *Catalogue of Hebrew Manuscripts in the Collection of Elkan Nathan Adler* (Cambridge, England: The University Press, 1921).
13. Richard Gottheil, William H. Worrell [eds.], *Fragments from the Cairo Genizah in the Freer Collection* (New York: Macmillan, 1927).
14. Benzion Halper*, Descriptive Catalogue of Genizah Fragments in Philadelphia* (Philadelphia: Dropsie College for Hebrew and Cognate Learning, 1924).
15. Agnes Smith Lewis, Margaret Gibson [eds.], *Palestinian Syriac Texts From Palimpsest Fragments in the Taylor-Schechter Collection* (London: C.J. Clay, 1900).

16. Neil Danzig, *Catalog of Fragments of Halakhah and Midrash From the Cairo Genizah in the Elkan Nathan Adler Collection of the Library of the Jewish Theological Seminary of America* (New York: Jewish Theological Seminary of America, 1997).

17. Solomon Schechter, *Facsimiles of the Fragments Hitherto Recovered of the Book of Ecclesiasticus* (Oxford: University Press, 1901).

18. It was later discovered that the E. N. Adler collection contained a fragment from a fifth codex (ENA 3597, as well as a fragment from a Hebrew paraphrase of Ben Sira (ENA 3053). These were published in: Joseph Marcus [ed.], *The Newly Discovered Original Hebrew of Ben Sira (Ecclesiasticus XXXII, 16-XXXIV, 1), the Fifth Manuscript, and A Prosodic Version of Ben Sira (Ecclesiasticus XXII, 22-XXIII, 9)* (Philadelphia: The Dropsie College for Hebrew and Cognate Learning, 1931).

19. "Targum." *Encyclopedia Judaica*, c1972.

20. *Index of References Dealing with Talmudic Literature [CD-ROM]* (New York: The Saul Lieberman Institute for Talmudic Research, Jewish Theological Seminary of America, 2003).

21. For past overviews see: Jay Rovner, "The Computerized Genizah Cataloguing Project of the Jewish Theological Seminary of America; its History, Current Status, and Future Prospects, With Some General Considerations of Bibliographic Control of Genizah Fragments," *Shofar* 8, 4 (1990): 37-58; Robert Brody, "Cataloging the Cairo Genizah," *Judaica Librarianship* 10, 1-2 (1999-2000): 29-30; Ezra Chwat, "Danzig's Catalog of Halakhah and Midrash Fragments in the E.N. Adler Collection and its Usage as a Research Tool," *Jewish Quarterly Review* 90, 3-4 (2000): 405-415.

22. David Solomon Sassoon, *(Ohel Dawid) Descriptive Catalogue of the Hebrew and Samaritan Manuscripts in the Sassoon Library, London* (London: Oxford University Press, 1932).

23. For the purposes of this paper, see the following definition: "The Dublin Core concentrates on describing intrinsic properties of the object. Intrinsic data refer to the properties of the work that could be discovered by having the work in hand, such as its intellectual content and physical form. This is distinguished from extrinsic data, which describe the context in which the work is used. For example, the "Subject" element is intrinsic data, while transaction information such as cost and access considerations are extrinsic data. The focus on intrinsic data in no way demeans the importance of other varieties of data, but simply reflects the need to keep the scope of deliberations narrowly focused." Dublin Core' is shorthand for the Dublin Metadata Core Element Set which is a core list of metadata elements agreed at the OCLC/NCSA Metadata Workshop in March 1995. The workshop report forms the documentation for the Dublin Core element set. (Stuart Weibel, Jean Miller, Ron Daniel, *OCLC/NCSA Metadata Workshop Report*. (OCLC, March 1995).

24. MARC 21 Format for Bibliographic Data, 1999 ed.

25. TEI website see: http://www.tei-c.org [accessed Jan. 12, 2005].

26. Digital Scriptorium website see: http://sunsite.berkeley.edu/Scriptorium [accessed Jan.12, 2005].

27. MASTER website see: http://www.cta.dmu.ac.uk/projects/master [accessed Jan.12, 2005].

28. *Anglo-American Cataloguing Rules. 2nd ed., 2002 rev.* (Ottawa: Canadian Library Association; Chicago: American Library Association, 2002).

29. Gregory A. Pass, *Descriptive Cataloging of Ancient, Medieval, Renaissance and Early Modern Manuscripts* (Chicago: Bibliographic Standards Committee, Rare Books and Manuscript Section, Association of College and Research Libraries, a Divi-

sion of the American Library Association, 2002) [not yet in print when template was developed].

30. Plural of *Piyyut* (from Greek *poietés*, poet; *poiésis*, poetry): postbiblical Hebrew liturgical poetry.

31. The development of the 16-bit Unicode character set has enabled the display of Hebrew and Arabic characters within the scope of this project. Vernacular transcriptions of Hebrew and Judeo-Arabic text appear within the records, most notably in the 500 fields. For the use of Unicode in MARC 21 see: *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media. CHARACTER SETS: Part 2. UCS/Unicode Environment*: http://www.loc.gov/marc/specifications/speccharucs.html [accessed Jan. 12, 2005].

32. Developed by Terry Reese: http://oregonstate.edu/~reeset/marcedit/html [accessed Jan. 12, 2005].

33. The MarcEdit release at this time could only handle 8 bit text; romanization symbols needed to be converted to MARC/ALA compliant characters. All Unicode Hebrew and Arabic characters were converted to 8-bit "junk." Thankfully, a more recent release of MarcEdit allows for the translation of 16-bit characters to UTF-8; this same release also includes utilities that convert MarcMaker files to XML according to Dublin Core and EAD criteria.

34. For MARC 21 specifications for Record Structure, Character Sets, and Exchange Media, see: http://www.loc.gov/marc/specifications/speccharmarc8.html [accessed Jan. 12, 2005].

35. Specifications for digitization and use of Dublin Core provided by Greg Bear (Manager, Schoenberg Center for Electronic Text & Image, University of Pennsylvania Library), E-mail message (July 16, 2004).

36. Cornell University Library MARC to Dublin Core Crosswalk see: http://metadata-wg.mannlib.cornell.edu/forum/2002-09-20/CUL_MARC_to_DC_Crosswalk.htm [accessed Jan. 12, 2005].

37. LizardTech company website see: http://www.lizardtech.com [accessed Jan. 12, 2005].

38. Princeton Geniza Project see: http://www.princeton.edu/~geniza [accessed Jan. 12, 2005].

39. S.D. Goitein, *A Mediterranean Society: the Jewish Communities of the Arab World as Portrayed in the Documents of the Cairo Geniza* (Berkeley: University of California Berkeley Press, 1967-1978).