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

Heidi G. Lerner  
*Stanford University*

Seth Jerchower  
*University of Pennsylvania, sethj@pobox.upenn.edu*

Follow this and additional works at: https://repository.upenn.edu/library_papers

**Recommended Citation**


Publisher URL: http://www.haworthpress.com/web/CCQ

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/library_papers/7  
For more information, please contact repository@pobox.upenn.edu.
The Penn/Cambridge Genizah Fragment Project: Issues in Description, Access, and Reunification

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 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

Disciplines
Archival Science

Comments
Publisher URL: http://www.haworthpress.com/web/CCQ
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
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. [Article copies available for a fee from The Haworth Document Delivery Service: 1-800-HAWORTH. E-mail address: <docdelivery@haworthpress.com> Website: <http://www.HaworthPress.com> © 2006 by The Haworth Press, Inc. All rights reserved.]

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 genizah2 (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 Kanīṣat al-Yerūšalmīyīn (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 8825). 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 nāgīd6 (in Islamic countries, the head of the Jewish community). Prior to this, the Kanīṣat al-Yerūshalmīyīn 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*[^1] [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[^2] (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*).[^3]

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*,[^4] 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.[^5] Neil Danzig’s 1997 catalog of Rabbinic fragments at the Library of the Jewish Theological Seminary of America,[^6] 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,[^7] 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) exhaus-tive bibliographies. The late Michael Klein’s catalogs of Palestinian Targumic fragments assess fragments within the context of the Targumim (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. 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**

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). 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. 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* 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), the Digital Scriptorium, or the European project known as Manuscript Access through Standards for Electronic Record (MASTER), 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), and Descriptive Cataloging of Ancient, Medieval, Renaissance, and Early Modern Manuscripts. 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 tannaitic, gaonic), etc.; bodies and genres of literary texts such as midrash, piyyutim, 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

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID:</td>
<td>NAFR200330971</td>
</tr>
<tr>
<td>VST:</td>
<td>2003-10-13</td>
</tr>
<tr>
<td>ST:</td>
<td>p</td>
</tr>
<tr>
<td>MS:</td>
<td>c</td>
</tr>
<tr>
<td>EL:</td>
<td>n</td>
</tr>
<tr>
<td>001</td>
<td>nr2003030971</td>
</tr>
<tr>
<td>003</td>
<td>DLC</td>
</tr>
<tr>
<td>005</td>
<td>20031011162018.0</td>
</tr>
<tr>
<td>008</td>
<td>030922n?·acannaabn----------?a-aaa----c</td>
</tr>
<tr>
<td>010</td>
<td>‡a nr2003030971 ‡z nr20030333038</td>
</tr>
<tr>
<td>040</td>
<td>‡a PU-CJS ‡b eng ‡c PU-CJS ‡d PU-CJS</td>
</tr>
<tr>
<td>100</td>
<td>‡a Shemaryah ben Aharon, ‡c ha-Kohen, ‡d 12th/13th cent.</td>
</tr>
<tr>
<td>400</td>
<td>‡a Shemariah ben Aaron, ‡c ha-Kohen, ‡d 12th/13th cent.</td>
</tr>
<tr>
<td>400</td>
<td>‡a Shemaryah, ‡c ha-Kohen, ‡d 12th/13th cent.</td>
</tr>
<tr>
<td>400</td>
<td>‡a Shemarya ben Aharon, ‡c ha-Kohen, ‡d 12th/13th cent.</td>
</tr>
<tr>
<td>670</td>
<td>‡a Collected liturgical poems in honor of the bridal room, 12th century - 13th century? ‡b (name not given)</td>
</tr>
<tr>
<td>670</td>
<td>‡a Ency. Judaica, c1972; ‡b v. 13, col. 595-596 (Shemariah b. Aaron ha-Kohen, Babylonia, 12th/13th cent.)</td>
</tr>
<tr>
<td>670</td>
<td>‡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])</td>
</tr>
<tr>
<td>670</td>
<td>‡a Institute of Microfilm Hebrew Manuscripts, online catalog, Sept. 23, 2003 ‡b (hdg.: Shemaryah ben Aharon ha-Kohen)</td>
</tr>
<tr>
<td>670</td>
<td>‡a Sefune shirah, c1967; ‡b p. 144 (Shemaryah ha-Kohen [also in acrostic]) p. 146, etc. (Shemarya ha-Kohen ben Aharon)</td>
</tr>
</tbody>
</table>
and guides used in cataloging may not have sufficient scholarly back-
ground 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 to-
gether 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, 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 in-
structed 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 indi-
vidual 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, 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 expe-
dited through global “find and replace” commands. The individual
MarcBreaker records were then uploaded into the University of Penn-
sylvania 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 cor-
rected MARC 21 bibliographic records were exported from Penn’s
Voyager database (Penn’s current version of Voyager only supports the
MARC-8/ALA character set 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 using a MARC to Dublin Core crosswalk model developed by Cornell University Library 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...
purposes, digital surrogates are converted from the original TIFF images using LizardTechs Multi-Resolution Seamless Image Database (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, is engaged in placing S. D. Goitein’s transcriptions online. Links could be made available from within the individual records to fully searchable, online transcriptions of the fragments in both plain and
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


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-XXIV, 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).


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.


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-
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].


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).


