

Marriage, cheese and pirates: Text-mining the Cairo Genizah

Ben Outhwaite
Cambridge University Library



The Ben Ezra Synagogue
Fustat, Egypt

A HOARD OF HEBREW MSS.

By DR. S. SCHECHTER.

The Genizah, to explore which was the object of my late travels in the east, is an old Jewish institution. The word is derived from the Hebrew verb "ganaz," and signifies treasure-house or hiding-place. When applied to books it means much the same thing as burial means in the case of men. When the spirit is gone, we put the corpse out of sight to protect it from abuse. In like manner, when the writing is worn out, we hide the book to preserve it from profanation. The contents of the book go up to heaven like the soul. "I see the parchment burning





Solomon Schechter at work in Cambridge University Library, 1898



MS. 24 VOL. 2
88-100

MS. 24 VOL. 2
88-100

MS. 24 VOL. 2
88-100

TS. Mac. 25 VOL. 1
1-94

TS. Mac. 25 VOL. 2
95-107

TS. Mac. 25 VOL. 2
95-107



T-5 AS R4 minutes

cm
in

לכתוב ואלו נמחוקתם ל' יוסף אל בן יוסף מני חתן משה
 דניאל מני בן ששון הקדוש להראותנו נביות רבות נתנו ביד משה
 והפיוני נבואה אנשים ונשים ונמכרנו הנה בקוסטנדינו ונראו
 בנביות הבורא ואמריו הקדוש עלינו ולא תהיה עלי לא בלתי
 אחרים כחיים מעשים אחרים אני לכתוב ואלו נמחוקתם ל' יוסף
 ג' י' מארע' נפרדנו מאתכם והשחתנו מארע' מארע' נמחוקתם ל' יוסף
 האומר נודונונו עלינו נמחוקתם ל' יוסף נבואה רבות
 אבות מועד והבת נשאה ל' ונשאה על בחור כהן ויפה בלוי ונשאה
 יש ל' עמי אולי נכח להשתחוותם ל' ונשאה על בחור כהן ויפה בלוי ונשאה
 היא ומכח חולמים אשר מנאנו ואמריו המרומה ממנו בלוי ונשאה
 הויאת מרע על בן חיל' ואי שח' מארע' עתה כמה אריות הודיענו
 ואנ' י' לתשובת שלום ל' ק' בלוי ממך והייתכן בזה עליך או הייתכן
 ואנ' בא המדרת' בנא אשר כתבת חלק העובדי עמי וילכו ה' י'
 יעבדו ועתה נראה ונבין אך הפעם עמי שמואל אם אתה חס ומחוס
 עלינו ויחשב עלינו מה שתעשה כי מכתובתע ייחזקת ונש' עזרת
 כול חכמות על יד שמואל אים נרומא שעשית לך חת מן ירוס' עזרת
 כידין שלום קוס אולי שובא ל' להיות הויסא ל' עמי אולי ונש' ונש'

שם שלום ממני אנו מעד ואנו יעד כיצד
לדאוג את כעך ולחזות את מליך
וקעלמסכע כהלוך לך המקום זה הוי בחן
קוד שא ברך הוי ודאפ פער רוב

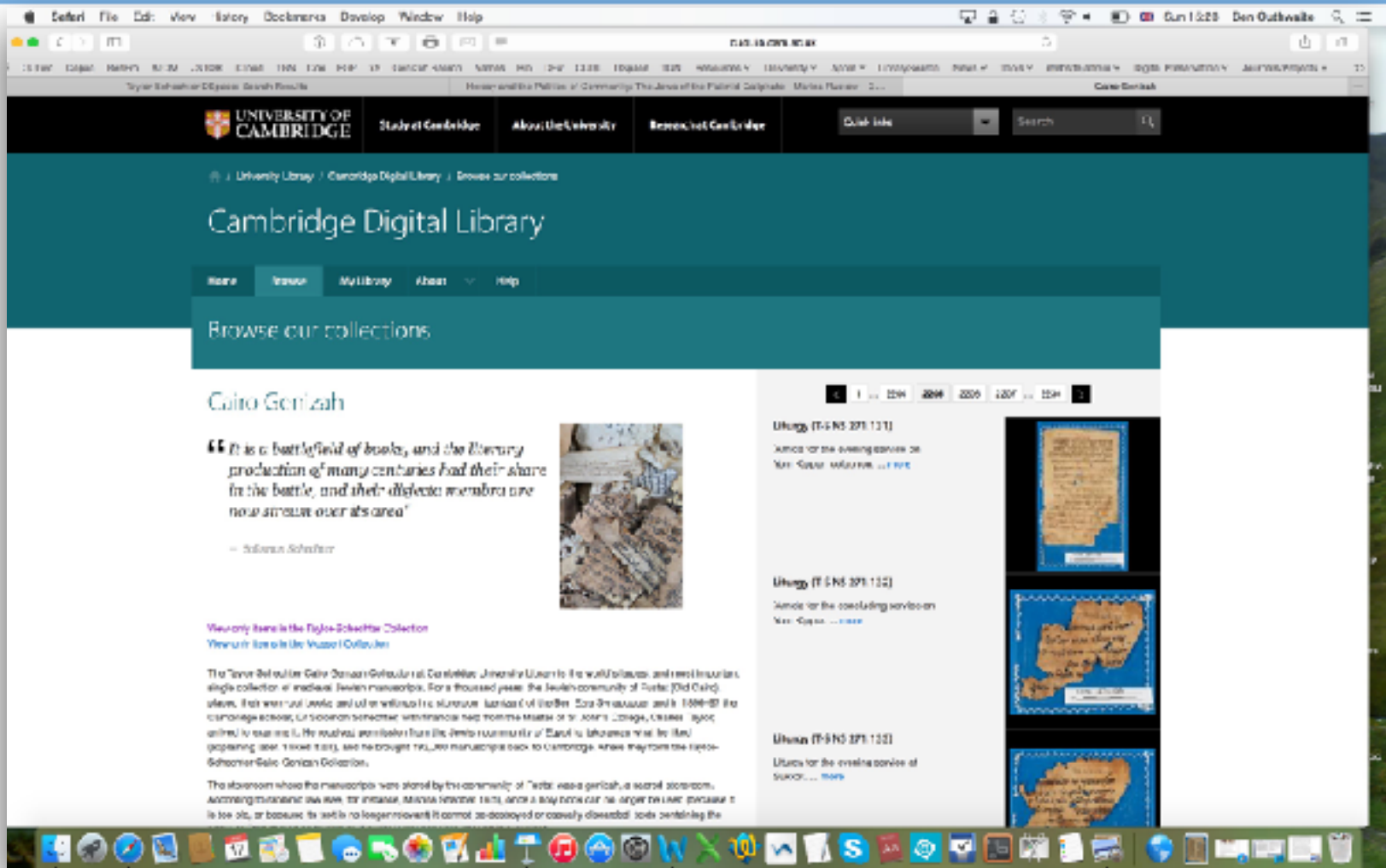
ולא התעבובי שאכוסב לך בכלשעה לאשמע
האוב ונינטערינו על אית דבר יאוד כן שמע
בחיוס כתבתי זה הכתב ואני בבקשה מן שוכי
וחדיע מה אתה עליו וכל צורך שוכי לך שלומך
ואנשי ביתי בשלום ואין בלבי דבר יכאיבני לא
מכאן ואודיעך שא שוקר בניע גדול עם
שוק והחזרה זה שלשוע שרים שנה ולא נכספתי

למה תעשה כח לא שתן ובצד ודע מה תעשה
בילדים שך ואמרו הודיעני בשבילי המים כג
ביום והשמים והלוח שנים עשר וטל בכסף
אלץ מי עינים ואמרו כי אנה הילך המלך
והשמים שנים עשר וטל בכסף

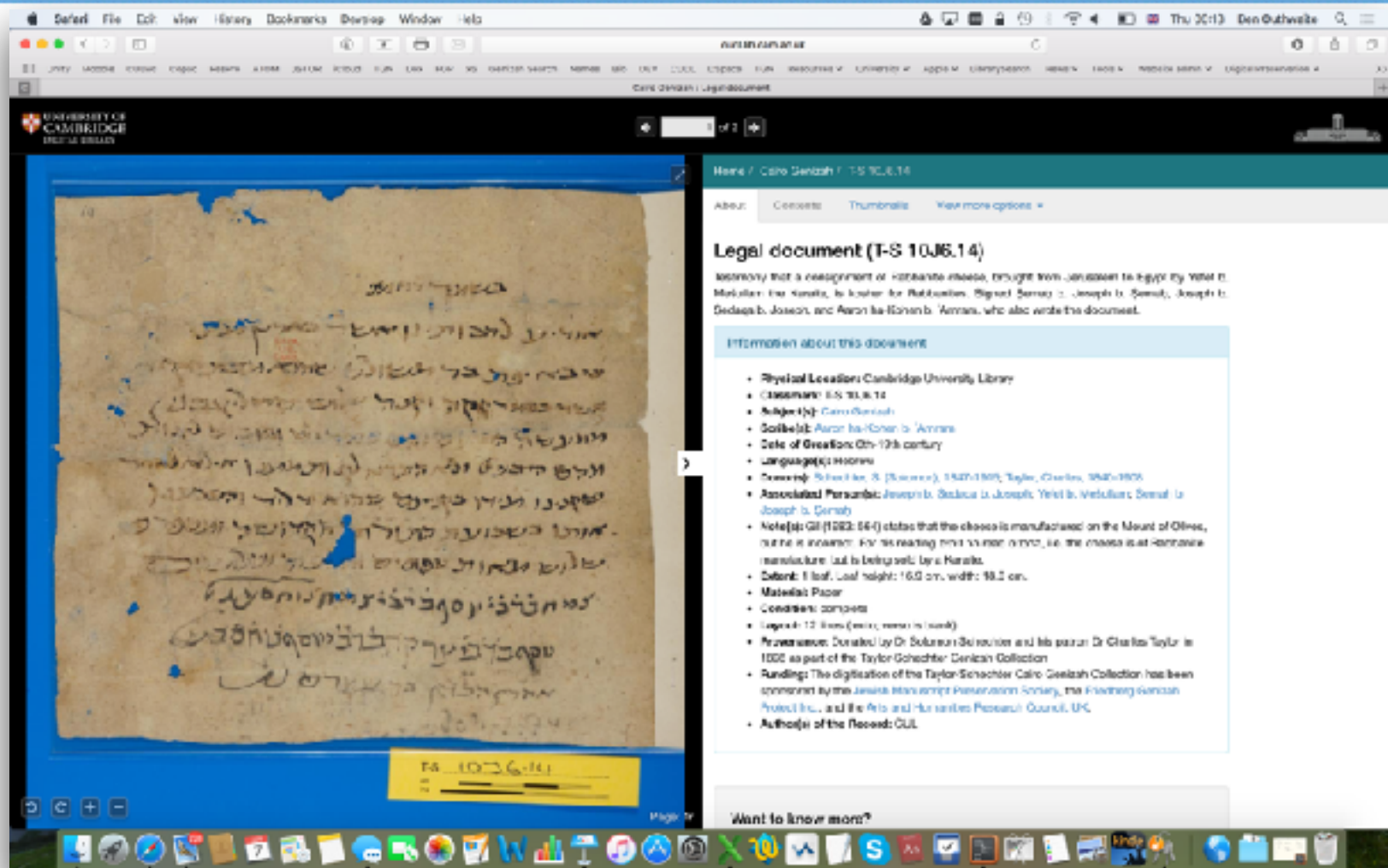
A research unit within the UL since 1974...



Cambridge Digital Library



Cambridge Digital Library



Documents

Images

Folio Genres

Libraries

Essays

People

Buildings

News Stories

Transcription

Index

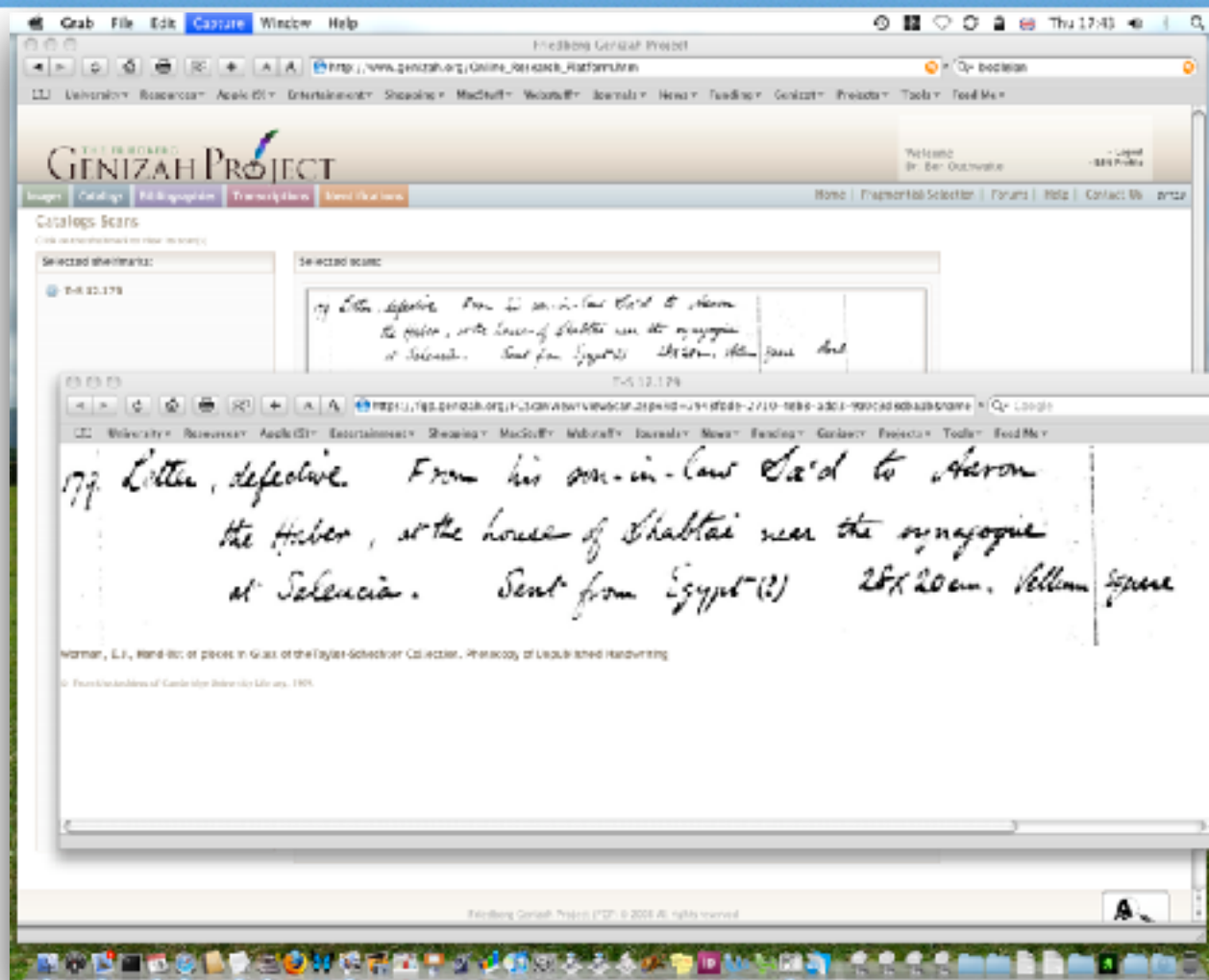
T-S 10 J 6.14

PDF

Alternate title: TS 10 J 6 14**Source:** Gil, Palestine, Pt. 2, p.584 (Doc. #309)**Attribution:** C.B. 12-09-87 (p)**Description:** Testimony signed by Aaron ha-Kohen b. Amram and two others regarding the kashruth (ritual lawfulness) of some cheese made by the Karaites (of Samariya), ca 1050.

בשמן החמץ	1
אודיע לרבותים אשר בארץ מצ[רים]	2
כי בא יפת בר משולם שהוא מכת הקראים	3
אשר במסרתקה וקנה שלושים רסלין גבמן	4
ממעשה הר דיתים והם כשרים וטובים לקנות	5
מהם הרבנים ולא התרנו לקנות ממם אילא לאחר	6
שקנים מיחד בקנינים שהוא יד ליד והשבעם	7
אותם בשבעת התורה הקדושה ומספרם	8
שלוש מאות ספוסים ות[ש]נה ושלושים	9
צמח ברבי יוסף ברבי צמח סחם ערן	10
יוסף ברבי צדקה ברבי יוסף סחם בע	11
אהרן הכהן בר עמרם נג	12

Making best use of legacy data



100 years of published scholarship

The screenshot shows a web browser window displaying the Cambridge Bibliography Editor interface. The header includes the University of Cambridge logo and the title "Bibliography Editor Cairo Geniza Collection". Below the header, there is a "Reference List" section. A search bar shows the fragment identifier "MS/TS/00010~00009-0014". Below the search bar, a table lists references referred to by the fragment. The table has columns for Reference Type, Reference Position, Reference, Title, Secondary Title, Year, and Volume.

Reference Type	Reference Position	Reference	Title	Secondary Title	Year	Volume
Full text	564	Gil, Moshe	Palestine during the first Muslim period 634-1099 (Hab.)		1993	2
Mention	370 (Index)	Rusakov, Marina	Heretic and the politics of community: the Jews of the Muslim Empire		2008	
Mention	32	Al-Buhārī, Ahmad	Ḍaḥḥā al-Fitana al-Fitana	3rd ed.	0	24
Mention	425	Gil, Moshe	634-1099: Palestine during the first Muslim period 634-1099	Cathaca	1998	1
Mention	500 (Index)	Gil, Moshe	Palestine during the first Muslim period 634-1099 (Hab.)		1993	2
Mention	424, 425	Gil, Moshe	A Mediterranean Society: The Jewish Communities of the Muslim Empire		1967	1
Mention	4, 345	Gil, Moshe	Palestine during the first Muslim period 634-1099		1998	
Mention	425	Khan, Yusef	The medieval Hebrew translations of Arabic into Arabic	Israel Oriental Series	1992	12

Text Mining the Cairo Genizah (*Manuscript Cultures 7*)

Article

In the Shadow of Goitein: Text Mining the Cairo Genizah

Christopher Stoker, Gabriele Ferrario, and Ben Duthwaite | Cambridge

Abstract

The widespread digitisation of manuscripts has brought about an era of unprecedented access to a range of important historical collections. However, the task of navigating these data associated with these online digital collections represents a significant barrier to those wishing to navigate them in order to identify manuscripts relevant to a particular research question or theme. We propose a novel solution to navigating based around text mining published editions, commentaries and other secondary literature in order to automatically generate a rich, semantic, electronic catalogue. This research explores a range of techniques from the fields of Information Retrieval (term-weighted vocabularies), Natural Language Processing (named entity recognition and Topic Analysis (topic models)). Our initial results demonstrate the potential for these approaches to produce significant volumes of descriptive metadata which, when evaluated in the context of retrieval effectiveness, provide suitable evidence on which to perform analysis and make discoveries. A search engine which recommends manuscripts based on the contents of our automatically derived catalogue achieves a Precision @ 10 of 0.54, which significantly beats a baseline strategy of random selection.

1. Introduction

The Taylor-Schlechter Genizah Collection at Cambridge University Library is the single most important collection of medieval Jewish manuscripts in the world.¹ As of June 2013, its 193,000 manuscripts have now been completely digitised and are in the process of being made available online as part of the Cambridge University Digital Library.²

¹ See <http://www.dlib.ox.ac.uk>.

² Cambridge University Digital Library (2014, URL: <http://cdl.lib.cam.ac.uk> accessed on March 4, 2014).

While mass digitisation has significantly improved access to the collection it is clear that discovery – the act of deriving resources or a particular manuscript – will remain a given information need – remains a key challenge. This is largely due to the sheer size of the collection coupled with the lack of any substantive metadata describing the content of individual manuscripts. The inability to navigate the collection by content presents a substantial roadblock to the diverse group of scholars looking to exploit this unique source of information spanning history since Maimonides and Near East. The following catalogue entry describes fragment T-S 24.64 (see Fig. 1), which aptly demonstrates the full extent of the problem:

T-S 24.64 – *Letter*

55 + 14, 74 lines + marginalia (Hebrew 7 + 1 line (verse))

Place: 1 Land, Tunis, Judaea, Arabia

A lengthy letter from Kalid b. Isaac to Abraham b. Yip, middle of 10th c.

Manual efforts to improve the quality of our catalogue by the Genizah Research Unit (GRU) are ongoing, but the scale (in terms of number of manuscript fragments) and complexity (manuscript condition, language constraints and required subject expertise) make the cost of full description, transcription and translation prohibitive. In light of this, we have elected to explore the potential for a technology-based solution.

2. Related work

Recent advances in technology have opened up the possibility of automatically deriving catalogue data from digital images of manuscripts. In particular, the Friedberg Genizah Project has experimented with extracting the shape, size and con-

³ Shulman, Chinitz, Wolf, and Dushkowitz 2013.

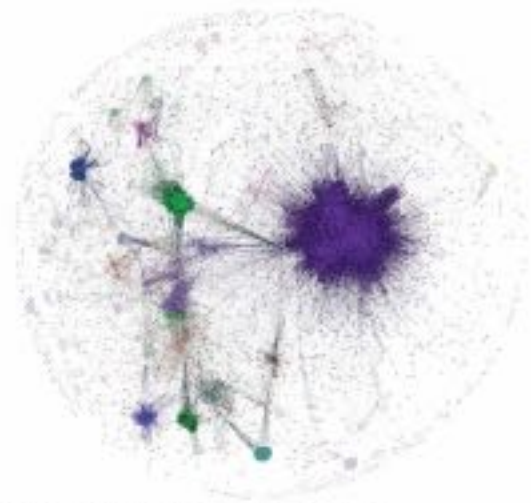


Fig. 2: Visualization showing Genizah fragments clustered based on co-occurrence in the secondary literature.

text the largest cluster of fragments belongs to a subset of the *Genizah* that the *Genizah* describes as documentary. These fragments consist of the everyday ephemera of life in the classical Genizah period, e.g. letters, accounts, *memoranda*, papers, *amot*, *depositions* and other top-level writings. Our analysis clearly shows that the largest single contribution to the literature, measured in terms of fragments discussed, comes from the writings of Shalom Dov Goitein, principally his six-volume work *A Mediterranean Society: the Jewish communities of the Arab world as portrayed in the documents of the Cairo Genizah*.³ In addition to the scholarship surrounding the Documentary Catalogue, there are several other distinct clusters that represent the breadth of Genizah studies, which include magic, literary works,

³ Friedberg 2010.

⁴ Friedberg 2010: 1991.

medicine, liturgy and religious law. In order to maximise our coverage of the collection, we have tried to target a cross section of works that encompasses all of these clusters.

If we consider the example of manuscript T-S 24.64, then our analysis of the bibliography identified 14 scholarly works that are known to have discussed this fragment to varying degrees. The full texts of seven of these were available to us for inclusion within our corpus.

3.2 Corpus construction

The process of building our corpus is ongoing, and scholarly works continue to be added and when we clear the rights. Automating new texts involves formal *biting* the source material into a machine-readable format (UTF-8) and then automatically segmenting the raw text according to its structure (e.g. page boundary, subsection, chapters). As of December 2013, our corpus contains 38 scholarly works by



The Mellon project: mining 100 years of publications

Genizah Fragment (Add.2586)

Information about this document

- **Physical Location:** Cambridge University Library
- **Classmark:** Add.2586
- **Subject(s):** [Cairo Genizah](#)
- **Date of Creation:** 6th-19th century
- **Author(s) of the Record:** CUL

The Mellon project: mining 100 years of publications

The screenshot displays the University of Cambridge Digital Library interface. On the left, a manuscript page is shown with Hebrew text and a map of the Middle East. On the right, a word cloud is displayed, featuring terms such as "marriage", "scholar", "clothing", "hezekiah", "abraham", "israel", "jewish", "evident", "preceding", "martaba", "sugar", "jewelry", "eleventh", "religious", "subject", "wished", "qinyan", "qadish", "trousers", "worn", "fatimid", "mosul", "type", "congregation", "descendants", "color", "father", "glass", "love", "yevamot", "importance", "accountance", "halla", "short", "imitated", "brought", "double", "sadana", "sonar", "type", "mosul", "marriage", "scholar", "clothing", "hezekiah", "abraham", "israel", "jewish", "evident", "preceding", "martaba", "sugar", "jewelry", "eleventh", "religious", "subject", "wished", "qinyan", "qadish", "trousers", "worn", "fatimid", "mosul", "type", "congregation", "descendants", "color", "father", "glass", "love", "yevamot", "importance", "accountance", "halla", "short", "imitated", "brought", "double", "sadana", "sonar". Below the word cloud, there are buttons for "Download" and "Print", and a section for "Export and Cite" with options for "Download" and "Print".

Rated tags


UNIVERSITY OF CAMBRIDGE
DIGITAL LIBRARY

1 of 2

Home / Cairo Geniza / TS 16.25

About Contents Thumbnails View more options

Keywords from text mining and user annotations



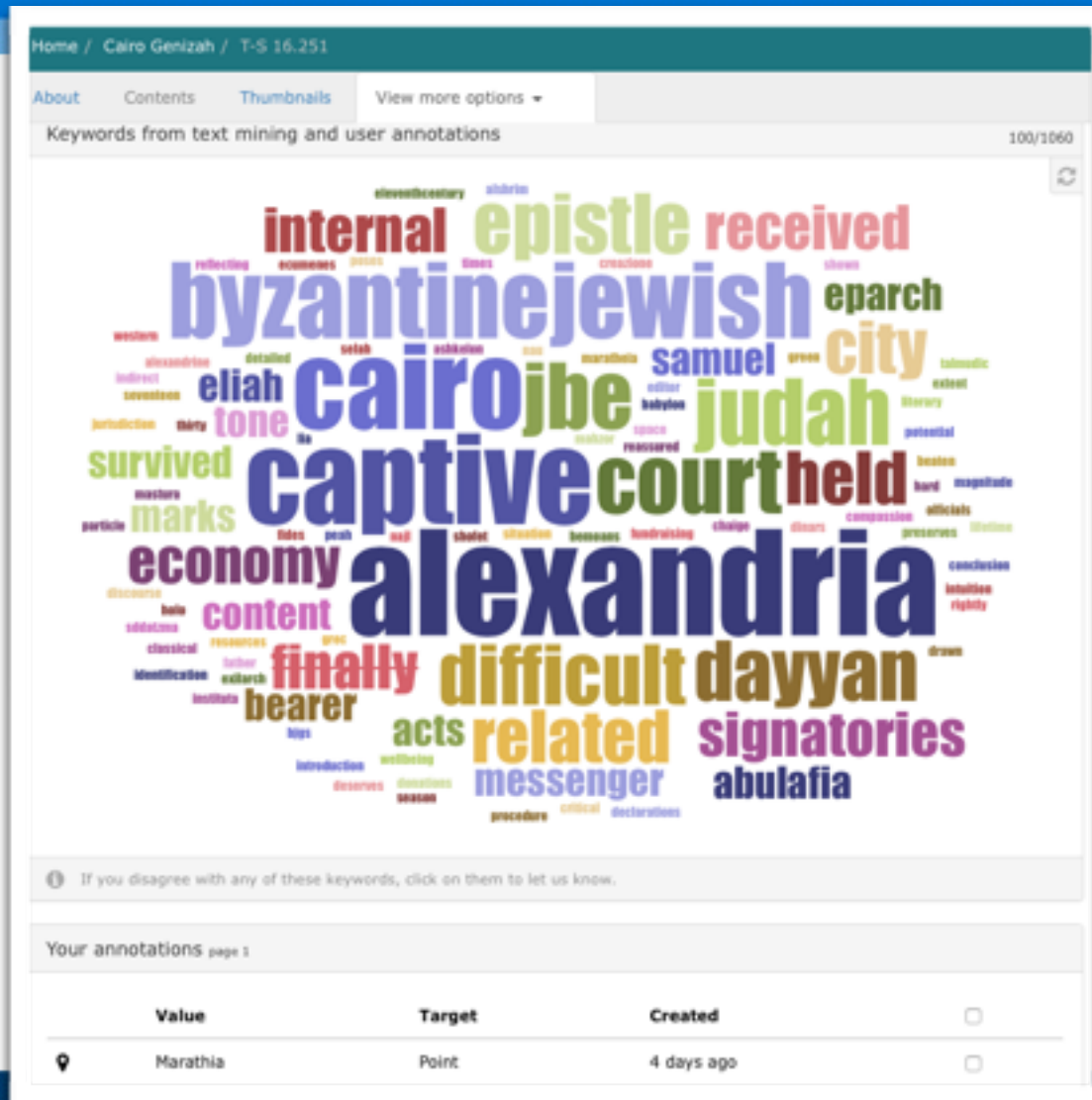
If you disagree with any of these keywords, click on them to let us know.

Your annotations page 1

Value	Target	Created
Marathia	Point	4 days ago

Page: 1r

Maturing tag cloud



Similar tags suggest related manuscripts

A screenshot of a web browser displaying the University of Cambridge Digital Library interface. The browser window shows the URL "sdf.ibe.cam.ac.uk" and the page title "Cairo Geniza: Order of payment; Babylonian Talmud". The page is titled "Cairo Geniza: Order of payment; Babylonian Talmud" and features a navigation bar with links: Home / Cairo Geniza / Mosseri L.6.2. Below the navigation bar, there are tabs for "About", "Contents", "Thumbnails", and "View more options". The main content area displays a large image of a manuscript fragment with Arabic script, labeled "MOSSERI L.6.2". To the right of the main image, there is a section titled "Similar items" which shows a grid of circular thumbnails of other manuscript fragments. A tooltip is visible over one of the thumbnails, titled "Letter and order of payment", with the text: "Kedar: Order of payment. Text opens after basmala. The text is in Arabic. Close with date (most of which is not) and signum characteristic of official orders of payment and receipts. *emo". The browser's address bar and various icons are visible at the top of the window.

User-derived data



Searching different qualities of data

Search the collection

Thanks to generous funding from the Andrew W. Mellon Foundation, over 6000 documentary Genizah manuscripts (e.g. letters and legal documents) have been associated with key terms - such as 'cheese', 'pirates', or 'gambling' - as well as names, dates and places drawn from over 100 years of published scholarship on the Collection. These associated terms provide an entry point into the Collection for those unsure of how to find documents of interest to their field of study.

Curated
metadata

Secondary
literature

Crowd-
sourced



- We have 310,000 images, but there is no catalogue of the Cairo Genizah Collection in Cambridge
- There is a large amount of legacy data of varying quality
- The size dictates that this will be a long-running project, and therefore we need a pragmatic approach to creating and sustaining the resource
- The aim is to put the best possible image in front of the person most qualified to assess it: **we should be helping people find things, not reading them for them**
- <http://www.lib.cam.ac.uk/collections/departments/taylor-schechter-genizah-research-unit/projects/discovering-history>