Journal of Open Humanities Data

Esten E 2020 Digital Second Edition of Judaica Americana: A Bibliography of Publications to 1900. *Journal of Open Humanities Data,* 6: 4. DOI: https://doi.org/10.5334/johd.15

DATA PAPER

Digital Second Edition of Judaica Americana: A Bibliography of Publications to 1900

Emily Esten

Data curation, Methodology, University of Pennsylvania Libraries, US

estenemily@gmail.com

This dataset, extracted from Robert Singerman's 1990 publication of *Judaica Americana*, contains bibliographic data of publications before the year 1900 pertaining to Jewish people, Judaism, and Jewish culture published in the United States, in any language. The data is primarily stored on the research data repository site ScholarlyCommons at the University of Pennsylvania, and it powers an Omeka site of the same name. Reuse potential of the data includes its use as a reference tool for early American Jewish history and history of the book, historical bibliometric data, book trade documentation, and as a support for the extant study of digitization efforts in research libraries.

Keywords: Bibliographic data; Judaism; Publications; USA

(1) Overview

Repository location

https://repository.upenn.edu/judaica_americana/2/.

Context

This dataset was produced as part of the creation of the Digital Second Edition of Judaica Americana, a pilot project within the framework of the Penn Libraries' Judaica Digital Humanities program [1].

This dataset is derived from the 1,500+ page PDF file of the second edition of Robert Singerman's *Judaica Americana*, the copyright of which was donated to Penn Libraries in 2019 [2]. In it, Singerman identified 6,500+ monographic and serial publications, and an additional 3,000 entries in the supplement, presented with bibliographical descriptions, classification explanations, and holdings information. Supplemental entries to the second edition were identified by Singerman through digitized, full-text books accessible from freely available online services like HathiTrust and Internet Archive, auction sites, and online marketplaces for booksellers.

The first edition of *Judaica Americana* was notable for several differences from its predecessors in the bibliographies of American Jewish literature. It extended bibliographic listings to 1900; it included location symbols for all known copies of a work; it identified each work with references to its listing in early Jewish and standard classic American reference works; it included transliterated titles; and its extensive index allowed for search not only of author, title, and broad subject areas, but of place of publication and publisher [3].

(2) Methods

Steps

Using the document version of *Judaica Americana*, I developed a Python script, "extract_singerman.py," to create a dictionary with the key as the Singerman ID number (or an assigned ID number if it was a new entry since the print edition) with the following values: year of publication, the full entry as listed in Singerman, notes, and holdings information [4]. I further extracted the author/editor of an entry, title, place of publication, printer/publisher, and asterisks through a combination of manual identification and OpenRefine, an open source desktop application for data cleanup, wrangling, and transformation activities.

With a scanned PDF of the print edition's index, I used a Python script developed by Digital Scholarship Librarian Jonathan Scott Enderle to perform Optical Character Recognition (OCR) via tesseract, an open source text recognition Engine [5]. This created three separate text files with each line displaying a heading followed by a list of locators (Singerman ID numbers). After extensive cleaning, a final Python script, "flip-index-headers.py," created another dictionary, with each locator as a key and headings as values [6]. This dictionary was written into a separate CSV and then linked with corresponding entries in the dataset based on the Singerman ID number.

Sampling strategy

Subject matter, not author ancestry, is the determining factor for inclusion in *Judaica Americana* and the dataset. The basic premise was to "include all separately

printed works issued under Jewish auspices including works issued under non-Jewish auspices relating to the Jewish people and their culture, from antiquity to modern times" [3].

Quality Control

To normalize the location and author data post-extraction, I used several clustering methods in OpenRefine. The dataset matches the document as exactly, where spelling variations exist. For example: "Philadelphia. Congregation Beth Israel" and "Philadelphia. Beth Israel Congregation" are considered unique authors. Assumed authors, denoted with square brackets, are considered unique from known authors, such as "[Scott, Walter, Sir]" and "Scott, Walter, Sir."

(3) Dataset description

Object name

"Dataset for Judaica Americana: A Bibliography of Publications to 1900".

Format names and versions

CSV

Creation dates

Start date: 2019-08-01; end date: 2019-10-22

Dataset Creators

Emily Esten
Roles: Data curation, Methodology, Writing (review and editing)
Affiliation: Judaica Digital Humanities Project Coordinator, University of Pennsylvania

Arthur Mitchell Fraas
Roles: Supervision
Affiliation: Senior Curator, Special Collections, University of Pennsylvania

Arthur Kiron
Roles: Conceptualization, Project administration, Writing (review and editing)
Affiliation: Schottenstein-Jesselson Curator of Judaica Collections, University of Pennsylvania

William Noel
Roles: Supervision
Affiliation: Associate Vice Provost for External Partnerships, Director of Kislak Special Collections Center for Rare Books & Manuscripts and the Schoenberg Institute for Manuscript Studies, University of Pennsylvania

Robert Singerman
Roles: Conceptualization, Data curation, Writing (original draft)
Affiliation: Emeritus University Librarian, University of Florida

Language

The dataset contains eight headers, as follows:

- **pid:** The Singerman ID number as assigned in *Judaica Americana* or with supp[0000] if part of the supplement.
- **asterisk:** Rows noted by means of asterisk mean the compiler has not seen all of the monographic items presented in this row under holdings.
- **year:** Year of publication. Years ending with a question mark are estimated by the compiler.
- **entry:** The full entry as listed in *Judaica Americana.*
- **author_editor:** Last-Name, First-Name OR Institution OR Company, when known.
- **location:** Geographic location of the **printer_publisher**, formatted as city, state abbreviation (e.g. Philadelphia, PA).
- **holdings:** A selected list of library symbols can be found on pages i-iv of *Judaica Americana.* Those not included are standard ones utilized by the National Union Catalog, maintained by the Library of Congress.
- **title:** Title of the publication as listed in *Judaica Americana.*
- **printer_publisher:** The printer or publisher, formatted as First-Name Last-Name OR Institution OR Company, when known.
- **notes:** Any additional information associated with the entry, typically regarding additional editions of an entry.
- **index:** The headers associated with this Singerman ID number according to the print edition.

The majority of the dataset is written in English. In the **title** field, the dataset includes Aramaic, Danish, English, French, German, Hebrew, Judeo-German, Norwegian, Spanish, and Yiddish text. Transliteration for Hebrew, Greek, and Russian is enclosed in brackets and follows the *ALA-LC Romanization Tables: Transliteration Schemes for Non-Roman Scripts* approved by the Library of Congress and the American Library Association. Transliteration for Yiddish is enclosed in brackets and follows the Weinrich/YIVO system. With the exception of the Hebrew *chet* (represented by "ח"), diacritical marks used for the Romanization of letters in the Hebrew alphabet have been ignored [3].

License

Repository name

University of Pennsylvania ScholarlyCommons https://repository.upenn.edu/.

Publication date

First published to the repository on 2019-10-22.

(4) Reuse potential

This data has reuse potential for scholarly research within studies of early American Jewish history and nineteenth-century American history in general. Like its physical counterpart, the chronological focus of the data allows for mining, clustering, and historical

context of texts. Its broad subject allows for exploration of Jewish communal activity through society and institutional constitutions, as well as non-Jewish attitudes towards Jewish people during this period, as Jewish communities spread with the territorial expansion of the United States in the second half of the nineteenth century.

For book historians, this dataset offers historical bibliometrics to view trends in book data, identifying macroscopic trends in the nineteenth-century book market. It will allow researchers to identify and visualize the connections among various printers, publishers, and places: a key area of study in Hebrew press history.

This dataset also has the potential for use by those involved in the book trade. Users can utilize the dataset as a reference record for extant copies and subsequent editions. One can also quickly generate counts of monograph holdings at various libraries. Furthermore, archivists and librarians can provide more complete descriptions of existing archival materials and books across private and institutional collections.

This dataset also has the potential for augmentation and expansion, being undertaken in the development of the project [7]. For example, we have linked entries listed in Yosef Goldman's *Hebrew Printing in America, 1735–1926: A History and Annotated Bibliography.* Goldman's bibliography contains overlapping information with Singerman's bibliography for these entries, but is unique for its inclusion of reproductions of many of the title pages, brief content and author biographical notes and vernacular Hebrew-character titles. The conception and design of this dataset will allow for easy reference and research between the two seminal bibliographic works, and for researchers to identify texts using Hebrew and English characters.

Virtually all of these texts are no longer restricted by US copyright, and many of them have been scanned for digital preservation by libraries. We are expanding the dataset by including links to digital facsimiles of the publications included in *Judaica Americana*, and intend to incorporate PDF files for integrated full text search and discovery across the corpus. For the development of the Digital Second Edition, we have added links to holdings in WorldCat, HathiTrust, GoogleBooks, and Penn Libraries. This will allow Singerman's bibliography to serve an additional purpose as an all-inclusive digital library for full-text searchable references: an annotated bibliography for the twenty-first century.

Finally, this dataset is complemented by the Union List of Nineteenth-Century Serials included in *Judaica Americana* in the same repository. Singerman's attempt was the first to arrange this multilingual collection of serials and their post-1900 issue publications in one bibliography [8]. It models the same data process as described above [9]. We are working to collaborate with The Ohio State University's project, "Union List of Digitized Jewish Historic Newspapers, Periodicals and e-Journals" in a mutual partnership to identify and provide access to digital facsimiles [10].

Additional File

The additional file for this article can be found as follows:

- ***Judaica Americana* Header Explanations.** This Markdown file contains background information on the dataset, including explanations for and descriptions of the headers within the dataset. DOI: https://doi.org/10.5334/johd.15.s1

Acknowledgements

Many thanks to the distinguished scholars, librarians, and staff who generously provided guidance with the project, including Nicky Agate, Laurie Allen, Michelle Chesner, Chris Clement, Camille Davis, Laura Newman Eckstein, Doug Emery, Jonathan Scott Enderle, Katherine Lynch, Emily Morton Owens, William Noel, Jordan Rothschild, and Kenny Whitebloom.

Competing Interests

The author has no competing interests to declare.

References

1. **Esten E, Singerman R, Kiron A, Fraas AM.** Digital Second Edition of Judaica Americana [Internet]. *Digital Second Edition of Judaica Americana*. Penn Libraries; 2020 [cited 2020 Apr17]. Available from: https://singermanja2.exhibits.library.upenn.edu/.
2. **Kaganoff N.** Judaica Americana: A Bibliography of Publications to 1900 by Robert Singerman. *American Jewish History*. 1990; 80(1): 136–141.
3. **Introduction.** In: *Judaica Americana: A Bibliography of Publications to 1900* [Internet]. 2nd ed. [cited 2020 Apr 17]. Available from: https://repository.upenn.edu/judaica_americana/1/.
4. **Esten E.** extract_singerman.py. 2020. Available from: https://github.com/judaicadh/ja2-scripts. DOI: https://doi.org/10.5281/zenodo.3894691
5. **Enderle JS.** tess.py. 2020. Available from: https://github.com/judaicadh/ja2-scripts. DOI: https://doi.org/10.5281/zenodo.3894691
6. **Esten E.** flip-index-headers.py. 2020. Available from: https://github.com/judaicadh/ja2-scripts. DOI: https://doi.org/10.5281/zenodo.3894691
7. **Esten E, Singerman R, Kiron A, Fraas AM.** Digital Second Edition of Judaica Americana [Internet]. *Digital Second Edition of Judaica Americana*. Penn Libraries; 2020 [cited 2020 Apr17]. Available from: https://singermanja2.exhibits.library.upenn.edu/.
8. **Esten E.** Union List of Nineteenth-Century Jewish Serials Published in the United States [Internet]. [cited 2020 May 27]. Available from: https://repository.upenn.edu/judaica_americana/3/.
9. **Esten E.** extract-singerman-serials.py. 2020. Available from: https://github.com/judaicadh/ja2-scripts. DOI: https://doi.org/10.5281/zenodo.3894691
10. **Galron-Goldshcläger J(Y).** Union List of Digitized Jewish Historic Newspapers, Periodicals, and e-Journals [Internet]. The Ohio State University; 2020 [cited 2020 Apr 17]. Available from: https://library.osu.edu/projects/hebrew-lexicon/Jewish-Press.htm.

How to cite this article: Esten E 2020 Digital Second Edition of Judaica Americana: A Bibliography of Publications to 1900. *Journal of Open Humanities Data* 6: 4. DOI: https://doi.org/10.5334/johd.15

Published: 02 July 2020

]u[*Journal of Open Humanities Data* is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS