JS/DH: Primary Sources and Open Data

Michelle Chesner
Norman E. Alexander Librarian for Jewish Studies, Columbia University Libraries

Librarians have been in the business of describing information for as long as the field has existed. Although the description can be different depending on the format of the item being described, most professionals working with digitized sources are familiar with the term "metadata." The digital shift has heightened the importance of data, meta- or otherwise, as a source for research in its own right. Vendors realized this immediately, and there are many "data packages" available for purchase or subscription, in areas ranging from agriculture to zoology. "Open data" is a newer concept, one has gained importance in the digital world in recent years. Open data means that datasets are shared freely, accessible to anyone with the ability to download a file, regardless of affiliation or research budget.

Some examples of Judaic applications that make use of open data are the Places of German-Jewish History, which uses open data from Wikidata and the European Holocaust Research Infrastructure (EHRI Project), and the New York Historical Synagogues Map site at the Center for Jewish history, which utilizes Open Street Maps and openly accessible maps at the New York Public Library (NYPL Map Warper). When Sefaria initially began, it was a crowdsourcing project that utilized data (i.e. texts and translations) that were already freely available or were submitted by users. Sefaria has now come full circle: its staff works directly with copyright holders to make more texts freely available. The now-accessible data can then be reused by Wikisource and other projects. The Open Siddur Project collects liturgical texts from all denominations to make them available for use and reuse by rabbis, congregants, and scholars.

When it comes to digitized manuscripts at a library, open data means providing access to all of the metadata—both for the document itself and to its digital surrogate. At the simplest level, this means that the images themselves are available for download in high resolution, for use and reuse with few or no limitations. More and more repositories are doing this, which has promoted the use of their materials in ways that may not have been previously imagined. While many institutions have digitized their manuscripts and placed them online, there is still a small number that have placed severe restrictions (and/ or fees) on their images in order to discourage reuse without permission, in some cases even adding watermarks to the images.[1]

However, more and more institutions are embracing open data and encourage the use and reuse of their collections. Two are particularly notable in the context of Judaica manuscripts: University of Pennsylvania's Openn project, and the British Library's Hebrew Manuscripts and Open

[1] Librarian Sarah Werner addressed many of the issues with locked-down digitization in a 2015 blog post entitled "How to Destroy Special Collections with Social Media in Three Easy Steps: A Guide for Researchers and Librarians." The post addresses far more than simply the issue of locked-down digitized collections, and is a very useful discussion for any institution considering if and how to digitize and post their images online.

Data Services, The Polonsky Foundation Catalogue of Digitised Hebrew Manuscripts.

The OPenn project contains “complete sets of high-resolution archival images of manuscripts from the University of Pennsylvania Libraries and other institutions, along with machine-readable TEI P5 descriptions and technical metadata.”[2] While there are only four collections listed under the Curated Collections page (one of which is of Penn’s Cairo Geniza collection), the site contains many, many more collections, which is not immediately apparent. A search for the word “Hebrew,” for example, yielded nearly 900 results, including manuscripts not only from Penn’s collection, but also from another Judaica collection—the Rylands Library at the University of Manchester. Penn’s repository, thus, is open not only to its own materials, but also to those of other institutions looking to make their digitized manuscript data completely accessible. Other collections in the repository include the Philadelphia Area Consortium of Special Collections Libraries (PACSCL), which is a collaborative of 14 heritage institutions, and Columbia University, which is among the institutions partnering with OPenn on its Muslim World Manuscripts Initiative. OPenn’s inclusion of materials from collections other than its own means that access restrictions vary from completely public domain (CC0) to requiring attribution (CC-BY and CC-BY-SA); nevertheless, there are no fees for use.

The British Library allows free access to its manuscripts for non-commercial use, but the website’s Terms of Use note that the library might “charge fees for the use of digital files for editorial or commercial purposes and may partner with third-party vendors that offer products for sale using the digital images. The Library may also charge fees for digital files that are in high-resolution format suitable for professional print reproduction.”[3] In other words, the British Library’s collection could be used in for-cost vendor products, and a scholar wishing to publish a book with its images may still have to pay fees to do so.

Because the British Library has a dedicated site for Hebrew manuscripts, it also has the opportunity to go beyond a repository and become an educational tool that the broader OPenn cannot match. The British Library website provides access to the manuscripts while including valuable additional context, such as scholarly essays by experts in their various fields. These essays are particularly valuable to those who teach classes on Jewish manuscripts, as well as to those who might have an interest in a specific topic, such as medieval illumination or kabbalah.

Even though the project is not yet complete, a sizable mass of data is already available on the British Library’s site—if one knows where to look. There is no clear list of all of the digitized manuscripts in the collection, and even the Collection Items page only shows 41 “highlights” from the collection. Notwithstanding its title as a “Catalogue of Digitised Hebrew Manuscripts,” the site serves primarily as a curated exhibit of a small portion of the collection. The above-men-

[2] Openn website, accessed April 17, 2019, Readme page.

[3] “Terms of use of digitised Hebrew manuscripts,” accessed April 29, 2019, https://www.bl.uk/hebrew-manuscripts/terms-of-use.

tioned essays are quite useful to the general user, but an advanced user looking for the data and images of the rest of the manuscripts may have a more difficult time with their research. Most of the pertinent information about the whole collection is buried in the project's "About" page.[4] Once accessed, however, there is a wealth of data available. A spreadsheet of comprehensive cataloging information (including such details as conservation status) is available for download, as are TEI XML records and JPEGs of the manuscripts themselves (available in batches by call number as a .zip file). Rather than the usual "we don't give copyright status information" that many institutions use as their default statement, the British Library generously notes that the images in the dataset are out of copyright (although they do recommend that users read their "Ethical Terms of Use" guide[5] before reusing the data). Digitized manuscripts are searchable only through the British Library's general Digitised Manuscripts site, but that information is hard to find as well. It would be helpful if the homepage of the Hebrew manuscript site had additional links or clarification on how to access the materials.

When comparing the British Library and OPenn websites, the latter seems to be geared more toward the advanced user, while the British Library's site was easier for a general user to navigate. While every page of every book was downloadable individually in the highest quality, finding a way to download an entire book on Openn was impossible. The "Technical Readme" page on OPenn[6] included instructions for downloading packages of images with data, but it required specific software as well as command line knowledge. (Advanced users will be pleased with the FTP and RSYNC options for data retrieval.) The British Library, on the other hand, provides a pre-packaged zip file of sets of manuscript images with metadata in .csv format available for download, which is easy to use for a scholar who is simply searching for a manuscript and its accompanying data.

OPenn's "openness" to including outside collections and projects means that it must remain a repository and only a repository. It thus does not have the additional luxury of the British Library Hebrew Project (a unique site for a specific subset of collections) to add many essays and other additional "bells and whistles" that makes the British Library's site such a good resource for those beginning in manuscript studies. Overall, both are excellent resources for scholars doing work in manuscript studies, since they provide not only the facsimiles of the manuscripts themselves, but the important context that comes with the data relating to each item -- both physical and digital.

[4] "About The Polonsky Foundation Catalogue of Digitised Hebrew Manuscripts," accessed August 26, 2019, https://www.bl.uk/hebrew-manuscripts/about-the-project.

[5] "Ethical Terms of Use," accessed August 26, 2019, https://www.bl.uk/help/ethical-terms-of-use.

[6] "Technical Readme," accessed August 26, 2019, http://openn.library.upenn.edu/TechnicalReadMe.html.