Morgan Papers: Exploring the Correspondence of California's First Female Architect

Zach Vowell, Ethan Kusters, Luca Soares, Samuel Verkruyse, Joey Wilson and Foaad Khosmood
California Polytechnic State University

Introduction

Descriptive metadata and full-text transcripts have long been valued for their roles in powering search engines and faceted browsing. But as the morganpapers.org web application demonstrates, such textual data (both structured and unstructured) can be leveraged to build a variety of tools which provide deeper and broader insight than simple searching and browsing.

Figure 1, William Randoph Hearst and Julia Morgan (right), Bison Archives

The Robert E. Kennedy Library at Cal Poly recently completed digitization of a unique body of correspondence between architect Julia Morgan and William Randolph Hearst, carried out during the construction of what is now known as Hearst Castle. The structure is a masterpiece and the crown jewel of Morgan's illustrious career throughout California, where she worked as the state's first female licensed architect. The collection consists of over 2,500 letters, telegrams, notes, and other documents (totalling over 3,200 pages), spanning the years 1919-1941. The pieces were written in several places across the United States and overseas. As each piece of correspondence was digitized, it was ingested in the library's archival repository along with its MODS-based metadata (used by Library of Congress), and full-text transcripts for both typescripts and manuscripts.

Usability Challenge

Shortly after making the collection available through the libraries website, it became clear we had a promotion challenge. The digitized content was difficult to find. The subject matter was described as obscure and uninviting to most college audiences. The content as a data corpus [1] was obscured by the approximately 60,000 other digital objects residing in the online repository. To foreground the correspondence, the library proposed a companion web application specific to this special collection, that would provide targeted access to the correspondence text through search and browsing, but preserve original scans for viewing from the library resources. Further, an automated recommender system [2] was proposed which would leverage the repository's REST API to connect users to image resources related to the correspondence text. A particular challenge was how to associate the correspondence pieces with available images. The task was given to four Cal Poly computer engineering students as a Capstone project closely supervised by the library's digital archivist.

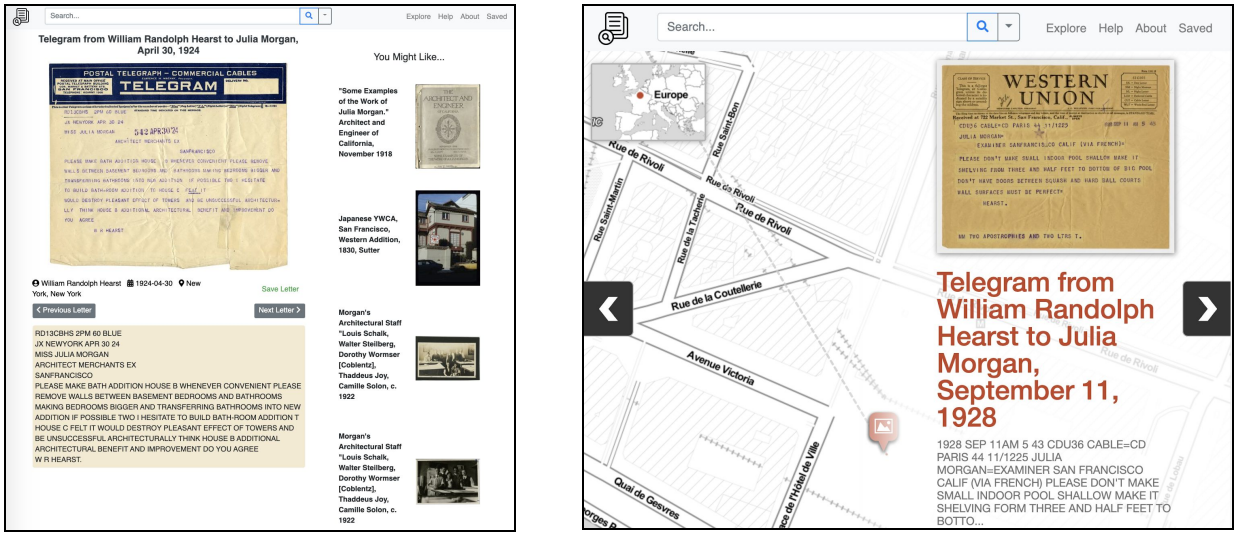

Figure 2, (Left image) Viewing a typical correspondence with the original document on top, transcribed text below, and recommended digital items on the right. (Right image) A 1928 Hearst telegram to Morgan sent from Paris, France.

Morganpapers.org

The website, now live at morganpapers.org, has 6 major components: 1) targeted advanced search of the Morgan Papers correspondence, 2) an automated recommender system which points to images related to a particular piece of correspondence, and which is powered by a few natural language processing (NLP) software packages, 3) an interactive map plotting each place where correspondence was written from, and which groups the correspondence by place, 4) an interactive timeline building tool, 5) curated groupings of letters based on author, buildings, and subject, and 6) a "letter of the day" feature which presents users with every piece

of correspondence written on the current day. It should also be noted that very little content is stored on morganpapers.org site; rather the system takes full advantage of the Islandora-based API to dynamically construct search results and item view pages, and will respond to any improvements to metadata, as well as new content additions made in digital.lib.calpoly.edu.

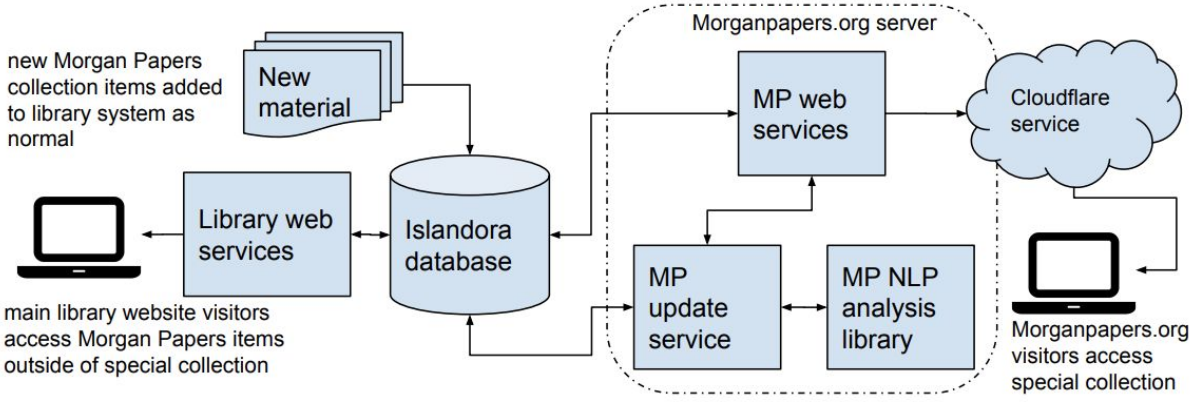

Figure 3, Morganpapers.org web architecture and its relationship to the Library web system. Arrows indicate flow of data. All source code is available at https://github.com/Castle-Computing .

The system is convenient for librarians, and offers some advantages for the visitor. It allows library staff to continue working with the familiar Islandora-based management system. New material is inserted into the system the traditional way. In the case of Julia Morgan papers, images and transcriptions are uploaded into Islandora as new material. Morganpapers.org then periodically accesses the information via API, seeds the search system, updates the timelines and interactive tools, and creates more immersive content pages, complete with related item recommendations.

The NLP-specific packages are NLTK and TextBlob (Python) and Stanford Core NLP (Java), but the site also employs several general purpose Python packages, such as Time Decorator, URLLib2, and ElementTree XML, to ensure the recommended content system is fully functional.

NLTK is a popular Python NLP library used to parse the contents of the correspondence into a list of phrases containing a list of words. For each noun, a TF-IDF [3] score is calculated within its respective piece of correspondence. The results are used both as indices for retrieval of contextual recommendation items to be displayed on the same page as the main content.

User Study

To assess the usability of the new site, a small preliminary user study was conducted. Personas and scenario-based tasks were developed, and given to five test subjects to attempt. Test subjects were recruited through the library's student advisory council, and included two women and two students of color. As members of the council, the test subjects were familiar with library resources in general, and provide advocacy for both library and the students that use the library, but they did not identify themselves as previous users of digital.lib.calpoly.edu, the Julia Morgan

Papers, or indeed any of the resources managed by the Special Collections and Archives department.

Figure 4, an example heatmap generated from user study indicating the most visited parts of the search results page by test subjects

Results from the preliminary testing, which employed analysis of mouse-click data, heat maps, and ethnographic observations, indicate that users were engaged with the added functionality, exploring subjects and areas with more enthusiasm compared to what had been observed from the plain-web version of the letters. User comments also indicate they need more context for things like inaccurate transcriptions and at-times confusing recommendations.

Table 1: Usage metrics by test subjects showing relative engagement ("friction" is a measure of frustration provided by Mouseflow software)

Display URL	Visit time (ms)	Clicks	User friction	Render time (ms)	Scroll (% of page)
/search	31902	989	0.26939654	801	57
/	15625	387	0.45528457	1722	97
/explore	16234	474	0.57425743	320	69
/cart	9808	91	0.40816328	815	89
/timeline	20685	113	0.33333334	166	100
/about	17880	23	0.21052632	386	87
/help	12587	24	0.4	284	73
/bibliography	16827	14	0.54545456	1028	100

References

1. Thomas, Will R., Benjamin Galewsky, Sandeep Puthanveetil Satheesan, Gregory Jansen, Richard Marciano, Shannon Bradley, Jong Lee, Luigi Marini, and Kenton McHenry. (2019) "Petabytes in Practice: Working with Collections as Data at Scale", Data and Information Management 3, 1: 18-25. https://doi.org/10.2478/dim-2019-0004

2. Konstan, J.A. & Riedl, J. (2012) “Recommender systems: from algorithms to user experience.” User Model User-Adap Inter 22: 101. https://doi.org/10.1007/s11257-011-9112-x

3. Peng, T. , Liu, L. and Zuo, W. (2014). “PU text classification enhanced by term frequency–inverse document frequency-improved weighting.” Concurrency Computat.: Pract. Exper., 26: 728-741. doi:10.1002/cpe.3040