Julie Carlson
INFO 628: Data Librarianship & Management
Final Project Check-In #1
October 27, 2021

Since 1966, the National Park Service has maintained the National Register of Historic Places, "the official list of the Nation's historic places worthy of preservation" (National Park Service, 2021). As of June 2021, this list contains more than 96,000 properties—data about these places can be found on the National Park Service's website. Although the data is freely available, there is a dearth of comprehensive analysis that examines the characteristics of listed properties. Where are they primarily located? What is the predominant category of the properties (e.g. building, district, site)? What "areas of significance" are the most represented? Using the National Park Services' dataset as my raw data, I will explore the National Register through quantitative analysis to consider what kind of places are deemed worthy of preservation.

The National Park Services' dataset is particularly suitable for quantitative analysis because of its large size. The dataset contains 96,644 rows and 21 fields; quantitative methods will assist me in "identifying broad patterns" from this vast collection of data (Lemercier & Zalc, 2019). After cleaning the dataset in OpenRefine, I will use Python scripts to perform statistical analyses. This work will result in .csv files, which I will use to visualize my findings with Tableau Public. Visualizing my results will help me explain my findings, but could also illuminate aspects of the data. As Lemercier and Zalc (2019) note, visualizations "give information on aggregate patterns […] and detect exceptions and 'outliers.'" Perhaps we'll learn that a certain state has far fewer properties than others have listed.

I also plan to perform quantitative text analysis using Voyant Tools to explore the National Register through a different lens. Certain fields in the dataset have a fixed set of values. For example, the "Category of Property" is only described as "Building," "District," "Object,"

"Site," or "Structure." While statistical analysis will allow me to count the instances of these particular property types, nuance about the properties is lost; a house, a school, and a post office are all reduced to a "building." The "Property Name" field contains more descriptive information that could be valuable to analyze, so I will use Voyant Tools and its built-in visualizations to see which terms appear most frequently there. Analyzing text in Voyant Tools has helped scholars "frame important terms through count, frequency, and relativity that ultimately gave way to new areas of desired inquiry" (Miller, 2018). Given that there has yet to be a thorough overview of the National Register of Historic Places, I hope that my methodology will both provide insight about the places and suggest future areas of inquiry.

References

Lemercier, C., & Zalc, C. (2019). *Quantitative Methods in the Humanities: An Introduction.* University of Virginia Press.

Miller, A. (2018). Text Mining Digital Humanities Projects: Assessing Content Analysis Capabilities of Voyant Tools. *Journal of Web Librarianship*, *12*(3), 169-197. https://doi-org.ezproxy.pratt.edu/10.1080/19322909.2018.1479673

National Park Service (2021, August 9). *National Register of Historic Places: Data Downloads.* https://www.nps.gov/subjects/nationalregister/data-downloads.htm

PLAN OVERVIEW

A Data Management Plan created using DMPTool

Title: Exploring the National Register of Historic Places

Creator: Julie Carlson

Affiliation: Pratt Institute (pratt.edu)

Project abstract:
Since 1966, the National Park Service has maintained the National Register of Historic Places, “the official list of the Nation's historic places worthy of preservation.” The National Register contains more than 96,000 properties as of June 2021. Although data on the properties is freely available as a dataset, there is a dearth of scholarship exploring the places' characteristics. Using quantitative analysis methods, I will examine the dataset to consider what kind of places are deemed worthy of preservation.

Start date: 09-17-2021

End date: 12-20-2021

Last modified: 10-27-2021

EXPLORING THE NATIONAL REGISTER OF HISTORIC PLACES

DATA COLLECTION

I will use the National Park Service's "Spreadsheet of NRHP Listed properties (listing up to 06/17/2021)" file as my raw data. This spreadsheet is freely available on the National Park Service's website in .xlsx format. The spreadsheet contains 96,644 rows, covering properties added to the National Register from its inception in 1966 through June 17, 2021. It is worth noting that the National Park Service maintains a separate spreadsheet of properties removed from the National Register, so my work will not be a comprehensive overview of every property that has ever been listed; I will only be exploring properties actively listed as of this project.

For my project, I will convert the .xslx file into .csv format to enable long-term access that does not rely on proprietary software. I plan to clean the data in OpenRefine and save my results as a separate .csv file. I will write and share Python scripts, which will be saved as .py files, to perform quantitative statistical analysis on the cleaned dataset. These Python scripts will produce additional .csv files, which I will use to create visualizations of my findings. I will create visualizations with the freely-available software Tableau Public. Additionally, I will use Voyant Tools—an open source application—to perform quantitative text analysis, resulting in additional visualizations generated by the application.

The project files will be organized by file type within one larger project folder. Per the structure standards discussed in class, raw data and metadata will be in a “data” folder, cleaned data and visualizations will be in a “results” folder, scripts will be in a “src” folder, and associated text documents will be in a “docs” folder. File names will be easily understandable, and will be prefixed with the date created in ISO 8601 format, as recommended by Kristin Briney.

DOCUMENTATION AND METADATA

Each folder will contain a README file with details about the contents. The “docs” folder will also contain a codebook with metadata about the raw dataset and the datasets that I create throughout the project.

ETHICS AND LEGAL COMPLIANCE

The raw data at the base of my project is government content in the public domain. My cleaned dataset and any visualizations I produce will remain in the public domain for free reuse.

STORAGE AND BACKUP

I will keep three copies of all materials associated with this project: one on my laptop, one on Google Drive, and one on an external hard drive. I will save my files to the hard drive each Thursday.

SELECTION AND PRESERVATION

The raw dataset, processed datasets, Python scripts, visualizations and all accompanying documentation will be saved on GitHub for the foreseeable future.

The spreadsheet containing my project's raw data is periodically updated on the National Park Service's website. For reproducibility purposes, the actual dataset I downloaded from their website (containing listings through 6/17/2021) would be particularly valuable to retain. Additionally, my documentation and Python scripts may be of use to future researchers. As the National Park Service updates its dataset, researchers could potentially run my script on the new spreadsheet and gain new, more timely insights.

DATA SHARING

All data, scripts, visualizations, and documentation will be shared on GitHub in open file formats, like .csv and .py. All data and analysis will be made available by the end of the project, December 20, 2021. There will be no restrictions on data sharing.

RESPONSIBILITIES AND RESOURCES

I will be responsible for all data management activities. I will take into account feedback received from my peers and Professor Vicky Rampin, and revise/implement the DMP accordingly.

My project will require freely available resources including OpenRefine, Visual Studio Code, Voyant Tools, and Tableau Public.