Dataset Documentation & Justification

Upon scraping the *College Women* website for each nodes' ID, FIELD, and VALUE elements on July 1, 2017, the text facets for the FIELD categories were the following:

- Cite This Item* - not a useful category for data visualization project
- Collection
- Collection Guides* - not enough to have a substantial amount of data
- Contributor
- Copyright and Use* - not a useful category for data visualization project
- Creator
- Date
- Description
- Format
- Institution
- Location
- Original URL* - not a useful category for data visualization project; web scraping removed the hyperlinks
- Physical Description
- Subject
- Themes
- Title

(FIELD categories marked with * are categories that were omitted from the dataset for all IDs.)

For the remaining FIELD categories, the following changes were made using OpenRefine:

- Collection
    - Minor clustering changes were made
- Contributor
    - No changes were made
- Creator
    - No changes were made
- Date
    - All dates are represented by year only. Any additional information (month, day, season) were omitted.
    - Time periods were changed to be represented by the starting year (ex: 1905-1910 as 1905). This will allow for a cleaner timeline.
    - Time periods that were estimates were represented by that year (ex: ~1905 or circa 1905 or c. 1905 as 1905).

        - For ID element 13678, the date given was “19—“. Upon further inspection of this node’s description on the College Women website, the date was found to be 1865 and corrected as such.
        - For ID elements 13584 and 13588, the date given was “undated”. Upon further inspection of these nodes’ descriptions on the College Women website, they were omitted. This will allow for a cleaner timeline.
    - Description
        - No changes were made
    - Format
        - Minor clustering changes were made so that all formats were placed into 4 types: Photograph, Correspondence, Diary, or Scrapbook. “Photo Albums” are a part of the Photograph category, and any type of letter or postcard is a part of the Correspondence category.
    - Institution
        - No changes were made. Note that Barnard College has no submissions to the College Women data set.
    - Location
        - Formatted into one of the following:
            - (City, State) for elements with given location in a specific city in the United States
            - (State) for elements with given location in a particular state
            - (City, Country) for elements with given location in a specific city outside of the United States
            - (Country) for elements with given location in a particular country
        - Vassar College’s letters’ locations were all “letters (correspondence),” and so the locations have been omitted
    - Physical Description
        - No changes were made
    - Subject
        - Minor clustering changes to standardize plurals and case were made
    - Themes
        - Minor clustering changes to standardize plurals and case were made
    - Title
        - No changes were made