## Issues encountered and changes made

### Geocoding Data

The team planned to augment centroid-based geocoding with street-level geocoding from Google’s geocoding web services, this in order to potentially produce more informative geospatial visualizations. Several issues were encountered:

1. street Addresses not always clean enough to pass through Google
2. street addresses that use non-standard designations (e.g. “barracks” rather than street)
3. street designations that Google silently passes to a centroid based on city, state, country

Addresses in these categories had to be manually flagged and manually inspected or modified. More generally, it’s not clear that street-level geocoding adds much to visualization (e.g. does a dot on South Denver inform more than a dot on centroid Denver on a small map). The team believes there is value in clean addresses in the overall dataset: the project curator might want a cleaned address scheme going forward: but this is not one of the deliverables asked for in the project scope.

### Gender Data

The team believed there was value, both as a learning exercise and to the overall dataset, of approaching gender identification of user names using a programmatic approach. The approach chosen was to take the Social Security Tables from 1880 – 1914 and use these to identify registration trends (this approach is mentioned in the client project list of questions). The tables were aggregated into a master table and sorted by occurrence based on gender. Issues encountered included:

1. ambigendrous names—names that can be either male or female
2. names predominant or occurring in other countries but not reflected in the US SSA table registrations

In the case of ambigendrous names, a programmatic mitigation is possible. The dominant count was used as the basis of projection: for example, in the case of the name ‘Billy’, the ambigedrous ratio is given by:

* Billy Male 381527
* Billy Female 5346

Based on historic registrations, ‘Billy’ was projected to be female in only 1.38% of occurrences. Therefore the code was designed to project the name as male (barring any obvious contraindications such as a title of ‘Mrs.’ There are other cases that are not as clear cut where manual intervention is necessary. For example, first name ‘Robin’ and middle name given as ‘Christopher’: in this case Robin indicates female but the addition of Christopher as a middle name clearly suggests male. These cases have to be flagged and mitigated manually—barring lots of programming logic to add middle names to the gender scoring process.

In the case of non-American names: these have to be mitigated manually based on Google searches of name origin, usage and meaning.

### Uniqueness Criteria

One aspect of analysis that came up in the midst of initial visualizations was the question of relating unique versus aggregated data points related to single writers. Many questions that one may attempt to visualize have an implicit assumption of uniqueness or non-uniqueness that must be addressed as part of the analysis. For example, in a geospatial visualization, if the same writer produces a significant number of letters from a single location, should that be part of the visualization or should that writer count as weight 1. As an extreme consider that one writer from a town in Asia wrote 100 letters—the visualization might lead someone to believe that a certain Series had a significant following in Asia, which would be misleading. Uniqueness and distinctness are important to understand and relate.