End of ACS Grant Report, February 2017
Project Title: Fighting Fever in the Caribbean: Medicine and Empire, 1650-1902
PI: Mariola Espinosa, Associate Professor of History; Matthew Butler, Senior Developer
HTRC Support: Eleanor Dickson, Boris Capitanu, Ryan Dubnicek

This collaborative project focused on using computational and digital tools to analyze a corpus of published works on yellow fever published between 1650-1900. The goal of this project was to generate data to visualize yellow fever knowledge in the Caribbean over time. This visualization and analysis is part of a larger research project which will result in a book on how yellow fever and medical understandings of the disease combined to affect the struggle for empire in the region. By bringing attention to the development and application of medical ideas of yellow fever into the study of how that disease shaped the struggle for empire at the intersection of the Atlantic world, this book will provide a more complete understanding of how the confluence of microbes and men produced the modern Caribbean. To that end I proposed to work with the English, French, and Spanish language materials related to yellow fever in the HathiTrust digital corpus using both the metadata and the full texts of these volumes. In the end we hope to generate, in addition to the traditional publications, an online component in the shape of a web-based presentation of the ways in which knowledge traveled and changed over time and through translations, which will be housed at the University of Iowa Libraries Digital Scolarship & Publishing Studio (the Studio) and available to the public. I anticipate this can include tables, graphs, network maps, and traditional maps, among others ways in which one can "see" the production and adoption of knowledge about yellow fever, the locations of where knowledge originates, and how that knowledge moves through the geographies of empires.

The Methods:
Throughout the 8 months we have been working together, Boris Capitanu has worked on writing and running the code to generate the different worksets, with the assistance of the rest of the team. Mariola Espinosa generated the historical questions and "sanity checked" the results, Eleanor Dickson advised on the content and structure of the HathiTrust corpus organization, and Matthew Butler assessed the workability of the results into visualizations.

Process and Results:
First Step: Generate a Yellow Fever Corpus Workset (1650-1900)
The first step in this ACS grant was to generate a workset comprised of all the accessible volumes in English, French, and Spanish related to yellow fever and published between 1650-1900. First, Espinosa created a dictionary of "yellow fever" terms comprised of any term historically referencing the disease in all three languages. For clarity purposes "yellow fever" is used here to mean any of these terms. With this list of terms Capitanu and Dickson then began to extract a list of volume ids that included >1 mentions of "yellow fever".
The first iteration of this workset was generated by mid-August, with 170,543 volumes (dictionary_matches.csv). In order to have a useful workset comprising only volumes that significantly addressed yellow fever, we started working on ways to fine-tune the workset building. Espinosa calculated the number of appearances for each volume and found that nearly half of the volumes only had one appearance of "yellow fever", making them not significant. In addition, a "sanity check" revealed that there were some volumes not counted which had significant mention of "yellow fever". Espinosa identified nnc2.ark:/13960/t9m33k80n as only

containing 4 mentions of "yellow fever", when a search within a different version of the same volume (using google books) revealed 67 appearances of "yellow fever". This example revealed that this volume is missing a significant portion of the book in HaThi. After exploring other volumes in the workset, we take this to be a rare occurrence.
Process of creating a manageable workset: After calculating the total number of appearances of "yellow fever" in each volume and doing some close reading of the texts Espinosa determined that any volume containing 10+ mentions of "yellow fever" would be included in the master workset. Capitanu then generated a new workset (above_thold_matches.csv)

As we began to work using this new workset to generate other worksets (Race and Citations, explained below), other issues issues appeared. For example, we needed to remove duplicates within the workset. By the end of September, this fine-tuning of the workset took place. Capitanu produced a new workset with unduplicated volumes containing >10 mentions of "yellow fever" (workset_ids_deduped.txt). We also determined that we would use extracted data from enumerationChronology when the publication date was not available in the metadata. By the end of October an additional issue was revealed. The workset contained a significant number of volumes with publication dates after 1900. Eliminating these volumes from the workset generated a more manageable list of volumes to work with (final_adjusted_workset). It is this final_adjusted_workset that the rest of the data was generated from.

Race Workset:
Espinosa created dictionary of "race" terms comprised of words and phrases historically meaning or indicating race in Spanish, English and French (case insensitive). After working out some of the glitches in the master workset as mentioned above, by November Capitanu had generated a workset of "yellow fever" volumes that mentioned "race". (race_matches.csv) This data will be used to trace the use of "race" as related to "yellow fever" over time and through empires. Butler and Espinosa are currently working on generating a graph and timeline with this information.

Citation Workset:
In order to trace networks of medical knowledge regarding yellow fever, Espinosa proposed that a citation workset be generated. This workset would look at what authors and works get cited over time. Using author names (ignoring corporate authors and missing authors) and volume short titles (everything up to the colon, or first six words) Capitanu would generate a workset of citations. By mid November we had a sample of this citation index and errors were corrected, because the initial run was taking a long time. (citations_sample.csv.zip) By the mid-November Capitanu had generated a citation workset (citations.csv.zip). Butler and Espinosa will work on visualizing this data using network mapping.

Entities Workset:
By early November, Capitanu had extracted entities from the "final_adjusted_workset" and a "final_adjusted_workset.entities.tar" file was generated.

Access to the corpus:
Espinosa requested access to the corpus contained in the final_adjusted_workset, a total of 13,368 volumes. Access to the corpus will allow not only the continuation of similar inquiries

into the sources on yellow fever, but also would provide access to reference the types of distant reading findings with close readings of the text. The access was requested in November and granted in February.