American Viticultural Areas Digitizing Project

Michele Tobias, GIS Data Curator, UC Davis Library, mmtobias@ucdavis.edu

Tom Brittnacher, Geospatial Data Curator, UCSB Library, tombritt@ucsb.edu

Project Team

Michele Tobias (GIS Data Curator, Data Management Program, UC Davis Library) leads the project. She manages the daily activities for the project, reviews data contributions, and manages contributors. Tom Brittnacher (Geospatial Data Curator, UCSB Library) leads the data creation team at UC Santa Barbara.

The project has team members at multiple institutions: Quinn Hart (Digital Applications, UC Davis Library), Vessela Ensberg (Program Director, Data Management Program, UC Davis Library), Axel Borg (Wine Librarian, Research Support Services, UC Davis Library), David Michalski (Geography Librarian, Research Support Services, UC Davis Library), Michael Colby (Librarian, Content Support Services, UC Davis Library), Yao Fang (Student Employee, UC Davis Library), Amber Reyes (Student Employee, UC Davis Library), Andrew Jessup (Collaboratory Services Specialist, UCSB Library), Maga Kim (Undergraduate, Collaboratory Staff, UCSB Library), Timothy Kwong (Undergraduate, Collaboratory Staff, UCSB Library), Alex Mandel (UC Davis Center for Spatial Sciences), Grant Miller-Francisco (non-UC contributor), Kyle W. Smith (non-UC contributor), Ouwxmaniac (non-UC contributor), Arielle Rose (Intern, UC Davis Center for Spatial Sciences), and Sierra Mabanta (Intern, UC Davis Center for Spatial Sciences).

Project Summary

The AVA Project uses a unique configuration of open source data creation and management tools to create a collaborative, publically-accessible, and freely available digital version of the American Viticultural Areas (AVA) boundaries from the official legal boundary descriptions registered with the US ATPF Code of regulations. This dataset provides researchers with an important tool as they examine the scientific, economic, and historical aspects of viticulture. The methods employed in this project have a broad applicability to other data creation projects. The project is housed on GitHub: https://github.com/UCDavisLibrary/ava

Project Narrative

Wine is an important industry in California as well as the US. The Wine Institute estimates that 325,000 Californians and 786,000 people in the US are employed in jobs related to the wine industry[i]. The US Federal Government defines specific areas of wine grape growing that produce a unique quality of wine called American Viticultural Areas (AVAs). These boundaries are defined and proposed by wine growing experts and are often used on wine labels to indicate a particular quality of the product. Wine specifically and agriculture in general are a large part of the University of California's contribution to this industry and academic thought, as is climate chance. Climate change is undoubtedly changing environmental conditions in wine-growing areas and factors such as rainfall, humidity, and sun exposure can all affect the wine grape crop. Digital boundary data allows for overlaying many datasets geographically, integrating any number and type of data into one analysis, and supporting both research and learning. Having access to the AVA boundaries allows researchers to begin investigating important questions about how predicted changes will affect the wine grape crop and plan for the future climate.

Having digital data on the boundaries of these AVAs would seem to be rather important for answering questions about the US wine industry, but there was no dataset that was free of cost, open, and made for use in research. The goal of the project was to create a freely available open data set usable in research that can be updated as changes are made to the federal legal descriptions. The scale of the project necessitated supporting collaboration to allow many people in a variety of physical locations to make contributions.

Our solution was to create the AVA boundary dataset, make it freely available in an open data format, and make the methods understandable and repeatable. Because the amount of data to be created was so large - 240 boundaries - we made the project collaborative and open to outside contributors. The version control and online access to code make GitHub, an online repository used commonly for group programming endeavors, ideal for collaborating with contributors anywhere on the globe. The UC Davis Library has applied these coding methods to a large collaborative data creation project to create digital geographic data designed to be used in research. We used open source software and open data formats to make the data easy to use, share, and visualize.

This project uses a fairly innovative methodology for employing GitHub to manage a collaborative academic data creation process (see Figure 1). The key technologies making this project possible were GitHub, QGIS, and R. GitHub's online repository[ii] and git version control made collaboration possible. We use GitHub for data version control and as an online repository for storing both data and documentation of methods. We adapted GitHub's issues list to manage the list of data to-be-created and prevent duplication of efforts. The project's documentation serves several purposes. The documentation catalogs the methods so researchers using the data have a clear understanding of how the data was created. The documentation housed on the site also provides an easily-accessible way to introduce new contributors to

the project's goals and methods and acts as reference for established contributors. We use QGIS, a powerful open source GIS (digital mapping) software, to bring together source data (typically scanned USGS topographic maps) and digitize the AVA boundaries from the Federal Register document descriptions. An R script provides an automated way to combine newly completed boundaries into the existing dataset.

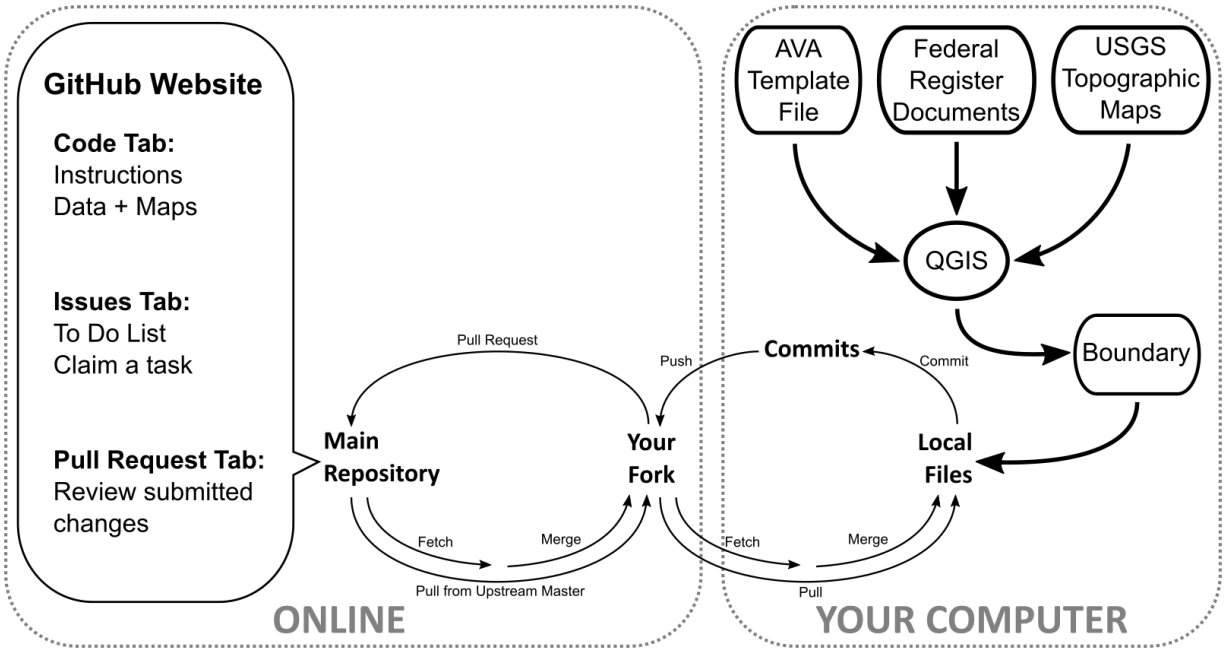

Figure 1: AVA Digitizing project GitHub workflow diagram

The AVA Project came together very quickly, thanks to the cooperation and skill of the team. In a meeting in early February 2017, Michele Tobias, David Michalski, & Jurry Stratford (UCD Librarian Emeritus) identified the initial problem – the lack of free digital access to this important dataset. Shortly thereafter, Michele met with UC Davis copyright expert Michael Wolfe to identify any potential issues with copyright; none were identified. The first meeting with initial UCD stakeholders and partners was at the end of February. After this meeting, Quinn Hart and Michele discussed and wrote the initial draft of the documentation for the digitizing methods. By the beginning of April 2017, Quinn had set up the GitHub repository and populated the issues list with the list of AVAs to complete. The first AVA was completed in April as well. In May 2017, UCSB Library's team joined the project, taking on the Southern California AVAs. In September, the first non-UC contributor, Kyle Smith, submitted a data contribution. In November, UCSB completed the southern California AVAs. In May 2018, UC Davis completed Arizona's AVAs and the project had taken on four new contributors.

One important measure of success for this project is the number of AVA boundaries completed. As a team, we've completed almost 60% of the AVA boundaries for the whole US, 87% of the 140 California AVAs, and all of Arizona's AVAs.

Another important measure of success is the number of collaborators contributing data. In just over a year, the project has amassed 18 contributors, with three more people currently working to learn the methods and eventually contribute as well. This project would be extremely time-consuming and labor-intensive for any one institution to undertake, which might have deterred UC Davis from attempting the task. The methodology, with its use of open data formats and version-control technology, was chiefly designed to reduce barriers to collaboration, especially with teams beyond UC Davis, while maintaining overall quality of the final product. The first collaborators to this project was the UCSB Library team. The initial partnership with UCSB provided an opportunity to test both the methodology and the ability of multiple campuses to collaborate on a project. Following that partnership's success, the project was easily expanded to include contributors in California and Texas, as well as other labs at UC Davis. A team at University of Wisconsin will be contributing soon as well. Collaboration has sped up the creation of AVA boundaries and spread the labor force among a larger pool of individuals, thus being a key to the project's success.

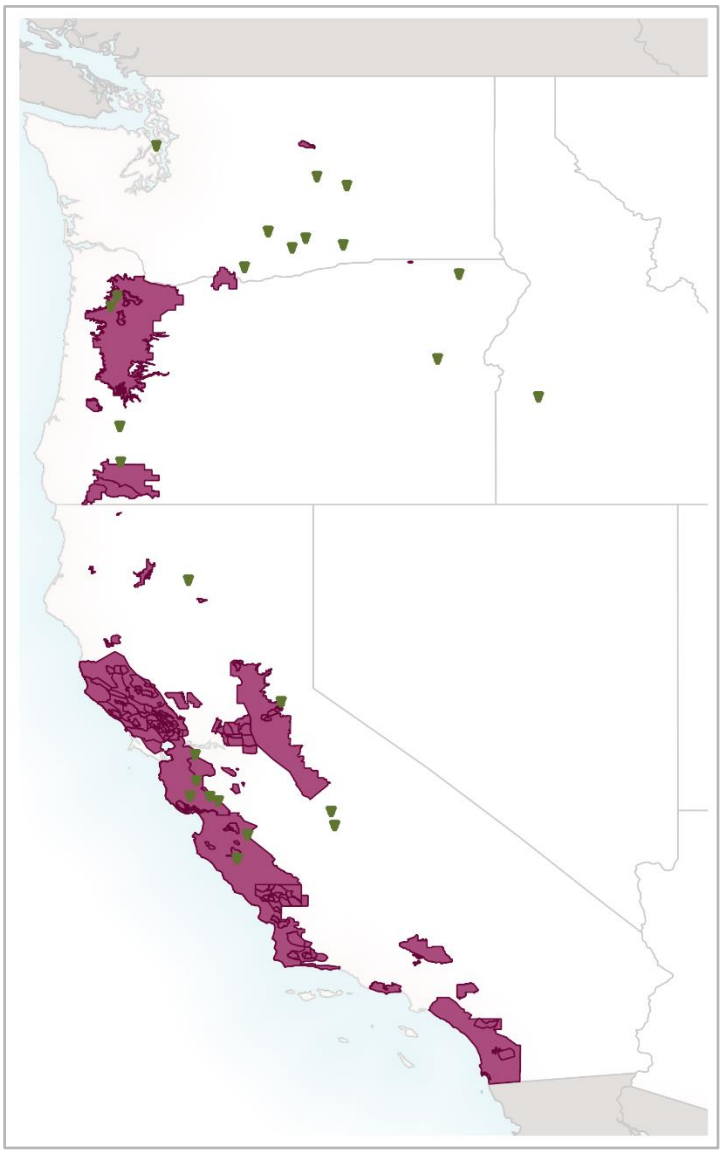

Figure 2: Completed AVAs (purple) and approximate locations of incomplete AVAs (green) on the west coast.

Collaborations also support education. UC Davis Environmental Science & Policy professor Robert Hijmans' lab uses the project to teach QGIS and collaborative tools. New interns work through the AVA project methods, learning skills they will need for other projects, and eventually contribute AVA boundaries as they learn.

The impact of the data produced by this project was broad and swift. The data has been in almost immediate demand, even before the whole dataset is complete, because it fills a significant gap in wine-related spatial data. Washington

Post cartographer, Lauren Tierney, inquired about using the data in an article about the 2017 California wine country fires. While she didn't end up using the dataset, the fact that it was of interest to a widely known publication is a good indication that the data is relevant outside the academic sphere. The AVA data for Arizona is currently being used for data analysis on wine geography of Arizona for a book published by Routledge. The analysis looks at the mismatch in Arizona's wine tourism industry and the places of wine production. The dataset provides an important point of analysis in this investigation.

The methods have also been impactful as a template for other large collaborative data creation projects. With guidance from Michele Tobias, UC Davis Economics professor David Rapson, PhD student Ben Dawson, and a team of interns were able to apply these methods to create electrical grid data from print maps produced by Southern California Edison for their entire market. The UC Davis Library has also applied the methods to a project to make digital guides to air photo coverage (called indexes) for a large number of photos that previously had no indexes. The project has applications outside of academia as well. These methods are currently under consideration for a California Department of Fish & Wildlife project to model the spread and impact of the recent nutria (a large destructive rodent from South America) invasion to California's Central Valley.

This project would not be possible without the contributions of our extensive team and the technologies that make this collaboration possible. We thank you for your consideration of this application and are excited to share this project with you!

---

[i] http://www.wineinstitute.org/files/Wine%20Institute%20Economic%20Impact%20Highlights%202016.pdf

[ii] https://github.com/UCDavisLibrary/ava