Future Development Guide for MAAPL's Data Collection and Management System

Developing a Data System to Address Foreclosures in Massachusetts

Team Members

John McCarthy, John Riley, Natalie Dionne, Sophia Noel, And Theodore Campbell

Project Advisors

Dr. Scott Jiusto

Dr. Gbetonmasse Somasse

Project Liaison

Ms. Grace Ross

Massachusetts Alliance Against Predatory Lending (MAAPL)

and Worcester Anti-Foreclosure Team (WAFT)

December 10, 2020

WPI Cape Town Project Center

Please view our website for more information on our project:
https://wp.wpi.edu/southafrica/projects/2020-projects/data/

Project Summary:

This project assisted the Mass Alliance Against Predatory Lending (MAAPL) in addressing the ongoing home foreclosure crisis by gathering data into an organized data system that could be used to analyze trends in foreclosures in the state. To this end, we devised a plan to accumulate data for our sponsor which resulted in a framework of a data system to support future analyses. The Data System includes instructional guides for gathering the information, a functioning database, and data that has been gathered from public sources. We also developed a guide for future development of the system. The project also served to develop an initial relationship between WPI and MAAPL.

Table of Contents

This document includes a description of the B-Term 2020 Cape Town Project Center's Foreclosure Data Team project and detailed recommendations for future projects and system development. It acts as a detailed guide to individuals and groups continuing this work, as it explains in depth the team's methods, recommendations, and notes that will benefit future teams.

Project Summary

The team has developed a data system to aid the Massachusetts Alliance Against Predatory Lending (MAAPL) in its fight against illegal foreclosures across the state of Massachusetts. This project encompassed several components, including the Data System itself, and other supplemental materials, which are summarized below, and described in further detail later in this guide.

1. Data System: The main deliverable of this project is a data system that can store data collected both manually by volunteers or using automated processes for data scraping. The system is structured as a relational database, which allows for data values collected from various legal documents to be organized and related in several tables, from which users can query for specific information related to a foreclosure. This information can be used to analyze illegalities in the foreclosure documents on a large scale, which is a major goal for the sponsor.

2. Future Development Guide: After the conclusion of this project, there were still be many opportunities to expand upon the team's work and further develop the Data System that was created. Because of this, the team has created this Future Development Guide, which outlines the current functionality of the system, gaps within the system that should be addressed going forward, recommendations for data analyses to take on in the future, and other useful information for future project teams, MAAPL, or others who may work with the system.

3. Guides for Public Data Collection: As mentioned above, there is the potential for volunteers at MAAPL to collect data from legal documents which will be inputted into the database. Because the government websites where this data is located are difficult to navigate, the team has created **Guides for Public Data Collection** to guide volunteers through the process of collecting data from the websites for each Massachusetts County's Registry of Deeds and the US Securities and Exchange commission website.

This project concluded with proof-of-concept code for a MySQL relational database and populating that database with data created to act as a placeholder for data that will be gathered in the future, as well as a repository of data on individual foreclosures from 7 of the 21 Registries of Deeds in Massachusetts that will be used to further populate the database. Further improvements will be required to make data collection and input more automated and streamlined, analyze, and visualize the data to prove trends in foreclosures, and develop a user-friendly way for MAAPL volunteers to interface with the database and access the data stored within it to help at risk homeowners.

Titles for Documentation for Future Teams

These are the titles of the documents that are referenced throughout the Future Development Guide. A short description is included for each of these documents at the end of the Future Development Guide.

- Guides for Public Data Collection
- Contact Info Spreadsheet
- Database Design Organized by Source Documents
- Email templates
- Indexed Data Standardization Guide
- Intake Form Consent Draft
- Proof of Illegalities from the Registry of Deeds
- SEC Request Documents

Project Methods

Assessing Data Sources

The team first accessed all the possible data sources to determine what data can be collected from each source and the methods that are required to create collections of that data. These sources were determined through interviews with Grace Ross of MAAPL, in which she told us about the most important documents and datapoints that MAAPL uses when trying to identify illegalities in foreclosures. Using this information, the team built a detailed Excel sheet describing the data fields to be collected, from what sources they are collected, and how they are collected and input into the database. By doing this, the team was able to organize their ideas about the overall flow of the Data System and what would be required to collect all the data points that MAAPL wanted. **This spreadsheet and a table including information on each main data source will be provided to future teams working with MAAPL.**

Data Collection

Once the team had a more detailed plan of the key components of the Data System, the data collection process began. Data was collected to populate the system using the following methods.

1. Collected data manually from the Registry of Deeds and US Securities and Exchange Commission

   The Registry of Deeds and US Securities and Exchange Commission (SEC) websites are difficult to navigate when trying to find specific documents and data, so it is important that the team and future teams understand what this process looks like. The **Guides for Public Data Collection** included in the published supplemental materials explain the process the team took to locate documents on these government websites and are a tool for others to utilize when searching for documents. These guides include instructions for each county in Massachusetts on how to perform searches, locate and open documents, and locate desired data values. They will be helpful in the future for volunteers to gather data manually that was not able to be collected automatically through the Data System. Manual

gathering of information consists of following the **Guides for Public Data Collection** to gather the public documents from the Registry of Deeds. From there, volunteers would use Amazon Web services to convert the document to a JSON file and then add it to the Database.

2. Reached out to the SEC, Massachusetts Registries of Deeds, and CoreLogic to acquire data in bulk

   Because the process described above is time-consuming and not user-friendly, the team reached out to the Registries of Deeds and US SEC directly through email to request the data in bulk. Templates of the emails sent to these sources are available to future teams and individuals working on this project in the **Email Templates**. This document provides a guide to future teams contacting similar organizations to request data. Seven out of the 21 Registries in Massachusetts responded to the inquiries, giving various amounts of information. These included instructions on how to navigate the county's website, PDFs of requested documents, and indexed data in the form of CSV files. Future project teams will receive the **Contact Information Spreadsheet** including contact information for each Registry, the SEC, and others who were contacted or participated in the project work throughout the term. This spreadsheet will also include information on what type of data, if any, each registry provided the team.

3. Converted documents into machine readable text and extracted useful information such as date, address, and auctioneer.

   The document conversion was completed using Amazon Web Service's Optical Character Recognition tool, called Textract (https://aws.amazon.com/textract/). This easy-to-use service is free for up to 1000 pages per month and $1.50 for each additional 1000 pages each month. The OCR conversion is completed by uploading PDFs of documents from the Registry of Deeds and other sources to the service. The OCR will output text files and JSON files which can be used to extract data and add it to the database. This tool also includes a Human Review step which would allow human volunteers to double check the accuracy of the OCR and add in any data fields that cannot be correctly read by the OCR, such as handwritten information. This human review option does seem to have potential for easily adding handwritten information into the database, but it has not been evaluated in depth. The team suggests that future teams test the ease of use of this option and whether it is feasible for MAAPL volunteers to help input handwritten information.

   For more information on this data collection step, we recommend talking with Steve Floridia, a MAAPL volunteer, who has completed lot of work with this service during our project term and has been extremely helpful during the database development.

4. Allowed for data input from Case Intake Form

   Although MAAPL's Data Intake System was not completed as part of this project term, the team created space in the database to allow information from the case intake form to enter the database. This information will not be entered into the

database unless the individual consents to their personal data being input into MAAPL's database. This section of the database will be expanded according to the final data fields in the intake form and populated once the form is completed by future groups and/or individuals.

Planning the Database: The ERD

In addition to understanding the process, it was helpful to understand how these data points all related to each other within the Data System. This part of the Data System planning was to be completed by creating an Entity Relationship Diagram (ERD). An ERD organizes the entities, attributes, and relationships in the Data System to show how the data collected in the system relate to each other within the database. Creating the ERD helped the team plan the individual data values that will be collected within the Data System, as well as how the data are connected, to create a robust database.

Detailed System Design Process

The team used the following

Step 1: Data Collection

Once the team had a more detailed plan of the key components of the Data System, the data collection process began. Data was collected to populate the system using the following methods.

- Convert documents into machine readable text and extract useful information such as date, address, and auctioneer.
- Create illustrated detailed step by step instructions (**Guides for Public Data Collection**)
- Allow information to come from the case intake form once it is completed later
- Reached out to the SEC, Massachusetts Registries of Deeds, and CoreLogic to acquire data in bulk

Step 2: Create Data Input Protocol:

The team created the **Indexed Data Standardization Guide** for how the specific information from the publicly available documents will be input into the Data System (Weber, n.d.). This was done to minimize errors in the data being input and streamline the data input process. This was completed by planning ways to standardize the format of the data input into the system, as well as creating streamlined code that worked to input data in the standardized form.

Step 3: Clean the data:

The data must be stored and presented in standardized formats so that the collected data is more organized in the database, making it easier for users to locate and analyze data. Our process was inspired by the Eviction Lab, which is doing similar work. Their steps are described in greater detail in the Eviction Lab's documentation (Desmond et al., 2018), and our adapted process is described below.

- Scan for and remove perfect duplicates
- Standardize names according to our data input protocol
- Complete another scan for duplicates after all reformatting
- Scan for outliers and determine if they are actually outliers or are incorrectly formatted or inputted

- Format into tables where all components of names/addresses/dates are broken into separate fields
- Pull data from these tables into the tables in the database

Step 4: Establish data connections:

It was important to map out how the various areas are going to connect to each other. This was achieved by creating a much more detailed model of where the information is going to be going and in what ways it would need to be processed before it gets there (Microsoft support, n.d.).

Step 5: Iterative Design:

After establishing connections between the data sets within the database, it was important to check in with MAAPL, to have them assess the current system implementation to make sure it meets their needs. After receiving sponsor feedback, the team repeated the implementation process to apply feedback.

Developing the Database

The SQL code we created was directly translated from ERD we created. We decided to use MySQL database for two main reasons. The first being the ease of setting up database MySQL. The next reason is because Steve Floridia, who is running the database, is familiar with MySQL so it would be easy for him to use.

The reasoning behind unique foreclosure identifiers

It's important to have a unique identifier to be able to link all the foreclosure documents together. The unique identifier for foreclosures we decided to use was recorded date and address combined.

Streamlining Input of Excel documents and CSVs

The team created a standardized protocol for how the specific information from the publicly available documents was entered into the Data System as mentioned earlier (Weber, n.d.). This was done to minimize errors in the data being input and streamline the data input process.

First, the team assessed the indexed data spreadsheets obtained from the Registries of Deeds. Nearly every Registry provided the data in different formats. To write a program to input the data from these indexed data CSVs into the database, the files had to be reformatted into a standard format. This standardization process required the following steps:

1. Standardize the data inputs according to naming and conventions described in the **Indexed Data Standardization Guide** (provided to future project teams). For example, denote all document types as either “FORECLOSURE DEED”, “CERTIFICATE OF ENTRY” or “POWER OF ATTORNEY”. Sometimes the Certificate of Entry is written as “Possession” or other names and the database code needs to have standard document names in order to keep track of where in the database each data point belongs.
2. Reorder the spreadsheet columns according to the standard CSV format found in the Data Standardization Guide. A standard column order is needed in order for the database code to effectively and correctly extract data from the CSV files into the database.

3. Consolidate the spreadsheet data based on document types. As described in more depth in the Data Standardization Guide, the way that grantors and grantees are listed in the indexed data vary by registry. Indexed data from most registries contain multiple lines for each document because there is often more than one party listed on each foreclosure document. In this step, we consolidated the spreadsheets into one line for each document, so that the information from individual documents can be more easily added into the database. This was achieved using Excel formulas and creating Document IDs from the book and page numbers, towns, and document numbers for each recorded document and consolidating lines with the same Document IDs into one line in the spreadsheet. Finally, Foreclosure IDs were created using the recorded date, address, and town for each document to link data from all recorded foreclosure documents in the database. The **Data Standardization Guide and Data Reformatting Excel Sheets** provided to future teams include more details on this process.

Finally, the data stored in the CSV files and JSON files from the Amazon Webs Services OCR program were inputted into respective data tables in the database using Python code to parse these files and add the data to the MySQL database.

Using the Data System

How to use our system

- How to rebuild database using foundational files stored on WPI server
    - First you will need to set up your MySQL environment so you can create your own tables. Then in MySQL you need to create a script and insert all of the code for the fill foreclosureDatabaseCreation.sql. Upon running this all of your tables will be generated and you next step will be to add data.
- Collecting data
    - How to use AWS (need pdfs): Steve
- Data input info
    - Instructional guides for CSV reformatting
        - Refer to **Indexed Data Standardization Guide**
    - How to input CSV files
        - To add data from a properly formatted csv file, setup an environment in which you can run python and run the script readIntoDatabase.py. This script will bring up a finder window and select the CSV you would like to import. After go to the SQLInsert.sql file that should be generated and select all and copy. Go over to MySQL and paste this into a SQL file. Next click run and all your data should be imported. NOTE:If any of the data in your CSV is already in the database you import will fail because the database doesn't allow duplicates.
    - How to add data from AWS: ask Steve
    - How to add additional data by hand

        - You can manually add data to the database through SQL insert statements in MySQL currently.
- How to make changes to the system as a developer (like a short developer's guide)
    - You can update the python script to import data with any python editor you would like. The SQL code can be edited in any text editor. To build on the SQL code you should reference the ERD to make sure additions continue and do not contradict with the current structure.

Recommendations for Future Work

WPI Project Teams

This project has created many opportunities for a partnership between WPI and future projects that will help develop MAAPL's data and case management systems. One way to complete these goals is to recruit WPI students through Major Qualifying Projects (MQPs), Interactive Qualifying Project (IPQs), and research groups. These teams will have the resources, time, and skills to continue developing MAAPL's data system and address other technology needs. We anticipate an IQP team to continue our work in the spring of 2021, so we have compiled a few ideas for meaningful and appropriate projects that will greatly benefit MAAPL in the short term.

<u>Interactive Qualifying Project (IQP) Teams</u>

1. Create a system to analyze the information in the database so that it can be used by the different persons that would have access to the database
    a. Create a Case Management System to help MAAPL volunteers store contact information and specific foreclosure case information for at risk homeowners. It could also include a user-friendly way to add data to and draw data from the database. This would greatly simplify MAAPL's current document-based operations and make it easy to interface with the database.

    b. Create a series of standardized analyses to help streamline the process of visualizing the data. There is a large quantity of information that the Data System will be collecting meaning there is potential for the information within it to be used in a variety of ways, such as for visitors to the website or for lawyers withing MAAPL. If this process could be streamlined it should allow for people to use the Data System for different things without it becoming confusing. Here are some examples of areas which specific visualizations and analysis could be used:

        i. **Data for At-Risk Homeowners:**
        The main purpose of this function of the Data System is to help people who cannot afford a lawyer create legal documents to stall or prevent the foreclosure process so they can get more help. This pause in the process will be key for MAAPL to be able to give them further support. Farther down the line if MAAPL chooses to collect data on how successful they have been at fighting illegal foreclosures are, they

could choose to display this to encourage people to use or make their data remain available to MAAPL to help others and themselves.

ii. **Data to Support Lawyers and Policy Makers:** The data lawyers from MAAPL will need to get policy changes and bring class action cases to court might change over time. This data will come primarily from the OCR data collected from public data sources. As of now important data will be how over time companies discriminated towards certain neighborhoods and groups. Examples of inflating housing prices of homeowners the year a homeowner takes out a new mortgage and comparing it to the municipal taxes and the actual home values (Ross, 2020). Not only showing that companies misrepresented the values of houses but also targeted specific groups will have a significant impact on cases.

iii. **Data Visualization for MAAPL/WAFT Website**
Data that will be presented on MAAPL's website is primarily going to be focused on providing insight into the current crisis. This data could come from public data sources like the Registry of Deeds, as well as more current information coming from the Data Intake System. The information that will be provided will be intended to convey the severity and widespread nature of the foreclosure crisis while also encouraging people to work with MAAPL to help fight their foreclosures and evictions. As such the information should generally be limited to graphics with short descriptions to go with them. The confidentiality of individuals is important and as such graphics should not contain information that in any way relates to an individual or other organization in a negative way. Some examples of graphics might be line graphs of the number of foreclosures in Massachusetts, or some sort of counter which displays how many people have made use of the Data Intake System to help fight foreclosures. That is not an exhaustive list of graphics but instead two examples of things that may be included.

2. Further develop the Case Intake Form to create legal documents and further populate MAAPL's database using data provided by at risk and foreclosed homeowners who are receiving help from MAAPL
    a. MAAPL currently has a similar system designed but not in place. The purpose is to provide homeowners with concrete help and help MAAPL gather further information on foreclosures in Massachusetts.
    b. Our team created a plan for this form and a consent form that was approved by the IRB. We suggest using our draft of the consent form as a guide for the consent form for the finalized Case Intake Form and editing it, as necessary**.** It has been provided as part of the documentation to future teams.

Notes and Recommendations:

Completing MAAPL's existing data intake system will help MAAPL by allowing new individuals to create legal documents to stall their foreclosure. MAAPL currently has a nearly complete version of the Case Intake Form, which a future team would acquire and update it so that it functions properly. The team would take this nearly complete form and finish it. In addition, the team would connect the intake form so that some of the information is added to the Data System so long as the person consents.

A Future IQP team should familiarize themselves with the Data System and create a means for information to be automatically be added to the Data System if the person consents. This will allow MAAPL to easily access significant documentation relating to that foreclosure through links in the Database. This could be done by storing the values in a CSV from the survey. Once someone creates the CSV, they will need to create an additional python script to insert this data into the database. This script will bring up an explorer window. Then they select the CSV that will be imported into the database.

Another thing the team can do is to take the existing data system and use the information that it collects to help MAAPL to use the information in a few general formats. One example would be the creation of a graph that shows the number of foreclosures in MA over a certain time period compared to the number of foreclosures that are legal, this chart could be used as a visual aid when speaking to the legislature in an attempt to change laws. Another example of an analysis document would be a pie chart displaying the foreclosures by auction performed by a certain auctioneer the chart would display when errors are present in the documents, they filed compared to the total number of foreclosures by auction that they performed. The creation of charts can be done by pulling certain sets of information from the database and inputting it either into excel or some other data visualization software. The team should plan to meet with a representative from MAAPL to develop a strong understanding of what information would be helpful in each area.

<u>Major Qualifying Project (MQP) Teams or Research Groups</u>

1. Utilize a coding language to automate the process of collecting data from public databases
    a. This will allow the database to function with significantly less user input which will reduce the amount of time required to add large datasets to the database.
    b. This may include developing or utilizing an OCR program that can automatically read handwritten signatures.
2. Expand the database to include data from other states, for example in New England, the Northeast, or the rest of the country
    a. This would provide MAAPL insight into the foreclosure crisis in regions outside of Massachusetts and would expand MAAPL's influence and reach in the future, if they wish to pursue legal battles outside of Massachusetts
    b. This would be a long-term project, once MAAPL's database and data collection systems for Massachusetts are completed and functional
3. Develop an OCR system to turn handwritten information in documents into text that can easily be entered into the database
    a. This is a complex process that will likely involve machine learning and artificial intelligence to "train" a computer to read handwritten information.

Because this is a complex problem that has not fully been addressed yet, this is a major project that would make a significant difference in not just MAAPL's operations, but potentially have many other applications.

Notes and Recommendations:

The team created a set **Guides for Public Data Collection** for a person to pull information from the registry of deeds. We recommend using that as a guide to gather the documents from the ROD using a computer program. Additionally, when the Data System is expanded to include the registries from multiple states there would be more new websites that documents would need to be pulled from.

- A challenge to this project is the laws relating to foreclosures differ between states and the documents are stored mostly on a county basis so there would be a large bulk of websites that documents would be pulled from (though these websites may follow similar general formats)
- The legal analysis would need to be done on a state-by-state basis so the team would need to program a logic table to ensure that the foreclosures are assessed according to the laws for the state they were in. This means that the team would need to develop some level of knowledge about the process of foreclosure in other states

Automating document retrieval will allow the Data System to gather the information very efficiently. By minimizing the need for a person in the data collection process the Data System will be able to gather significantly more information in a shorter period. This will be extremely helpful when it is expanded to a larger region as the number of documents will increase dramatically.

The use of logic to assess the legality of each foreclosure by state will be helpful because the individuals at MAAPL are familiar with the legal process of foreclosure in Massachusetts, however they are less familiar with the process in other states. As such the team would need to do research into how foreclosures work in the various states in the region that they are expanding the system to work in.

General Recommendations

- Automate and improve the data cleaning and standardization processes:

  The data must be stored and presented in standardized formats so that the collected data is more organized in the database, making it easier for users to locate and analyze data. The data spreadsheet reformatting process should also be more completely automated using Excel or other programming languages. This will streamline the data input process and reduce the chance for human error, which could potentially destroy any data successfully collected.

  Also, complete the following steps to more thoroughly clean and standardize the data:

  1. Scan for and remove perfect duplicates
  2. Standardize names according to our data input protocol
  3. Complete another scan for duplicates after all reformatting

4. Scan for outliers and determine if they are actually outliers or are incorrectly formatted or inputted
5. Format into tables where all components of names/addresses/dates are broken into separate fields
6. Pull data from these tables into the tables in the database

- Consider Google forms or instructional videos to help guide volunteers through manual data collection and inputting the data collected into correct CSV/Excel file columns, if the human review step of Amazon Web Services does not work to easily input handwritten data into JSON/text files output by the OCR system and then added to the database
- Create log files for the database to document everything input into and queried from the database. This will help in the future if the system ever breaks, by assisting in troubleshooting and debugging efforts.
- Continue compiling data from all data sources or interest. This includes other Registries of deeds throughout Massachusetts, the SEC, CoreLogic, etc. Use this new data to complete new analyses as needed by MAAPL.
    - We suggest using the email examples which are available to future WPI project teams as templates for when requesting this kind of data.
    - Also helpful when contacting the SEC would be the SEC Request Documents that describes our most recent data request (see the Useful Information section below for more details on this document)
- There are some additional documents related to foreclosures that are of interest to MAAPL. Decide what documents are most important to MAAPL and important data points in those documents and request these documents and the associated indexed data from Registries of Deeds to continue building out MAAPL's data system. This may also involve editing the existing SQL database code to create more tables to store this additional data in the database.

Useful Information

The team has compiled many files that help detail and visualize our work throughout the term and include valuable information that will be helpful to future teams and individuals working with MAAPL on similar projects. These files will be provided to future project teams. They include the following:

- **Guides for Public Data Collection.pdf:** contains the instructional guides for accessing and collecting data from documents through the SEC and Registries of Deeds. This is useful to both understand the organization of these systems and think of ways to automate the data collection process, as well as for training volunteers to collect this data manually if needed.
- **Contact Info.xlsx:** Contact information for all individuals that the team contacted and corresponded with throughout the project term, including information on what those individuals assisted the team with and who they are
- **Database Design Organized by Source Documents.xlsx:** a detailed spreadsheet that lists all important data fields included in the database, as well as methods for

how the team collected or suggested collecting the data to input it into database, and where in the document the datapoints are stored

- **Email templates**.docx: drafts of emails the team sent to the Registries of Deeds and SEC to request data. These may be useful guides for making similar data requests in the future.
- **Indexed Data Standardization Guide.docx:** describes the steps to reformat and standardize indexed data spreadsheets to prepare for inputting them into the database
- **Data Files Heading Descriptions.xlsx:** contains information on the indexed data provided by each Registry of Deed, as well as the standardized format for files that are entering the database
- **Intake Form Consent Draft.docx:** draft of possible consent for to be used when building out the Data Intake System. Even though we did not work with this system, we had this section approved by the IRB, so this consent draft may be a good place to start when building out the intake form
- **Proof of Illegalities from RoD.docx:** descriptions of proof of illegalities that can be found from the Foreclosure Deeds, Powers of Attorney, and Certificates of Entry recorded in the Registries of Deeds. A good place to start understanding the illegalities that commonly occur in foreclosures and a good start when thinking about analyses to pursue.
- **SEC Request Documents.pdf:** contains the response we received when we requested data from the SEC. We followed up but did not receive a response. This document provides proof of our communications with the SEC and could be used to continue our request without making a completely new one. Also includes information on getting data from the SEC using the EDGAR search described earlier in this document.