White Paper

Report ID: 115572

Application Number: HD-228942-15

Project Director: Sidney Irwin Dobrin

Institution: University of Florida Board of Trustees

Reporting Period: 5/1/2015-4/30/2016

Report Due: 7/31/2016

Date Submitted: 7/3/2016

Type of report: Final performance report

Grant number: HD-228942-15

Title of project: MassMine: Collecting and Archiving Big Data for Social Media Humanities Researchers

Name of project director(s): Sidney Dobrin (Project Director); Matthew Gitzendanner (Co Project Director); Laurie Taylor (Co Project Director)

Name of grantee institution: University of Florida

Date report is submitted: May 11, 2016

Project Activities

The MassMine project team representing participants from the Department of English, George A. Smathers Libraries (Libraries), and Research Computing at the University of Florida (UF) were awarded $60,000 to finish the version 1.0 release, develop a robust training program, and promote the MassMine open source software. MassMine enables researchers to collect their own social media data archives and supports data mining, thus providing free access to "big data" for academic inquiry. MassMine further supports researchers in creating and defining methods and measures for analyzing cultural and localized trends, and developing humanities research questions and data mining practices. The primary aims of this project were to: 1) refine the MassMine tools to support collection, acquisition, and use of available social media and web data; and, 2) develop a training program and corresponding online resources for supporting the broad use of MassMine by humanities researchers, regardless of experience.

MassMine was intentionally designed by and for humanities research as a social media data collecting and archiving tool. The project began with several proposed project results and products: 1) Adding additional data sources; 2) Export & Processing Module; 3) Full Training Program and Documentation, and 4) a Graphical User Interface (GUI). The GUI was planned to complement the console interface, with both interfaces served by the same engine, and with parallel functionality accessible through both. In developing the code with multiple releases for use on personal and central/high performance computing centers, as well as implementing the full training program, which includes consultation with the scholarly advisory board, the 4th step of the plan was changed. As proposed, the project team expected a high demand for a GUI to support and enable humanities researchers in their work with social media data collection and archiving. Through the consultative process in collaboration with the project team and scholarly

advisors, the recommendation from these and new users was to refine the documentation—adding videos whenever possible—for using the command line directly and not adding a GUI. This request was based on multiple trainings which included elements of GUI-ified actions, which participants found to be more difficult and less efficient (directly and for value-add) than using the command line. This change reflects the changes more broadly with the full training program, which is still underway, and which has moved to a combination of group and individual trainings based on the need for many participants to engage with MassMine as a new resource and tool to think through the process of data collecting and archiving for their research.

The additional project deliverables are proceeding as planned and proposed, with any changes determined in consultation with the project team and scholarly advisors. Since receiving startup funding in May of 2015, MassMine has undertaken a complete rewrite of its core software. The original prototype used R libraries to access Twitter, but these libraries had inherent limitations that were holding back further development of the software. Since then, Nicholas Van Horn wrote new, original libraries for accessing APIs and for converting raw JSON data pulled from social network APIs. This new technology has allowed MassMine to efficiently and sustainably add several new data sources. In addition to improved access to Twitter's Rest and Streaming APIs because of the rewrite of the core software, MassMine now accesses Tumblr's API—which provides complete historical access to all of Tumblr's social network data. Wikipedia and Google Trends were also added as data source. Finally, MassMine has added a general web scraping tool—allowing users to pull the text, images, and URLs from any web page or lists of web pages. Beyond the improvements to access and data curation, the rewrite of MassMine's core software has greatly improved MassMine's portability and

installation for users. With the original R prototype, users had to install many dependent programs and additional software in order to use MassMine.

The full rewrite of MassMine's core programming, accomplished during the grant, now allows MassMine's code to compile to the C language, greatly improving the portability of the tool for research. The newly rewritten software now requires no dependencies or additional software in order to be installed and ran by users. This upgrade greatly reduced the amount of technical knowledge required to install MassMine (users only need to know how to download and unzip a file to install), and it allows MassMine to start supporting the OS X operating system. Eventually, this change will also allow MassMine to start supporting Windows as well. Following open source trends in software development, Linux and OS X were the first operating systems supported by MassMine (because their underlying systems are similar—allowing for more efficient development).

For export and processing, the JSAN tool was built to help users process raw JSON data from social networks, and easily turn this data into an accessible spreadsheets form. Currently, only conversion to CSV files are supported, but JSAN also has functionality that allows users to specifically select only the columns of data they need for their research. Work is underway to add analytics and data visualization to the Export & Processing Module—which will add aid scholars in researching social media. Basic text mining functionality, like determining the most frequent words, #hashtags, and @users occurring within a stream of data. Time series analysis and sentiment analysis will be added as well. Eventually JSAN will become a full analysis suite to complement MassMine's data collection capabilities. However, for now, JSAN's current processing and exporting functionality fulfills the project's initial aims.

Additionally, new trainings specific to the grant project, to another grant project (funded as a seed grant from the UF Informatics Institute for embedded use and training with MassMine in Journalism and Communications classes), and at different events have led to greater interest and involvement on MassMine, with events including:

- Presentation/training at UF's first-ever DH Bootcamp on January 28-29, 2016
- Presentation for UF's Informatics Institute on November 17, 2015
- Presentation on MassMine for the Conference on College Composition and Communication (CCCC) in April 2016
- Online Presentation by Shelby Miller on March 4, 2016 (https://prezi.com/lcuw3nbvwr_c/massmine/)
- MassMine online presentation on February 6, 2016 (https://prezi.com/vm64txg1oovj/massmine-is-a-social-media-mining-and-archiving-application/)
- MassMine Facebook page with 61 likes for promoting team activities (https://www.facebook.com/MassMine-1474163546155816/)

Publications include:

- Van Horn, N. M., Beveridge, A., & Morey, S. (*In press*). Attention Ecology: Trend Circulation and the Virality Threshold. *Digital Humanities Quarterly*.
- Van Horn, N. M., & Beveridge, A. (2015). Writing eScience: Using Data Science Tools to Study Networked Writing Ecologies. *In Proceedings of the 2015 conference on college composition and communication*.

- Beveridge, A. Looking in the Dustbin: Data Janitorial Work, Statistical Reasoning, and Information Rhetorics. *Computers and Composition Online*. Fall 2015. http://casit.bgsu.edu/cconline/fall15/beveridge/

Other:

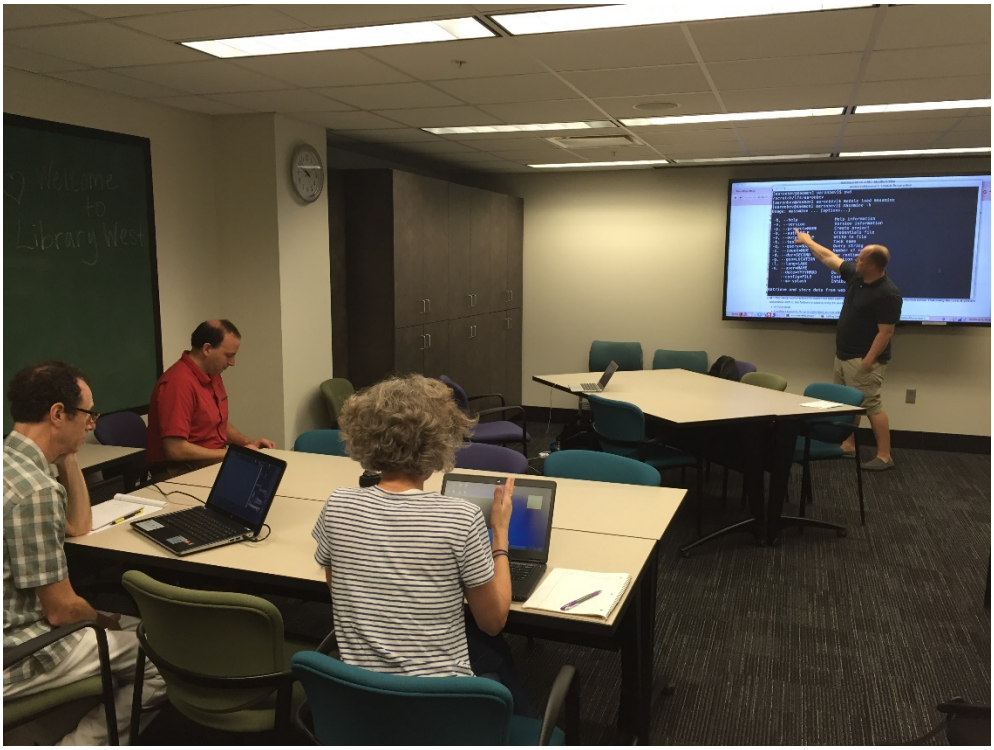
- Included in courses for several classes.
- Included in syllabi for spring 2016: http://english.gmu.edu/people/sholmes9 (http://webcache.googleusercontent.com/search?q=cache:pq50Ec-Ux2kJ:s3.amazonaws.com/chssweb/syllabuses/26350/original/Engh_308_Syllabus_(spring_2016)_official.docx%3F1454608048+&cd=33&hl=en&ct=clnk&gl=us)
- News covered in the *Independent Florida Alligator*: (http://www.alligator.org/news/campus/article_d1e94ccc-ceb7-11e4-957c-e3d4b73f4ec3.html)

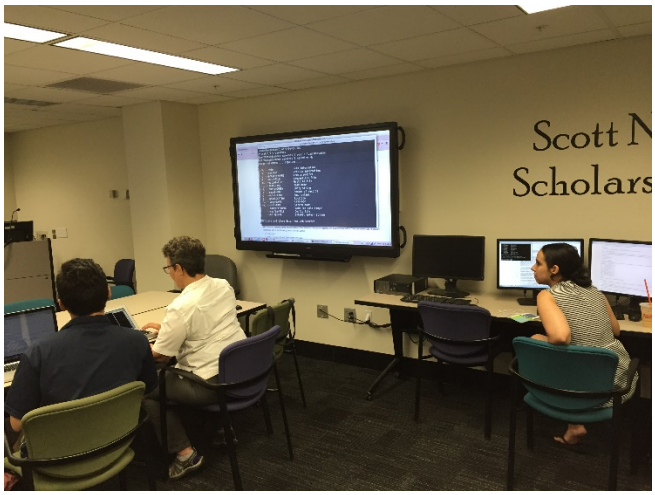
Training sessions included the group trainings, trainings with project team members, as well as additional consultative sessions:

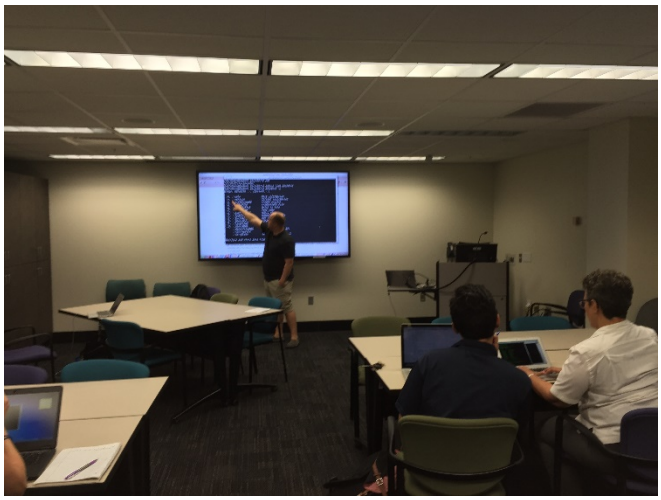
- Susan O'Brien, African Studies and History, UF
    - Using MassMine to track social movements in Nigeria using Twitter data.
- Antii Miihkinen, visiting scholar, College of Business, UF
    - Using MassMine to study the public reception of accounting documents and financial reports for publicly traded companies as displayed on Twitter.
- Norman Lewis, Journalism and Mass Communication, UF
    - Used MassMine in his big data and journalism course for graduate students in the Journalism and Mass Communication program
- Sean Duke, PhD candidate in English, UF

    - Using MassMine to track the Twitter conversation about the Hugo Awards. Specifically, the project looks at how particular conservative campaigns have tried to "hack" the voting process of the Hugo awards, and the conversations and responses to these campaigns.
- Denise Toled, MA student in JMCC, UF
    - Using MassMine to study feminism and presidential politics on Twitter.
- Wang Muxi, MA student in JMCC, UF
    - Using MassMine to study the Twitter practices of professional athletes and their use of Twitter for fan engagement and marketing.
- Danielle Sharpe, Department of Epidemiology, UF
    - Using MassMine to study public health issues on Twitter.
- High school student, project research

Photos from Trainings







Accomplishments

The primary aims of this project have been accomplished:

1) Refine the MassMine tools to support collection, acquisition, and use of available social media and web data
    a. MassMine tools available as linked from http://www.massmine.org/
2) Develop a training program and corresponding online resources for supporting the broad use of MassMine by humanities researchers, regardless of experience.
    a. Training materials available on www.MassMine.org
    b. Numerous humanities scholars and researchers trained in a series of onsite and online trainings, with full training materials online for others

Audiences

Most immediately, the audiences could be said to be the computational community of Digital Humanities; however, that would be too limited in scope. In reality, MassMine is useful to any who are interested in adding social media archiving and analysis to humanities research

for areas as varied as politics, trending topics, styles, fashions, popular culture, social movements, reception of artworks, etc. For more on audiences, see the list of participants under project activities. For an example of impact, the video “What is MassMine?” has been viewed 1,092 times as of May 11, 2016 (https://www.youtube.com/watch?v=1J2ywTHhGvU). The new training videos will continue to gather statistics. The http://www.massmine.org/ website has received increasing traffic over time, demonstrating consistent use outside of the core development and testing/training team. To date, the site has received traffic from 11,384 unique visitors, for a total of 17,311 visits with 30,301 pages served. The MassMine software has been downloaded approximately 992 times as of the date of this writing. These numbers are reflective of human users, and exclude traffic generated by robots, worms, or replies with special HTTP status codes (errors, etc.).

Evaluation

The MassMine project was evaluated on an ongoing basis by the project team for the project goals, activities and deliverables. As covered in the Accomplishments section, the project has been successful based on evaluation of the actual accomplishments with the goals established for the report period. The ongoing, iterative evaluation activities have informed the next stages for the continuation of the project through additional funding sources.

Evaluation by scholarly users has been positive, with several scholars using MassMine in their research, MassMine embedded in journalism courses, and MassMine collaborations for future grant proposals.

Continuation of the Project

MassMine is continuing as embedded within courses at the University of Florida and Capital University. MassMine has also collaborated with Epidemiology researchers at UF to submit a grant proposal to the National Institutes of Health. The proposed research will develop new algorithms built on MassMine's functionality that will support the study of alcohol use in underage drinkers, demonstrating the value of humanities collaboration and perspectives for the use of humanities' tools and research in the sciences.

MassMine is also a major project of the TRACE Innovation Initiative (http://trace.english.ufl.edu/about/), which is a research endeavor developed and maintained by the University of Florida's Department of English. TRACE works at the intersection of ecology, posthumanism, and writing studies. Providing an interdisciplinary forum for scholars, TRACE focuses on the ethical and material impact of media. TRACE acts as a hub for several distinct projects including an online journal and MassMine for which we are always seeking submissions.

Long Term Impact

MassMine's long term impact is already being demonstrated with:

- New research made possible by and generated using MassMine and the attendant training resources
- Greater awareness of the importance of the humanities for other research
- Greater awareness of the potential and an established relationship for ongoing collaboration with the humanities and high performance computing at UF, and as a model for other institutions
- Classroom integration of MassMine at multiple institutions, in multiple classes

Best practices and lessons learned from MassMine include highlighting the:

- Importance of inclusive, diverse engagement to bring a variety of disciplinary and other perspectives to bear on the needs and concerns for the software development of research tools
- Importance of iterative and engaged development to ensure tools meet user needs, and to ensure time is not expended on unneeded development (e.g., stopping development of the GUI, and replacing it with more extensive development of core MassMine functionality (portability, processing, exporting, analysis); moving from more group trainings and text based tutorial to video tutorials and individualized trainings and consultations to meet researcher needs)
- Importance of highlighting the contributions of perspectives and methods from the humanities for other research areas

Grant Products

All grant products are available on multiple sites, and all are as linked on the MassMine website: www.MassMine.org. The Full Training Program has been a core focus of work to date. For the full training program, in addition to the in-person and online trainings, as well as the consultations with individuals, the project has produced many training resources:

- All materials, teaching resources, training videos documentation, code, etc.: http://www.massmine.org/
- Grant project materials: http://ufdc.ufl.edu/AA00025642/00001/allvolumes
- MassMine team email list with 20 members, discussion archives: https://lists.ufl.edu/cgi-bin/wa?A0=MASSMINE-L