

Data Management Plan

MassMine is being developed specifically to support data research needs in the humanities. This includes the ability to access and engage with all levels of tools and data research. Open source code is essential to support external review for reproducible research, support ongoing open development to support data research in the humanities, and enable and foster collaboration among humanists for data research. In developing MassMine, one of the Project Team goals is for MassMine to exemplify open and collaborative approaches for software development and training in the process of improving access to data research. The same overall alignment will be used in making all technical decisions including those related to the GUI for MassMine. Like MassMine's other components, the GUI will be based on open standards and compliant code to support use on any operating system with an open standards compliant web browser (Windows, Apple, Linux). MassMine code is already publicly available through GitHub and will be released to GitHub on an ongoing basis. MassMine is released under the GPL license as open source for download by anyone. Using GitHub others can also "fork" a copy of the code. Forking is the term for creating a new version of the software where developers can continue development on a separate trajectory to submit new changes and additions to the software. Versioning and debugging will be controlled through [GitHub's update/submissions system](#), and new changes will be developed and released through that same system under the supervision of the MassMine team at UF.

For success, all materials for the project need to be shared openly and as widely as possible. The investigators commit to openly sharing all data in a timely manner. The proposed MassMine project focuses on software development and training which do not involve any private or otherwise restricted data, and do not involve any data that would present a risk to disclosure. The team does not anticipate any privacy issues, ethical issues, or intellectual property issues. Because MassMine enables other research projects, for the research data collections created by MassMine which could potentially have privacy and other concerns, the project training program will explicitly include data privacy and IRB approvals as supporting resources for data research.

In addition to MassMine code on GitHub through regular releases, each major release version of the code also will be archived to the broadly accessible [IR@UF](#). Project documentation, tutorials, and training materials will be hosted in the IR@UF. Materials will include documentation, project examples, sample data sets, guides on additional resource articles and related open source analysis software, etc.

The Smathers Libraries at UF commit to archiving and making materials accessible on an ongoing basis and at project end. This is in keeping with normal practices of the Libraries' commitment to open and expedient dissemination of grant products and grant materials (e.g., ["Unearthing St. Augustine" grant materials](#)) to support research needs and to assist in building a culture of grantsmanship. UF dedicates staff time to digital preservation and access from the Digital Production Services staff, IR@UF Manager, Digital Development & Web Services Team, Digital Librarian, and others.

The project will generate a variety of data materials, with the majority being code, training resources, and documentation. Specific forms include: whitepaper, planning materials, reports, webinar videos, training materials, and meeting notes. Programming for MassMine was developed in and will continue to use the R language as the underlying technology for MassMine. R is an open source language, as are all of the development environments for coding in R, so technical resources for the programming and software development of MassMine are freely available and well supported. Documentation will be embedded in source code, in separate ASCII files (e.g., plain ASCII, AsciiDoc, HTML, XML) and/or in formatted files (e.g., PDF, DOCX, PPTX). Training and support materials will be stored in standard formats (e.g., HTML, PDF, AVI, PPTX, etc.). Researcher datasets and accompanying files will be made available in their original and normalized formats ([brief list of selected, recommended formats](#)).

The Project Team will use GitHub for sharing code and code documentation, with all data openly available for anyone through GitHub. For permanent and findable support, all grant data materials will be openly accessible and preserved in the [IR@UF](#), powered by the SobekCM software, which provides metadata for all materials (at the item, group, and aggregation levels), permanent identifiers and URLs, multiple file formats and digital object packages (preservation and access copies), and more. All materials for this project will be openly accessible and will be made available as soon as possible, with the supporting metadata for findability and usability, with all project data made available at minimum twice each year and the majority of the project data made available in regular releases each week or more frequently.

The Libraries are committed to long-term digital preservation of all materials in the UF Digital Collections (UFDC), including the IR@UF, and in UF-supported collaborative projects as with the [Digital Library of the Caribbean \(dLOC\)](#). Redundant digital archives, adherence to proven standards, and rigorous quality control methods protect digital objects. Through UFDC, the Libraries provide a comprehensive approach to digital preservation, including technical support, reference services for both online and offline archived files, and support services by providing training and consultation for digitization standards and long-term digital preservation. The Libraries support locally created digital resources as powered by and hosted with the [SobekCM Open Source Repository Software](#), including the [UFDC](#) which contains over 381,000 digital objects with over 30 million files (as of February 2014). The Libraries create METS/MODS metadata for all materials. Citation information for each digital object also is automatically transformed by the [SobekCM software](#) into MARCXML and Dublin Core. These records are widely distributed through library networks and through search engine optimization to ensure broad public access to all online materials.

In practice consistent for all digital projects and materials supported by the Libraries, redundant copies are maintained for all online and offline files. The digital archive is maintained as the [Florida Digital Archive \(FDA\)](#) which was completed in 2005 and is available at no cost to Florida's public university libraries. The software programmed to support the FDA is modeled on the widely accepted Open Archival Information System. It is a dark archive and supports the preservation functions of format normalization, mass format migration and migration on request. As items are processed into the UFDC for public access, a command in the METS header directs a copy of the files to the FDA. The process of forwarding original files to the FDA is the key component in UF's plan to store, maintain and protect electronic data for the long term. If items are not directed to load for public access, they do not load online and are instead loaded directly to the FDA ([more information](#)).