

7. Data Management Plan

Expected data. The primary data to be produced, maintained and distributed by the *Visual Page* is the software we create. This will consist of C++, Python and Java source code, configuration files, generated source code documentation, and higher-level system and user documentation. We will refer to this collectively as the software. The software will be developed as a publicly available open source project from the outset. We will use GitHub as our source code repository and version control system.

In addition to the software, we will also collect digital facsimiles of books of poetry to use in developing and testing our application. This collection will, as permitted by copyright law and licensing restrictions, be made available to the public for download via our project web site. Where copyright law or licensing restrictions prevent such distribution, we will provide sufficient detail to enable those who are interested in replicating our work to locate and retrieve these documents. We will maintain metadata that we associate with these documents (e.g., author, publisher, publication date) internally using a relational database. This metadata will be made publicly available as a CSV file available for download at our web site. Internally, we may use a number of derived forms including high-resolution and thumbnail page images along with metadata to link these derived forms to their sources. Instead of distributing this derived data, we will provide the tools required to create it along with detailed instructions for doing so.

Throughout the project we will create and use data files that describe the visual features extracted from the documents. To enable analysis and evaluation of our work, we will prepare public versions of these data sets to be distributed under the Creative Commons Attribution 3.0 license. These data sets will be available from the *Visual Page* project web site.

Period of data retention. All software and data will be made accessible throughout the course of the project as it is developed. The visual feature data-sets will be prepared and made available once we anticipate that no more work will be performed to extract visual features from the documents. Upon completion of the project, the web site and all related data will be transferred to servers maintained by Natalie Houston's institution or another hosting provider. The project's software will be maintained on GitHub by DARTS for a period of not less than 5 years.

Data formats and dissemination. The poetry books will be stored and disseminated as PDF documents, compressed as a gzipped tar file for convenience. Data sets consisting of extracted visual features will be stored and disseminated as ARFF (Attribute-Relation File Format) and CSV (Comma Separated Values) files. All resources will be made publicly available. With the exception of the digital facsimiles for poetry books, we do not anticipate any privacy, confidentiality, security, intellectual property or other rights or requirements that will impact our ability to store and disseminate this data. For digital facsimiles that we are not permitted to re-distribute, we will provide documentation for where we obtained our sources. All digital copies of restricted data will be destroyed upon completion of the project.

The software will be distributed as source code under the terms of the Apache 2.0 and the documentation will be distributed under the terms of the Creative Commons Attribution 3.0 license. Within the scope of the Start-Up phase of this project, we do not anticipate distributing compiled, executable binaries of our software.

Data storage and preservation of access. All data (excluding software) will be stored and managed for the duration of the project on a server provisioned by Digital Archives, Research & Technology Services (DARTS) using Amazon Web Service (AWS). Data will be stored on an Elastic Block Store (EBS) device which will be automatically backed up on a weekly basis to AWS Simple Storage Solution (S3). All software created for the project will be stored using GitHub and accessible through the Git source code control application and via the GitHub web site.