

Data Management Plan

Project Title: Active OCR: Tightening the loop in human computing for OCR correction

Institution: University of Maryland **Project Director:** Travis Brown

Budget: \$41,906 **Beginning:** 06/01/2012 **Ending:** 05/31/2013 **Duration:** 12 months

This data management plan was created on September 12, 2011, for submission to the Office of Digital Humanities (ODH), National Endowment for the Humanities as required by ODH Guidelines in the interest of securing funding for this project. This is the first version of the data management plan associated with this data.

Types of Data: The data produced by this project will consist of uncorrected transcriptions of texts in a variety of genres from the Text Creation Partnership Eighteenth Century Collections Online (ECCO) corpus published by Gale Cengage. Project research will also involve analysis of image files of pages from transcribed works. Small scripts or programs will be used to transform the text transcriptions into forms suitable for processing. Data that results from processing will comprise “character box data,” coordinate information describing how an optical character recognition (OCR) engine has divided images of pages from texts into boxes representing characters, and also, corrected, proofread versions of the original texts which incorporate corrections generated by the interaction between human curators and the proposed OCR software. The software enabling this interaction between volunteers and the OCR engine as well as the core OCR system itself will exist as machine images (specialized sets of software designed to run on a virtualized computing platform).

Data Standards and Capture: Data for this project will be created in number of forms. The raw, unprocessed transcriptions exist as XML files encoded using a custom version of the Text Encoding Initiative (TEI) schemas. The page images exist as high-resolution image files in open formats. These data will be received from the publisher of the ECCO corpus. Scripts written in open-source programming languages, annotated with comments to explain their functions, will transform data into a simple plain text format suitable for analysis by the OCR engine. This simple ad-hoc format will be documented in a text file included with the data. During the course of processing, the character box data will exist as plain-text coordinate data. Since the form of this data is specific to the OCR processor, the columns of output will be labels and a text file will be included to serve as a “code book.” The improved algorithm for OCR correction of 18th-century and other early modern texts will be significant data for this project. Finally, the corrected texts will be encoded as unicode text with basic structural markup conforming to level 1 of the Digital Library Federation Guidelines for interoperability.

Metadata: Descriptive metadata in the form of the TEI Header and MODS records received with the original text files will also be added to corrected output files with appropriate documentation of the corrections made to the texts. Metadata for the intermediate data forms (pre-processing and character-box data) will be provided as column labels or full-text description of the data files. These will not adhere to any formal standard but will be sufficient to allow other users to parse and reuse data created by the project. Machine images will have basic descriptive and technical metadata to allow them to be run in virtualized computing environments on supported platforms.

Legal Policy: The original source texts for this project exist in the public domain. Researchers are free to release modified versions of these texts after processing by the Active OCR system. Volunteer contributors who correct texts will be asked to license their corrections to the project for display and redistribution as a condition of using the site/software. Images used during analysis will remain copyrighted by the publisher and will not be redistributed. In the absence of policies stipulated by the funding agency, researchers will release datasets and software under licenses identified in policies developed by the Maryland Institute for Technology in the Humanities (MITH) in consultation with the university libraries. For software, the Maryland Institute for Technology in the Humanities (MITH) uses an Apache License Version 2.0.

Datasets such as the character box data produced by this project will be released to the public domain through the use of a Creative Commons license (CC0).

Data Storage, Security, and Backup: Data will be physically stored on a password-protected server maintained by the Information Technology Division of the Maryland University Libraries. Servers are housed in a secure machine room with redundant power. No data will reside on any other portable or external media. Since page images will be on loan from the publisher Gale Cengage, data security will be a particular focus of project researchers. Server logs will be monitored periodically to ensure no unauthorized access of the publisher's files takes place. Data is backed up incrementally through a service provided by the Maryland Office of Information Technology, which has a proven record of and commitment to secure data archiving for the university. In addition to backups hosted on university servers, data will be copied to tape and stored at a geographically-distant site through a service provided by Iron Mountain information services. The specific volume of storage for this project is not anticipated to exceed 1 TB.

Access, Sharing & Re-use: All data from this project will be made available for download from either the dedicated site hosted by MITH or from the Digital Collections of the University of Maryland within 6 months of the end of the grant period. There will be no additional permissions required to download or reuse data except for those specified above.

Long-Term Preservation: Within three years from the end of the grant period, data will be permanently archived with the University Libraries at Maryland. No data will remain on servers controlled by the Maryland Institute for Technology in the Humanities (MITH). Data will remain publicly available through the libraries' digital collections. The University of Maryland has an interest in investing in the management of data created by researchers affiliated with Maryland and to providing digital preservation services, such as file validation, integrity checks, and, if needed, format conversion.