Data Management Plan

Products of the Research:
All the data used for this project comes from newspapers and periodicals published in the era 1889 to 1893. All of the newspapers are either in the public domain because of their date of publication (prior to 1922) or they are available to researchers affiliated with Virginia Tech through subscriptions through the University Libraries. The articles used for this research are available in one or more of the following formats and databases:

Digitized newspapers, available through a public database:
- Chronicling America, hosted by the US Library of Congress
- State Library of Berlin
- University of Bonn
- Bavarian State Library
- Austrian National Library (ANNO)
- University Library of Freiburg

Digitized newspapers in a subscription database, from VT University Libraries:
- Proquest Historical Newspapers (select titles)
- America's Historical Newspapers

Digitized newspapers in subscription databases, but **not** from VT University Libraries:
- Proquest Historical Newspapers (Library of Congress or other libraries)

Digitized medical periodicals:
- HathiTrust, consortium of university libraries, including Virginia Tech
- Medical Heritage Library, Wellcome Trust and National Library of Medicine
- Internet Archive, publicly accessible library materials

Periodicals from the Medical Heritage Library allow access to public domain digitized collections through the Internet Archive. Periodicals from the Hathi Trust are available to Virginia Tech researchers through the partnership agreement. Separate agreements to allow access to the German scholars from Hannover University will be negotiated as needed.

Data Formats:
All of the research data for this project began as printed editions of newspapers or periodicals. In all cases, newspapers were filmed for preservation as microfilm. Newspapers that have been digitized are already available in pdf formats. Periodicals were scanned from the bound copies held by university libraries. Articles will be saved in pdf format. Databases that allow readers to access the OCR text make it possible to save articles as text files. The saved articles will be stored on a project site, using Scholar, an open source software adopted at Virginia Tech for instructional and research uses, or the university's licensed google folders. Only registered users of the site will have access to these materials during the research process. Data that is already publicly available, such as articles from the Chronicling America collection, can be made available to the public by linking to these online versions. Articles from newspapers in subscription databases can be made publicly available as needed for research purposes. The estimated amount of data secured through these methods will be less than 200 gigabytes.

Access to Data and Data Sharing Practices and Policies:
All data for this project originated in the public domain, and therefore confidentiality, privacy, security, and intellectual property issues are not relevant. In the case of materials available through subscription, data sharing arrangements will be developed in consultation with University Libraries and the vendors, on

terms consistent with the subscription and licensing agreements already in place. Source code for algorithms and format converters, plus preprints of papers, presentations, technical reports, and educational material will be posted on a dedicated project website, which will be updated regularly during the project. Preprints will be posted soon after acceptance; software will be posted after successful testing in trial version. The software will be provided as-is, requiring only proper acknowledgment; limited technical support will also be provided by the co-PIs and their students to qualified research groups, at no cost. Resources needed for web page creation and maintenance are minimal and readily available. Students and PIs have the necessary experience, since they design and maintain their own webpages.

Policies for Re-Use, Re-Distribution, and Production of Derivatives.
This research will generate derivatives in forms appropriate to the fields of humanities and computer sciences, including publications in scholarly journals, online research updates, and online postings. Access to these research materials will take forms appropriate to their form, including subscriptions to journals, the book will be available online, in libraries, and for purchase, and online postings will be freely available. That said, we believe in no-cost access to publicly-funded research data, software, and preprints, and we will go the extra mile to ensure the widest possible dissemination of our research results and the means to (re)produce them. Following completion of the project, we will consider offering a mature version of the code under a GNU or Apache license.

Archiving of Data:
Data will be archived in the Scholar site, hosted by Virginia Tech Information Technology services. Additional copies will be preserved by University Libraries. Virginia Tech Libraries is a member of MetaArchive Cooperative, a digital preservation consortium that uses a LOCKSS (Lots of Copies Keeps Stuff Safe) preservation strategy. Archival copies of files are contributed to a secure, closed-access network of dark archive servers set up between the MetaArchive Cooperative's institutional members. All servers are stored in different geographic locations and maintained by different systems administrators. When content is added to the institutional repository, VTechWorks, it is visited by seven of the network's servers, each of which replicates and preserves a copy. Servers are selected and assigned to content on the basis of their widespread geographical location. All seven servers revisit the content source on a regular basis to find content that has been added. The seven servers also check in with each other regularly to make sure that all copies are identical. If a mismatch is detected, the servers come to quorum regarding which copies are correct and which do not match, and then the network repairs the files. Repaired files are stored alongside the originals so that no file version is ever lost/replaced within the system. Content in MetaArchive is regularly migrated to new storage media in order to maintain its integrity. All of the material produced by this project (including source code, documentation, preprints, technical reports and other non-copyrighted publications) will be preserved for at least three years beyond the end date.