9. Data Management Plan

Expected Data

The project expects to produce six types of data: (a) software source code, (b) website source code, (c) reports and articles, (d) metadata from case studies, and (e) closed captioning data and (f) still images extracted from television episodes.

Data elements (a) through (d) will be saved as git repositories and published on GitHub. All of these data will be publicly available as they are built by the team. The commit history in the git repository provide a complete summary of the process used to construct all of the data elements. Also, the code for creating the extracted metadata (d) is included in the examples directory of the source code in (a).

Data elements (e) and (f) are being stored using cloud storage provided by Box under an institutional license from the University of Richmond. These elements are under copyright protection and we will therefore not distribute them outside of our core team. If other researchers require access to this material, there is an easy approach. We provide reproducible code in (a) for extracting the copyrighted material from commercially available media formats (typically DVDs). External groups can then recreate (e) and (f) after acquiring a copy of the raw source material.

Period of Data Retention

All of the data that will be made public, elements (a) through (d), will be accessible on GitHub as it is created. There will be no embargoing of the data and no need for a separate process of publishing the data.

At the end of the grant phase, a machine readable version of elements (a) through (d) will also be placed on the University of Richmond Scholarship Repository. The UR Scholarship Repository is a service of the University Libraries at the University of Richmond. Published materials and data included in the digital repository reflect the research and scholarly work of the university community and are openly available to the general public for download and use. The University Libraries provide this service to University of Richmond faculty, staff, and students free of charge and are committed to providing perpetual access to deposited content.

Data Formats and Dissemination

The source code (a) and (b) will be written in code readable by any plain text editor. The DVT Toolkit will utilize Docker for ease of use, but could also be run without Docker by manually installing all software dependencies. The research reports and articles (c) will be written in markdown; this is also readable by any plain text editor and can be converted into a large number of other formats using the open source software *pandoc*. Markdown is also seamlessly converted to HTML by the GitHub platform.

Extracted metadata (d) is saved as plain text files using a comma separated value scheme. This format can be read and parsed by any programming language and most statistical and visualization software. The closed captions (e) are also stored in plain text files. Still images (f) are stored in the JPEG format, which is readily readable and convertible by all major web browsers, image programs, and email clients. We anticipate no need to convert these to another format inside of the next 10 years. All text files will be saved using the UTF-8 encoding.

Data Storage and Preservation of Access

Elements (a) through (d) will be stored on GitHub, with a duplicate copy of the entire repository stored on our internal Box account and on GitLab. The copyrighted material will continue to be hosted on our institutional account with Box. Migration to a new cloud platform in the long-term, if necessary, will be managed with the help of the University of Richmond's IT department. A machine readable version of elements (a) through (d) will be placed on the University of Richmond Scholarship Repository for long-term storage.