8. DATA MANAGEMENT PLAN
T-AP DiD "Responsible Terrorism Coverage (ResTeCo): A Global Comparative Analysis of News Coverage about Terrorism from 1945 to the Present"

This data management plan will be implemented and managed by PI/Althaus for data stored or analyzed at the Cline Center, by PI/Wessler for data stored or analyzed at Mannheim University, and by PI/van Atteveldt for data stored or analyzed at Vrije Universiteit Amsterdam. Overall responsibility for compliance with the data management plan will be overseen by PI/Althaus.

Types of Data and Software to Be Used, Produced and Distributed

Two types of data will be used in the proposed project: (1) copyrighted full-text news data, and (2) non-copyrightable metadata and extracted features derived from full-text records (examples include word frequency tables, lists of named entities appearing in news texts, sentiment scores, etc.). All copyrighted full-text news data that will be used in the proposed research are already in the physical control of collaboration team members, with all required permissions already secured. The proposed project will also develop software tools and algorithms that can be deployed by other researchers on other textual corpora to replicate the analyses generated by the proposed research activity. These tools will be publicly released and disseminated under a permissive open-source licenses such as the MIT license. As far as possible, individual modules will be published separately to enhance community participation. All code will be published on github or a comparable platform.

Raw Textual Data: Format and Content

The Cline Center text data includes over 85 million news articles, most of which are protected under US copyright law. All these articles have metadata associated with them, indicating news source, title, date, etc., as well as extracted features stored and available as part of this project, such as named entities and geocoded place references (see Appendix Table 3). Cline Center full-text news data is stored in MongoDB and is documented with metadata files stored in SOLR indices. Exportable metadata files and extracted features will be distributed in ASCII text, CSV, or JSON files. Metadata attributes drawn from the Dublin Core Metadata Initiative will be used to define the schemas associated with the data stores.

The Cline Center corpora are complemented by over 10 million Dutch news articles stored in AmCAT (only a subset of these articles will be used in the ResTeCo project), an open source text analysis infrastructure. Internally, text and metadata are stored in a PostgreSQL relational database and are indexed using an Elasticsearch cluster. All data is accessible through an API that can export to CSV, JSON, and other formats. Metadata is based on Dublin Core and includes a URL that points to the original source, which for the longitudinal Telegraaf links to the freely accessible Dutch Royal Library (KB) Delpher application that contains both the OCR'd text and a scan of the original newspaper page. The Mannheim data are saved in a relational MySQL database. They are saved in fulltext together with metadata such as the origin URL, the date the article was published and crawled, the article title and, if they exist, the article authors and categories. These data can be exported into a CSV or TXT/JSON format.

Metadata and Extracted Features Derived from Full-Text News Data

Exportable data for public release will include metadata elements that identify the specific source article as well as extracted features that were derived from the full-text article at the levels of documents, actors within documents, and statements within actors. All metadata will be published under a permissive license (CC-BY or comparable) and will be accessible through an API linked from the project web site. Moreover, all data will be published as Linked Open Data, allowing easy access and combination with other data sources. Finally, all metadata will be linked to the original textual data, both from the original source (e.g. URL for online news or open archives) and the textual data entry stored in the respective

institution as detailed above. As far as permissible, headlines and snippets will be accessible through the project web site API to enhance validation and qualitative understanding of the metadata.

Copyright limitations prevent the project PIs from redistributing or making publicly available the copyrighted news text from which metadata and extracted features are drawn. However, much of the original source texts are either directly available on the Internet (in the case of web-crawled news stories) or available to the research community through standard news aggregation vendors such as Nexis-Lexis or ProQuest Historical Newspapers. In order to accommodate the needs of researchers who want to validate the quality of extracted features against the original copyrighted text, the ResTeCo project is committed to publishing metadata adequate to tracking down the original source material in a large number of cases (e.g., URL, title, source publications, date of publication, etc.). Such metadata cannot itself be copyrighted, so distribution of this metadata at a level of detail that allows researchers to track down original source records on their own should satisfy the validation needs of most users.

Three categories of metadata and extracted features will be publicly distributed:

1. *Document-level metadata/extracted features*: (a) date of publication, source of publication, title, URL, etc.; (b) PETRARCH events derived from the document; (c) classifier output on relevance for containing information about terrorism; (d) topics and subtopics derived from LDA analysis; (e) presence of episodic / thematic framing elements; (f) named entities / referenced actors; (g) ingroup / outgroup cues.
2. *Additional actor-level metadata/extracted features within documents:* (a) originating document; (b) associated organizations / groups; (c) actor labeling (terrorist, freedom fighter, militant, insurgent, etc.).
3. *Additional statement-level metadata/extracted features within actors:* A semantic network analysis consisting of (a) originating document; (b) source of statement (actor, or journalist if none); (c) target of statement (actor); (d) evaluation (sentiment); (e) ingroup / outgroup cues; (f) statement topic derived from LDA analysis.

Provisions for Archiving, Preservation, and Distribution of Data and Software

Since legal restrictions prevent the copyrighted full-text news data from being publicly released by the research team, only the non-copyrightable metadata and extracted features will be publicly released and disseminated after the conclusion of the proposed research activity. All derived metadata and extracted features will be publicly distributed through the Illinois Data Bank (https://databank.illinois.edu/), which assigns DOIs to all data files and maintains a stable and policy-compliant environment for preservation and public distribution of a wide range of data forms for a minimum period of five years past the date of original publication. All software tools and algorithms developed for this proposed research will be publicly distributed via GitHub.

The Cline Center has been running and developing cyberinfrastructure for preservation of news data since 2006. The Cline Center data store is actively updated, managed and maintained by dedicated Cline Center staff on an ongoing basis. These activities are independent of the proposed budgeted activities and will continue long after the proposed research has been completed.

Gold Standard Data

The final data deliverable of this project is the gold standard data that will be used to validate the text analysis methods. These data will be published under a permissive license (CC-BY or comparable) As far as possible, these manual annotations will be conducted on publicly available source material so the raw text can be published or linked together with the annotations.