Data Management Plan

We will make all the data and methodological procedures generated in the proposed investigation easily accessible to the research community via the web, using standard formats. This will be done by posting the data and methodology on the websites of the USC Shoah Foundation, the UCREL website at Lancaster University, the Spatial History Lab at Stanford University, the Holocaust Geographies website at Texas State University, and the Maine Dataverse Network. These universities will provide basic support and data storage. We will also post our publications on those websites, as well as on the Digital Commons at the University of Maine.

A. Expected data:
This project will propose and test analytical and interactive graphical exploratory tools for understanding and gaining insights into the oral histories contained in video interviews of Holocaust survivors from the USC Shoah Foundation's Visual History Archive (VHA). These capabilities will be harnessed through an extensible framework and toolset geared especially at the humanities research community. The system and its ancillary data will result in datasets and analytical results that will be published in research publications and made available through the project websites. The investigation will also produce digital and non-digital visualizations that we will make available and will use as figures in publications.

The following significant artifacts will be produced in the course of this research:

- A working dictionary of spatial and relational terms that we will use for further analysis of the spatiality of Holocaust survivor testimony. The dictionary will be developed for release to the larger research community and interested parties (such as the USC Shoah Foundation) through research publications and the above websites. Data and query processing capabilities will be made available to the research and education communities while ensuring that users' privacy is preserved. We will release a detailed description of our methodology, including the names of particular corpus linguistics (CL) and natural language processing (NLP) software and analytical modules proved useful in our exploratory analysis, through the above websites. New or custom modules that will be developed as part of this project may also be released with an open source license.
- Any new software or methodological procedures that are used to generate the data dictionary, and the pilot project's hybrid methodology combining those methods and procedures with close listening with manual and computer visualization, will be shared with the broader research community through the above websites, research publications, and research presentations. We will also share the names of CL and NLP methods that were developed by other researchers, along with how scholars and the public can access those methods, to preserve the rights and authorship of those who created and who maintain and update the websites and other means of access to those software programs.
- Results produced by the proposed research will be made available to participating government agencies (e.g., the United States Holocaust Memorial Museum) and to the broader research community through the above websites, research publications, and research presentations.

The following types of data will be retained, utilized, and archived, including:

1) Analyzed data (e.g., digital information that is published, including digital images, published tables, and tables of the numbers used for making published graphs).
2) Metadata that define how these data were generated (e.g., data that will be published in theses, dissertations, refereed journal articles, supplemental data attachments for manuscripts, books and book chapters, and other print or electronic publication formats), as well as full documentation of our source materials, including unique identification information for each VHA interview used in the investigation, and full bibliographic information about all other sources consulted in this investigation.
3) Raw data (e.g., data derived from VHA video interviews in order to create the data dictionary) and

transcripts of the interviews. Due to the nature of the project, the collected data will not need to be anonymized before publication. The VHA interviews used in the study are available to other researchers and the public through the USC Shoah Foundation.

For the **evaluation** of different phases of research, we will prepare and use the following datasets:

1) Datasets to test the validity of the data dictionary will be generated from the VHA video interviews: Such datasets will be constructed by considering as many available properties as possible, such as the location of the speaker and other individuals, time, locations of points of interest, location of concentration camps, railway stations, etc.
2) Preprocessed offline datasets from the Shoah Foundation testimonies and the Holocaust Historical GIS applications generated during the NSF-sponsored research (award nos. 0820487 and 0820501) (for Budapest, Italy, Auschwitz, and the SS camps system): Both GIS and oral histories data, as well as other contextual historical data, will be used to develop, validate and demonstrate the validity of our approach.

These data will be made available to interested researchers and educators in the field upon reasonable request.

B. Standards to be used for data and metadata format and content:

Data formats that will be used to make data available to others, including any metadata, will include ASCII, XML, JPEG, etc., selected according to the standards in publishing each specific type of data. Collected or used video and image data will be available in standard formats such as MPEG and JPEG, respectively. Geographical data sets will be described according to standards endorsed by the Federal Geographic Data Committee (FGDC), including the Content Standard for Digital Geospatial Metadata (CSDGM), and others as applicable.

C. Access to data and data sharing practices and policies:

All participants in this proposal will conduct research and publish the results of their work. Papers will be published in peer-reviewed academic journals or as peer-reviewed conference proceedings. Beyond the data posted on the above websites, primary data, samples, and other supporting materials created or gathered in the course of work will be shared with other researchers upon reasonable request, at no more than incremental cost and within a reasonable time of the request. The project websites named above will be maintained by the participants for a minimum of three years after the conclusion of the grant.

1) Policies and provisions for re-use, re-distribution, and the production of derivatives:
   It is the policy of all institutions participating in and supporting this proposal to encourage, wherever appropriate, research data to be shared with the general public through Internet access. Public access will be regulated by the partners to protect privacy and confidentiality concerns, and respect proprietary or intellectual property rights. Administrators will consult with the University's legal office to address any concerns on a case-by-case basis, if necessary. Terms of use will include requirements of attribution along with disclaimers of liability in connection with any use or distribution of the research data, which may be conditioned under some circumstances.

2) Plans for archiving and for preservation of access:
   Data and other research products will be made available immediately after publication. Final peer-reviewed journal manuscripts, and supplemental information such as data tables for graphical information in manuscript figures, which arise from NEH funds, will be posted at the websites named above no later than twelve months after publication (if publishing agreements allow it; if not, the above websites will provide a reference to the journal articles coupled with the supplemental information.) These records will be durable, accessible through web protocols, and made safe from tampering or falsification. The storage media will be updated as necessary to keep it current.