7. Data management plan

Introduction

This plan for Level II Start-Up funding describes the management, dissemination, retention, and archiving of **unique** research data produced during the proposed project. Collaborating with Indiana University (IU) Libraries, this project will provide for sustainable discovery, access to, and preservation of these data for use by other researchers, instructors, and interested members of the public for the length of this project and beyond. This will be facilitated through data and publication deposits in existing open-access disciplinary and/or institutional repositories. A second repository for the data, when fully aggregated, will be the National Archive of the Republic of Tatarstan (Russia).

Responsibilities

The PD will oversee the DMP with responsibility for ensuring that all requirements are fulfilled and will supervise the creation and management of the CEMPP public web platform in both its English and Tatar-language versions. For the latter, Iuri V. Pivovar, Chief Technical Specialist at the National Archive of the Republic of Tatarstan (NART), will assist him. Technical Director, Vincent Malic, will create and manage the project database and English-language public web platform; he will also work with Indiana University's Data Management Service for the successful migration of data to its repository and public server.

Expected Data

The data generated by this research, for which there are no precedents, are derived from a massive set of **MBs** compiled in the Russian Empire between 1828 and 1918 containing demographic data about approximately 25,000 Muslim inhabitants in the city of Kazan. The information recorded in these registers is organized in tabular format, with rows documenting particular demographic events (e.g., birth, death, marriage, or divorce) and with columns containing feature information per row. Experts in Kazan will transcribe these tables into digital format while retaining their original Tatar language. Transcription is regulated by a set of explicit guidelines that standardize the migration of physical data from source to database and the notation of exceptions and document defects. A complete transcription of an annual metrical book produces a single Excel file containing all of that book's tabular data.

Data Formats

Using a lightweight Python script, we will transfer the data contained within the Excel files to a MySQL *preprocessed* relational database that preserves the data in the form they take on the physical pages of the **MBs**. Doing so will ensure a high level of fidelity. Though the data are divided into table cells, for many column headings the data in a cell contain multiple, distinct pieces of information. For this reason, we will run the data in the preprocessed database through a processing pipeline that sifts discrete pieces of data from complex table cells. This processing pipeline will depend on the development of a Tatar Natural Language Processing framework that uses machine learning techniques supported by domain expertise to extract entities and relations from table text. We will develop this Tatar-language framework by expanding on existing NLP tools, such as Stanford's CoreNLP suite and the Natural Language Toolkit. Though this framework will be calibrated to meet the needs of the CEMPP, it will also constitute a valuable contribution to the advancement of NLP technology for Tatar, a threatened language. We will make all Tatar-language NLP tools so developed freely available under the Creative Commons Attribution 3.0 license.

The processing pipeline output, or *processed* database, will be organized around entities identified in tabular form. Most of this information will concern identified individuals and the circumstances of their births, deaths, marriages, and divorces, or their status as third-party witnesses to or participants in these events. The processed database will be a MySQL relational database implementing the Intermediate Data Structure (IDS) for Longitudinal Historical Microdata v. 4. Using this robust and widely adopted format for demographic data ensures that the processed database will be fully accessible to researchers world over and that the CEMPP database can be seamlessly integrated with other demographic databases employing the same model. A suite of Python scripts will periodically synchronize the processed MySQL database with an online-facing RDF/XML Semantic Web database that preserves the IDS schema but makes the data available for SPARQL queries and integration with other Semantic Web databases.

In the later stages of the project, team members will develop a set of rules for inferring whether two items in different records refer to the same entity. The software for making these inferences will be written in R and Python and will also be made publicly available under the CCA 3.0 license. Potential entity matches and their associated level of confidence will be stored in a separate table in the processed MySQL database and will also be made available as supplementary triples in the online RDF/XML dataset. Metadata regarding the transcription, digitization, and processing of all data, such as the date the data were created and the framework used for processing, will be recorded and stored alongside the data in dedicated MySQL tables.

Data Storage and Preservation of Access

Indiana University provides cloud storage systems through the Research Technologies division of IU Information Technologies Services as well as open access repository services through the University Libraries. The initial Excel files, the preprocessed database, and the processed database will all be stored in IU's Research File System (RFS). RFS data are regularly backed up and stored in physically secure environments in Bloomington and Indianapolis, ensuring robust data longevity. While the project is being built and expanding, access to the files may be granted to researchers beyond the project team. The data will be managed and synced in consultation with IU Data Management Service**.** For long term preservation, the database will be deposited and made accessible through the open access IUScholarWorks Repository. The IU Webserve publishing service, providing server space and scripting environments, will host the public website and online SPARQL endpoint.

Data Dissemination

The preprocessed and processed MySQL databases will be made available in the form of downloadable MySQL dumps on the project website. Both of these databases grow with the input of transcribed **MB** data; as a result, all versions of the database will be listed on the website alongside timestamps and basic content statistics. The RDF/XML mirror of the processed database will be available as downloadable triples and accessible via a SPARQL endpoint to the latest version. All code developed for processing and analysis of the data will be available with extensive documentation on the project's GitHub repository.

Property Rights, Ethics, and Privacy

There are no copyright issues for this project nor is protection of human subjects of concern, because all of the persons identified as being born in the last year of the metrical books examined (1918) or before are more than likely to be dead.