Austro-Hungarian Maps Dataset Data Processing Plan

Section A: General

A1: Goals of experiment

To make available an experimental project from the 2015 Library of Congress GIS Research Fellows program, accessible in a format that is both comprehensive and more easily digestible for computational and non-computational re-use.

A2: Describe the scope of the intended workflow or pipeline.

This dataset was originally developed as part of a 2015 project of the Library of Congress GIS Research Fellows program. Paper materials were digitized, georeferenced, and a derivative shapefile was generated for quick display of each sheet's georeferenced bounding box. The set was then lightly refined for public dissemination in 2022. Details of methods can be found below and in the dataset's README.md file.

A3: Data delivery format and specifications for data generated in the experiment.

The following file formats are used in this dataset:

- TIFF, uncompressed (.tif) - Uncompressed TIFF files are used to store the original images of the paper map sheets.
- GeoTIFF, uncompressed (.tif) - GeoTIFF files are used to store the georeferenced sheet images. These are TIFF files with additional geospatial metadata added to their headers.
- Esri shapefile - A single "shapefile" is comprised of multiple files. The shapefile in this dataset is comprised of one each of .shp, .shx, .prj. and .dbf files.
- CSV (Comman Separated Values, .csv) - Descriptive metadata files are included as CSV files. The dataset's README.md file contains data dictionaries for column headings.
- Markdown (.md) - Documentation files are included as .md files. Markdown files use simplified notation to create webpages that can also be more easily opened in plain text

editors. Although you can read .md files in simple text editors, the .md files in this dataset include notation that is meant to be rendered in webpages. If the files are being downloaded directly, the best way to view them is to open them in any markdown editor. Many free markdown editors exist online.

A4: Description of intended use.

The data package will be made publicly available through a S3/Cloudfront distribution on data.labs.loc.gov.

Section B: Data Doumentation

B1: Description of dataset.

Title of dataset

Austro-Hungarian Maps Dataset

Composition

- *1 cover-sheet.md*: a markdown file that describes the original map set from which this dataset is derived
- *1 README.md*: a markdown file that describes the dataset, including details of its production and interpretation
- *4,998 digitized map sheet images*, not georeferenced (TIFF image format)
- *4,877 georeferenced map sheet images*, with the map collars removed (GeoTIFF image format)
- *3 digitized index map sheets*, not georeferenced (TIFF image format)
- *3 digitized legend sheets*, not georeferenced (TIFF image format)
- *1 shapefile*: represents the bounding boxes of all the georeferenced TIFFs (comprised of 4 files)
- *1 metadata.csv*: a CSV file containing the metadata for all images in the dataset
- *1 manifest.txt*: a text file listing the image id, MD5 hash, and location of the images in the dataset

Provenance

This dataset has a layered provenance, outlined below.

- **Dataset preparation**: The dataset itself was prepared with minimal refinements from a 2015 project of the Library of Congress GIS Research Fellows program. More details on these refinements can be found in the "Preprocessing steps" section below and in the datasets's README.md file.

- **Digitization**: In 2015, all sheets held at the time by the Library of Congress were digitized and are included in this dataset's 4,998 digitized map sheet images. Various legend and index sheets were also digitized and included in the set. The sheets were then georeferenced using methods described in the dataset's README.md. This work was completed under the Library of Congress GIS Research Fellows program with contributions from John Hessler, Amanda Brioshe, Erin Kelly, Evan Neuwirth, and Michael Schoelen.

- **Original materials acquisition by Library of Congress**: The Library of Congress has a large but incomplete copy of this map set, meaning that some sheets are missing from the collection. The set was collected over time from various sources, rather than as a one-time acquisition from a single source. The collection was generally acquired prior to the 21st century, and may have been acquired by international exchange, U.S. government deposit, donation, or other sources. Although additions may continue to be made to the physical collection in the future, this dataset is a point-in-time capture of the set as of 2015. Additional provenance information may be found on markings on the original sheets. For example, some sheets will have stamps from previous custodians, such as the U.S. Army Map Service. Most sheets will also have a Library of Congress stamp that indicates the date of processing. Note that this date may not align with the sheet's original acquisition (particularly for historically-acquired materials).

- **Original materials production**: The map set is a composite of multiple sets produced by varying government entities over time. The set was initially prepared and issued by the Austro-Hungarian Monarchy's Milita■rgeographisches Institut beginning around 1875, as part of its Third Military Survey. After the dissolution of the Austro-Hungarian empire in 1918, parts of the set were continued by successive governments including Third Reich Germany. Thus, the sheets represented by this dataset had multiple creators over time, who would have had varying schedules for production. Some sheets were wartime products. See also the Content Advisory, found in the data package's README.md and cover-sheet.md files.

Compilation methods

For the dataset preparation, the source data was taken directly from the output of the 2015 Library of Congress GIS Research Fellows program. No sampling methods were used.

Preprocessing steps

In preparation for the public release of the dataset in 2022:

1. Filenames were standardized for clarity.
2. The shapefile attribute table was updated with standardized filenames and re-exported using Esri's arcgis Python library.
3. The "sheet_img" files were cropped and deskewed (rotated) for clarity of presentation.
4. 104 images in the "geo" directory that used approximate transformations stored in Esri .tfwx and .aux files, rather than GeoTIFFs, were removed.
5. A metadata.csv file was generated, based on metadata in filenames.

Additional details of the filenaming scheme and of the original methods used for the 2015 project can be found in the dataset's README.md file.

Potential risk to people, communities, and organizations and strategies for risk mitigation.

The data package provides the following Content Advisory, found in the README.md and cover-sheet.md files:

Maps, like other archival materials, are produced from the particular socio-political perspective of their creators, expressed through the use of labels, choice of language, drawing of borders, and other visual symbology. Maps can contain contested land claims, offensive place names, and partial perspectives on political and cultural conflict. The historical maps contained in this dataset were initially prepared and issued by the Austro-Hungarian Monarchy's Milita■rgeographisches Institut beginning around 1875. After the dissolution of the Austro-Hungarian Empire in 1918, parts of the set were continued by successive governments. Some portions of this set were produced by Third Reich Germany in the 1930s and 1940s during and after annexing areas to its south and east. Most of this set is produced by various governmental military mapping agencies.

The regions covered by this set of maps are linguistically and culturally diverse. Generally, place names in the set are identified by their German cognates, but they are also known to sometimes be in Hungarian or Slavic languages.

How will the experiment team address outdated or potentially offensive terms or elements of data that may be harmful if encountered by human users?

The content of the original maps has not been altered except to crop for georeferenced display. Please also refer to the Content Advisory above.

Copyright, licensing, rights, and/or privacy restrictions