

# Historical Map Annotations for Text Detection and Recognition

Archan Ray<sup>1</sup>, Ziwen Chen<sup>2</sup>, Ben Gafford<sup>2</sup>, Nathan Gifford<sup>2</sup>, Jagath Jai Kumar<sup>1</sup>, Abyaya Lamsal<sup>2</sup>, Liam Niehus-Staab<sup>2</sup>, Jerod Weinman <sup>\*2</sup>, and Erik Learned-Miller<sup>1</sup>

<sup>1</sup>College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, Massachusetts 01003

<sup>2</sup>Department of Computer Science, Grinnell College, Grinnell, Iowa 50112

October 2, 2018

## 1 Introduction and scope

This document describes a data set designed for testing the performance of text/graphics separation and character recognition algorithms on text in scanned historical map images. Thirty one maps from the nineteenth and early twentieth centuries (1866–1927) were chosen from nine atlases in the David Rumsey Map Collection.<sup>1</sup> Most maps are of individual states, though some are regional and one is of the entire U.S.; many are engraved with occasional handwritten text. The original MrSid files are converted into uncompressed TIFF images. The images and annotations are available from the following URLs.

**MrSid Images** doi:11084/10473

**TIFF Images** doi:11084/19499

**JSON Annotations** doi:11084/23296

## 2 JSON data format

Each map is annotated in a JSON file. The top-level structure is an array of *words*, each of which is a dictionary containing two keys: `text` and `items`. The value for `items` is an array of one or more *segments*. A segment is itself a dictionary of two keys `text` and `points`.

The value of `points` is an array of pairs of  $(x,y)$  (or equivalently, `column,row`) coordinates of the bounding polygon for the image region containing the string represented by the `text` value. The first point in the `points` array corresponds to the bottom-left point of the region as given by reading order (left-to-right). The points then continue in a counter-clockwise ordering (beginning by following the baseline). Any number of points may be given to indicate the bounding polygon.

The `text` of the *word* is simply the concatenation of the *segments* constituting it.

Files are UTF-8 encoded, so that non-ASCII symbols appearing in the maps can be fully represented. (The most common example being from the copyright notice.)

The following gives a simple example (excerpted from map D0042-1070007) of a single word consisting of a single segment:

```
[ { "text": "Grinnel",
  "items": [ { "text": "Grinnel",
    "points": [ [4103, 4023],
                [4235, 4024],
                [4235, 3995],
                [4103, 3994] ] } ] } ]
```

The thirty-one annotated maps have a total of 33,868 segments in 32,659 words.

---

\*Corresponding author, [jerod@acm.org](mailto:jerod@acm.org)

<sup>1</sup><http://davidrumsey.com>

### 3 Annotation methods and criteria

Generally speaking, a *word* is a character string without a space. Some *words* are divided into multiple *segments* when separated by a large tracking or other layout factors (e.g., appearance across the center fold or book binding or bisected by a large highway). In particular, when a word is crossed by other segment's text (i.e., between its characters), the individual characters are annotated as individual segments of the word. If a place name was hyphenated to clearly indicate a break (e.g., "Jacks-" and "onville"), these two segments are left in separate words. The segment(s) for any text where a hyphen forms part of a place name (e.g., "Winston-Salem") are collected in a single word.

Spelling in the `text` fields is intended to match that present in the original map (rather than modern geonomy). When there is ambiguity, the annotator compared to character exemplars from other known words in the same document, informed by known modern or historical geonomy to arrive at the best judgment of what the map actually says.

If a character is not visible, it is omitted from the ground truth text. However, most partially visible characters are included when they may be confidently inferred (e.g. by toponym geography). Segments with some other partially visible or otherwise illegible characters are marked as having unreadable text, as indicated by a null value for the `text` field.

Punctuation (particularly ".") has been included in the bounding polygons and transcriptions. However, character descenders are consistently excluded from bounding polygons, which typically track the main character baseline. The bounding polygon includes ascenders and stretches to encompass the left and right extents of the underlying transcribed text.

While every effort was made to use the Unicode representation of the underlying text, some text layouts of the period do not make this possible. In particular, abbreviations were often made with punctuation appearing directly beneath characters with elevated baselines (though not necessarily in a smaller font). For example, the word "Fort" might have been abbreviated as  $F^t$  on the map. The transcription places such "underdots" immediately after the character, so this would be transcribed as

```
"text": "Ft."
```

while a string appearing as  $3^{rd}$  would be transcribed

```
"text": "3r.d."
```

While some character shapes clearly indicated the raised character to be one of either the upper or lowercase form, other characters did not clearly allow it. The most common example being M<sup>c</sup>NALLY. Because the letter case of the raised "c" letterform was not clearly distinguishable, the lower case version was used in the transcriptions:

```
"text": "Mc.NALLY"
```

In other cases where a "small caps" type was used, the entire transcription is rendered in capital letters. For example HIGHWAY is transcribed


```
"text": "HIGHWAY"
```

### 4 Revision History

Ravi Chande, Dylan Gumm, and Jerod Weinman originally curated the collection of map images and published twelve annotated maps in 2013 (doi:11084/3246). The same authors also initially annotated the full set map images, which was subsequently revised for accuracy and consistency by Larry Boateng Asante, David Cambronero Sanchez, and Jerod Weinman (doi:11084/19349). While this collection annotates the same maps, the annotations have been made independently, only using the previous annotations to automate part of the verification process.

This version of the annotations represents verifications of and revisions to an initial version created by Archan Ray with assistance from Jagath Jai Kumar.

**Acknowledgments** Development of this resource was made possible with the support of the National Science Foundation under Grant Numbers 1526350 and 1526431.

 Copyright 2018. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. Sections 1 and 3 are derived from “A Data Set of Annotated Historical Maps” (Boateng Asante et al., 2017, doi:11084/10469) under CC-BY-NC-SA.