# A Data Set of Annotated Historical Maps

Larry Boateng Asante, David Cambronero Sanchez, Ravi Chande, Dylan Gumm, and Jerod Weinman
Department of Computer Science
Grinnell College
Grinnell, Iowa 50112
jerod@acm.org

August 18, 2017

## 1 Introduction and scope

This document describes a data set designed for testing the performance of optical character recognition algorithms on text in scanned historical map images. Thirty maps from the nineteenth and early twentieth centuries (1866–1927) were chosen from nine atlases in the David Rumsey Map Collection.[1] We focus on maps of the United States because the U.S. Board of Geographic Names[2] maintains the Geographic Names Information Service (GNIS), a database of all officially-named geographic entities in the United States, both current and historical.[3] Most maps are of individual states, though some are regional and one is of the entire U.S.; most are manually typeset, with occasional handwritten text. The original MrSid files were converted into uncompressed TIFF images for annotation and recognition.

The images, annotations, and example processing code are available from

- `http://www.cs.grinnell.edu/~weinman/research/maps.shtml` and
- `http://digital.grinnell.edu` via `http://hdl.handle.net/11084/19349`.

## 2 XML data format

Maps are annotated in XML as four primary nested entities: `map`, `window`, `label`, and `word`.

Each XML file contains one `map` entity. The sole attribute of a `map` entity is `src`, which indicates the filename (but not path) of the TIFF image described by the XML file. The filename (sans extension) is the unique portion of the originating MrSid file's URL on the Rumsey website.

The `map` entity contains one or more `window` entities to separate the sets of `label`s for each geographic coordinate frame of reference present in the image (i.e., for images with an inset map).

A `label` entity annotation corresponds to a piece of recognizable text on a map. A `label` entity has a `text` attribute for the ground truth label, which could include the full name of a city (e.g., "Des Moines"), body of water (e.g., "Rock Creek"), county, or sometimes even the text in a map's legend. A map entry should at least have as many label entries as there are toponym labels on the map, though not all maps have been completely annotated (discussed further below).

Some maps have an additional `gnis_id` attribute of each `label` entity, which gives the numerical GNIS Feature ID associated with the item. When none can be determined, the value "-1" is used.

If a map text label is associated with an object at a specific image location denoted by the map, a `point_location` entity (with numerical attributes `x` and `y`) within the `label` annotation indicates the image coordinate of the label. Cities, mines, and mountain peaks often feature such a point location.

Each label's text is manually segmented into words. We annotate each space-separated word in the text of a `label` with a `word` entity. Each `word` entity is nested within a `label` and has a `text` attribute for the word string (which may have punctuation, but no spaces).

[1] `http://davidrumsey.com`

[2] The U.S. BGN is part of the federal U.S. Geographical Survey created to "maintain uniform geographic name usage throughout the Federal Government". See `http://geonames.usgs.gov`.

[3] GeoNames (`http://www.geonames.org`) maintains a worldwide equivalent of the GNIS.

Several entities nested within `word` contain the information necessary to create a tight bounding box around that portion of the image. A word's baseline is stored as a series of `point` entities within a single `baseline` entity. The order of these points follows the letters of the word in reading order from "left" to "right" (along the word's baseline curve), regardless of the word's absolute orientation on the map. The left and right word boundaries are represented as `point` entities inside `leftbound` and `rightbound` entities. Image coordinates used to measure x-height or capital letter height are stored in `x_height_points` or `caps_height_points` entities. This data is recast as a series of `x_height` and `caps_height` entities, summarized by `average_x_height` and `average_caps_height` entities, each with a single numerical `height` attribute. If no instances of capital or lowercase points can be found, the averages are given as `NaN` values.

All this data allows us to draw tight bounding boxes around words and eventually normalize them.

# 3 Annotation methods and criteria

Each map was manually annotated by first marking the label on the map, marking its point location if one existed, and finally marking and transcribing all of the words that made up that label. Each word was annotated individually with its baseline, lowercase letter heights, capital letter heights, its left and right boundaries, and the text of the word itself.

Linear baselines were marked under the first and last characters of words and curved baselines were approximated by marking multiple characters within the curve. Points for the lower and upper case letters were marked at the topmost point of those respective letters. The left and right boundaries were marked on the outermost pixel of the leftmost and rightmost letters.

The annotation program calculated the letter heights by projecting the selected points above a letter onto the baseline line segment closest to each point.

While `label` entities are designed to cover recognizable text, not all text has been labeled (e.g., some legends and mileages are excluded) but the goal was to annotate all toponyms.[4] As a result, the data may not be appropriate for text/graphics separation tasks (i.e., text detection).

Generally speaking, a `word` is a character string without a space. Conversely, `word`s within a `label` are to be separated by pixel distances interpreted as spaces (an intercharacter distance larger than the overall letter tracking of the label text). However, there may be some punctuation-driven deviations (see below on abbreviations). Conversely, some multi-character segments are gathered together into a single `word` when they represent a single token from the gazetteer, even though separated by a space much larger than tracking due to layout (e.g., appearance across the center fold or book binding or bisected by a large highway).

Spelling 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`.

Abbreviation punctuation (particularly ".") is inconsistently annotated in the `text` and inconsistently included in the bounding box (e.g., `rightbound`) of the `word`. As a result, it has not been a factor in evaluation (only alphanumerics have been processed). As a general rule, it seems that when the `rightbound` includes the period, it should appear in the `text` field.

When a placename word (in the colloquial sense) is intentionally hyphenated on the map and has its substrings appearing with different apparent baselines, these are usually labeled as two separate `word` entities belonging to the same `label`. The hyphen might not be included (as with the abbreviation punctuation).

Cities, towns, and similar placenames have several related GNIS entries, often belonging to different GNIS Feature Classes.[5] In particular, there may be a *Civil* entry for the City or Town, corresponding to the political or legal entity. More generally, there is typically also a *Populated Place* entry corresponding to the "area with...a permanent human population." Unincorporated towns may have such entries without a corresponding *Civil* entry. Another possibility is a *Civil* entry for a Township by the same name. These legal entities are typically subordinate to a county. When annotating the GNIS Feature ID, we prefer features belonging to the *Civil* Feature Class for a town or city first. If that does not exist, we use the *Populated Place* entry. When neither of those exist we resort to the *Civil* Township entry.

Further complications arise when a map marks a *Post Office* which existed before any other settlement. In such cases, we use the corresponding class precedence listed above, rather than utilize the *Post Office* GNIS entry. The exception

[4] Two maps have not had toponyms completely labeled: D5005-5028102 and D0006-0285025.

[5] USGS GNIS Feature Class Definition, `https://geonames.usgs.gov/apex/f?p=gnispq:8`

is when no settlement arose. Such cases are typically marked "(historical)" in the GNIS, and when no matching *Civil* or *Populated Place* entry can be found, we use the *Post Office* entry.

# 4 Example

The following is an abbreviated example containing one label for a map.

```
<map src="D0042-1070007.tiff">
  <window>
    <label text="Grinnel" gnis_id="467979">
      <point_location x="4063.9419" y="4014.7475">
      </point_location>
      <word text="Grinnel">
        <leftbound>
          <point x="4105.0485" y="4005.9115"></point>
        </leftbound>
        <rightbound>
          <point x="4226.4475" y="4007.4482"></point>
        </rightbound>
        <baseline>
          <point x="4118.11" y="4020.51"></point>
          <point x="4224.91" y="4022.05"></point>
        </baseline>
        <x_height height="11.0653"></x_height>
        <x_height height="10.6729"></x_height>
        <x_height height="12.0049"></x_height>
        <average_x_height height="11.2477"></average_x_height>
        <x_height_points>
          <point x="4139.62" y="4009.75"></point>
          <point x="4165.75" y="4010.52"></point>
          <point x="4204.93" y="4009.75"></point>
        </x_height_points>
        <caps_height height="22.2821"></caps_height>
        <average_caps_height height="22.2821"></average_caps_height>
        <caps_height_points>
          <point x="4118.11" y="3998.23"></point>
        </caps_height_points>
      </word>
    </label>
  </window>
</map>
```

# 5 Revision History

Chande, Gumm, and Weinman originally published twelve annotated maps in 2013. The same authors also initially annotated the full set given here, which was subsequently revised for accuracy and consistency by Boateng Asante, Cambronero Sanchez, and Weinman.

**Acknowledgments** Development of this resource was made possible with the support of Grinnell College, HHMI, and the National Science Foundation under Grant Number 1526350.