Rainer Simon*, Peter Pilgerstorfer, Leif Isaksen**, Elton Barker***

Towards semi-automatic annotation of toponyms on old maps

Keywords: map digitization; toponym detection; image processing

Summary: Present-day map digitization methods produce data that is semantically opaque; that is to a machine, a digitized map is merely a collection of bits and bytes. The area it depicts, the places it mentions, any text contained within legends or written on its margins remain unknown - unless a human appraises the image and manually adds this information to its metadata. This problem is especially severe in the case of old maps: these are typically handwritten, may contain text in varying orientations and sizes, and can be in a bad condition due to varying levels of deterioration or damage. As a result, searching for the contents of these documents remains challenging, which makes them hard to discover for users, unusable for machine processing and analysis, and thus effectively lost to many forms of public, scientific or commercial utilization. Fully automatic detection and transcription of place names and legends is, likely, not achievable with today's technology. We argue, however, that semi-automated methods can eliminate much of the tedious effort required to annotate map scans entirely by hand.

In this paper, we showcase early work on semi-automatic place name annotation. In our experiment, we utilize open source tools to identify potential locations on the map representing toponyms. We present how, in next steps, we aim to extend our experiment by exploiting the spatial layout of identified candidates to deduce possible place names based on existing toponym lists. Ultimately, or goal is to combine this work with a toolset for manual image annotation into a convenient online environment. This will allow curators, researchers, and potentially also the general public "tag" and annotate toponyms on digitized maps rapidly.

Introduction

Collections of high-resolution digitized old maps are increasingly being made available on-line, through initiatives such as those by major libraries or private collectors (e.g. the British Library,[1] the National Library of Scotland,[2] or the Institut Cartogràfic de Catalunya;[3] or the David Rumsey collection,[4] respectively), as well as through federated search portals such as OldMapsOnline.[5] However, present-day map digitization methods produce data that is *semantically opaque*; that is to a machine, a digitized map is merely a collection of bits and bytes. So while users may view the map once they have found it, searching for written content contained within the image – such as place names (toponyms) or legends – remains impossible, unless a human expert appraises the image and manually transcribes this information to its metadata.

* AIT – Austrian Institute of Technology, Vienna, Austria [rainer.simon@ait.ac.at]
** University of Southampton, Southampton, United Kingdom [l.isaksen@soton.ac.uk]
*** The Open University, Milton Keynes, United Kingdom [e.t.e.barker@open.ac.uk]

[1] http://www.bl.uk/maps/

[2] http://maps.nls.uk/

[3] http://cartotecadigital.icc.cat/cdm/

[4] http://www.davidrumsey.com/

[5] http://www.oldmapsonline.org

Since place names form the underlying semantic content of almost all geographic documents, the ability to identify them is essential in any attempt to work with, compare or interpret them. For early maps and geographic documents this ability is especially important, because while they rarely conform to standard geometries, they often provide the earliest attestations to towns, peoples, and other spatially localized phenomena. Tools, infrastructure and resources for collating, aligning, and exploiting toponyms in early maps and geographic documents would therefore have a broad and significant impact across a range of fields, including Archaeology, History, Classics, Genealogy and Modern Languages.

For modern printed documents, the challenge can be addressed to some extent by the use of Optical Character Recognition (OCR), which extracts machine-readable text from scans. However, state of the art OCR technology fails when faced with old maps, which are typically handwritten, may contain text in varying orientations and sizes, and can show varying levels of deterioration or damage. As a result, the actual content of these documents remains unsearchable, thus hard to discover for users, and unavailable to machine processing and analysis. Such content is effectively lost to many forms of public, scientific or commercial utilization.

In this paper, we showcase early work on the detection of possible toponyms on scanned old maps, using existing open source tools. Fully automatic identification and transcription of place names is, arguably, not achievable with today's technology. Our goal is therefore to devise semi-automated methods, which eliminate as much of the tedious manual effort required to annotate map scans entirely by hand as possible. In this paper, we provide a brief overview of related work, introduce our technical approach, and present example results produced with our first prototype implementation, as well as the major challenges encountered. We conclude the paper with an outlook on the *Pelagios 3* research project, one part of which will aim to refine the methods described in this paper.

Related Work

Toponym recognition in scanned maps is an area of active research. The vast majority of this work, however, focuses on contemporary maps. Cao and Tan (2002), for example, present an approach that separates text and graphics in scanned maps, and subsequently feeds the extracted text into state of the art OCR software for toponym identification. In similar work, Velázquez and Levachkine (2003) propose a refined approach to enhance separation between overlapping text and graphics, including curvilinear text. Pouderoux et al. (2007) present an automatic method for extracting the toponym layer from scanned maps, based mainly on image segmentation and connected component processing. This method includes *empirical filtering* steps that prune and correct intermediate results. It also relies on state of the art OCR software for the final processing step. Chiang and Knoblock (2010) present an improved OCR-based approach, with enhancements regarding the detection of toponym orientation and separation of overlapping labels.

Our research shows that prior work that specifically concentrates on old maps is scarce. Most closely related to our use case is the work of Weinman (2013), who presents a word recognition system that was tested on a collection of 19th century U.S. state and regional maps. His work focuses on the alignment between toponym images on geo-referenced, pre-annotated maps (i.e. maps where the toponym locations have been marked up a priori) and place names from a gazetteer. As Weinman does not go into the details of how the toponym images are being annotated, his work is complementary to ours.

Technical Approach

The goal of our initial experiments was to research approaches for annotating maps, which identify potential toponyms in terms of their *location* and *extent*, as well as in terms of their *orientation* on the map image. To identify possible toponym candidates, we experimented with a sequence of processing steps on a set of sample maps. We continuously refined the steps and tuned their parameters with each map type, incrementally. The three key phases that now form our processing workflow are: (i) background-foreground segmentation, (ii) feature detection and (iii) feature linking. Image processing operations were implemented using *OpenCV*,[6] a general-purpose computer vision toolkit.

Background-Foreground Segmentation

The first processing phase, background-foreground segmentation, generates a black-and-white mask image, which separates the "background" areas of the map from the "foreground" which is used for further processing. This step is the most crucial one as far as the quality of the final result is concerned. Unfortunately, it is also the step whose results vary most widely, depending on the type of map used. This made a certain amount of manual pre-selection and tuning necessary. In this manual step, we either defined certain color ranges to be treated as background colors; or built a background mask by applying a strong median filter to the map. Median filtering has the effect of eliminating thin structures from the image (such as fine lines and text), while keeping the overall color distribution of the map intact. Subtracting the filtered image from the original then yields a good base mask for background segmentation.

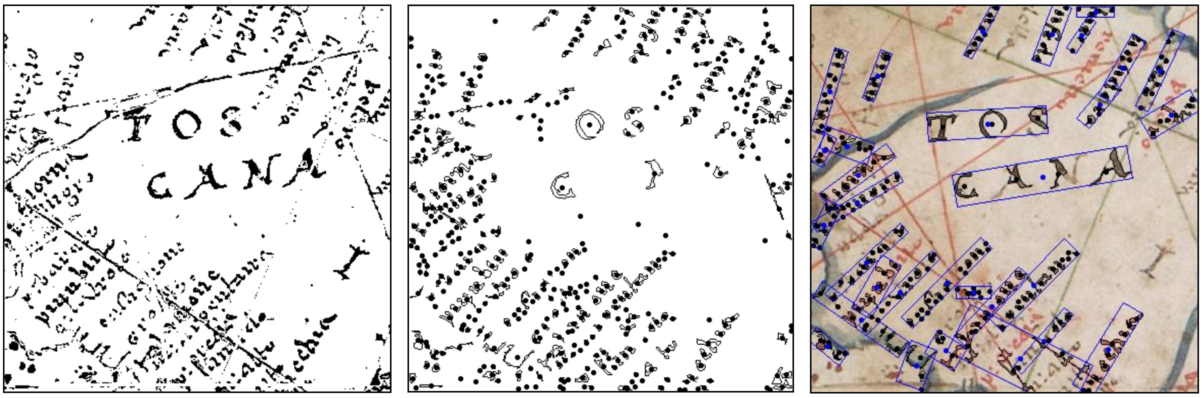

Figure 1: processing phases: background-foreground segmentation (left), detected contour features (center), features linked to "feature groups" denoting toponym candidates, overlaid on top of original image (right).

After initial color-based segmentation, we cleaned up the mask for further processing by removing lines (graticule, rhumb lines) via Hough transform filtering; eliminated parts with low color gradient (because text usually has strong edges); and then employed morphological image processing operations, which are well-known to lessen the effects of imperfections introduced through color-threshold masking (e.g. by dilating and closing objects eroded as a consequence of color masking). For an example output from this phase, refer to Fig. 1, left.

[6] http://opencv.willowgarage.com/

Feature Detection

After generating the mask image, the next phase locates and characterizes *features* – in our case, connected objects – on the foreground image. An approach that provided good results was to employ an algorithm that detects contours. The advantage of this approach is that it is computationally relatively light-weight. The drawback, however, is that it produces errors whenever two objects flow into each other (e.g. in case of a toponym merging with a line segment of similar brightness). So these situations need to be resolved as much as possible in the background-foreground segmentation step. Additionally, we introduced rules to filter out invalid objects based on heuristics concerning: covered area, width, or aspect ratio. Example output from the feature detection phase is shown in Fig. 1, center.

Feature Linking

Feature detection will not identify toponyms directly, but individual connected objects on the image. Toponyms will usually consist of any number of features. The next phase therefore post-processes the detected features, so that they are linked to groups that likely represent a single toponym. Our general strategy to feature linking is based on a set of empirical constraints and heuristics. Pairs of features that satisfy the constraints are assigned to the same group (and, hence, considered to belong to the same toponym). Constraints include thresholds on:

- distance: how does the distance between two features' centroids compare to the distance between their bounding boxes?
- increase in area: is the area of the bounding box that encloses both features together within specified limits, as compared to the area covered by the two separate bounding boxes for the features individually?
- direction: does the linking lead to an orientation that is plausible? Either compared to a pre-set standard orientation or toponyms already detected nearby? Do some directions lead to more favorable configurations? (I.e. is there a direction where a specifically high number of features satisfy the linking constraints?)
- absolute bounds on width and height.

The linking algorithm is run iteratively, with toponyms being "grown" out of individual features by successive linking.

Preliminary Results

We conducted trials of our approach on a set of sample maps, chosen according to (subjectively) three different levels of technical complexity:

- A section of a Ptolemaic map of the British isles[7] which, for the purposes of toponym detection, represents the least challenging case because of a uniform background.
- A sheet from a mid-18th century Austrian surveying campaign, which was more challenging due to a low separation between toponyms and background.
- A section taken from a 17th century portolan chart,[8] which was chosen as the most challenging example, featuring areas of low color contrast and bad readability, as well as re-

[7] http://www.bl.uk/onlinegallery/onlineex/unvbrit/p/001hrl000007182u00060vrb.html

gions with a high number of (technically) problematic structures such as intersecting lines and ornamental features.

A larger study of our approach, and a detailed analysis of the results, is still outstanding. But our initial test yielded good results for the first sample map: this map contained 41 toponyms. 38 of those were located correctly, although there was a slight misalignment of the detected bounding box on two toponyms. As for the three remaining toponyms: two of those were erroneously identified as a single, merged toponym; the last toponym (cf. Fig. 2, left, "ALVION INSVULA BRITANNICA") was erroneously split into two feature groups. On the most challenging sample (cf. Fig.2, right), we could visually discern 323 toponyms for places (in small font) and 10 for regional areas (in large font). The test produced 532 possible detections, with a recall of approx. 50%, and a precision of approx. 31%.

For detailed inspection, we have made the output of this test available online as interactive Web presentations: the full-resolution sample maps, with overlaid annotations can be found at http://rsimon.github.io/toponym_identification.

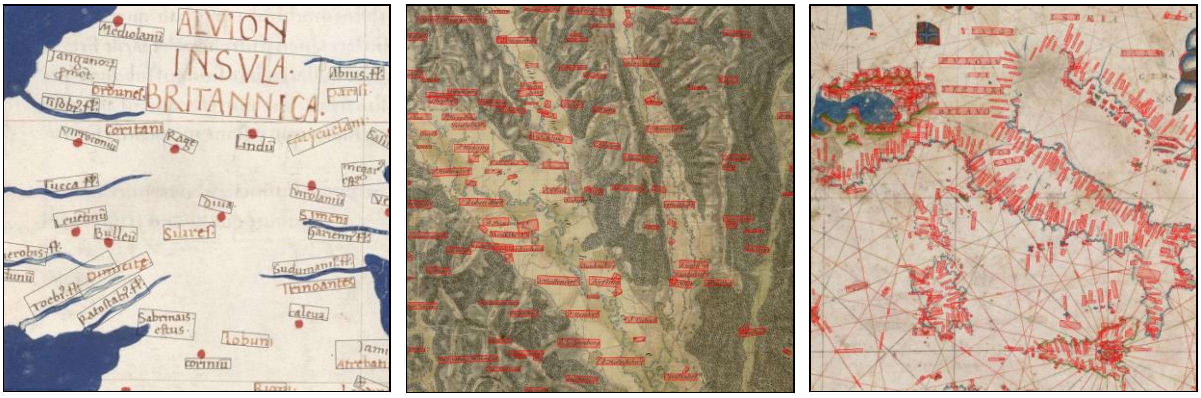

Figure 2: detected toponyms: 15th century map of the British Isles (left), 18th century Austrian land survey map (center), 17th century portolan chart (right). Full resolution available at http://rsimon.github.io/toponym_identification

Challenges & Possible Improvements

As mentioned previously, our approach requires human intervention. Additionally, it faces the challenge that parameters have to be tuned towards a certain optimum average toponym size. This limitation was not particularly severe in the case of our current sample maps: these would typically have toponyms in two font sizes – one standard, small size for individual places, and a large size for regions and larger areas (compare Fig. 1 and 2). Both sizes we resolved in two separate processing passes. For more general use, however, this limitation is certainly one that needs to be addressed in the future. A further insight gained from our experiments concerns some typical errors scenarios we observed recurrently. A more thorough examination of these errors will also have to be postponed to future work. Nonetheless we would like to list some exemplary error scenarios here:

- **Ornament irritation.** Symbols and decorative elements that have structures in size and density (and color) similar to toponyms frequently cause false positive detections. We expect that additional heuristics may be able to alleviate this problem, as these false detec-

[8] Salvator Oliva, Mediterranean. HM 2515. PORTOLAN ATLAS. Marseilles, 1619. http://commons.wikimedia.org/wiki/File:Salvator_Oliva._Mediterranean._HM_2515._PORTOLAN_ATLAS._Marseilles,_1619.B.jpg

tions usually exhibited different clustering and overlap behavior. An example is shown in Fig. 3, top left.

- **Line bleed.** Toponyms that intersect with (or are located nearby) lines, either those denoting geographical features, or graticule or rhumb lines, can distort the recognition result. Two examples can be seen in Fig. 3, top right: the detected bounds for the 5th toponym from the left are misplaced due to the coastline. Slightly further right, a false detection was caused due to another segment of coastline. We expect that proper tuning of processing parameters may be able to lower the number of such errors somewhat. However, it is unlikely that they can be avoided altogether. Human verification (combined with a good user interface for rapid correction) is possibly the only way to address this challenge.

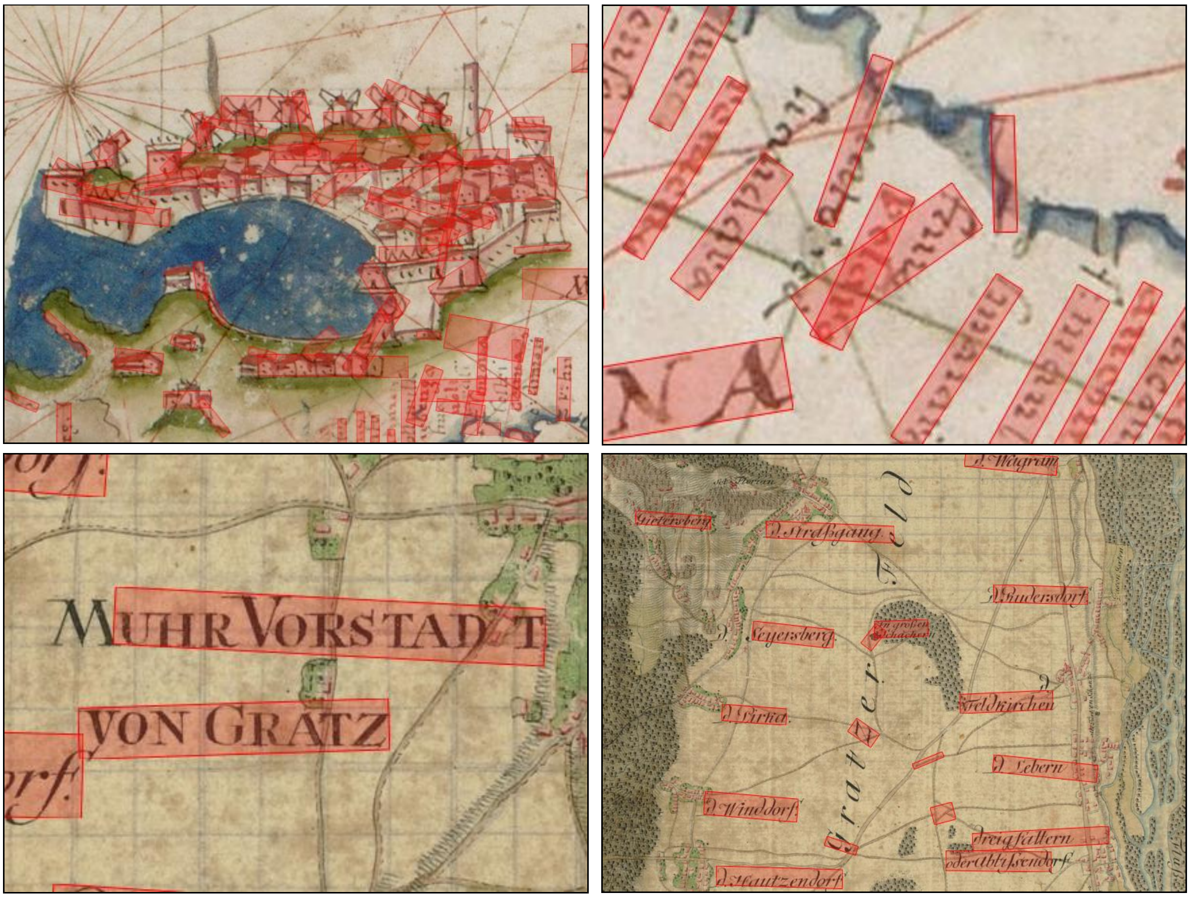

Figure 3: common error situations encountered during experiments: irritation due to symbols and ornamental features (top left); irritation due to line features flowing into toponyms, and “toponym cross-talk” (top right); split toponyms (bottom left), undetected large-area toponym (bottom right).

- **Toponym crosstalk.** Especially in the presence of distracting elements such as lines, our heuristics would erroneously lead to toponym bounds that run across two actual, neighboring toponyms. An example of this can also be seen in Fig. 3, top right (center of the image, the diagonal bounding box crossing over another bounding box). Like in the case of errors caused by line bleed, it is unlikely that these can be avoided. However, in cases where they cause overlap, they can at least be detected, and flagged to a human operator for verification.
- **Split toponyms.** Our current processing approach does not specifically deal with toponyms that are split across multiple lines. An example is shown in Fig. 3, bottom left.

- **Large area & curvilinear toponyms.** Likewise, our heuristics are ill-suited to detect toponyms that cover large areas, which are oriented significantly different from other toponyms on the map, or which run along a curved baseline. An example for a large-area case is shown in Fig. 3, bottom right (the toponym running from bottom to top).

In general, we expect that the amount of manual tuning and intervention can be further reduced by refining the processing workflow. Nonetheless, we don't expect that toponym identification on old maps can be fully automated any time soon. Therefore, we also plan to prototype user interfaces and graphical tools that can help non-technical users to easily experiment with different filter settings and combinations, by providing instant visual feedback on their actions.

Future Work: The Pelagios 3 Project

This paper presented early work on detecting the number, location, extent and orientation of toponyms on digitized old maps. We intend to build upon this work as part of the upcoming *Pelagios 3* research project.[9] Overall, this project aims to:

- provide an index of toponyms attested in a large corpus of early geospatial documents (maps and geographic writing), and the places they refer to
- create an open toolset that allows the scholarly community to enhance and refine the index incrementally, by annotating for themselves toponyms in further historical sources, as and when they are digitized.
- develop a freely available "analysis workbench" that will enable researchers to bring together spatial documents in new and innovative ways, e.g. to conduct visual and statistical comparisons between properties of different collections, or geospatial documents.

In addition to continuing the work outlined in this paper, we also intend to investigate how we can actually identify – rather than just locate – toponyms (semi-) automatically. To this end, we intend to leverage the Pelagios place index. The index itself will build upon existing toponym lists and digital historical Gazetteers, such as *Pleiades,*[10] the *China Historical GIS,*[11] and *A Vision of Britain Through Time.*[12] A further source of toponyms for Pelagios 3 will be the significant work on portolan chart toponymy undertaken by Pujades (2007) and Campbell (2012). We envision a system where a human user is annotating by hand, and is supported by a "recommender system" that suggests possible toponyms from the index, by exploiting knowledge about nearby places and their spatial arrangement. An early version of a similar system has been presented in previously by Simon (2011a) and Simon (2011b). *Pelagios 3* is due to commence in September 2013. It will continue for two years, supported by grants from the Andrew W. Mellon Foundation.

[9] http://pelagios-project.blogspot.co.uk

[10] http://pleiades.stoa.org

[11] http://www.fas.harvard.edu/~chgis/

[12] http://www.visionofbritain.org.uk/

References

Cao, R. and Tan, C. L. (2002). Text/Graphics Separation in Maps. In *Graphics Recognition Algorithms and Applications*, Lecture Notes in Computers Science, vol. 2390, Springer Berlin Heidelberg: 167 – 177.

Velázquez, A. and Levachkine, S. (2003). Text/Graphics Separation and Recognition in Raster-Scanned Color Cartographic Maps. In *5th International Workshop on Graphics Recognition Algorithms and Applications (GREC2001)*, Barcelona, Catalonia, Spain, July 2003.

Pouderoux, J., Gonzato, J.-C., Pereira, A. and Guitton, P. (2007). Toponym Recognition in Scanned Color Topographic Maps. In *Proceedings of the 9th International Conference on Document Analysis and Recognition*, vol. 1. IEEE Computer Society, Washington, DC, USA: 531-535.

Chiang, Y. and Knoblock, C.A. (2010). An Approach for Recognizing Text Labels in Raster Maps. In *20th International Conference on Pattern Recognition (ICPR 2010)*: 3199 – 3202.

Weinman, J. (2013). Toponym Recognition in Historical Maps by Gazetteer Alignment. In *International Conference on Document Analysis and Recognition* (ICDAR).

Pujades i Bataller, R. J. (2007) *Les cartes portolanes: la representació medieval d'una mar solcada*. Barcelona: Institut Cartogràfic de Catalunya.

Campbell, T. (2012) The Introduction and Abandonment of Toponyms on Portolan Charts 1300 to 1600 http://www.maphistory.info/ToponymyMenu.html (online reference, last accessed July 17, 2013).

Simon, R., Haslhofer, B., and Jung, J. (2011a). Annotations, Tags and Linked Data. Metadata Enrichment in Online Map Collections Through Volunteer-Contributed Information. In *e-Perimetron* vol. 6, no.3: 129-137.

Simon, R., Haslhofer, B., Robitza, W., Momeni, E. (2011b). Semantically Augmented Annotations in Digitized Map Collections. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL '11)*. ACM, New York, NY, USA: 199-202.