Querying Historical Maps as a Unified, Structured, and Linked Spatiotemporal Source (Vision Paper)

Yao-Yi Chiang
University of Southern California
Spatial Sciences Institute
3616 Trousdale Parkway, Los Angeles, CA 90089-0374
yaoyic@usc.edu

ABSTRACT

Historical spatiotemporal datasets are important for a variety of studies such as cancer and environmental epidemiology, urbanization, and landscape ecology. However, existing data sources typically contain only contemporary datasets. Historical maps hold a great deal of detailed geographic information at various times in the past. Yet, finding relevant maps is difficult and the map content are not machine-readable. I envision a map processing, modeling, linking, and publishing framework that allows querying historical map collections as a unified and structured spatiotemporal source in which individual geographic phenomena (extracted from maps) are modeled with semantic descriptions and linked to other data sources (e.g., DBpedia). This framework will make it possible to efficiently study historical spatiotemporal datasets on a large scale. Realizing such a framework poses significant research challenges in multiple fields in computer science including digital map processing, data integration, and the Semantic Web technologies, and other disciplines such as spatial, earth, social, and health sciences. Tackling these challenges will not only advance research in computer science but also present a unique opportunity for interdisciplinary research.

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous—*Geographic Information Systems*; H.2.8 [**Database Management**]: Database Applications—*Spatial Databases and Geographic Information Systems*

General Terms

Algorithms, Design

Keywords

historical maps, data integration, digital map processing, semantic web, GIS, historical GIScontent

SIGSPATIAL'15 November 03 - 06, 2015, Bellevue, WA, USA
Copyright is held by the author. Publication rights licensed to ACM.
ACM 978-1-4503-3967-4/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2820783.2820887.

1. VISION

Historical spatiotemporal datasets and historical Geographic Information System (GIS) support a variety of studies such as cancer and environmental epidemiology, urbanization, and landscape ecology (e.g., [11]) but existing data sources (e.g., gazetteers) typically contain only contemporary information. Historical maps are a great source of geographic information in the past (e.g., historical place names, landmarks, and transportation networks) and are often the only source that provides professionally surveyed historical data. Today, map archives such as the USGS (United States Geological Survey) National Geologic Map Database,[1] USGS Topographic Maps,[2] David Rumsey Map Collection,[3] OldMapsOnline.org,[4] and the National Library of Scotland,[5] store a large amount of historical maps in either paper or scanned format. However, only a small portion of these historical maps is georeferenced and event fewer of them have machine-readable content or comprehensive metadata. This prevents the maps from being indexed and searched and limits the opportunity for both researchers and the general public to access valuable historical information.

Even with the recent advance in map processing techniques [5], making a large number of historical maps searchable (by keywords, locations, and time) and their content accessible in an analytic environment (e.g., in a GIS) is still prohibitively expensive and time consuming. As a result, studies that require geographic information in the past often approximate historical information using contemporary datasets. For example, the Yellow-Star Houses project identified 1,944 designated compulsory residences in Budapest from historical decrees (circa 1944), geocoded these historical addresses with contemporary street datasets, and mapped them on Google Maps. While Google Maps provide convenient visualization tools, a historical map can contribute richer geographic information in the past, such as nearby transportation hubs at the time (Figure 1).

Studies that require accurate historical information are usually limited to process only a few historical maps and examine a small area or a short time period for which manual data curation is possible. Beattie [2] created a three-dimensional historical topography of the Ballona Creek watershed (Marina del Rey, California) from two historical USGS topographic maps (circa 1896 and 1902). This histor-

[1] `http://ngmdb.usgs.gov/ngmdb/ngmdb_home.html`
[2] `http://ngmdb.usgs.gov/maps/TopoView/`
[3] `http://www.davidrumsey.com/`
[4] `http://OldMapsOnline.org`
[5] `http://maps.nls.uk`

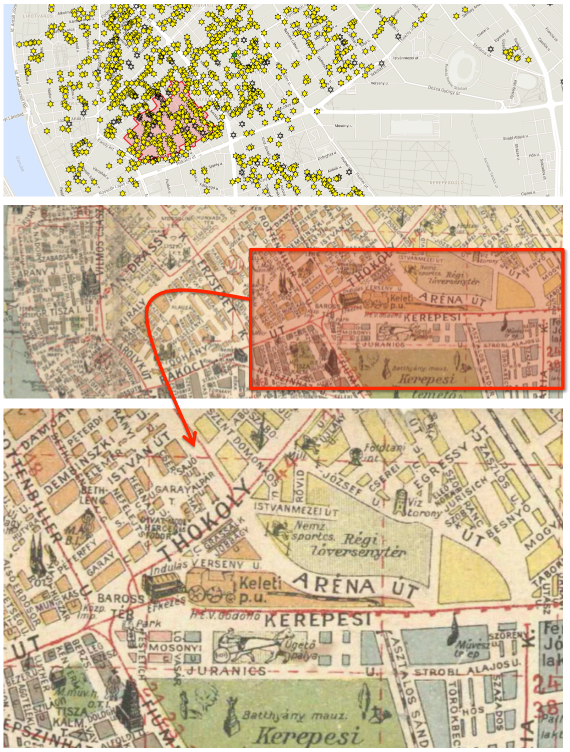
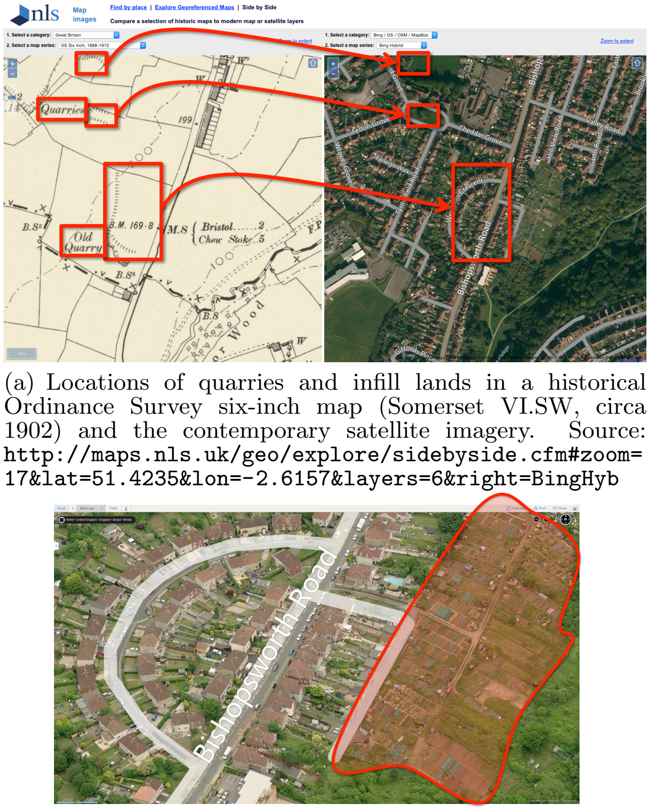


Figure 1: Top: Yellow-star houses shown on Google Maps. Sources: `http://www.yellowstarhouses.org`. Middle and Bottom: A section of a historical map of Budapest (circa 1941) shows the transportation hub. Source: `http://riowang.blogspot.com/2011/02/after-siege.html`

ical topography enables environmental planners to compare historical and current conditions of the Ballona Creek watershed to identify landscape changes. Kurashige [13] used historical census data, voting records, and precinct boundaries extracted from a 1920 map to study "who" (e.g., occupations and political parties) in Los Angeles voted for the 1920 California Alien Land Law that discriminates against Japanese.

I envision a map processing, modeling, linking, and publishing framework that allows querying historical map collections as a unified and structured spatiotemporal source in which individual geographic phenomena (extracted from maps) are modeled with semantic descriptions and linked to other data sources (e.g., DBpedia). These semantic descriptions will capture the intra-relationships between geographic phenomena within a map (e.g., an infill land near a quarry could be a waste disposal site) and the inter-relationship between historical map data and the huge amount of Linked Data already published on the Internet. This framework will make it possible to efficiently study historical spatiotemporal datasets on a large scale (both in time and space) and solve problems that cannot be easily answered now.

Figure 2 shows an example use case if the proposed framework is successful. Figure 2(a) shows a portion of a historical Ordinance Survey six-inch map (circa 1902) and a contemporary satellite imagery of the same area (Bristol, United Kingdom). The historical map shows two quarry locations and infill lands (the red rectangles). Quarries are a common pollution source (at which the polluted materials could be dumped at nearby infill lands). This type of contamination could make the soil not suitable for growing editable plants. Figure 2(b) shows a current farming area on the potential contaminated land (the red polygon). The question at hands is whether or not it is safe to grow grapefruits in this area. The modeled historical spatiotemporal data from the proposed framework will support the following reasoning process. Possible contamination materials from a 1902 quarry are heavy metals M. The infill land in the historical map is modeled as a probability surface of M (i.e., the target region). The accumulated rain precipitation over the target region from 1902 to 2015 is R. The main soil type of the target region is S. The probability of that the target region still contains M given the probability surface, R and S is so low so the top soil can be used to grow editable plants. Since grapefruit trees have a shallow root system, growing a grapefruit tree is safe in the target area.[6]


nls Map images
Find by place | Explore Georeferenced Maps | Side by Side
Compare a selection of historic maps to modern map or satellite layers
Quarries
Old Quarry
Bristol
Chew Stoke
Bishopsworth Road

(a) Locations of quarries and infill lands in a historical Ordinance Survey six-inch map (Somerset VI.SW, circa 1902) and the contemporary satellite imagery. Source: `http://maps.nls.uk/geo/explore/sidebyside.cfm#zoom=17&lat=51.4235&lon=-2.6157&layers=6&right=BingHyb`

(b) Current farming areas (approximated by the red polygon) that could affected by the historical contamination site.

Figure 2: Using historical maps to identify historical contamination sites

Realizing the proposed framework poses significant research challenges in multiple fields in computer science, including digital map processing, data integration, and the Semantic Web technologies. Next section explains the challenges and proposes future research directions for overcoming these challenges.

2. CHALLENGES

Figure 3 shows an example implementation of the proposed framework. The challenges for realizing this framework include three interrelated challenges: (1) how to make

[6] This is just an example. By no means the author is a soil contamination expert.

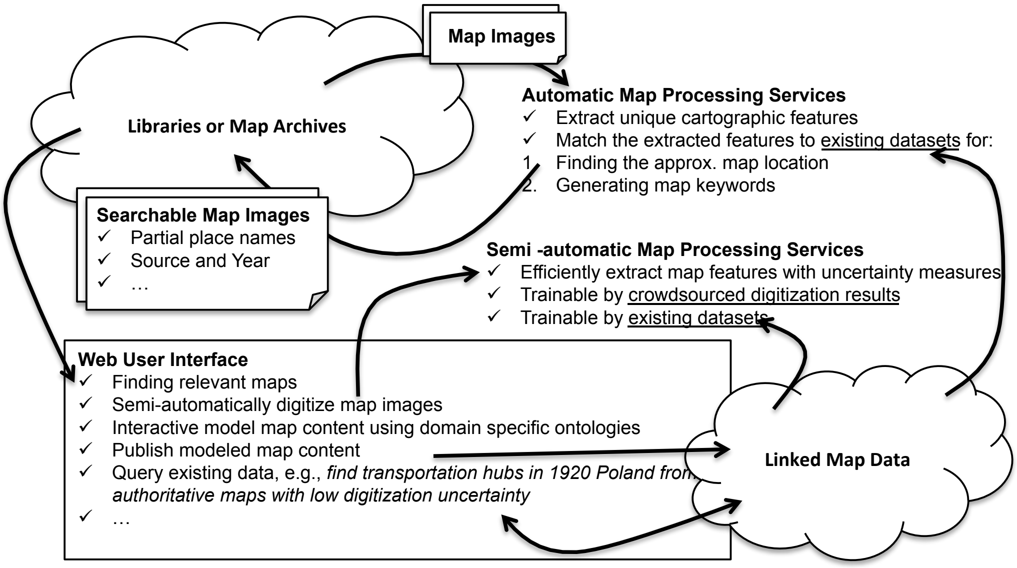

Map Images
Libraries or Map Archives
Automatic Map Processing Services
✓ Extract unique cartographic features
✓ Match the extracted features to existing datasets for:
1. Finding the approx. map location
2. Generating map keywords
Searchable Map Images
✓ Partial place names
✓ Source and Year
✓ …
Semi -automatic Map Processing Services
✓ Efficiently extract map features with uncertainty measures
✓ Trainable by crowdsourced digitization results
✓ Trainable by existing datasets
Web User Interface
✓ Finding relevant maps
✓ Semi-automatically digitize map images
✓ Interactive model map content using domain specific ontologies
✓ Publish modeled map content
✓ Query existing data, e.g., find transportation hubs in 1920 Poland from authoritative maps with low digitization uncertainty
✓ …
Linked Map Data

Figure 3: An example implementation of the proposed framework

historical maps easily searchable, (2) how to efficiently and accurately convert map content to machine-readable format and record provenance information and potential conversion errors in the process, and (3) how to generate semantic descriptions for historical maps, their content, and digitization related information (e.g., uncertainty) and link them to other data sources.

2.1 Finding relevant historical maps

Finding relevant maps for a solving the problem at hands is never easy.[7] This is because most of the historical maps in the libraries or map archives are usually just scanned images with limited metadata. The challenge here is how to automatically generate comprehensive metadata to support effective map searches. One possible solution is to automatically link map images to existing datasets to then use the linked datasets to generate map metadata. Weinman [16] built a text recognition approach that demonstrates linking historical map content to existing datasets can be achieved automatically. His approach recognizes text labels in maps to then match the recognized text to a gazetteer for georeferencing the maps. While this approach shows promising results with small-scale maps (for which most of map text exist in the gazetteer), many interesting maps are large-scale maps. Instead of matching text, another direction is to automatically extract distinctive geographic features (i.e., geometry) from maps (e.g., road intersections and contour lines [4, 14]) and then search for matching features in existing datasets to link map images to other datasets (e.g., [3]). The challenge for this feature-matching approach is twofold. First, the automatic approaches could be sensitive to the quality of the input map and hence would not generate enough distinctive features [6]. Second, the matching process could be computational expensive if the search space is large. A combination of text and feature matching could help to overcome these difficulties by providing different types of matching candidates and reducing the search space (e.g., only search for a match in the regions that contain at least 50% of the recognized text).

After the map is linked with other datasets, the next challenge is how to generate metadata that will enable effective map searches. Gelernter [8] demonstrated an automatic approach towards this goal. She developed a text mining approach to find maps in journal articles. Her approach also classifies the maps by years and themes using their companion text. This work shows that if we have enough data linked to the map images, we can generate a comprehensive set of metadata.

2.2 Converting map content to machine-readable format and record uncertainty

Once a user identifies the maps of interest, the next challenge is to efficiently convert the map content to machine-readable format. The mainstream approach for this purpose still heavily relies manual work with some help from raster-to-vector conversion software. Beattie [2] spent more than 70 hours on manual tasks for extracting contour lines from the two USGS historical maps (which also requires the knowledge of a variety tools in image processing and GIS). Godfrey and Eveleth [9] demonstrated a GIS workflow for digitizing a 1986 Idaho map for displaying the map information in a Web environment. Both the British Library and the David Rumsey Map Collection held events to georeference their map collections by crowdsourcing. The New York Public Library‘s crowdsourcing approach for map digitization went one step further to provide semi-automatic tools for extracting parcel polygons from US insurance maps.[8] They also noted that fully manual approach would not scale to process their map collections in a reasonable time [1].

To handle the vast variety of historical map types, crowdsourcing with semi-automatic approaches for map digitization is more robust (than fully automatic approaches) in producing accurate results [5]. The challenges here are how to build adaptive semi-automatic techniques that improve the level of automation as more maps are processed and to eventually eliminate manual work once the enough samples are processed. As crowdsourcing for map digitization is used, approaches for cross-validating user generated content (e.g., the location, size, and shape of the infill lands in Figure 2(a)) (see [15] for more discussions on challenges

[7] To appreciate this difficulty from experience, I encourage the readers to explore how long it would take to find a large-scale map of 1941 Budapest.

[8] `https://github.com/NYPL/map-vectorizer`

in user-generated geocontents) and the provenance information needs to be properly recorded and passed to the final datasets (e.g., [7]).

In addition, while there exists an abundant work on map processing techniques [5], none of the work go beyond raster-to-vector conversion to record the processing "uncertainty" during the extraction. As noted in an earlier technical report from Aeronautical Chart and Information Center [10], accuracy of the source material, intermediate and final products needs to be considered to achieve the optimum utilization of a map product. To estimate the accuracy of the final datasets, the challenge is how to build systematic and objective evaluation methods for individual steps in a map processing tool and produce a final accuracy estimate.

2.3 Modeling and publishing map content

Having machine-readable map content, the next challenge is how to add semantic descriptions to the datasets and link the datasets to other sources. Geographic information is already a crucial link connecting entries on the Linked Data Web and various ontologies for modeling geographic datasets are developed (e.g., NeoGeo Vocabulary Specification[9]) [12]. Adding semantic descriptions to the map content will enable ontology-supported searches beyond search by place, time, format, and keyword. Modeling the extracted dataset as linked data could also promote data sharing and support studies that require large historical spatiotemporal datasets. The challenge here is how to build the techniques that allow a user to easily model and publish their geographic datasets. Existing tools for data integration such as the interactive Web application, Karma[10] has the basic functionality for modeling geometries (i.e., points, lines, and polygons) but not complex geographic phenomena (e.g., a probability surface). Once the data can be easily modeled with semantic descriptions and linked to other data sources, simply hosting the data on a webpage will greatly help the data to be indexed by search engines and used by other researchers.

3. SUMMARY

This paper described the vision of a map processing, modeling, linking, and publishing framework that enables querying historical map collections as a unified and structured spatiotemporal source. This framework supports users to answer important questions that require spatiotemporal datasets in the past. The challenges include fundamental research in efficient and effective methods for converting map content to machine-readable format, recording provenance and uncertainty information during such digitization processes, modeling map content, provenance, and uncertainty information, and linking the modeled data to other data sources. The resulting tools and datasets will enable a much wider utilization of historical maps and support a variety of studies in contrast to the existing capabilities rely heavily on manual work and is lack in data sharing.

References

[1] M. G. Arteaga. Historical map polygon and feature extractor. In *ACM SIGSPATIAL International Workshop on MapInteraction*, pages 66–71, 2013.

[2] C. S. Beattie. *3D Visualization Models as a Tool for Reconstructing the Historical Landscape of the Ballona Creek Watershed*. Master thesis, University of Southern California, 2014.

[3] C.-C. Chen, C. A. Knoblock, C. Shahabi, Y.-Y. Chiang, and S. Thakkar. Automatically and Accurately Conflating Orthoimagery and Street Maps. In *ACM International Conference on Advances in Geographic Information Systems*, pages 47–56. 2004.

[4] Y.-Y. Chiang. *Harvesting Geographic Features from Heterogeneous Raster Maps*. PhD thesis, University of Southern California, 2010.

[5] Y.-Y. Chiang, S. Leyk, and C. A. Knoblock. A Survey of Digital Map Processing Techniques. *ACM Computing Surveys*, 47(1):1–44, 2014.

[6] Y.-Y. Chiang, S. Leyk, N. H. Nazari, and S. Moghaddam. The Impact of Graphical Quality on Automatic Text Recognition in Digital Maps. In *International Cartographic Conference*, 2015.

[7] D. Garijo, Y. Gil, and A. Harth. Challenges for provenance analytics over geospatial data. In B. Ludäscher and B. Plale, editors, *Provenance and Annotation of Data and Processes*, volume 8628 of *LNCS*, pages 261–263. Springer, 2015.

[8] J. Gelernter. *MapSearch: a protocol and prototype application to find maps*. PhD thesis, Rutgers, The State University of New Jersey, 2008.

[9] B. Godfrey and H. Eveleth. An adaptable approach for generating vector features from scanned historical thematic maps using image enhancement and remote sensing techniques in a in a geographic information system. *Journal of Map & Geography Libraries*, pages 18–36, 2015.

[10] C. R. Greenwalt and M. E. Shultz. Principles of error theory and cartographic applications. Technical Report ACIC Technical Report No. 96, Aeronautical Chart and Information Center, 1962.

[11] I. N. Gregory and P. S. Ell. Historical GIS: technologies, methodologies, and scholarship. volume 39. Cambridge University Press, 2007.

[12] K. Janowicz, S. Scheider, T. Pehle, and G. Hart. Geospatial Semantics and Linked Spatiotemporal Data - Past, Present, and Future. *Semantic Web*, 3(4):321–332, 2012.

[13] L. Kurashige. Rethinking Anti-Immigrant Racism: Lessons from the Los Angeles Vote on the 1920 Alien Land Law. *Southern California Quarterly*, 95(3):265–283, 2013.

[14] A. Pezeshk. *Feature Extraction and Text Recognition from Scanned Color Topographic Maps*. PhD thesis, Pennsylvania State University, 2011.

[15] D. Pfoser. On user-generated geocontent on user-generated geocontent. In *Advances in Spatial and Temporal Databases*, volume 6849, pages 458–461. Springer, 2011.

[16] J. Weinman. Toponym Recognition in Historical Maps by Gazetteer Alignment. In *International Conference on Document Analysis and Recognition*, pages 1044–1048, 2013.

[9] http://geovocab.org/doc/neogeo

[10] http://usc-isi-i2.github.io/karma/