Title

Geospatial Informatics Key to Recovering and Sharing Historical Ecological Data for Modern Use

Journal

Photogrammetric Engineering & Remote Sensing, 83(11)

ISSN

0099-1112

Authors

Kelly, Maggi
Easterday, Kelly
Koo, Michelle
et al.

Publication Date

2017-11-01

DOI

10.14358/pers.83.10.779

Peer reviewed

APPLICATIONS PAPER

Geospatial Informatics Key to Recovering and Sharing Historical Ecological Data for Modern Use

Maggi Kelly, Kelly Easterday, Michelle Koo, James H. Thorne, Shruti Mukythar, and Brian Galey

Abstract

Many scientific disciplines need to locate, digitize, and integrate the collections of historical ecological data that often remain hidden in paper archives. Synthesizing historical and contemporary ecological data with ecosystem models can help researchers understand how species, communities, and landscapes are changing across space and time. Since these data are often stored in multiple collections and in various formats, data integration can be challenging. This paper presents a case study of the digitization of a large historical vegetation survey, i.e., the Wieslander Vegetation Type Mapping (VTM) project in California which highlights the importance of recovering and sharing such datasets. The protocol developed to digitize, georeference, visualize, and share these data is founded in geospatial concepts, and the VTM project showcases the increasingly important role of geospatial experts in the fields of ecology, history, and other sciences. These methods are flexible, transferable and broadly applicable to other fields.

Introduction

As current and future challenges around climate change, disease and pest management, and land cover change move to the forefront of policy, planning and management, there is a need to combine and synthesize diverse streams of historical and contemporary ecological data to better understand the patterns, drivers, and consequences of species, community, and landscape change over time (Beller *et al.*, 2017; Bürgi *et al.*, 2017; Rapacciuolo *et al.*, 2014). Enabling such a long-term perspective in data analysis can be challenging because for most of the 20th century, geographical and ecological records were developed and maintained by individuals or small academic groups, and focused in place and time (Frehner and Braendli, 2006; Michener, 2006). This has been a standard scientific norm but it makes multi-scale, cross-disciplinary research more challenging because important datasets can be difficult to find, retrieve, evaluate and use in multi-scalar ecosystem models (Borgman, 2012; Hampton *et al.*, 2013; Jones *et al.*, 2006; Morrison *et al.*, 2017; Tenopir *et al.*, 2011). Working with historical ecological data to answer pressing contemporary ecological challenges such as climate and land use change will require developments in data curation, integration, and sharing. Many researchers argue that such activities will require open technologies that facilitate sharing including web services, open standards, and application programming interfaces (APIs), which have the potential to create emergent knowledge through novel combinations of information (Carpenter *et al.*, 2009; Kelly *et al.*, 2016; McIntyre *et al.*, 2015; Peters, 2010; Tingley and Beissinger, 2009).

Methods for the reconstruction of past vegetation abundance and pattern can include biological methods such as dendrochronology or palynology (Egan,2005), but when biological samples are not available, many researchers use historical map and plot data, as well as other cultural references (Beller *et al.*, 2017; Grossinger *et al.*,2007; Stein *et al.*, 2010; Whipple *et al.*, 2011) to understand past conditions. In the Eastern and Midwestern United States of America there are extensive archival records, including the General Land Office surveys and other early land surveys (Galatowitsch, 1990; Mladenoff *et al.*, 2002; Schulte and Mladenoff, 2001), but in the North American West these data are less common and the reconstruction of past conditions can be hampered by a relative paucity of vegetation data.

The recovery of historical geographic and ecological data for modern use requires a suite of key concepts including georeferencing, error and uncertainty estimation, spatial data management, cartography and visualization, and integration of data into spatial modeling platforms. These concepts are well known to geospatial experts yet maybe new to other researchers and data managers (Golledge, 2002; Goodchild, 2009). Detailed workflows underpinned by geospatial knowledge and expertise are needed to ensure key concepts and information are preserved. These workflows can then be reproduced to make various historical data collections digital. This "applications" article describes a case study of one important historical vegetation data collection and the protocol and technology used to recover data and make them available for modern ecological analyses. The case study is the California Wieslander Vegetation Type Mapping (VTM) collection (Wieslander, 1935).

The VTM collection, created in the 1920s and 1930s, has been described as "the most important and comprehensive botanical map of a large area ever undertaken anywhere on the Earth's surface" (Jepson *et al.*, 2000). It was pioneered by Albert E Wieslander, an employee of the Forest Service and

Maggi Kelly is with the Department of Environmental Science and Policy, University of California Berkeley, 130 Mulford Hall, #3114, Berkeley CA 94720; Geospatial Innovation Facility, University of California Berkeley; and University of California Division of Agriculture and Natural Resources.

Kelly Easterday is with the Department of Environmental Science and Policy, University of California Berkeley, 130 Mulford Hall #3114, Berkeley CA 94720.

Michelle Koo is with the Museum of Vertebrate Zoology, University of California Berkeley.

James Thorne is with the Department Environmental Science and Policy, University of California, Davis, CA 95616.

Shruti Mukythar, and Brian Galey are with the Geospatial Innovation Facility, University of California Berkeley.

Photogrammetric Engineering & Remote Sensing
Vol. 83, No. 11, November 2017, pp. 779–786.
0099-1112/17/779–786A

doi: 10.14358/PERS.83.10.779

Forest and Range Experiment Station in Berkeley, California. Overall, the collection covers about 16 million ha (160,000 km^2) or just over a third of California exclusive of the deserts and large agricultural areas (Kelly *et al.*, 2005). The collection includes over 200 vegetation maps, 18,000 vegetation plots, 3,000 photographs, and over 23,000 plant voucher specimens (Figure 1). It is a detailed, extensive (although not complete), and multi-modal description of the vegetation of California in the early 20th century, and its availability in digital form (*vtm.berkeley.edu*) presents multiple opportunities to examine, characterize, and understand changes to California landscapes. This paper uses the VTM as a case study for illustrating several key themes including the importance of rescuing historical data; the need for best practices for data digitization, the value of data visualization; data fusion and integration; and the critical role of web based infrastructures for sharing scientific data.

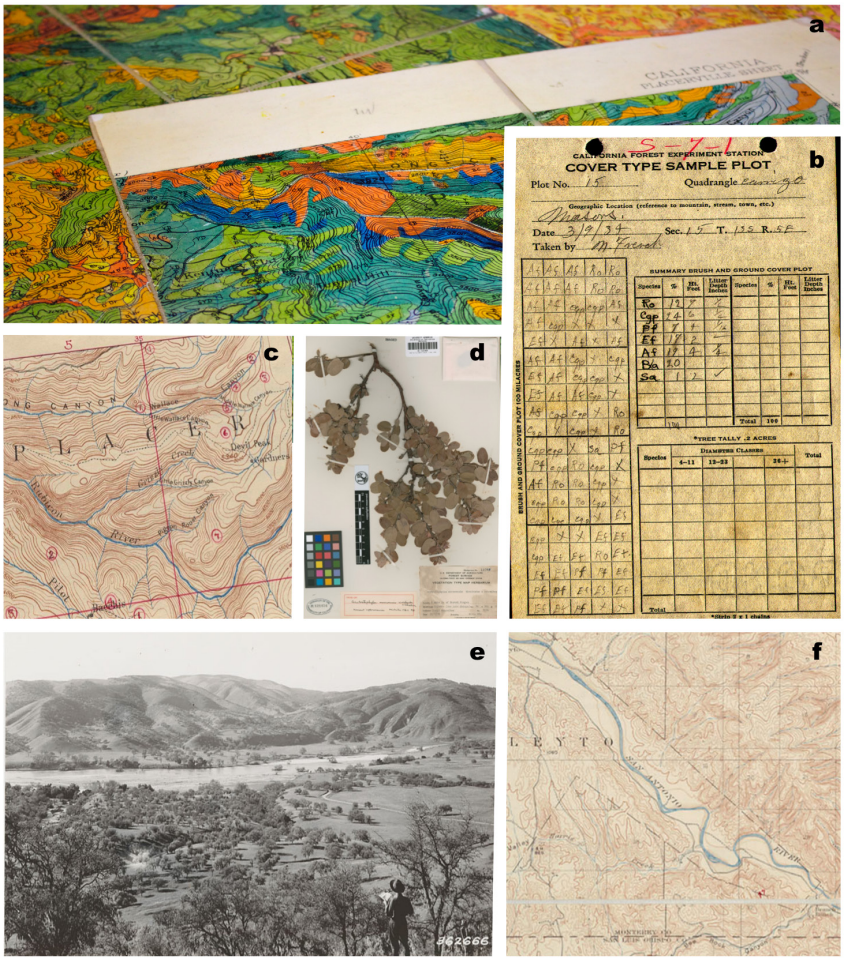

Figure 1. Examples of the components of the original VTM collection: (a) a vegetation map from Placerville, El Dorado County; (b) a plot card from San Luis Obispo County,(c) a plot map, showing numbered locations of plots in Placer County (maps associated with the photography collection and herbarium specimens are similar); (d) an image of an herbarium specimen (*Arctostaphylos morroensis*) from San Luis Obispo County; (e) a landscape photograph looking NW across San Antonio River in Monterey County, 1938. The photograph shows grassland, Douglas oak woodland, and *Artemisia californica*, *Eriogonum fasciculatum* (information taken from photograph notes); and (f) a portion of the key map associated with landscape photograph 1e showing the mark (lower right) indicating the photographer's vantage point looking over the San Antonio River.

VTM Collection Digitization

During the 2000s, several research groups in California began complementary and comprehensive efforts to digitize the individual parts of the VTM collection; these efforts have since been joined. The methods used to georeference and estimate error and uncertainty although slightly different are standard and comparable. The general workflow developed (scanning analog material, georeferencing, estimation of error, creation of digital database, visualization, and serving of data using a web API) made use of best practices pioneered and enhanced by the geospatial discipline. We describe the process for each part of the VTM collection here.

Vegetation Maps

The original VTM maps were drawn on 1:62,500-scale and 1:125,000-scale U.S. Geological Survey topographic quadrangles from the late 19th and early 20th centuries (Figure 1a). The maps were cut into tiles and mounted on linen canvas for use in the field. The maps were carefully scanned one or two tiles at a time at a standard 300 dpi resolution. VTM tiles were primarily registered to scanned versions of the same topographic maps on which they were drawn, typically first edition 30′ quadrangles from the early 1900s U.S. Coastal Geodetic Survey. Subsequently locational accuracy was tested by registering the VTM maps to 1:24,000-scale USGS Digital Raster Graphic quadrangles on a per-quadrangle basis using multiple tie points per tile to identify a Root Mean Square Error or RMSE (Thorne *et al.*, 2008). Vegetation polygons were traced manually at a resolution in which the traced line was finer than the boundary line it was copying (typically 1:6,000) and the plant species codes recorded within each polygon were transcribed into a GIS database. An automated method for polygon extraction (e.g., Soille and Ansoult, 1990) was not possible due to the variable quality of the vegetation maps. The original VTM plant species codes were linked to historical plant species names and then converted to current scientific nomenclature, and the species in each polygon were assigned to current vegetation and habitat types. This process is detailed in Thorne *et al.* (2008). The project used the Manual of California Vegetation Types (Sawyer and Keeler-Wolf, 1995), and the California Wildlife Habitat Relationships Models (WHR) (Mayer and Laudenslayer, 1988) for land cover classifications. Once these attributes were added to the maps, they were error-checked and finalized as GIS layers (Thorne *et al.*, 2008). Map georeferencing errors were investigated using the RMSE metric. The RMSE for each quadrangle ranged from 21.7 to 189.6m.

Plot Data

Each VTM plot was rectangular in shape with the longer axis running upslope. The plots were 800 m^2 in size in forests and 400 m^2 in scrub and chaparral communities. Information on dominant species, ground surface characteristics, average height of the dominant species, and trees greater than 10 cm in Diameter at Breast Height, (DBH)

were tallied by species and diameter class. Information about each plot was recorded manually on paper cards (Figure 1b) and the location of the plot marked on a map. Each plot has a unique identifier based on its topographic quadrangle name, map section, and unique plot number that allows the plot data to be linked to its location on the digitized plot maps. This relational structure was maintained as the plots and plot data were manually entered into the database. To test the plot database for accuracy, 200 randomly selected plots from the digital database were verified with the original datasheets in each field. Incorrect or missing values were counted as errors. Errors in the transcription ranged from 0 to 11.0 percent with an average rate of 1.4 percent. The majority of fields (27 out of 32) had low errors (e.g., less than 2.0 percent).

Plot Maps

Plot locations were marked on 1:62,500-scale and 1:125,000-scale U.S. Geological Survey topographic quadrangles from the late 19th and early 20th centuries (Figure 1c), which had been tiled and mounted on linen canvas for field transport. These were scanned at 600 dpi, one cut tile at a time, and georeferenced to 1:24,000-scale USGS Digital Raster Graphic quadrangles using multiple tie-points per tile. All plot locations were digitized manually and their location attributed to the plot ID. The plot ID was used to join the plot location with the plot data. The total error produced by each step of the process was estimated using the total error formula (Wieczorek *et al.*, 2004). The total error (the square root of the sum of the squared errors) produced by the georeferencing process ranged from 126.9 m to 462.3 m, and each plot is attributed with the total uncertainty as a combination of error resulting from the digitization process and the original quality of the maps. The process is fully described in Kelly *et al.* (2008).

Herbarium Specimens

The VTM field teams collected herbarium specimens (Figure 1d) for every species recorded on the vegetation maps or in the sample plots (Ertter, 2000). Over 23,000 of the VTM specimens have been digitized and georeferenced through the efforts of the Jepson Herbarium and the Consortia of California Herbaria (CCH). Specimens were originally digitized primarily to township, range, and section centroids, corresponding to an average uncertainty of 805 m (or one-half of a mile). Refinement of VTM specimen mapping based on label locality data beyond township, range, and section is ongoing through digital curation efforts of the holdings of the CCH and the Berkeley Natural History Museum consortium (*https://bnhm.berkeley.edu/informatics/collection-databases/*).

Photographs, Key Maps, and Metadata

Black and white photographs (9.2 × 13.6 cm) were taken from 1920 to 1941 (Figure 1e and 1f), most of which were keyed directly on to USGS topographical maps using red pen, and marking an arrow to show the vantage point and view of each photo. Each photograph was scanned, and information from captions was entered into a database that is searchable by keyword, genus, and species. The locations of each photograph were georeferenced by (a) measuring the distance and bearing of the known southwest corner of each USGS topographic key map to the marked photograph location on the map and deriving its location; or (b) for photographs lacking a corresponding key map, locations were georeferenced based on their written locality description using the same point-radius methodology as employed for museum specimens (Wieczorek *et al.*, 2004). Uncertainty estimates were based in the first case on the scale of the key map, the size of the location marker, and in the second case on the known uncertainties associated with location data (Wieczorek *et al.*, 2004). These uncertainties ranged from 10 m to over 19 km depending on the method used.

The now-digital version of the VTM collection includes the vegetation maps, vegetation plot locations and associated plot data, photograph locations and, herbarium specimens. There are VTM data features in every county in California, and several areas have relatively high overlap (i.e., density) of VTM data types: for example the central Sierra Nevada forests in Alpine, Tuolumne, and Calaveras Counties, and the central coast woodlands in Santa Barbara, San Luis Obispo, and Ventura Counties (Figure 2e). The range of data types captured through the type specific process outlined above (maps, plots, specimens, photographs) represent a wide variety of

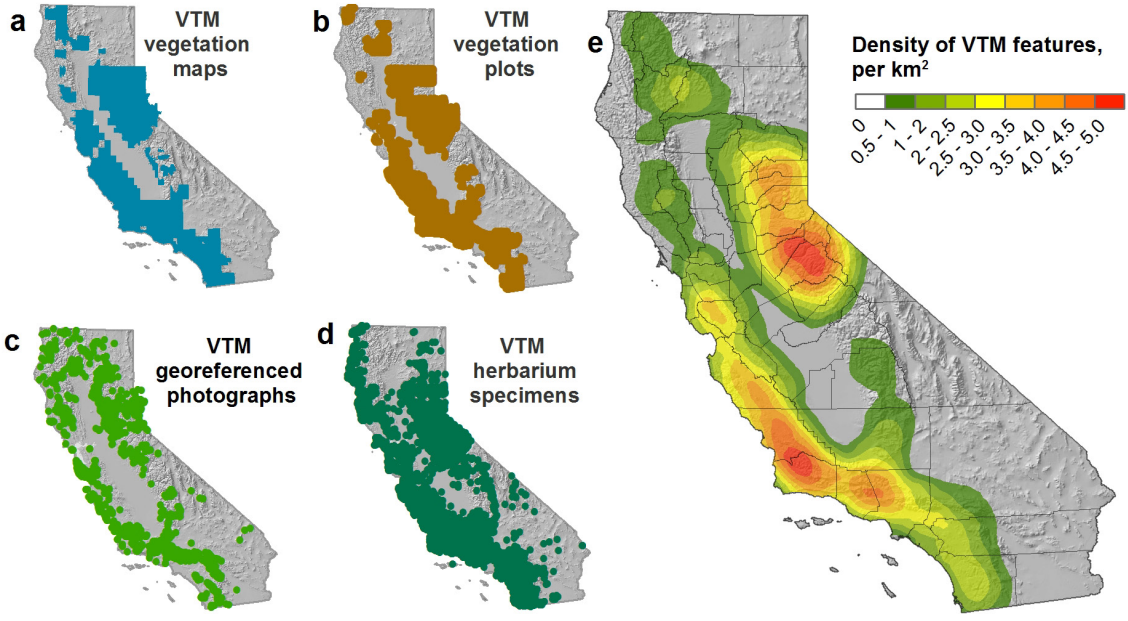

Figure 2. Digital VTM data in California: (a) vegetation maps; (b) vegetation plots; (c) photograph locations; (d) herbarium specimens; and (e) the density per km² for all VTM features.

workflows that cover a large range of common historical ecological data formats (Vellend *et al.*, 2013).

API and Website Development

The first web application that made parts of the VTM collection available to a wider audience was developed in 2004 (Kelly *et al.*, 2005) when online mapping technologies were in their infancy and server performance and hosting costs were large limiting factors. The map portion of the site was originally built using MapServer.org technology. Since then, web mapping has made tremendous technological strides with several key advances (e.g., the release of Google Maps™ API in 2005), enabling more sophisticated searches, visualizations and application development through open Application Programming Interfaces, known as APIs. A web API is an application that serves machine-readable data and functionality to applications that represent the data to users.

All digital spatial VTM data are made available via an open source web-mapping application (*http://vtm.berkeley.edu*) developed using an open source software stack. The VTM website was built by the Berkeley Geospatial Innovation Facility (*http//:gif.berkeley.edu*) using the Berkeley Ecoinformatics Engine Application Programming Interface (EcoEngine API) (*https://ecoengine.berkeley.edu/*); a RESTful API approach for serving data in general and the VTM data in particular (Figure 2). A RESTful API allows for common data exchange formats with interoperability between data sources and applications. The EcoEngine API is a gateway to explore and compare species and geospatial data to understand biotic responses to global change. It is a portal to many of the diverse biological and environmental collections housed at UC Berkeley, and part of a network of digitized museum collections used for the study of biodiversity, climate change, and ecosystem change (Pyke and Ehrlich, 2010). In addition to the VTM data, the EcoEngine API serves about 5 million records of museum specimens, field station records, soil and pollen data, sensor readings, as well as biophysical base layers such as past and modeled future climate and land use.

The VTM website accesses the individual parts of the VTM collection – VTM photographs, vegetation maps, plots, and specimen locations (Figure 3) which are stored using PostgreSQL, a relational database that supports storage and analysis geospatial vector data through the PostGIS extension. The website presents a user interface for data exploration, search, visualization, and download. Exploration and search are facilitated through an interactive map interface that allows a user to click on a feature, after which a pop up window displays greater detail about the feature. Data can also be downloaded directly from the site.

Visualizations are critical to our ability to process complex data and to engage users in data understanding and several key open source tools allow for this seamless engagement. The VTM map interface was built using Leaflet (*leafletjs.com*), a lightweight JavaScript mapping library within which Open Street Map (*https://www.openstreetmap.org*) base layers are provided (Figure 4). The original vegetation maps display

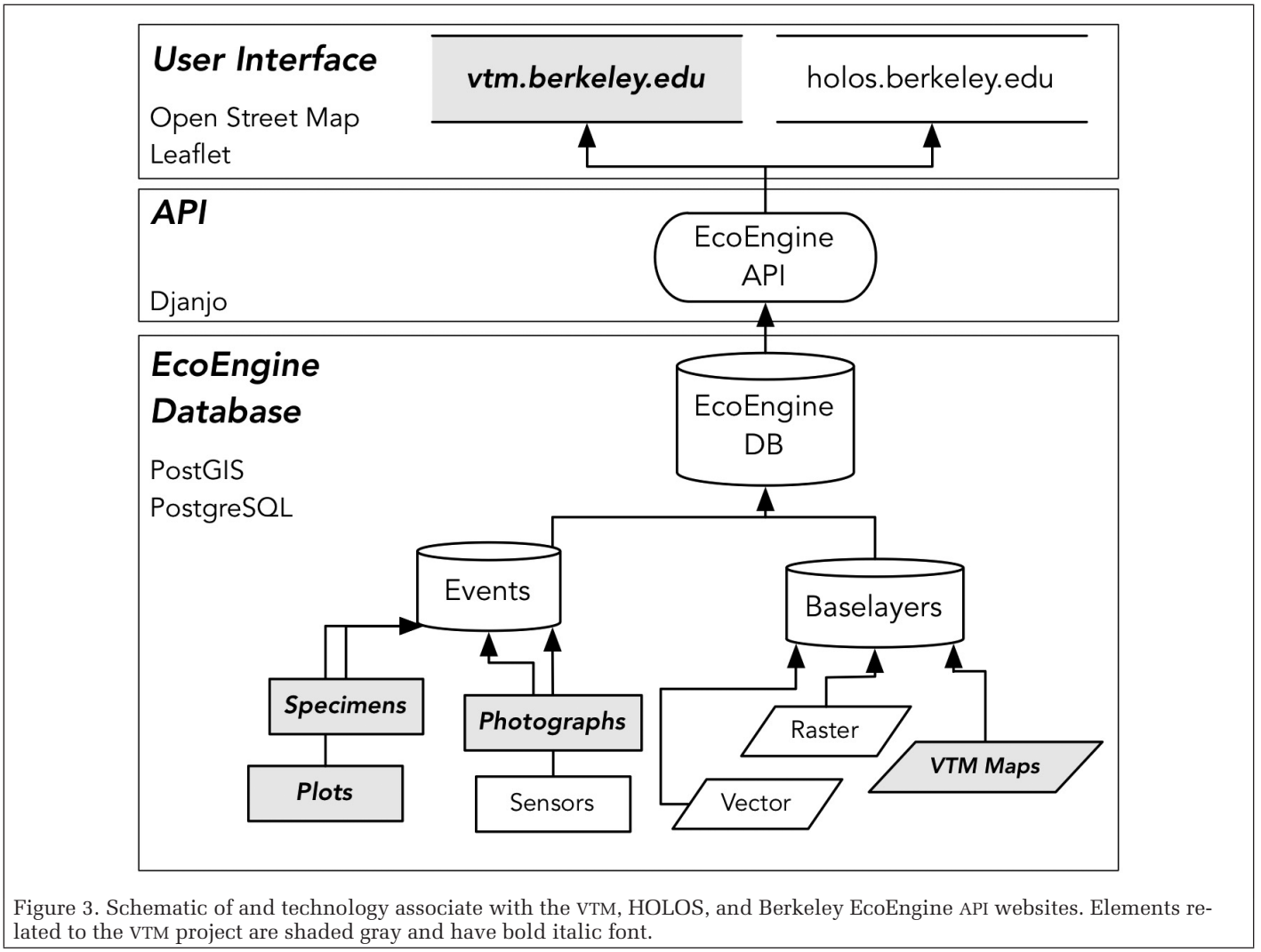

Figure 3. Schematic of and technology associate with the VTM, HOLOS, and Berkeley EcoEngine API websites. Elements related to the VTM project are shaded gray and have bold italic font.

vibrant color schemes (Figure 1a), but rather than try to replicate this variety, we rendered the GIS version using a simplified standard USGS NLCD land cover color scheme (Fry *et al.*, 2008).

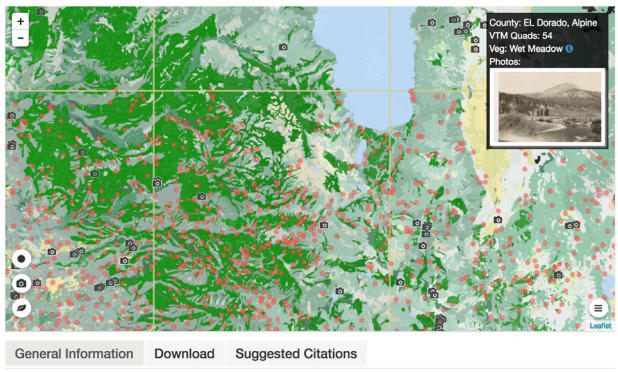

Figure 4. A screen capture of the VTM website showing an area covering part of Lake Tahoe and the Tahoe National Forest: background colors are vegetation polygons, red dots are plot locations, and black icons are locations of georeferenced photographs. One photograph has been queried by clicking. The map legend is also available by clicking on the legend icon in the lower right corner of the window.

Analysis Using the VTM Data

The VTM maps and associated plot data were an important precursor to prominent state classification systems and mapping surveys such as plant community distributions as described in the California Manual of California Vegetation classification (Sawyer and Keeler-Wolf, 1995), and the State Cooperative Soil-Vegetation Surveys of California (Colwell, 1977). Prior to 2005, selected portions of the original VTM collection were used only when they could be located in libraries and personally digitized. Plot resurvey efforts were conducted in the Tehachapi Mountains (Minnich 1995), San Diego County (Franklin *et al.*, 2004), and Lassen National Park (Taylor, 2000) to study changes in forest composition and changing fire interactions within both chaparral and forest communities. Non-georeferenced plot data was used to develop vegetation community classification schemes in oak and rangeland communities (Allen *et al.*, 1991; Allen-Diaz and Holzman, 1991), and VTM vegetation maps were used in the interpretation of historical aerial photographs over a 50 year period in the Los Angeles Basin (Freudenberger *et al.*, 1987). Each of these efforts was confined to particular regions or vegetation communities in part due to the difficulty of tracking down and digitizing relevant data.

However, since the digitization and sharing of various parts of the collection the scope and scale of research has expanded considerably, demonstrating that web-based infrastructure is key for sharing scientific data. Kelly *et al.* (2016) provides an overview and map of recent research using the collection. For example, since the release of the digitized plot dataset researchers have been able to perform analysis at the regional (Dobrowski *et al.*, 2011; Dolanc *et al.*, 2013a; Dolanc *et al.*, 2013b) or statewide scale (Fellows and Goulden, 2008; McIntyre *et al.*, 2015). Data digitization and release has also allowed for comparisons of plot data with contemporary ecological data to examine large-scale changes to a range of vascular plants (Crimmins *et al.*, 2011; Crimmins *et al.*, 2013; McIntyre *et al.*, 2015). The availability of the digital dataset has enabled integration with climate scenario modeling to understand biogeographic responses to climate change (Rapacciuolo *et al.*, 2014), and numerous studies have examined possible futures for California vegetation using the VTM plot data (Conlisk *et al.*, 2012; Dobrowski *et al.*, 2011; Swanson *et al.*, 2013) and sophisticated modeling techniques such as Maxent (Phillips *et al.*, 2006) as well as other multi-scalar ecosystem models.

One recent example highlights the way in which the open VTM data can be used in ecological research. McIntyre *et al.* (2015) used the VTM plot data in comparison with modern-day forest inventory data collected by the Forest Service through its Forest Inventory and Analysis (FIA) program. They compared the digitized VTM data to modern FIA data, and examined how tree size class distributions, basal area estimates, and species composition have changed in California over 80 years around the state. They observed changes in tree density, in size class, and in species composition that were correlated with a range of possible climactic explanatory factors, including climate water deficit (CWD).

We have identified underutilized parts of the collection that remain open avenues for investigation. The vegetation maps and plots are the most commonly used parts of the collection with only two published reports focusing on the photography collection. Repeat photography of the now-georeferenced photographs might add to our understanding of landscape and plant community change, and new technologies may aid in this process (Babic *et al.*, 2008; Hannula, 2016). Additionally, there is potential in the synthesis of VTM data: plots and maps, maps and photographs, plots, and specimens. There are several areas in the state where the collections coincide (Figure 2): including in particular the Central Coastal Ranges and the Northern Sierra Nevada.

Due to the historic nature of the dataset, which includes incomplete records and metadata, as well as shifts in protocol and nomenclature, these data present a number of challenges for contemporary researchers who use them. The challenges result from the terminology, methods and technology used at the time of data collection, as well as errors introduced through the digitization process. There are three types of processing required to bring these kinds of historical ecological data to light.

First, use of historical ecological data requires consideration of the historical taxonomy used, as well as consideration of the necessary cross-walking between historical and modern data. The species codes on the maps were created by the Wieslander project and are therefore not standard taxonomic codes which can make identification cryptic (Thorne *et al.*, 2008). Ecological taxonomy has also changed since the early 20th century (Barbour *et al.*, 2007), and confusion resulting from name changes can occur. Further, loss of information can result from deciphering hand written script as well as problematic duplicate coding (e.g., does 'R' = Redwood, Red fir, or Rock?). All of these challenges need to be dealt with before direct comparisons to modern data can be trusted, and may require scaling adjustments. Second, historical field protocols should be thoroughly understood and may require scaling adjustments in order to compare historical data with data derived from modern field collection methods. For example, the VTM crew used four size classes to bin the diameter at breast height of trees recorded in the plot data which requires a simplification of modern Forest Inventory and Analysis (FIA) data to allow comparison (sensu Dolanc *et al.*, 2013a; McIntyre *et al.*, 2015). In contrast, the VTM vegetation maps have a much higher species detail than is provided on modern land cover maps, which requires simplification of the historical data for contemporary analyses (Thorne *et al.*, 2008; Thorne *et al.*, 2013).

The third analytical challenge is the focus of this paper: the georeferencing that underpins the spatial records. The protocols used to digitize and georeference the analog data depended on the lab undertaking the process, and each part of the collection provide some measure or estimate of uncertainty in the final product, based on their protocol. For example, the plot and vegetation data include RMSE measure per quad included in the data as an additional field (Table 1). Such measures are critical for the user (Goodchild *et al.*, 2012). These measures or estimates of uncertainty can guide researchers on the use of the VTM data and other historical datasets. Some uses require highly accurate and precise georeferencing, and others have less stringent requirements; thus having a means to record the uncertainty allows a researcher to assess fitness of use. For example, the total error associated with each plot has been used in guiding field relocation efforts: researchers can use the error buffer to target field searches that match the text description provided in the plot description. Additionally, the RMSE value associated with each digital vegetation map has been used as a basis for determining the minimum threshold for raster resolution size for a change product comparing VTM mapped data with contemporary land cover maps (Thorne *et al.*, 2008). Grid cells larger than the RMSE assures that the grid cells being compared through time overlap (Thorne *et al.*, 2008). McIntyre *et al.* (2015) used a method similar to this when they chose a raster resolution to grid the VTM plot data in a comparison with modern USDA Forest Service Forest and Inventory Analysis (FIA) data.

Discussion

We currently face a dilemma and opportunity in science: we need to locate, digitize and integrate into larger data streams the collections of historical ecological and geographical data that often remain hidden in archives. These challenges require data interoperability, data integration, and frameworks for web-based retrieval, analysis, and visualization of spatially related environmental data based on the integration of distributed data repositories (Frehner and Braendli, 2006; Goodchild *et al.*, 2012; Hampton *et al.*, 2013; Wright and Wang, 2011). Such data and collaboration frameworks have been called for since at least the 1990s (e.g., Davis, 1995). There are many recent examples of frameworks that integrate biological data and museum specimen collections with publicly available scientific data such as climate scenarios, land use, and remotely sensed imagery, and provide tools to promote visualization and in-depth analysis (e.g., Abbott and Broglie, 2005; Beaman and Cellinese, 2012; Rapacciuolo *et al.*, 2017). The EcoEngine is one example of this kind of framework designed to address climate and land cover and land use change which span spatial and temporal dimensions. These kinds of data, visualization and analysis infrastructures expand the potential for large-scale research through the integration and synthesis of data drawn from local data sources as well as global data networks such as GBIF (*http://www.gbif.org*) or VertNet (*http://www.vertnet.org*). They also require continued focus on the development of data synthesis methods applied to disparate data from numerous sources and of varying quality (Goodchild *et al.*, 2012). However, once digitized and shared, these historical collections can make key contributions in evaluating change and planning for the future (McClenachan *et al.*, 2015; Vellend *et al.*, 2013).

The digital VTM collection (the plots, maps, photographs and specimens) is still incomplete. The journey from paper collection to open digital data has been a decades-long process that included many dedicated collaborators working in isolation and increasingly together. Parts of the collection were nearly destroyed numerous times over the 20th century (Wieslander, 1986), and locations of portions of the collection remain unknown today. The VTM digitization project is emblematic of many common challenges facing geographers and ecologists who work with data that exist in libraries, field stations, and universities that is not carefully indexed and stored. Much scientific effort is conducted in relatively small projects by individual investigators and research groups, sometimes without adequate indexing or preservation of data, and so these datasets become nearly invisible to scientists and other potential users and is more likely to remain underutilized and eventually lost (Heidorn, 2008; Michener *et al.*, 1997). An examination of the VTM project can contribute to the ongoing discussions in the academy and elsewhere about several key themes: (a) the importance of rescuing historical collections and dark data; (b) the need for best practices for data digitization, including uncertainty estimation; (c) the value of data visualization; (d) the need for data fusion and integration in ecological and spatial modeling; and (e) the critical role of web based infrastructures for sharing scientific data. All of the challenges raised by these themes require geospatial theory as a technological integrator and key analytical platform.

The VTM case study is a cautionary yet encouraging lesson on the importance of recovering and sharing historical data in ecological and geographical analysis: cautionary since parts of the collection have been nearly lost or disposed of several times, yet encouraging since the methods used here might be used to bring more dark data to light. For example, the maps from the State Cooperative Soil-Vegetation Survey of California (Colwell, 1977) that follow from the scope and time period of the VTM surveys have to our knowledge not been digitized and may add to the understanding of California flora in the post-war period. The general workflow developed for the VTM project (e.g., scanning analog material, georeferencing, estimation of error, creation of digital database, visualization, and serving of data available on the WEB API) is broadly generalized to historical ecological data collection and analysis. Parts of this workflow can be conducted through automation (e.g., Verhoeven *et al.*, 2012; Morgan and Gergel, 2012) or manual (Thorne *et al.*, 2008) processes depending on the quality of the original data. Potential for further automation of the workflow we present here rests on the exploration of emerging technologies (e.g., machine learning) from cross-disciplinary fields such as computer science. The VTM surveyors certainly introduced error in marking plots on maps

Table 1. Summary of spatial error and uncertainty in the VTM digitization processes.

VTM Component	Georeferencing Method	Error Metric	Reported/Estimated Uncertainty (m)
Vegetation Maps	Georeferenced to USGS DRG topographic quadrangles, using collections of tie-points.	RMSE	21.7 - 189.6
Plot locations	Same as above.	Total Error	126.9 - 462.3
Photograph locations	Calculated from distance and bearing of each marked point from the southwest corner of basemap.	Estimate based on distance, bearing and size of photo mark	10 - 19,401
Herbarium specimen locations	Digitized centroid of recorded township, range and section.	Estimate based on size of Section	0 - 805

and delimiting vegetation polygons in the field. Such error is impossible to quantify since no one from the VTM crews remains alive. And while we can never know what is "true" or "correct" with historical data, by providing a transparent estimation of the uncertainty inherent in the final product, we can allow other users to evaluate if and how these data will serve science. In some cases the estimated error might be too significant for a particular purpose. For example, vegetation plots originally located on coarse-scale maps produced before 1898 are particularly problematic (Kelly *et al.*, 2008) and likely will not be accurately relocated. But several contemporary ecological researchers (e.g., Easterday *et al.*, 2016; McIntyre *et al.*, 2015) have found valid uses for the plot data by incorporating the measures of error provided with the data.

Conclusions

There is ample evidence that the rescuing, digitizing, and sharing of historical ecological data is an important scientific endeavor, and we have a "generational imperative" to do so (Morrison *et al.*, 2017). These data provide benchmarks from which to compare change; they can be linked to contemporary ecological data to understand land use legacies and to help predict future changes. The VTM collection has played a significant role in understanding the California flora: past, present and future through its use in vegetation classification, vegetation and land cover change analysis, and modeling of possible future conditions. Rescuing and sharing this historical collection has led to scientific advancement across several disciplines, and serves as an example of the potential that other dark data collections may have. Georeferencing, uncertainty estimation, cross-walking between collection protocols, open source web mapping and API infrastructures as described in this paper are all critical methods that ensure data preservation, quality, reliability, transferability, and synthesis. Each of these pieces have their own associated challenges and necessitate particular skill sets that often require teams of people to complete, and geospatial concepts serve as a technological integrator. Flexible infrastructure such as APIs and web mapping encourage novel combinations of data by lowering the barriers to data synthesis and provide analytical collaborative frameworks for multidisciplinary research. The numerous challenges that confront society will require that historical collections, like the VTM, are preserved, shared, and linked with other novel data to gain valuable insight to past ecological processes and lend weight to potential future trajectories of landscapes or landscape processes.

Acknowledgments

The authors would like to thank several funders, including: the William M. Keck Foundation, Berkeley Institute for Global Change Biology, UC Division of Agriculture and Natural Resources Informatics and Geographic Information Systems Program, and the US Forest Service. Much of the work was conducted in UC Berkeley's Geospatial Innovation Facility. Several individuals have worked on the digitization process early on; we are indebted to Barbara Allen-Diaz, Ken-ichi Ueda, Brian Thomas, Tim De Chant, Esther Zeledon, Lisa Schile, Karin Tuxen-Bettman, and Anne Huber, among others. We thank Joyce Gross for use of her photo retakes. We also thank Sarah Hinman and the many undergraduate students who georeferenced the photographs and specimens in the Museum of Vertebrate Zoology and the University and Jepson Herbaria.

References

Abbott, J.C. and D. Broglie, 2005. OdonataCentral. com: A model for the web-based delivery of natural history information and citizen science, *Dragonflies and Damselflies (Odonata) of Texas*, 1:8.

Allen, B.H., C.A. Holzman, and R.R. Evett, 1991. A classification system for California's hardwood rangelands, *Hilgardia*, 59:2.

Allen-Diaz, B.H., and B.A. Holzman, 1991. Blue oak communities in California, *Madroño*, 38:80–95.

Babic, B., N. Nesic, and Z. Miljkovic, 2008. A review of automated feature recognition with rule-based pattern recognition, *Computers in Industry*, 59(4):321–337.

Barbour, M.G.,T. Keeler-Wolf, and A.A. Schoenherr, 2007. *Terrestrial Vegetation of California*, University of California Press.

Beaman, R.S., and N. Cellinese, 2012. Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science, *ZooKeys*, 209:7.

Beller, E., L. McClenachan, A. Trant, E.W. Sanderson, J., Rhemtulla, A., Guerrini, R. Grossinger, and E. Higgs, 2017. Toward principles of historical ecology, *American Journal of Botany*, 104(5):645–648.

Borgman, C.L., 2012. The conundrum of sharing research data, *Journal of the American Society for Information on Science and Technology*, 63(6):1059–1078.

Bürgi, M., L. **Östlund**, and D.J. Mladenoff, 2017. Legacy effects of human land use: Ecosystems as time-lagged systems, *Ecosystems*, 20(1):94–103.

Carpenter, S.R., E.V. Armbrust, P.W. Arzberger, F.S. Chapin, J.J. Elser, E.J. Hackett, A.R. Ives, P.M. Kareiva, M.A. Leibold, and P. Lundberg, 2009. Accelerate synthesis in ecology and environmental sciences, *BioScience*, 59(8):699–701.

Colwell, W.L., 1977. The status of vegetation mapping in California today, *Terrestrial Vegetation of California*, John Wiley & Sons, Sacramento, California.

Conlisk, E., D. Lawson, A.D. Syphard, J. Franklin, L. Flint, A. Flint, and H.M. Regan, 2012. The roles of dispersal, fecundity, and predation in the population persistence of an oak (Quercus engelmannii) under global change, *PLOS One*, 7(5):e36.

Crimmins, S.M., S.Z. Dobrowski, J.A., Greenberg, J.T. Abatzoglou, and A.R. Mynsberge, 2011. Changes in climatic water balance drive downhill shifts in plant species' optimum elevations, *Science*, 331(6015):324–327.

Crimmins, S.M., S.Z. Dobrowski, and A.R. Mynsberge, 2013. Evaluating ensemble forecasts of plant species distributions under climate change, *Ecological Modelling*, 266:126–130.

Davis, F.W., 1995. Information systems for conservation research, policy, and planning, *Bioscience*, 45:S36–S42.

Dobrowski, S.Z., J.H. Thorne, J.A., Greenberg, H.D. Safford, A.R. Mynsberge, S.M. Crimmins, and A.K. Swanson, 2011. Modeling plant ranges over 75 years of climate change in California, USA: Temporal transferability and species traits, *Ecological Monographs*, 81(2):241–257.

Dolanc, C.R., H.D. Safford, S.Z. Dobrowski, and J.H. Thorne, 2013a. Twentieth century shifts in abundance and composition of vegetation types of the Sierra Nevada, California, *US. Applied Vegetation Science*, 17(3):442–455.

Dolanc, C.R., J.H. Thorne, and H.D. Safford, 2013b. Widespread shifts in the demographic structure of subalpine forests in the Sierra Nevada, California, 1934 to 2007, *Global Ecology and Biogeography*, 22(3):264–276.

Easterday, K.J., P.J. McIntyre, J.H. Thorne, M.J. Santos, and M. Kelly, 2016. Assessing threats and conservation status of historical centers of oak richness in California, *Urban Planning*, 1(4):65–78.

Egan, D., 2005. T*he Historical Ecology Handbook: A Restorationist's Guide to Reference Ecosystems*, Island Press.

Ertter, B., 2000. Our undiscovered heritage: Past and future prospects for species-level botanical inventory, *Madroño*, 47(4):237–252.

Fellows, A.W. and M.L. Goulden, 2008. Has fire suppression increased the amount of carbon stored in western U.S. forests?, *Geophysical Research Letters*, 35(12):L12404.

Franklin, J., C.L. Coulter, and S.J. Rey, 2004. Change over 70 years in a southern California chaparral community related to fire history, *Journal of Vegetation Science*, 15(5):701–710.

Frehner, M. and M. Braendli, 2006. Virtual database: Spatial analysis in a Web-based data management system for distributed ecological data, *Environmental Modelling & Software*, 21(11):1544–1554.

Freudenberger, D.O., B.E. Fish, and J.E. Keeley, 1987. Distribution and stability of grasslands in the Los Angeles Basin, *Bulletin of the Southern California Academy of Sciences*, 86(1):13–26.

Fry, J.A., M.J. Coan, C.G. Homer, D.K. Meyer, and J.D. Wickham, 2008. *Completion of the National Land Cover Database (NLCD) 1992-2001 Land Cover Change Retrofit Product*, U.S. Department of the Interior, U.S. Geological Survey.

Galatowitsch, S.M., 1990. Using the original land survey notes to reconstruct presettlement landscapes in the American West, *Great Basin Naturalist*, 50(2):181–191.

Golledge, R.G., 2002. The nature of geographic knowledge, *Annals of the Association of American Geographers*, 92(1):1–14.

Goodchild, M., 2009. NeoGeography and the nature of geographic expertise, *Journal of Location Based Services*, 3(2):82–96.

Goodchild, M.F., H. Guo, A. Annoni, L. Bian, K. de Bie, F. Campbell, M. Craglia, M Ehlers, J. van Genderen, and D. Jackson, 2012. Next-generation digital earth, Proceedings of the National Academy of Science,109(28):11088–11094.

Grossinger, R.M., C.J. Striplen, R.A. Askevold, E. Brewster, and E.E.Beller, 2007. Historical landscape ecology of an urbanized California valley: Wetlands and woodlands in the Santa Clara Valley, *Landscape Ecology*, 22: 103–120.

Hampton, S.E., C.A. Strasser, J.J. Tewksbury,W.K. Gram, A.E., Budden, A.L. Batcheller, C.S. Duke, and J.H. Porter, 2013. Big data and the future of ecology, *Frontiers in Ecology and the Environment*, 11(3):156–162.

Hannula, M., 2016. *Time Series Analysis: Gap Filling and Feature Recognition*.

Heidorn, P.B., 2008. Shedding light on the dark data in the long tail of science, *Library Trends*, 57(2):280–299.

Jepson, W.L., R. Beidleman, and B. Ertter, 2000. Willis Linn Jepson's "Mapping in Forest Botany", *Madroño*, 47(4):269–272.

Jones, M.B., M.P. Schildhauer, O. Reichman, and S. Bowers, 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere, *Annual Review of Ecology, Evolution and Systematics*, 37:519–544.

Kelly, M., B. Allen-Diaz, and N. Kobzina, 2005. Digitization of a historic dataset: The Wieslander California vegetation type mapping project, *Madroño*, 52(3):191–201.

Kelly, M., K. Easterday, G. Rapacciuolo, M.S. Koo, S. Myukthar, P. McIntyre, J. Thorne, and B. Galey, 2016. Rescuing and sharing historic vegetation data for ecological analysis: The California Vegetation Type Mapping project, *Biodiversity Informatics*, 11:40–62.

Kelly, M., K. Ueda, and B. Allen-Diaz, 2008. Considerations for ecological reconstruction of historic vegetation: Analysis of the spatial uncertainties in the California Vegetation Type Mapping dataset, *Plant Ecology*, 194(1):37–49.

Mayer, K.E. and Laudenslayer, W.F., 1988. A guide to wildlife habitats of California, State of California, Resources Agency, Department of Fish and Game, Sacramento, CA.

McClenachan, L., A.B. Cooper, M.G. McKenzie, and J.A. Drew, 2015. The importance of surprising results and best practices in historical ecology, *BioScience*, 65(9):932–939.

McIntyre, P.J., J.H. Thorne, C.R. Dolanc, A.L. Flint, L.E. Flint, M. Kelly, and D.D. Ackerly, 2015. Twentieth-century shifts in forest structure in California: Denser forests, smaller trees, and increased dominance of oaks, *Proceedings of the National Academy of Science*, 112(5):1458–1463.

Michener, W.K., 2006. Meta-information concepts for ecological data management, *Ecological Information*, 1(1):3–7.

Michener, W.K., J.W. Brunt, J.J. Helly, T.B. Kirchner, and S.G. Stafford, 1997. Nongeospatial metadata for the ecological sciences, *Ecological Applications*, 7(1):330–342.

Mladenoff, D.J., S.E. Dahir, E.V. Nordheim, L.A. Schulte, and G.G. Guntenspergen, 2002. Narrowing historical uncertainty: Probabilistic classification of ambiguously identified tree species in historical forest survey data, *Ecosystems*, 5:539–533.

Morrison, S.A., T.S. Sillett, W.C. Funk, C.K. Ghalambor, and T.C. Rick, 2017. Equipping the 22nd-Century Historical Ecologist, *Trends in Ecology & Evolution*.

Peters, D.P., 2010. Accessible ecology: Synthesis of the long, deep, and broad trends in ecology and evolution, 25(10):592–601.

Phillips, S.J., R.P. Anderson, and R.E. Schapire, 2006. Maximum entropy modeling of species geographic distributions, *Ecological Modelling*, 190:231–259.

Pyke, G.H., and P.R. Ehrlich, 2010. Biological collections and ecological/environmental research: A review, some observations and a look to the future, *Biological Reviews*, 85(2):247–266.

Rapacciuolo, G., J. Ball-Damerow, A. Zeilinger, and V. Resh, 2017. Detecting long-term occupancy changes in Californian odonates from natural history and citizen science records, *Biodiversity and Conservation*, 1–17.

Rapacciuolo, G., D.B. Roy, S. Gillings, and A. Purvis, 2014. Temporal validation plots: Quantifying how well correlative species distribution models predict species' range changes over time, *Methods of Ecological Evolution*, 5(5):407–420.

Sawyer, J., and T. Keeler-Wolf, 1995. *A Manual of California Vegetation*, California Native Plant Society Press, Sacramento, California, 412 pp.

Schulte, L.A., and D.J. Mladenoff, 2001. The original US Public Land Survey records: Their use and limitations in reconstructing presettlement vegetation, *Journal of Forestry*, October: 5–10.

Soille, P.J., and M.M. Ansoult, 1990. Automated basin delineation from digital elevation models using mathematical morphology, *Signal Processing*, 20(2):171–182.

Stein, E.D., S. Dark, T. Longcore, R. Grossinger, N. Hall, and M. Beland, 2010. Historical ecology as a tool for assessing landscape change and informing wetland restoration priorities, *Wetlands*, 30(3):589–601.

Swanson, A.K., S.Z. Dobrowski, A.O. Finley, J.H. Thorne, and M.K. Schwartz, 2013. Spatial regression methods capture prediction uncertainty in species distribution model projections through time, *Global Ecology and Biogeography*, 22(2):242–251.

Taylor, A.H., 2000. Fire regimes and forest changes in mid and upper montane forests of the southern Cascades, Lassen Volcanic National Park, California, USA, *Journal of Biogeography*, 27(1):87–104.

Tenopir, C., S. Allard, K. Douglass, A.U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame, 2011. Data sharing by scientists: practices and perceptions, *PLOS One*, 6(6):e21101.

Thorne, J.H., B.J. Morgan, and J.A. Kennedy, 2008. Vegetation change over sixty years In the Central Sierra Nevada, California, USA, *Madroño*, 55(3):223–237.

Thorne, J.H., M.J. Santos, and J.H. Bjorkman, 2013. Regional assessment of urban impacts on landcover and open space finds a smart urban growth policy performs little better than business as usual, *PLOS One*, 8(6): e65258.

Tingley, M.W., and S.R. Beissinger, 2009. Detecting range shifts from historical species occurrences: New perspectives on old data, *Trends in Ecology and Evolution*, 24(11):625–633.

Vellend, M., C.D. Brown, H.M. Kharouba, J.L. McCune, and I.H. Myers-Smith, 2013. Historical ecology: Using unconventional data sources to test for effects of global environmental change, *American Journal of Botany*, 100(7):1294–1305.

Whipple, A.A., R.M. Grossinger, and F.W. Davis, 2011. Shifting baselines in a California oak savanna: Nineteenth century data to inform restoration scenarios, *Restoration Ecology*, 19(101):88–101.

Wieczorek, J., Q. Guo, and R.J. Hijmans, 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty, *International Journal of Geographic Information Science,* 18(8):745–767.

Wieslander, A.E., 1935. A vegetation type map of California, *Madroño*, 3(3):140–144.

Wright, D.J., and S. Wang, 2011. The emergence of spatial cyberinfrastructure, *Proceedings of the National Academy of Science*, 108(14):5488–5491.