Literature Mapper: A QGIS Plugin for Georeferencing Citations in Zotero

Authors: Tobias, Michele M, and Mandel, Alex I

Source: Air, Soil and Water Research, 14(1)

Published By: SAGE Publishing

URL: https://doi.org/10.1177/11786221211009209

BioOne Complete (complete.BioOne.org) is a full-text database of 200 subscribed and open-access titles in the biological, ecological, and environmental sciences published by nonprofit societies, associations, museums, institutions, and presses.

Your use of this PDF, the BioOne Complete website, and all posted and associated content indicates your acceptance of BioOne’s Terms of Use, available at www.bioone.org/terms-of-use.

Usage of BioOne Complete content is strictly limited to personal, educational, and non-commercial use. Commercial inquiries or rights and permissions requests should be directed to the individual publisher as copyright holder.

BioOne sees sustainable scholarly publishing as an inherently collaborative enterprise connecting authors, nonprofit publishers, academic institutions, research libraries, and research funders in the common goal of maximizing access to critical research.
Literature Mapper: A QGIS Plugin for Georeferencing Citations in Zotero

Michele M Tobias and Alex I Mandel

1UC Davis DataLab, University of California, CA, Davis, USA. 2University of California, Davis, CA, USA.

ABSTRACT: Many studies in air, soil, and water research involve observations and sampling of a specific location. Knowing where studies have been previously undertaken can be a valuable addition to future research, including understanding the geographical context of previously published literature and selecting future study sites. Here, we introduce Literature Mapper, a Python QGIS plugin that provides a method for creating a spatial bibliography manager as well as a specification for storing spatial data in a bibliography manager. Literature Mapper uses QGIS spatial capabilities to allow users to digitize and add location information to a Zotero library, a free and open-source bibliography manager on basemaps or other geographic data of the user's choice. Literature Mapper enhances the citations in a user's online Zotero database with geo-locations by storing spatial coordinates as part of traditional citation entries. Literature Mapper receives data from and sends data to the user's online database via Zotero's web API. Using Zotero as the backend data storage, Literature Mapper benefits from all of its features including shared citation Collections, public sharing, and an open web API usable by additional applications, such as web mapping libraries. To evaluate Literature Mapper's ability to provide insights into the spatial distribution of published literature, we provide a case study using the tool to map the study sites described in academic publications related to the biogeomorphology of California's coastal strand vegetation, a line of research in which air movement, soil, and water are all driving factors. The results of this exercise are presented in static and web map form. The source code for Literature Mapper is available in the corresponding author's GitHub repository: https://github.com/MicheleTobias/LiteratureMapper

KEYWORDS: GIS, QGIS, Zotero, Python, georeference, citations


TYPE: Technical Advances

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Michele M Tobias, University of California Davis, 1 Shields Ave., Davis, CA 95616, USA. Email: mmtobias@ucdavis.edu

Introduction

Analyzing the body of published studies in any given field continues to be an important process in academia. Academics use the traditional literature review, whether published alone or as a part of a larger paper, to describe in text form the literature that has been published about a particular topic and how the papers relate to each other. Often a literature review describes the geographic extent of the study site locations in the collected papers, but written descriptions are more commonly included in the review than a map (personal observation). As maps are tools for understanding the spatial relationships between entities, creating a map of the locations associated with the papers in a literature review allows a researcher to see spatial relationships in the published literature.

There are several recent examples of mapping the physical locations in which the plot of works of fiction took place—“Shakespeare on the Map,” “Placing Literature,” “The Literary City,” “No Sweat Shakespeare” are all examples of web maps of real-world locations that happen in fictional works. Another map, “Mythical Creatures of Europe” is a print map that shows the locations of mythical creatures as described in historical written documents giving specific publications that map to specific locations.

Despite the potential utility of visualizing the locations of scientific studies, academic journal articles are not often the subject of maps. In the few examples that exist, the data are usually aggregated at the country or regional level to obtain counts of publication, rather than to display individual publications.

There are a few examples of mapping individual publications that demonstrate the utility of this exercise for gaining a deeper understanding of a collection of publications. Lowry et al wrote a systematic review of biological invasion literature and produced a map of study site locations as part of the analysis showing that publications about biological invasions are concentrated in North America, Europe, and New Zealand. Goldstein et al is one of the few examples of mapping study sites directly. They mapped literature citing locations of 2 beach grasses on the U.S. Atlantic coast to understand the distribution of the species and combined these data with the Global Biodiversity Information Facility (GBIF) and iNaturalist observations. RubeI et al mined published literature to map tick locations in Germany. These 3 examples all synthesize spatial information to produce knowledge that would otherwise not exist.

In the realm of science, scholars and interested people are more likely to find maps showing the spatial extent of data than of a published study. It has become common practice to map data footprints in map-based search tools for data sets like aerial photos or satellite images. Beginning with projects like Alexandria Digital Library in the 1990s, data seekers increasingly expect to use a map interface to find georeferenced data products. However, this has not translated into other forms of data that have a spatial component such as ecological studies. As geographic searching becomes a more expected option for searching for entities with a location, such as restaurants, businesses, or even geospatial data, the idea of geographic searching for publications is a next logical step that can provide
place name disambiguation and flexibility in searching for published literature. JournalMap allows for geographic searching of published journal articles. As it relies on publishers to submit bounding boxes for articles, we can only expect new articles to be added to this data set. If a user wants to work with previously published work, they will need another option. In addition, this tool was designed for searching for articles, not for analyzing the geographic distribution.

Several citation management software packages are available to help users manage collections of citations, including Zotero, Endnote, and Mendeley. These products store information about each citation such as author name, title, and publication year, but none natively store location information beyond keywords that may contain location names. Mapping citation data often requires that the data be transferred to a geographic information system (GIS) program. There is currently no easy way to do this and no way to update changes between the citation manager and the mapping software. With the current citation management software options, combining spatial information with references is cumbersome, and the spatial data are largely separate from the citation information. One solution to this problem is to store spatial information in a citation database rather than in a GIS and access the spatial information for mapping purposes. The resulting database can be used to create traditional or web maps that can be easily shared.

The objective of this article is threefold. First, we introduce the concept of storing spatial data in a personal Zotero account and offer a specification for the format of the spatial data that makes it accessible to many programs. Second, we describe Literature Mapper, the QGIS plugin that we designed to read and write this specification. And finally, to demonstrate its application for environments related to air, soil, and water, we created a graphical user interface with the QGIS desktop tool or the web browser plugins, all of which can sync data to the user's online account. Users can insert citations and additional data storage is available for a fee for users needing more than 300 MB. Users can enter citations through the desktop tool or the web browser plugins, all of which can sync data to the user's online account. Users can insert citations and build a bibliography while writing a paper using the plugins for word processors such as Word, LibreOffice, and GoogleDocs.

The authors have no current plans to develop tools for other citation managers but would welcome the expertise of other developers to make this possible.

Zotero’s database was not explicitly designed to store spatial data in the bibliographic records; however, Literature Mapper leverages the flexibility included in the database schema to include spatial data. Zotero stores an individual bibliographic record as an Item. Items can be grouped by the user into Collections to organize citations. Each Item contains various standard bibliographic fields and, additionally, an extendible field called Extra which contains text strings. Literature Mapper records spatial data at the Zotero Item level by extending the Extra field.

Literature Mapper extends the Extra field with the storage of spatial data in GeoJSON format (IETF rfc7946). GeoJSON text is wrapped in an xml/html friendly geometry tag so that Extra field can be shared with other uses. Only the information between an opening <geometry> tag and closing </geometry> tag is used to store and update spatial information on an Item record so users can continue to store other text in this field in addition to the location data. Per the GeoJSON specification, geometry data are stored in a geographic coordinate reference system, decimal degrees of Longitude and Latitude, under datum (i.e. WGS84, EPSG:4326). Any valid GeoJSON geometry may be used, including Points, Lines, Polygons, and Multi* variants of basic geometries. Multi* type geometries allow for multiple locations per Zotero Item. At the current time, we have only implemented Points and MultiPoints in the Literature Mapper. In the future, other geometry types are possible; however, while Geometry Collections are an option in the GeoJSON format, we do not recommend Geometry Collections for use in this standard, as few geospatial software handle mixed geometries.

Example Extra field containing spatial information:

```
<geometry>{
  "type": "Point",
  "coordinates": [-140.504592652, 7.1942567162]
}</geometry>
```

We chose GeoJSON as the format to store the spatial data for several reasons. The first is that GeoJSON is a human- and machine-readable plain text standard format common to many geospatial software and programming languages including Python and JavaScript. Being text-based makes it easy to proof for errors with the human eye and allows for parsing with the base tools in most programming languages. As the data are stored as text, the concept should be extendable to any other citation software that has a custom free-form text entry. Simple text also ensures that the data will be stored and retrievable via the Zotero API correctly.

To access the GeoJSON stored in the Extra field of a Zotero Item, a client application should query the Zotero API for a Collection. Items typically belong to Collections which allow users to group Items. Within a collection, each Item then contains spatial data in the Extra field.

**Graphical Tool**

To facilitate the creation of geographic data in Zotero records, we created a graphical user interface with the QGIS desktop application. Literature Mapper is a plugin for QGIS that allows users to add geographic data to Zotero records in the specified GeoJSON format inside the Extra field. The current
Literature Mapper releases works with QGIS 3.x series and is written in Python 3. The plugin uses the Zotero Web API to retrieve a user’s existing citations in the Collection the user specifies and to save geospatial edits on Items back to a user’s online Zotero account. Creating the initial location and editing of location data are done using the QGIS map canvas interface. Figure 1 illustrates the workflow.

To use the Literature Mapper plugin, a user must have a Zotero online account, have their citations of interest stored in a public Collection, and have created an API key with read and write permissions through their online account. For full usage instructions please see the online manual.

On start up, the plugin queries the user’s Zotero account via an API call to retrieve all records from the specified Collection in JSON format, and extract the contents of the Extra field storing the geometry. Each record is parsed into a table where...
the user will see the Key, Year, Author, Title, and Geometry fields for each Item (Figure 2). If the user has previously added location to any records, they will be added to the map as a new temporary layer (memory layer) based on their geometry type. Point and Multipoint geometries are currently implemented. A point layer stores one point location per record. A Multipoint layer stores 1 or more point locations per record. In keeping with the GeoJSON specification, Literature Mapper defines the coordinate reference system for stored locations as Longitude and Latitude (in decimal degrees), World Geodetic System 1984 datum (i.e. WGS84, EPSG:4326) (i.e. QGIS requires spatial data have a coordinate reference system defined so data can be displayed in the proper location and the user should set the coordinate reference system of the project to WGS 84. Records that do not have a location stored yet appear in the table with a blank geometry attribute and are not added to the memory layer.

The user references each record to the map, clicking on the map to select where each record should be located. To add or edit the locations, users can select a record in the table window then activate the appropriate Digitizing Tool—Point or Multipoint—which jumps the user to the map canvas. After making edits on the map, the table will update the Geometry column of the selected record. When finished editing geometries, pressing the “Save to Zotero” button saves the results, which are pushed back to Zotero via the web API, and not saved locally.

Once data are uploaded with the Literature Mapper plugin, the Zotero API and Literature Mapper storage format make it possible to directly implement web maps using various JavaScript map libraries. An example implementation can be seen with Leaflet at http://micheletobias.github.io/maps/LiteratureMap.html and is discussed further in the Case Study section of this article.

Using the QGIS Literature Mapper plugin provides users with several key opportunities, particularly interoperability with other data sets and software. Locations can be referenced to existing maps and geographic data, including various base-maps provided through plugins like QuickMapServices or any other geographic data a user has access to. In addition to editing data, users can visualize the citations in an interactive manner and produce cartographic products using QGIS tools. QGIS implements an open plugin system where users can write plugins to create new tools, and because QGIS is a free and open-source cross-platform application, like Zotero, it is available to anyone.

As an open-source project, the full code for Literature Mapper is available, and community contributions are welcome at https://github.com/MicheleTobias/LiteratureMapper/ and each release of the software is documents and archived with a DOI at http://doi.org/10.5281/zenodo.595050. The code is licensed under GNU General Public License v2, which is the same as QGIS. As there are many differing use cases and possible additional features needed, the online issue tracker contains a list of issues as well as proposed enhancements and potential users are encouraged to make suggestions: https://github.com/MicheleTobias/LiteratureMapper/issues.

Figure 2. Georeferenced data loaded into the QGIS canvas as Temporary Scratch Layers with the Literature Mapper table georeferencing interface.
Case Study

Study site

To illustrate some potential uses of this tool, we will discuss a case study. We chose to map scientific literature related to California’s strand vegetation (Figure 3), not only because it is the first author’s area of expertise but because it provides an excellent example of why this tool is useful and perhaps necessary for research. The California strand is defined by the highwater line on the seaward side and the top of the foredune on the landward side30 and the vegetation communities found there vary with the latitude at which they occur.31,32 Owing to the linear north-south orientation of California’s coast, a study published about southern California is unlikely to be applicable to a site in northern California in terms of the species present. Mapping the study sites associated with each published study can help a researcher or environmental manager immediately identify studies relevant to a particular location and see trends in the locations and time periods studied.

Application

For clarity of the discussion, we offer the following definitions of terms. We define publication as original research published as a journal article, thesis, dissertation, or book. A study site is a location at which ecological samples or observations were made in the course of research leading to the publication in which it was described. A publication can contain one or multiple study sites.

The publications related to California strand vegetation were collected in the course of regular preparation for ecological field research and writing academic papers over the course of the last 10 years. The publications were entered into a Zotero collection, recording at minimum the author names, publication year, publication title, and publisher information for each publication. We reviewed the text of the publications to identify the study site locations mentioned within. These locations were digitized using Literature Mapper to store the locations in the Extra field of each citation. The locations were then available with the citation information through the Zotero API and the memory layers created by Literature Mapper.

The resulting data set containing the study site locations and publication information housed in Zotero was used to explore the spatial and temporal distribution of the studies. From this data set, Literature Mapper created a Point and a MultiPoint memory layer which we exported to GeoPackage files saved locally. We created maps to explore the spatial distribution of the study sites. We also created an online web map with Leaflet with the locations parsed directly from the Zotero API.

Results

The results of this exercise are a data set of publication information for published studies pertaining to California’s strand vegetation and their associated study site locations stored in one database housed in a Zotero Collection, as well as Point and MultiPoint vector data sets. The data set contains a total of 37 papers—27 single study sites stored in a Point vector layer and 10 with multiple study sites as a MultiPoint vector layer. In total, there are 230 study sites represented. The number of study sites per published paper ranges from 1 to 68, and the range of the years of publication is from 1904 to 2015.

From these rich geographic data sets, we produced a map of the study sites paired with a dot histogram pinned by latitude with dots styled by color to indicate the year the site was studied (Figure 4). The study sites are distributed in a patchy distribution along the California coast. The number of sites in each patch of studies varies as well. Southern Ventura County has the most study sites, 44, while other areas have as few as 2.

We also created a multimap depicting the study site locations separated into 20-year time windows to explore the spatial patterns in study sites over time. The spatial distribution also varies over time (Figure 5). In earlier years, there are fewer sites studied and the number of sites increases over time. There is a gap from 1937 to 1962 when no studies were published. The study site distribution between 1960 and 1999 were fairly similar, although not identical.

In addition, we produced a web map drawing geometry data directly from the online Zotero collection (Figure 4). The web map is available to view online at http://micheletobias.github.io/maps/LiteratureMap.html, and the code to make the web map is available at https://github.com/MicheleTobias/maps/blob/gh-pages/LiteratureMap.html. The web map shows all of the study sites. Each site has a popup associated with it that displays the citation for the publication that sampled at that site upon clicking on the site marker. The map allows zooming and panning so the user can explore any given area more closely. Context for the point markers is provided by an OpenStreetMap tile layer that loads beneath the point markers. This layer contains roads, place names, and points of interest (Figure 6).
Discussion

Mapping this data set allows us to better understand the spatial distribution of study sites and publications along the California coast. Spatial coverage of the ecosystem is fairly comprehensive, due in large part to one study, Peinado et al., who remarkably sampled at 68 sites in California. The geographic gaps along the coast that have not been studied appear to be mainly areas where the coastline is rocky or cliff-backed, so the strand ecosystem is not present, but we cannot confirm this speculation without comparing our data set to other data for sandy beach locations. By far, the most studies area of the coast falls around the southern portion of Ventura County, but the majority of those were studied in the last decade (Figure 2).
The spatial distribution of study sites is variable over time. There is a trend toward publications with an increasing number of study sites over time. Earlier publications discuss 1 or 2 sites but later studies tend to draw data from many study sites. There were no studies published between 1937 and 1962 and none of the studies published after that time period mention sampling at this time. This is an unfortunate loss in the ecological record. The 1970s and 1980s time period look rather similar in distribution. Barbour published multiple publications based on the same data set, so the sample sites are repeated over several decades, but the original data were created between 1973 and 1974.\textsuperscript{34} Most of the sites that have been studied have been studied recently—since 2011—and those are from 2 field studies that produced 3 papers.\textsuperscript{33,35,36}

We should be careful in how we interpret these maps. More publications does not necessarily indicate more knowledge gathered about a place. While the map looks rather filled in, we should be careful not to interpret this ecosystem as “well studied.” Several publications were based on a shared data set exploring different aspects of the data. For example, many of the publications with shared co-authors published papers about a single data set. In our map, these points are repeated, but do not represent additional data collection.

There are other questions that we can ask now that we have mapped these data and analyzed the publication attributes: why were no studies published between 1940-1959? Was it due to World War II or other political events or something else entirely? Is there a shift in academia to favor larger studies? Why are some areas more attractive for study? For example, why are there so many sites sampled in southern Ventura County? Is it the number of sandy beaches, access to the site (either physical access or permission to perform research), proximity to researchers, or something else? These questions illustrate the place for mapping published studies in the process of the scientific method. Exploration to answer one question leads to more questions.

The web map is useful for a number of reasons that make it a worthwhile endeavor in addition to static map figures. The map updates as the Zotero collection is updated so it can share the latest information with consumers of the data. As new studies are found, the authors can add them to the web map by adding them to the Zotero collection with the study site geometries. This means viewers do not have to wait for a new journal article to have access to the data. Maps in journal articles are tailored to the needs of the specific article and the data used to produce the map is not often easily accessible to other researchers. Here the data are accessible. As it uses the data from an online Zotero account, the data does not need to be re-processed or stored in a different place, so there is no issues with versioning or maintaining data in more than one location.

Online access to visualizations is an easy and fairly accessible way to share information in this point in time. It allows users to explore the data set in their own way. It is particularly useful to zoom in to larger map scales for exploring regional distributions. Users can easily find out which publications are about a given site without having to search through academic literature themselves. Given the variable nature of beach names, this tool could save other researchers or environmental managers quite a bit of time in identifying research. Antaya's\textsuperscript{37} “Rangeland Literature Map of Arizona” is another example of an application of this tool to help rangeland managers identify publications about studies at specific sites in the state of Arizona.

Like the publications mapped in this case study, many studies in air, soil, and water research involve observations and sampling of a specific location. Knowing where studies have been
previously undertaken can be a valuable addition to future research, including understanding the geographical context of previously published literature and selecting future study sites.

Final remarks

The Literature Mapper tool aided the process of digitizing study site locations, storing the geometries, and creating a spatial data set that could be used as the foundation for visualizations. The digitizing tools varied in their ease of use. Digitizing single points performed as one would expect, but digitizing Multipoints is challenging for points separated widely in space. This is an issue the authors plan to address in future versions of the tool. The creation of vector layers from the stored geometries is seamless. The format of the geometry strings makes it easy to work with both in developing the graphical Literature Mapper tool and also in programming the web map. Using the format to enter a list of coordinates by hand from lists of coordinates printed in published papers was easy. Having demonstrated the value, and implemented a spatial bibliography system, we see the future potential for publication metadata and citation software to natively support spatial referencing.

Mapping study sites offers insight into an aspect of the published academic literature that would be difficult to otherwise understand. Given that our case study was from an ecological system with physical drivers such as wind and water erosion and deposition, it would be easy to see that these methods can be applied to other related ecosystems. For example, Literature Mapper could be used to understand the distribution of studies of air, soil, and water pollution, landslides, rare soil types, or limnology. It also has applications in other fields. For example, mapping economic research could provide insight into assumptions about where economies are studied and where we apply the knowledge gained from these publications and mapping the publication sites of works of fiction could show shifts in publishing hotspots over time. The tool can be applied to any line of study that has a physical location on the planet. These and other spatial questions related to publications are all now more easily explored with Literature Mapper.

Author Contributions

MMT: Conceptualization, Methods development, Data visualization, Paper writing and revision. AIM: Methods development, spatial technology. It also has applications in other fields. For example, Literature Mapper could be used to understand the distribution of studies of air, soil, and water pollution, landslides, rare soil types, or limnology. It also has applications in other fields. For example, mapping economic research could provide insight into assumptions about where economies are studied and where we apply the knowledge gained from these publications and mapping the publication sites of works of fiction could show shifts in publishing hotspots over time. The tool can be applied to any line of study that has a physical location on the planet. These and other spatial questions related to publications are all now more easily explored with Literature Mapper.

Author Contributions

MMT: Conceptualization, Methods development, Data visualization, Paper writing and revision. AIM: Methods development, Paper revision.

ORCID iDs

Michele M Tobias https://orcid.org/0000-0002-2954-8710
Alex I Mandel https://orcid.org/0000-0003-2024-6801

REFERENCES