[CV-02-003] Vector Formats and Sources

Abstract

In the last ten years, the rise of efficient computing devices with significant processing power and storage has caused a surge in digital data collection and publication. As more software programs and hardware devices are released, we are not only seeing an increase in available data, but also an increase in available data formats. Cartographers today have access to a wide range of interesting datasets, and online portals for downloading geospatial data now frequently offer that data in several different formats. This chapter provides information useful to modern cartographers working with vector data, including an overview of common vector data formats (e.g. shapefile, GeoJSON, file geodatabase); their relative benefits, idiosyncrasies, and limitations; and a list of popular sources for geospatial vector data (e.g. United States Census Bureau, university data warehouses).

Keywords: data, geodatabase, shapefiles, sources, topology, vector

Author & citation

Diamond, L. (2019). Vector Formats and Sources. The Geographic Information Science & Technology Body of Knowledge (4th Quarter 2019 Edition), John P. Wilson (ed.).
DOI: 10.22224/gistbok/2019.4.8.

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Source Materials for Mapping. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers.

Explanation

1. Definitions

- **attribute**: quantitative or qualitative information about the features in a dataset
- **coordinate system**: a system for determining the relationship between coordinates and their location on another surface
- **datum**: a coordinate system defined on a specific three-dimensional surface
- **feature**: an individual geographic location or list of locations defining a discrete geospatial phenomenon
- **geometry**: the spatial information associated with a feature or set of features
- **GIS software**: a web-based or desktop-based software tool for creating, editing, and analyzing spatial data
- **legacy**: outdated, old, and/or no longer in use or supported
- **Simple Style Specification**: an accepted specification for geospatial vector data; endorsed by the Open Geospatial Council (OGC) and International Organization on Standards (ISO)

- **spatial database**: a database that contains spatial information with an index for spatial queries
- **static map**: a map that does not have interactive functionality
- **topology**: the geographic relationships between features in a dataset
- **vector data**: geospatial data that has a location component represented as features with finite coordinates stored as points, lines, and polygons
- **web map**: A map that is published and accessed via the internet, usually as part of a web page. Web maps fall into one of two categories:
    - **static web maps** are map images rendered in the browser that do not change given user input. These include map images that the user can increase or decrease in scale via zoom functionality in the browser without changing the image itself.
    - **dynamic web maps** are web maps that change appearance as they are viewed by the user. There are two sub-types of dynamic maps:
        - **animated web maps** change frequently and automatically, using time to represent one or more data attributes. Animated maps may include interactivity, but at least include controls allowing the user to pause, play, and adjust the starting point of the animation.
        - **interactive web maps** change in response to user input. Because of the ubiquity of this type of map on the internet today, many people casually think of “web maps” as synonymous with interactive web maps.

2. Vector Data Formats

2.1 Vector Data Basics

Vector data is composed of spatial features that are assigned a geographic location. The model of vector data is a set of features, each represented as a point , line, or polygon with an associated coordinate pair to mark each location. Points are represented as a single coordinate pair; lines consist of an ordered list of point coordinate pairs; and polygons are created from an ordered list of point coordinate pairs where the first point and the last point are the same, making a closed loop (Fang 2014). Compared to raster data (see raster data section), vector data can have larger file sizes, but has near infinite precision of geographic features; this lends itself to the common aphorism, “raster is faster but vector is correcter.” (Goodchild 1998) Geospatial data also contains attributes . Attribute data contains observed or calculated attributes related to each feature, often including results of spatial analysis that contribute to the theme of the map. This information can be quantitative or qualitative; for example, a polygon representing a park may include an attribute for name, open hours, amenities, etc. This information is typically stored in a database or table.

Combined, the geometry and attributes of a vector dataset provide a range of options for analysis and visualization. As examples, see GIS&T Body of Knowledge entries on Symbolization and Visual Variables, Scale and Generalization, and Statistical Mapping (Enumeration, Normalization, Classification).

2.2 Vector Data Formats

Each vector format represents a different way of storing geospatial data on a computer. Like most aspects of the mapping process, each data format is more or less suited to

specific types of projects. The main organizational dichotomies for geospatial vector data include how data is stored (georelational or object-based), what information is stored (topology), and whether the data is programmatically associated with other data (simple or composite) (Chang 2015).

Georelational data stores the geographic and geometric data separately from the attribute data. Shapefiles are an example of georelational data, as the attribute information is stored separately (in the .dbf file) from the geometric data (in the .shp file). Object-based data stores this information together in a single system. Object-based geospatial vector data models include GeoJSON, which stores attribute data and geometric data in the same way.

Topology describes the geographic relationships between features in a dataset, such as when two polygons share a border. This information is not included in every vector data format; some formats use a “spaghetti data model” (Dangermond 1982), which stores each feature as a list of discrete coordinate pairs. Most common vector data formats, however, do include topology information.

Simple data describes a specific type of feature and is the most common way to store geographic data. Composite data combines simple features to create an additional set of information.

Each vector dataset contains one or more features and describes each feature’s geometry/geographic location and attributes. But in addition to these basic characteristics, each format is also optimized for certain uses, including data types, data complexity, software programs, and visualization needs. This is by design, as geospatial data spans a wide range of uses (Longley et al. 2015). Additionally, some providers of geospatial data only provide data in a specific format. Table 1 contains a list of the most common geospatial vector data formats and their best uses.

Table 1. Common Geospatial Vector Data Formats

Format	Description	File extension(s)	Best Uses	Notes
Shapefile	One of the most common vector data formats; an open source format designed originally to be used with ArcGIS software	.shp, .shx, .dbf, .prj	Static map design and spatial analysis with desktop GIS software; spatial analysis with OGR2OGR; some web mapping services can consume and convert shapefiles	• Has indexing • Preserves geometric accuracy • No topology • Good for static/print cartography • Not usable in web/mobile maps • Has table-based data storage • Supports all geometries • Supports all projections • Open source
KML/KMZ	Stands for "Keyhole Mark Language"; XML-based storage system	.kmz, .kmz	Static map design with desktop GIS software; can be used with some web mapping frameworks; Google Earth visualization	• No indexing • No topology • Usable in web/mobile maps • XML-based data storage • Supports all geometries • Does not support projections • Open source

Format	Description	File extension(s)	Best Uses	Notes
CSV	Stands for "comma-separated values"; can be visualized as a table; can only represent point data; does not support projections; does not store	.csv, .tsv (tab-separated values)	Static map design with desktop GIS software; some web mappiong services can consume and convert CSVs	• No indexing • No topology • Usable in web/mobile maps • Table-based data storage • Only supports point geometries • Does not support projections • Open source
GPX	An XML-based format for storing tracked GPS coordinates			• No indexing • No topology • Usable in web/mobile maps • XML-based data storage • Only supports points and polylines • Does not support projections • Open source
GeoJSON/TopoJSON	A key-value format written in JavaScript Object Notation	.js, .json, .geojson	Web map display and design; spatial analysis with OGR2OGR	• No indexing • Topology in TopoJSON • No topology in GeoJSON • Usable in web/mobile maps
Geodatabase	An ArcGIS-specific format with several implementation methods	.gdb	Statis map display and design with ArcGIS desktop software; interactive map display and design with ArcGIS Online	• Indexing • Topology • Only usable with ArcGIS
PostGIS	A geospatial database extension for PostgreSQL databases	N/A	Storage for large amounts of geospatial data	• Indexing • Topology • Requires setting up a database
SpatiaLite	A geospatial database extension for SQLite	.sqlite	Storage for large amounts of geospatial data	• Indexing • Requires setting up a database
Vector tiles	A lightweight storage format for tiled vector data using protocol buffers	.pbf, .mvt	Web map display and design	• Is not feature-based
OpenStreetMap	A key-value format written in XML	.osm	Source data for analysis; must be converted to	• Has its own geometry system • Does not support projections

2.3 Data Characteristics

The wide variety of geospatial vector data formats have been formulated in response to the wide range of potential geospatial needs. As such, each format is optimized for a particular purpose, whether it be specific analyses, cartographic outputs, data types, or something else. Defining a set of common data characteristics and their best uses is helpful in determining which data type should be used.

Indexing
An index is an access structure for a database that allows for speedy retrieval of individual records in that database. The index exists separate from the database table itself and references specific records in that table so a search can be conducted without traversing the entire table. Once the target of the search has been identified in the index, it provides a pointer to that specific record (Elmasri 2011).

Indexing is helpful when querying large amounts of data or conducting complex analysis tasks. This makes indexed data most useful for static map output (print or digital). Vector data formats that provide indexes include Shapefile, File Geodatabase, PostGIS, and SpatiaLite.

Topology
As defined above, topology describes the geographic relationships between features in a dataset, such as when two polygons share a border. Vector data formats that provide topology include TopoJSON, File Geodatabase, PostGIS, and SpatiaLite.

Projections
Geospatial data can be encoded using one of infinite coordinate systems. But not every data format supports projected coordinate systems, with some only able to store geographic coordinates (latitude and longitude). If you require data in a specific coordinate system or a map that needs a specific projection, it is best to use a data format that supports projections. Data formats that support projections include Shapefile, File Geodatabase, PostGIS, and SpatiaLite.

Geometric accuracy
Almost all geospatial data formats can store data without simplification; all points, lines, and polygons in the dataset are preserved with every vertex in place. The exception here is the Vector Tile format, which can remove vertices from lines and polygons and/or drop point features as data is created at each individual zoom level.

Vector Tiles are a lightweight data format optimized for visualization in web and mobile mapping applications. Many data types can be tiled, but the Mapbox Vector Tile Specification is the industry standard (Quinn 2018). Vector data geometry and attributes are stored in tiles rather than as individual features, which makes them fundamentally different from all other vector data formats. Vector tiles are typically used in web and mobile mapping, as the vector tile data can be served to and rendered on a user's device, increasing the speed and versatility of the map as the user pans and zooms. For more information on using vector tiles and tiling for web mapping in general, see web mapping chapter.

2.4 Data Use Cases and Conversion

In addition to the inherent characteristics of data types, it is important to choose data that is compatible with your desired tasks and output.

2.4.1 Use Cases

- **Interactivity on web and mobile**. When publishing maps for web and mobile

consumption, the data must be in a format that the software displaying the maps can understand. With web and mobile maps, users can interact with the map via panning, zooming, and clicking to get more information about a particular feature or location (Longley et al. 2015). Whether or not data can be used in web and mobile maps depends on the software being used to create those maps. Note that some software tools will allow you to provide data in non-web-compatible formats, but once uploaded the software converts to a web-compatible format. Vector data formats that are optimized for use in web and mobile maps are KML/KMZ, GeoJSON, and Vector Tiles.

- **Analysis**. If you are conducting additional analysis on the data once it has been created or obtained, you must ensure that the format is usable with the analysis tools you are using (Boden 2018). For example, if you are using QGIS to conduct spatial analysis, a proprietary format like a File Geodatabase that can only be used by ArcGIS would not be your best choice.
- **Data sharing**. If you are creating derivative data that will be shared with others, it is wise to use a data format that is compatible with a variety of software systems. It is also important to consider how that data will be distributed, as certain formats require more disk space than others and cannot easily be shared via the internet. Vector data formats that are common for sharing data are shapefile, GeoJSON, OpenStreetMap (for OpenStreetMap data), and CSV (for point data).

2.4.2 Data Conversion

As the diversity of geospatial data uses and needs has grown, so too have the number of data formats. In response to this trend, tools for automatically converting data between systems are being integrated into the most common GIS software today (Longley et al. 2015).

While the data you need to use for your project may not be in the correct format, there are many tools for converting data between one format and another. Users should use caution when converting data, however, as some data formats do not contain enough information to be useful when converted; for example, GeoJSON data can support multiple geometries within the same dataset, whereas shapefiles are limited to a single geometry type per dataset. Converting a multi-geometry GeoJSON file to shapefile may cause issues, either with the conversion software or with the resulting dataset.

There are several tools for converting data between formats. The most popular conversion tools are ArcGIS, ogr2ogr, and uploading data to Mapbox (converts data to Vector Tiles).

3. Vector Data Sources

Geospatial data has been available for consumers to access on the World Wide Web since at least the mid-1990s (Crampton 1995). The scope and variety of data available online has grown significantly over time, providing a cost-effective means of transmitting data to many users and yielding a wide range of available resources (Longley et al. 2015).

Data is available at a variety of scales (global, country, state, local) from governments, universities, private companies, and individuals. Some sources are aggregates of data from other sites, so the same data may be present in multiple locations. Please note that this list

is not comprehensive and represents a small subset of available data repositories. The following were selected based on the number of accessible data sources contained in each resource as well as the recency with which the repository was updated.

3.1 National / Federal Government Sources

This data is useful for large-scale analysis across a wide geographic area and is typically available in a range of common data formats. When using federal data, be sure to note the date the data was collected, the level of aggregation in the dataset, the completeness of the data, and the attribute information embedded (see Metadata, Quality, & Uncertainty). Sometimes federal data sources separate the geometry information from the attribute information, which requires joining a table of attributes to your data.

Several United States federal government agencies maintain and regularly publish geospatial data, both at national and local scales (see Table 2 for a selected list of data sources).

Table 2. Selected Geospatial Data Sources from the United States Federal Government

Source	Description	Sample datasets
National Historic Geographic Information System	National geographic boundary and demographic information for the United States from 1790 to present; funded by the National Institute for the Humanities and the National Science Foundation.	County, state, census tract, metropolitan area boundaries from 1790 to present; American Community Survey and Census demographic data; agricultural census data
National Archives	Every geospatial dataset owned, provided, or maintained by the National and regional data from the Department of Energy, federal government. Has an associated API.	National and regional data from the Department of Energy, Department of the Interior, and any other agency that produces geospatial data
Data.gov	Data (geospatial and non-geospatial) aggregated from a variety of government sources (including local governments).	Local government, agriculture, education, climate, disasters, manufacturing, maritime, etc.
GeoPlatform	Developed under the auspices of the Federal Geographic Data Committee (FGDC). Includes an API for accessing federal geospatial data, including real-time data	NOAA, USDA, NASA, and other agency datasets. E.g., hourly precipitation data for the United States from NOAA
U.S. Census Bureau, Maps and Data	Canonical datasets for the United States, including TIGER/Line shapefiles denoting political geographic boundaries	Political boundary data for state, country, metro, and census tracts.
American Fact Finder	Tabular data from the U.S. Census Bureau. This data can be cleanly joined to other geospatial data from the U.S. Census Bureau	CSVs at the state, county, metro, and census tract level for all demographic data
USDA Geospatial Data Gateway	High-resolution raster and vector data compiled from several different agencies	National agricultural data

Other countries also maintain and publish their own geospatial data sets (Table 3).

Table 3. Selected Geospatial Data Sources from National Government Agencies (non-USA)

Source	Country	Description
Government of Canada Open Data Portal	Canada	National datasets aggregated by the Government of Canada. The platform contains both geospatial and non-geospatial data in a variety of formats, both proprietary and open.
data.gov.uk	United Kingdom	Datasets from central government and local governments in the UK.
Open Government Data Platform India	India	National and regional data for India
Portal Satu Data Indonesia	Indonesia	National and regional data for Indonesia
Instituto Brasileiro de Geografia e Estatistica	Brazil	National and regional data for Brazil
data.go.jp	Japan	National and regional data for Japan
EU Open Data Portal	European Union	Continental data for countries in the European Union. Includes current and historical data.

3.2 State and Local Data

Every state in the United States (and many regional and local governments internationally) have an online repository of state-related data, and many counties and municipalities provide their own data as well. Some states have an aggregated collection of state-wide data, whereas some states publish data directly from each agency separately. While certainly not every city has its own data online, major cities like San Francisco, New York, and Chicago provide local data online as well.

Table 4. Selected State and Local Government Data Sources

Source	Locality	Description
Oregon Spatial Data Clearinghouse	Oregon, USA	Repository for spatial data created for and generated by the State of Oregon
State of California Geoportal	California, USA	Spatial data aggregated and provided by the California Department of Technology
BYTES of the Big Apple	New York City, New York, USA	GIS data provided by the New York City Planning Department
Little Rock's Open Data Portal	Little Rock, Arkansas, USA	GIS data provided by the City of Little Rock

3.3 University Sources

Universities collect and provide data (either via links to other sources or by hosting it themselves) either through their geography departments, libraries, or some other university department. Data may be global, national, or specific to the university's region.

Table 5. Selected University-provided Data Sources

Source	Description
Data Sources and Repositories, Robert E. Kennedy Library, California Polytechnic Institute	Thematic and basemap data for the state of California
PennState Unviersity Libraries, Maps and Geospatial	Thematic and basemap data, focused on Pennsylvania

Source	Description
American University Library	Thematic and basemap data
Washington State University Library	Thematic and basemap data, focused on Washington and the Pacific Northwest

3.4 Organizations and Data Projects

Organizations that use geospatial data in their work occasionally provide access to their derivative data. Organizations whose purpose is to provide geospatial data are also included in this section. The two most notable geospatial data creation projects are Natural Earth and OpenStreetMap. Natural Earth Vector data was introduced to provide a global dataset of cohesive, small-scale, high-quality data for cartographers to use beneath their thematic data (Kelso 2009). OpenStreetMap aims to be a digital, crowd-sourced map of the world, accepting edits and additions from its users (Haklay et al. 2008).

Table 6. Selected Organization and Data Project Sources

Source	Description
The Nature Conservancy	US-based and global conservation data
World Wildlife Fund	Global conservation science data
OpenStreetMap	Global, crowd-sourced basemap data; does not include historical data
Natural Earth	Global physical and political data at a variety of scales
Open Geospatial Consortium 3D Data	3D spatial datasets

3.5 Individually-compiled lists and repositories

There are several individuals and trade organizations that provide lists of links to geospatial data. These individuals have not created or produced this data, but provide the links as a service to cartographers and those involved with GIScience.

Table 7. Selected and Repositories of Geospatial Data Sources

Source	Description
Free GIS Data, in a list curated by Robin Wilson, University of Southampton	An organized list of more than 300 links to free geospatial data; includes sections for physical geography, human geography, and area/regional data
American Association of Geographers Historical GIS Data Clearinghouse	A list of time-based GIS data from organizations around the world; most data is regional/national
Resources from Robin Tolochko	An aggregated list of geospatial resources and datasets, mostly focused on the United States

References

Boden, S. (2018). Use the Five-Step GIS Analysis Process. Esri Blog posting.

Chang, K.-T. (2015). Introduction to Geographic Information Systems, 8th Edition. New York: McGraw-Hill Education.

Crampton, J. (1995). Cartography Resources on the World Wide Web. Cartographic Perspectives, 22: 3-11.

Dangermond, J. (1982). A Classification of Software Components Commonly Used in Geographic Information Systems. In Proceedings of the U.S.-Australia Workshop on the Design and Implementation of Computer-Based Geographic Information Systems, 70–91. Honolulu, HI.

Elmasri, R. (2014). Fundamentals of Database Systems. 5th Edition. Boston: Addison-Wesley.

Fang, Y.; Shandas, V., and Arriaga Cordero, E. (2014). Spatial Thinking in Planning Practice: An Introduction to GIS. PDXOpen: Open Access Textbooks. Last accessed November 8, 2019.

Goodchild, M. F. (1998). Geographic Information Systems. Santa Barbara, CA: Center for Spatial Studies and Department of Geography, University of California, Santa Barbara.

Haklay, M., Singleton, A. and Parker, C. (2008). Web Mapping 2.0: The Neogeography of the GeoWeb. Geography Compass, 2 (6), 2011-2039.

Kelso, N. V. & Patterson, T. (2009). Natural Earth Vector. Cartographic Perspectives, 64: 45-50.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). Geographic Information Science and Systems (4th Edition). England: John Wiley & Sons.

Quinn, S. (2018). Vector tiles: the next generation of tiled maps. GEOG 585 Open Web Mapping. Last accessed November 8, 2019.