Geospatial Vector Data

Standards for Improved Data Sharing

Kyle Onda • July 2022

In this blog, I am going to describe some practices that the Internet of Water Initiative at the Center for Geospatial Solutions recommends when it comes to sharing geospatial vector data. But first, what is geospatial vector data, and why is it important as a component of water data?

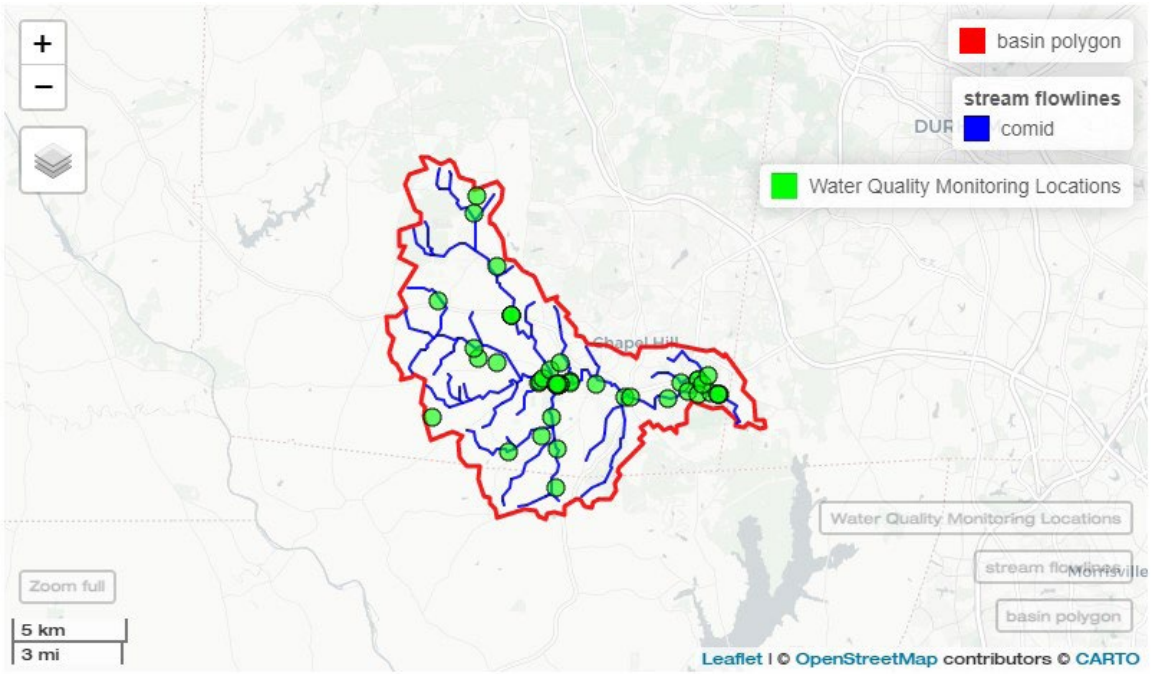
Geospatial vector data represents particular features on Earth (things with locations in GIS speak) and attributes about them. In general, vector data uses combinations of X-Y coordinates to describe the location, which might be a discrete **point**, **line**, **polygon**, or a collection of any of them. Vector data is important for water science and management, because it is at the foundation of how we generate, analyze, and communicate water data.

> ***"...data should be shared in ways that are easy for scientists and water professionals to analyze and for developers to work with to make decision support tools and public communication materials."***

For example, to analyze and describe a region's water quality, we may need to represent a watershed as a polygon, within which are stream segments represented as lines, near which are water quality sampling stations represented as points. All of these features in turn have various attributes that may include identifiers, names, and summary statistics regarding observed and modeled water quality metrics. The figure on the next page is composed of just such a combination of geospatial vector data around Morgan Creek near Chapel Hill, NC

It is important to be able to share geospatial water data – about watersheds, streams, aquifers, wells, streamgages, bridges, dams, culverts, conduits, effluent outfalls, administrative and management areas, etc. This data should be shared

in ways that are easy for scientists and water professionals to analyze and for developers to work with to make decision support tools and public communication materials. This is where **Data Standards** and **API Standards** come into play. This blog post focuses on Data Standards, and another blog post in this series will focus on API Standards.



GEOSPATIAL DATA STANDARDS

WHY?

Geospatial vector data standards are rules regarding the structure and format of data. They address concepts like how points, lines, and polygons can be expressed using coordinates (and coordinate reference systems), and how associated attributes for a given geometry are allowed to be structured and named. Data files that comply with a given data standard can then be used by the same software to be viewed, analyzed, edited, or written, which is great for accessibility! It means that your data can be easily used by those you share it with, as long as they have software that can work with the standard.

The Internet of Water Initiative at the Center for Geospatial Solutions recommends that water-related data should be shared using open (as in well-documented and with a license that allows anyone to use the standard without paying) and widely-supported standards so that more people can use the data. There are many

geospatial vector data standards, but not all are open or widely adopted. Thus, it is also generally considered good practice to provide the same data in multiple formats where feasible to maximize accessibility.

OK, SO WHICH STANDARDS?

Standards are constantly under development, and different data standards are suited to different uses. Thus, we recommend different standards for different anticipated uses.

EXPLORATION AND ANALYSIS OF AN ENTIRE DATASET

Many users wish to download an entire dataset and use their own tools and skills to analyze the data. We recommend the following data standards for this type of use case.

CURRENT RECOMMENDATIONS

- **OGC GeoPackage** (.gpkg) is a file database format based on SQLite. It can store raster imagery as well, but it can store all kinds of vector features in any coordinate reference system and associated attributes of many data types (numerical, string, dates). It combines attractive features of binary files (a single-file format that does not require dedicated software to run independently of any analysis tools) and databases (multiple tables and collections of features can be included, and individual rows of tables can be created, read, updated, deleted using SQL).
    - » **How to use:** GeoPackage files can be read and written by ESRI ArcGIS Pro, QGIS, Python, R, and many other geospatial analysis tools.
    - » **Best practices for sharing:** GeoPackage files are simple to share since they are one file with no particular size limits. In most cases, it is good practice to include a table in the GeoPackage as a data dictionary that defines the column names and data values for any feature tables included in the GeoPackage. GeoPackages can also be compressed if desired (e.g. .zip, .gz) to reduce file size.
    - » **Example in Water Data:** Here is an example GeoPackage. In this case, it includes river flowline features and a data dictionary table.
    - » **Caveats:** While GeoPackages can be viewed and written by ArcGIS Pro, they cannot be used in the ESRI ArcGIS ecosystem as a complete replacement for shapefiles or personal geodatabases (see below). This

is because they cannot be used directly (yet) for certain geoprocessing tasks or to host feature layers in ArcGIS Online/ Enterprise Portal. They must be converted to feature classes first.

- **Shapefile** (.shp, .shx, .dbf, .prj*) is the most widely used vector data format, although it is quite old and has several disadvantages (see Caveats) that GeoPackage addresses. Due to the Caveats below, we recommend sharing GeoPackages alongside shapefiles. It is a multifile format distributed in a single folder, minimally with .shp storing the geometry, .shx holding an index to allow fast querying, and .dbf storing the attribute table. There are many optional files, notably .prj storing the coordinate reference system.
  - » **How to use:** shapefiles can be read and written directly by almost all widely used geospatial data analysis tools. All files should be in the same folder when working with shapefiles. The .dbf attribute table can be opened by Microsoft Excel and other spreadsheet software.
  - » **Best practices for sharing:** To allow downloads of shapefiles, including all constituent files, it is best to place all the files into a compressed .zip folder so that they can be downloaded in one transaction. The .zip file would then be unarchived by data users before using in most cases, although some tools can read zipped shapefiles directly. In most cases, it is good practice to include a separate .csv file as a data dictionary that defines the column names and data values for the feature attribute table in the shapefile.
  - » **Example in Water Data:** Here is an example shapefile. In this case, it includes river flowline features. Note that it is compressed into a .zip archive that includes a .csv data dictionary.
  - » **Caveats:** Shapefiles, while widely supported, have several weaknesses:
    - multifile formats are vulnerable to incomplete sharing which makes the data useless due to missing files
    - limited attributes (255 column limit in the attribute table)
    - limited data values, including text limited to 254 characters
    - limited column names (limited to 10 characters)
    - limited to 2GB in size
    - there is no NULL value, so 0 cannot be distinguished from missing data in numerical fields, requiring constructs like using “-9999” as a missing code.
    - limited to one layer/table of one geometry type.

- **ESRI file geodatabase** (.gdb) is a proprietary file database developed by ESRI with similar capabilities to GeoPackage which does not have the Caveats associated with shapefiles. It is optimized for use within the ESRI ArcGIS ecosystem. It can be distributed as a folder with the extension .gdb, within which are many binary files. Due to being optimized for use with ESRI software, it is recommended to use this standard only if only ESRI users are intended to be supported. Otherwise, it is recommended to distribute GeoPackages alongside geodatabases.
    - » **How to use:** geodatabases can be read by ESRI ArcGIS Pro, QGIS, Python, and R, however, they can only be written using software that includes a licensed ESRI SDK.
    - » **Best practices for sharing:** geodatabase folders are best shared compressed. In most cases, it is good practice to include tables in the geodatabase as data dictionaries that define the column names and data values for any feature tables included in the geodatabase.
    - » **Example in Water Data:** Here is an example file geodatabase. In this case, it includes a variety of natural resource areas associated with groundwater. Note that it is distributed as a .zip compressed file.
    - » **Caveats:** While geodatabases can be read by many geospatial softwares, the process is generally more complex for non-ESRI software, and write or edit capabilities are severely limited or impossible without an ESRI license.

EMERGING STANDARDS

- **OGC GeoParquet** (.parquet) is an emerging bulk vector data format based on Apache Parquet. It is being designed to enable geospatial data workflows that involve cloud data warehouses, by providing a standard way to represent vector data in a highly efficient columnar data format that is fast to read and write, even for large volumes of data.

WEB VIEWING AND DATA STREAMING

Sometimes it is not desirable to download an entire dataset and use it in analysis software. Rather, users might only need to visualize data in a web mapping application in a browser window. Or perhaps, users may only need to access a small subset of data for a particular area. We recommend the following data standards for this type of use case.

- **GeoJSON** (.geojson or .json) is a vector data format based on JSON. As such, they are plain text files, very easy to parse by most web development tools and web browsers, and easy to work with for web developers. For many use cases, it can be considered a default standard for publishing geospatial vector data on the web.

  » **How to use:** .geojson files can be read and written by any text editor, as well as ESRI ArcGIS Online or Enterprise Portal, QGIS, Python, R, and many other geospatial analysis tools that use GDAL. However, an important use case for them is embedding them in web maps using web development languages like JavaScript, specifically in frameworks like Leaflet or OpenLayers.

  » **Best practices for sharing:** geojson files can be shared as single files, and can be compressed if desired (e.g. .zip, .gz) to reduce file size. However, a more appropriate use case for geojson is to be hosted on the web and accessible via HTTP request, which may involve being returned by a geospatial data web API (look out for the next blog post on Geospatial API Standards). In either case, it should be noted that the coordinate reference system of any shared geojson data must be WGS84.

  » **Example in Water Data:** Here is an example geojson hosted on the web, which can also be visualized directly in a web map. In this case, it includes dam location data.

  » **Caveats:**

  - GeoJSON files tend to be space inefficient, increasing in size with increasing amounts of data faster than many other data formats. This can make web maps based directly on geojson files very slow to load if there are many and/or complex geometries.
  - WGS 84 latitude/longitude is the *only* supported coordinate reference system
  - For ESRI users, geojson files can be used easily by ArcGIS Online but NOT ArcGIS Pro without the interoperability extension

- **FlatGeoBuf** (.fgb) is a developing binary vector data format designed to allow streaming, random access to the data. That is, the file can be queried partially from a remote web location.

  » **How to use:** .fgb files can be read by QGIS, Python, R, and many other geospatial analysis tools that use GDAL. However, an important use case for them is embedding them in web maps using web

development languages like JavaScript, specifically in frameworks like Leaflet or OpenLayers.

» **Best practices for sharing:** .fgb files can be shared as single files, and can be compressed if desired (e.g. .zip, .gz) to reduce file size. However, the most appropriate use case for this format is for it to be hosted at a web location accessible via HTTP request, so that web maps built in Leaflet, OpenLayers, and similar frameworks can request data from the .fgb.

» **Example:** Here is an example FlatGeoBuf hosted on the web, which is directly accessed to create this web map.

» **Caveats:**

- .fgb files are complex and not suited for transactional data management processes. They should be treated as static datasets that need to be replaced entirely with new versions when updates are needed.
- For ESRI users, there is currently no support for this format.
- .fgb is an open-source project but is not part of any community standards development process, so its development trajectory and long-term support are uncertain.

For additional guidance on geospatial vector data in a water data context, feel free to reach out to Kyle Onda (konda@lincolninst.edu).

Header Photo, John Thomas on Unsplash

Footer Photo, Andreas Sjovall on Unsplash