Professur für Geodäsie und Geoinformatik
Prof. Dr.-Ing. Ralf Bill
M.Sc. Markus Berger

Tutorial: Data Formats

Data Formats and their Use: Tips and Tricks

In this course unit common data formats shall be described and their use in the processing of open geodata shall be explained.

In addition, tips and tricks are collected to facilitate the joint processing of the data.

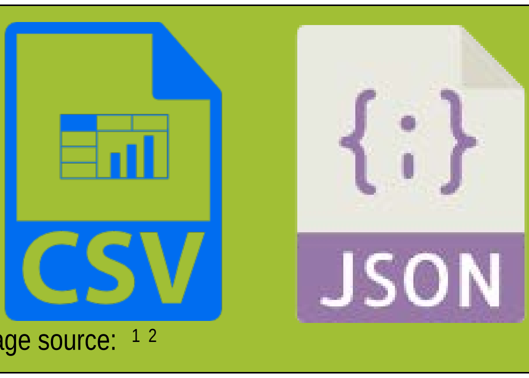

Image source: [1] [2]

Introduction

A wide variety of formats are used for processing open geodata. This depends on the one hand on which software is to be used, but on the other hand also on the format in which the provider of open data delivers it. Figure 1 illustrates this diversity of formats using GovData (www.govdata.de/), the open data portal of the German National and Federal Government, as an example.

It can be clearly seen that almost half are text-based formats such as csv, html or xml. A good 10% are provided in Excel format, Microsoft's proprietary format. Together with the text-based form csv, it can be guessed that about one third of the data offered is factual data/attributes in the sense of further processing in geographic information systems (GIS). Almost 10 % are pdf documents (e.g. legal texts, reports etc.), which are rather poorly accessible for further processing.

Almost a further third of the openly provided data is geoinformation in the broadest sense, which can be downloaded in a wide variety of processing maturity, with about 25% mostly only as a map service, i.e. suitable for viewing or as a background map. Only 7% of geodata formats are probably available to be used directly in GIS.

[1] Bildquelle: <div>Icons made by <a href="http://www.freepik.com" title="Freepik">Freepik</a> from <a href="https://www.flaticon.com/" title="Flaticon">www.flaticon.com</a> is licensed by <a href="http://creativecommons.org/licenses/by/3.0/" title="Creative Commons BY 3.0" target="_blank">CC 3.0 BY</a></div>

[2] <div>Icons made by <a href="https://www.flaticon.com/authors/smashicons" title="Smashicons">Smashicons</a> from <a href="https://www.flaticon.com/" title="Flaticon">www.flaticon.com</a> is licensed by <a href="http://creativecommons.org/licenses/by/3.0/" title="Creative Commons BY 3.0" target="_blank">CC 3.0 BY</a></div>

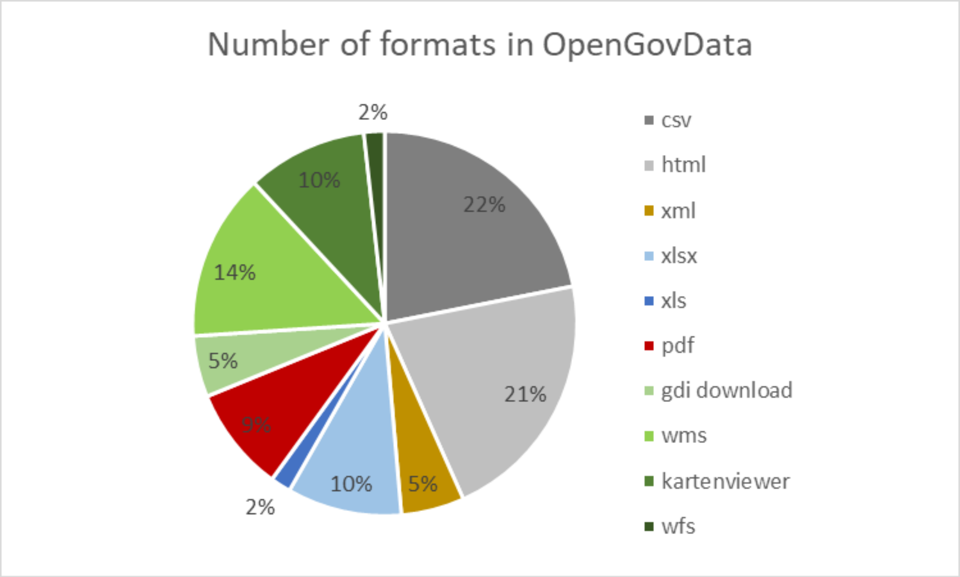

Fig. 1: Formats in OpenGovData (status 28.9.2017, 20683 hits, 42761 documents due to multiple format offers).

Basic formats

Binary files

These files do not follow a standardized character set, and often have no file extension. They contain arbitrary sequences of bits, without a directly associated interpretation. This is only brought along by the program with which they are read. Many programs define their own file extensions in order to mark the affiliation. If you open a binary file in a text editor, you will often find only unreadable characters.

Text files (TXT)

These files are usually stored as .txt files under Windows. TXT text files have no prescribed structure or notation, but are little more than a hint for the operating system to interpret the binary file contents as a sequential sequence of characters from a certain character set. Which character set is assumed depends on the operating system and the selected editor, and can often be changed in the same. The most common examples are ASCII or UTF-8. If the character set is interpreted incorrectly, unreadable misrepresentations often occur.

In the context of data processing a certain structure is often assumed or dynamically determined in .TXT files, since any other text-based format can be stored in such a file. In such cases, the file contents are often formatted like a CSV file.

Structured text formats

CSV format

CSV stands for comma-separated values and describes the structure of a text file for the storage or exchange of simply structured data. In CSV files, tables or a collection of lists of different lengths can be mapped. The file name extension is .csv.

A general standard for the CSV file format does not exist, but it is fundamentally described in RFC 4180. The character encoding to be used is also not specified; 7-bit ASCII code is widely regarded as the lowest common denominator.

The following excerpt shows a worldwide country record with abbreviations for nation names as well as the full name, area size, population, a regional subdivision from continent to region, and center coordinates in geographical longitude and latitude, as can be seen from the first row in the sense of the column headings. With regard to interpretation, it is usually necessary to refer to an accompanying text that describes the dimensions, abbreviations, etc. in detail.

Excerpt 1: Appearance of a csv file in the editor using the example of a worldwide country data record

```
FID_;FIPS;ISO2;ISO3;UN;NAME;AREA;POP2005;REGION;SUBREGION;LON;LAT
;AC;AG;ATG;28;Antigua and Barbuda;44;83039.000000000000000;19;29;-61.7830009;17.0779991
;AG;DZ;DZA;12;Algeria;238174;32854159.000000000000000;2;15;2.6320000;28.1630001
;AJ;AZ;AZE;31;Azerbaijan;8260;8352021.000000000000000;142;145;47.3950005;40.4300003
;AL;AL;ALB;8;Albania;2740;3153731.000000000000000;150;39;20.0680008;41.1430016
```

Tips and tricks

csv files can be opened with different text editors (Editor, Wordpad) and modified easily (search for certain characters, change of separators, change of special characters and German umlauts), if processing programs have problems with it.

XML format

The Extensible Markup Language (XML) is a markup language for structuring documents and data. Markup Languages provide rules by which parts of a text can be marked in a certain way in order to add additional semantics and properties, usually with the aim of making a text machine-readable.

XML is a simplified, extended version of SGML (Standard Generalized Markup Language), and is therefore just like SGML a meta language: a language that can be used to define a more specific language depending on the application.

XML is composed mainly of pairs of tags that behave like parentheses. If a tag A is opened in tag B, the tag must be closed before tag B is closed. Some tags stand alone and do not open a "bracket". Attributes can be specified directly in the tags, the rest of the content is between the pairs.

Excerpt 2: Sample XML file

```
<person>
        <name> John </name>
        <isAlive> true </isAlive>
        <age> 25 </age>
        <address>
                <cityStreet> New York, 21 2nd Street </cityStreet>
                <postalCode> 10021-3100 </postalCode>
        </address>
        <children> </children>
        <spouse> </spouse>
</person>
```

JSON format

The JavaScript programming language was designed with a focus on web applications and object orientation. Since the web was to be largely based on text-based protocols, a generally valid, structured data format for objects was required. From this idea the JavaScript Object Notation, short JSON, was born. As an exchange format, JSON is today the largest competitor to the XML format.

It is not only text-based, but also easily readable by humans. Objects consist of sets of attribute-value pairs. The objects themselves do not have a fixed structure according to the tradition of JavaScript, further attributes can be named and added at will. The attributes are represented textually, for the values there are the following types:

- Number: There is no difference between integer and float.
- String: A sequence of characters, marked by quotation marks.
- Boolean: The possible values are True or False.
- Array: An ordered list of values, marked with square brackets and separated by commas.
- Object: An unordered set of attribute/value pairs, marked by curly braces and separated by commas, the pairs themselves are divided by a colon. Same notation as the parent objects.
- null: An empty value, indicated by the word 'null'.

> ***Tips and tricks***
>
> JSON files are human readable, often clearer than XML files with the same content. However, errors are sometimes more difficult to detect than in XML, since the content and notation are less strict.

Excerpt 3: Sample JSON file

```
“Person”: {
        "name": "John Smith",
        "isAlive": true,
        "age": 25,
        "address": {
                "cityStreet": "New York, 21 2nd Street",
                "postalCode": "10021-3100"
        },
        "children": [ ],
        "spouse": null
}
```

Document formats

HTML format

A markup language for hypertext documents based on SGML (Standard Generalized Markup Language). Hypertext documents are structured, digital documents that can be linked to each other. HTML includes support for hyperlinks, images, meta information and other content by default. Documents are divided into HTML headers and HTML bodies. The header contains general information such as titles and metadata, the body contains the actual content displayed.

HTML mainly defines the semantics, but does not contain any information about formatting. The formatting is often defined in separate CSS files (Cascading Style Sheets), and can also vary depending on the browser. If dynamic behavior and more complex functions are required, JavaScript code can also be added.

Due to its importance in today's Internet, the language specification is constantly being expanded. Thus, versions of the language have been designed that are not based on SGML, but on XML. The latest extension is HTML5, which no longer strictly follows SGML or XML, and among other things includes greatly enhanced support for dynamic and multimedia content.

Excerpt 4: Sample HTML file

```
<!DOCTYPE HTML>
<html>
        <head>
                <title> Webkarte <title>
                <meta charset=”utf-8”>
        </head>
        <body>
                <div id=”map”> Load map here  </div>
        </body>
</html>
```

PDF format

The Portable Document Format (PDF) was developed by Adobe System to enable documents to be transported between different platforms so that they retain their original formatting. Typical conversion pitfalls are bypassed by not implicitly specifying structures such as "heading" or "paragraph here", but by explicitly specifying positions of texts and images in a vector format. Even used fonts can be provided directly in the .pdf files. It is also possible to append further data such as tables of contents and comments.

Because PDF was developed as an exchange format for finished documents, it is difficult to change them. Some programs allow minor changes such as correcting typos, but working on PDF files should always be done via the original file and re-exporting to PDF format.

Table and database formats

XLS/XLSX/XLSM format

The Microsoft Excel spreadsheet program is suitable for processing data organized in tables. The file extension .xls is the proprietary file format of versions 97 up to and including 2003, after which the extension is .xlsx (or .xlsm if macros are built in).

Excel provides extensive calculations with formulas and functions, an organization in workbooks that correspond to files, sheets that appear in tabs, and cells that contain the data. Up to 65,536 spreadsheets are possible per workbook. The cells of a spreadsheet are divided into rows and columns and can be accessed via a cell reference system. Each cell can be uniquely identified by a combination of letter and number, the so-called cell reference, which consists of row and column specifications. The first cell in the upper left corner is A1, where A is the first column and 1 the first row, and the reference also includes the worksheet name, since formulas in different sheets and folders can have the same reference, such as Table1!C4 + Table3!C4.

A number of formats are available for displaying values. In addition to ready-made formats such as date and time and special formats for postal codes, user-defined formats can be specified. Cells can also contain text in Excel.

Tips and tricks

When using Excel, you should pay attention to meaningful formats of the characteristics/attributes specified column by column (instead of standard, e.g. a number for which decimal places can be defined, text for descriptions, date formats for time specifications, etc.).

Once you have opened the Excel file, you should check whether all cells are filled with values. Data gaps should be provided with unique values that do not lie within the value range of the characteristic. If, for example, one has measured outside air temperatures in °C, the value range will usually be between -25° and + 40°. If no value is available, this can be clearly indicated by the value -99. This can then be filtered out more easily in a later evaluation.

When importing Excel files into GIS software, problems often occur when transferring the file formats in the cells. The GIS software then displays e.g. <NULL>. A simple export from Excel to csv format often helps here. The csv file can then be read into the GIS program.

DBF format

The file format is defined for dBase databases, one of the oldest database management systems. Although dBase itself is hardly in use, its .dbf files are still found in some long-lived applications. They are together with ArcGIS shapefiles and can still be opened by MS Excel, LibreOffice and Open Office Calc.

It is a relatively complex binary format, but basically it allows saving a table, with headers for the overall table and for the individual columns. The contents of the fields are stored in ASCII (a simple text format). The column header contains information about the data type of the ASCII code in the fields of the column, i.e. how it should be interpreted.

Tips and tricks

Many of the limitations encountered when using ArcGIS shapefiles are due to the age of the DBF format. Only a very limited number of bytes is provided for column names. There are also problems with German umlauts. These are not provided in the ASCII standard, and must be interpreted afterwards by means of some ASCII codes, which were left free for international special characters. If this interpretation does not work correctly, problems will occur.

When using shapefiles, it is best to avoid long column names and umlauts.

Vector data formats

SVG format (based on XML)

The Scalable Vector Graphics (SVG) format is the web standard for the description of vector graphics, i.e. graphics which are not defined by their pixels but consist of primitives such as lines, polygons and curves, and which can be dynamically redrawn (rendered) depending on screen size, zoom level and resolution.

They are not suitable for storing photos, but for the many illustrations and graphics that can be found on almost every website. Since SVG is based on XML, it is a textual image format that can be read and changed by humans. Also animations are provided in the standard, without changing dynamically the XML file. There are different time scales on which changes can be fixed, e.g. "Rotate for 5 seconds" or "Change your color with a mouse click".

The format is mainly oriented towards display, and does not store any additional information such as spatial reference. Points cannot be defined in a superordinate coordinate reference system (CRS, see Tutorial Coordinate Reference Systems), they have to be drawn relative to the currently given "canvas" by means of a filled circle.

Excerpt 5: Sample SVG file (header incomplete)

```
<svg version=”1.0” …>
        <circle cx=”10” cy=”20” r=”5”>
        <polyline points=”5,20 10,25 15,20 10,15”>
</svg>
```

DXF format

The Drawing Interchange File Format (DXF) comes from the context of CAD programs and was developed by Autodesk. Today it is the industry standard for data exchange between different CAD solutions. .dxf files are normally not used internally, but are imported and exported into the formats of the program used. The format can be stored and used both textually and binary.

Raster data formats

JPEG format

The JPEG file format (actually JFIF, for "JPEG File Interchange Format") is a standard for storing raster images according to the JPEG standard. The JPEG standard goes back to the Joint Photographic Experts Group of ISO. There are other such standardized file formats, but .jpg is the most commonly used, even if it does not implement the complete standard.

The standard, and with it the file format, focuses on offering different compression methods, depending on the intended application. Classically, JPEG files are used today to compress and save photos lossy. The often seen blocky artifacts of photos spread on the Internet are mostly due to too frequent or too drastic JPEG compressions. The step-by-step improvement of the image resolution, which is rarer nowadays but still visible with weak internal reception, is also a JPEG function.

Since the storage is often carried out without further configuration offers by the image processing program used, many users are not even aware that JPEG also supports lossless compression, the selection of different color spaces and different image composition methods.

Tips and tricks

JPEG should, if possible, only be used to store photos, not graphics with many monochrome, planar image parts. The compression methods are tailored to photos and the human perception of complex images, and at lower levels provide drastically reduced image sizes with barely noticeable quality degradation. The compression methods are less effective for graphics, and quickly have a massive impact on their quality. PNG is better here.

PNG format

PNG (Portable Network Graphics) is another widely used data format for images. In contrast to JPEG, however, it is lossless by default. It focuses primarily on graphics, not photos, and supports functions such as transparency, different color modes, and metadata.

PNG was designed as a free and open replacement for the GIF format because it initially used patented algorithms and allowed little color precision. However, GIFs still enjoy some popularity, as they allow animations unlike PNG.

TIFF format

Less common on the Internet than JPEG and PNG, is TIFF, short for Tagged Image File Format. The reason for this is that it mainly focuses on professional users such as graphic designers, photographers and publishers.

TIFF allows you to precisely customize the image file to the intended use by using different tags that can be added to the image and interpreted by programs. Some of these tags are predefined, such as defining used color spaces, compression methods, and many other steps that are required in other image formats. Tags can also be defined at will for any special application, and then only need to be interpreted correctly by the appropriate specialized software.

Geodata formats

GeoTIFF format (based on TIFF)

This is less a separate file format than a collection of additional tags for the TIF format. These add coordinate reference systems and projection information to the normal TIFF file. Combined with the ability for high color accuracy, saving multiple images per file and lossless saving in TIF format, GeoTIFF has become a standard in raster geospatial processing. It is often used, for example, in the processing of satellite and aerial photographs where accuracy is of the utmost importance.

GeoJSON format (based on JSON)

GeoJSON defines the structured storage of spatial information within the JSON format. For this purpose, certain attribute/value pairs must be assigned to the spatial objects, e.g. what type of geometry the object is as a string, and the geometry data itself as an object.

Excerpt 6: Sample GeoJSON file

```
{
        "type": "Feature",
        "geometry": {
                "type": "Point",
                "coordinates": [125.6, 10.1]
        },
        "properties": {
                "name": "Dinagat Islands"
        }
}
```

SHP format

A binary format for spatial geometries developed and openly available by ESRI. It is one of the most widely used GIS data formats and is the de facto standard, at least outside the web.

Shapefiles, in contrast to other formats, do not exist only as .shp files. Several other files with the same name but different endings must be available for correct use, such as .dbf (attribute tables), .shx (index tables) and .prj (projection data). .shp itself only covers the individual geometries, but the SHP format also supports attribute values, indexing, projections and much more. These are stored in their own files. Due to the binary formatting shapefiles are not human readable, but can be opened in nearly all programs that allow the import of geodata.

Tips and tricks

Many programs will only display the .shp file in the file browser when importing shapefiles. The remaining required files will be loaded automatically. If there is a problem with a shapefile, you should first check whether the other files also exist and are stored in the same location.

GML format (based on XML)

The Geographic Markup Language (GML) is both defined as a standard and specified by the OGC. This is a specialization of XML, with a focus on spatial objects. It allows the specification of objects, geometries and coordinate reference systems, but also of other concepts such as map styles. The main purpose of GML is to act as an exchange format for web-based geo infrastructures.

KML format (based on XML)

Similar to GML, the Keyhole Markup Language (KML) is a geographic specialization of XML, but originally focused on use in Google Earth. It therefore contains not only geometric information, but also extended information to visualize the data. GML can even be extended with style information using KML: Both are part of the OGC standard.

GeoRSS format (based on XML)

RSS feeds are used to publish changes in the content of a web service. These are detected and stored on the server, and then actively retrieved by the recipient. There is no need to maintain server-side subscriber lists. However, users need a program for querying and subscribing to feeds. The publications are formatted in XML, so they can be read directly by humans as well as processed automatically.

Based on RSS GeoRSS then allows to publish georeferencing in a standardized way. This is realized among other things by the use of GML instead of XML.

WKT format

An OGC (Open Geospatial Consortium) file standard for text-based storage of geodata such as vectors and coordinate reference systems.

Excerpt 7: Examples for object representation in WKT format

```
POINT X (4468298 5333781)
LINESTRING(1 12, 10 20, 5 6, 1 12)
GEOMETRYCOLLECTION( POINT(1 12), LINESTRING(1 12, 10 20))
```

Literature

Bill, R. (2016): Fundamentals of Geo-Information Systems. 6th edition. Wichmann Verlag. Offenbach-Berlin. 867 pages. Chapter 8.6.

Seip, C., Korduan, P., Zehner, M.L. (2017): Web-GIS. Basics, applications and implementation examples, Wichmann Verlag, Offenbach-Berlin. 552 pages.