Using the statistical language R as a Geographic Information System

Introduction: Data Management & Geospatial Data

Stefan Jünger / GESIS – Leibniz Institute for the Social Sciences

November 23, 2021

DOI: 10.5281/zenodo.5717830

cessda.eu @CESSDA_Data

About This Course

This (short) workshop will teach you how to exploit R and apply some of its geospatial techniques.*

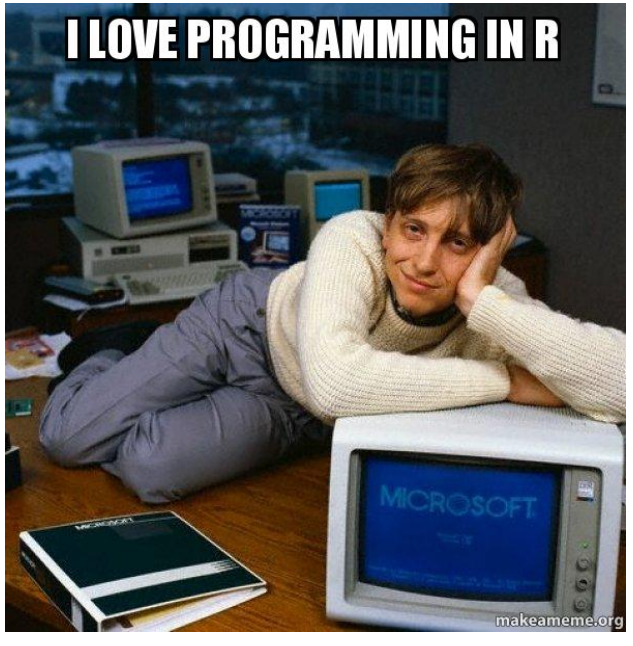
By the end of this course, you should...

- be less frightened with using geospatial data in R
  - including importing, wrangling, and exploring geospatial data
- be able to create (simple) maps based on your very own processed geospatial data in R

*Some materials are part of a previous workshop, for which I'd like to thank Anne-Kathrin Stroppe. Libby Bishop helped me with the information about CESSDA.

Prerequisites for This Course

- At least basic knowledge of `R`, its syntax, and internal logic
  - Affinity for using script-based languages
  - Don't be scared to wrangle data with complex structures
- Working versions of `R` (ideally `Rstudio`) on your computer



Source

About Me

- Postdoctoral researcher in the team Data Augmentation at the GESIS department Survey Data Curation
- Ph.D. in social sciences, University of Cologne

- Research interests:
  - quantitative methods
  - social inequalities & attitudes towards minorities
  - data management & data privacy
  - reproducible research

stefan.juenger@gesis.org | @StefanJuenger | https://stefanjuenger.github.io

Preliminaries

- The workshop consists of a combination of a few lectures and hands-on exercises
- Feel free to ask questions at any time
- Slides and other materials are available at:

https://github.com/StefanJuenger/CESSDA-R-GIS

The workshop will also be recorded. All slides, other materials and the recording will be shared on CESSDA channels.

Course Schedule

Time	Title
09:00-09:30	Introduction: Data Management & Geospatial Data
09:30-09:35	Exercise 1: R Warm up
09:35-10:00	Data Processing & Spatial Linking
10:00-10:30	Exercise 2: Geospatial Data Wrangling
10:30-10:45	Break
10:45-11:15	Easy Maps
11:15-11:45	Excercise 3: Build your own map
11:45-12:00	Closing, Q & A

Now

Time	Title
09:00-09:30	Introduction: Data Management & Geospatial Data
09:30-09:35	Exercise 1: R Warm up
09:35-10:00	Data Processing & Spatial Linking
10:00-10:30	Exercise 2: Geospatial Data Wrangling
10:30-10:45	Break
10:45-11:15	Easy Maps
11:15-11:45	Excercise 3: Build your own map
11:45-12:00	Closing, Q & A

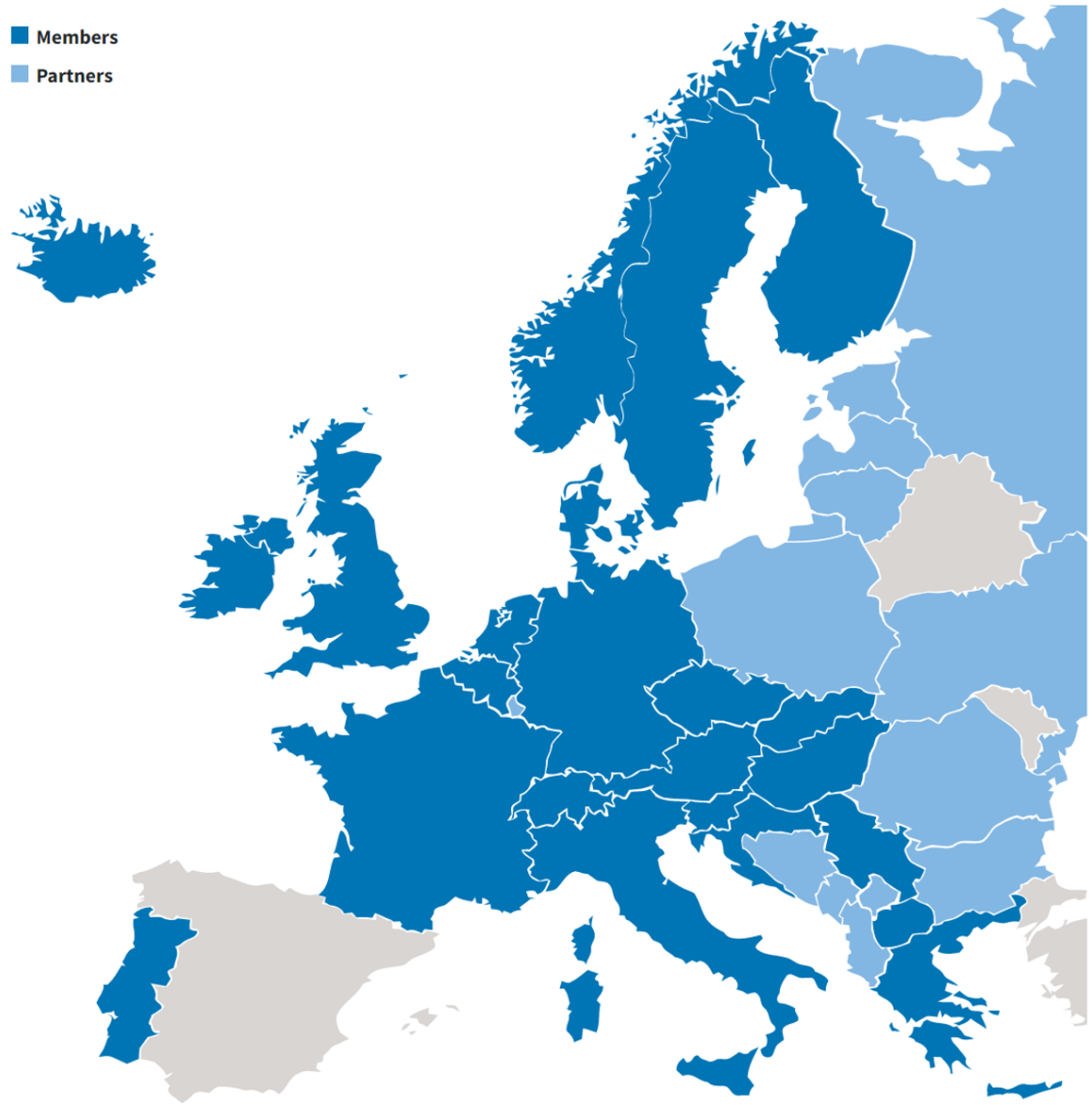
CESSDA

Our vision is that the provision of **access to social science data** and metadata is vital – for both science and society.

For this we must offer **services to data producers** to easily describe and store their data – if needed in a secured environment.

We will adhere to the **FAIR** (Findable, Accessible, Interoperable, Reusable) data principles to make data findable and provide information about the data, where they are, how they can be accessed.

We will also focus on providing **training** and enabling the transfer of expertise and sharing of knowledge on data, as well as relevant rules and regulations.



Tools & services

cessda
DC Data Catalogue

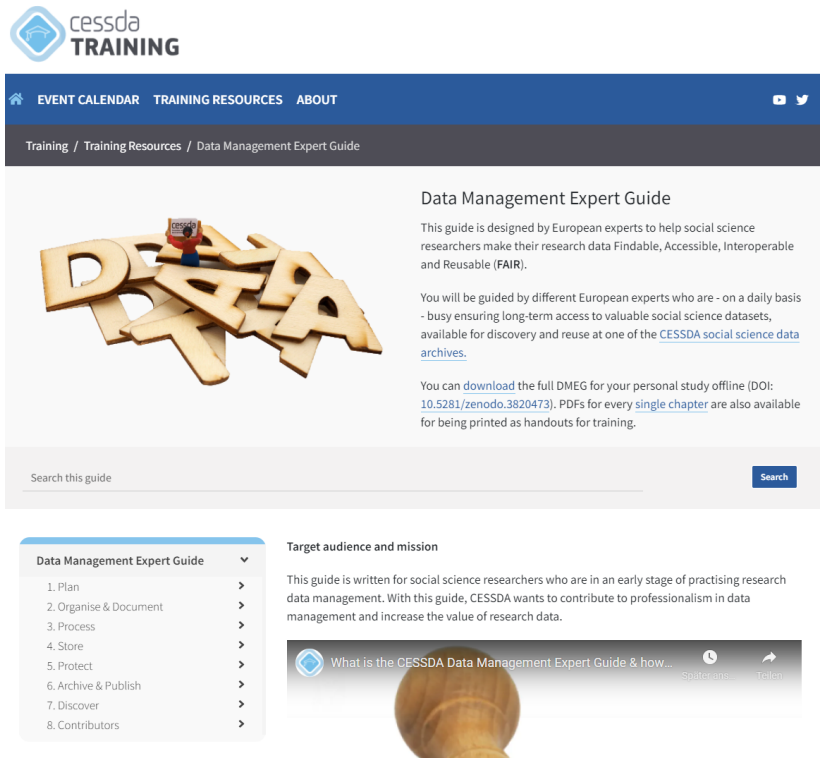
cessda
DMEG Data Management Expert Guide

cessda
VS Vocabulary Service

cessda
ELSST Thesaurus

cessda
TRAINING

CESSDA Data Management Expert Guide



https://www.cessda.eu/DMEG

Data Processing

« Previous Next »

Main take-aways

After completing your journey through this chapter you should:

- Be familiar with strategies to minimise errors during the processes of data entry and data coding;
- Understand why the choice of file format should be planned carefully;
- Be able to manage the integrity and authenticity of your data during the research process;
- Understand the importance of a systematic approach to data quality;
- Be able to answer the DMP questions which are listed at the end of this chapter and adapt them to your own DMP.

The content of this chapter was inspired by research data management manuals, guidelines, online courses and methodological texts published by several data organisations and experts, in particular the information provided by the UK Data Service (2017a), the "Guide to Social Science Data Preparation and Archiving" by the US-based data organisation ICPSR (2012), the online course Research Data MANTRA (EDINA and Data Library, University of Edinburgh, 2017), A guide into research data management by Corti, Van den Eynden, Bishop and Woollard (2014), Krejčí's "Introduction to the Management of Social Survey Data" (Krejčí, 2014), Gibbs (2007) and Data Management Guidelines produced and published by the Finnish Social Science Data Archive (Finnish Social Science data Archive, 2017).

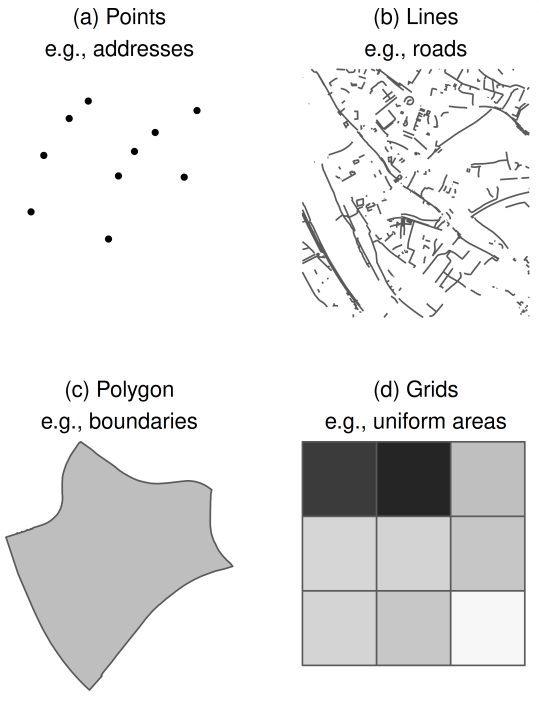
What Are Geospatial Data?

Data with a direct spatial reference → **geo-coordinates**

- Information about geometries
- Optional: Content in relation to the geometries

Can be projected jointly in one single space

- Allows data linking and extraction of substantial information



Sources: OpenStreetMap / GEOFABRIK (2018), City of Cologne (2014), and the Statistical Offices of the Federation and the Länder (2016) / Jünger, 2019

Geospatial Data in This Course

In the folder called `./data` in the same folder as the other materials for this workshop, you can find the data files prepped for all the exercises and slides.

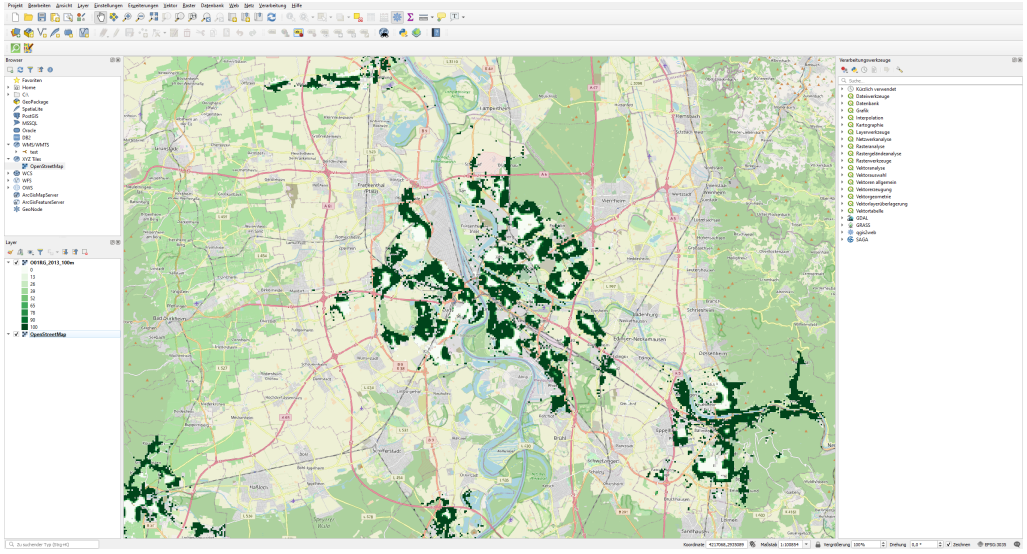
- Covid-19 cases for Cologne across the city's districts available at the Open Data Portal of Cologne
- Hospital locations in Cologne are also distributed via the Open Data Portal of Cologne
- Cologne's road network; guess from where
- Number of immigrants and inhabitants from German Census 2011 data are provided by the Federal Statistical Office Germany, Wiesbaden 2020

Please make sure that if you reuse any of the provided data to cite the original data sources.

What Is GIS?

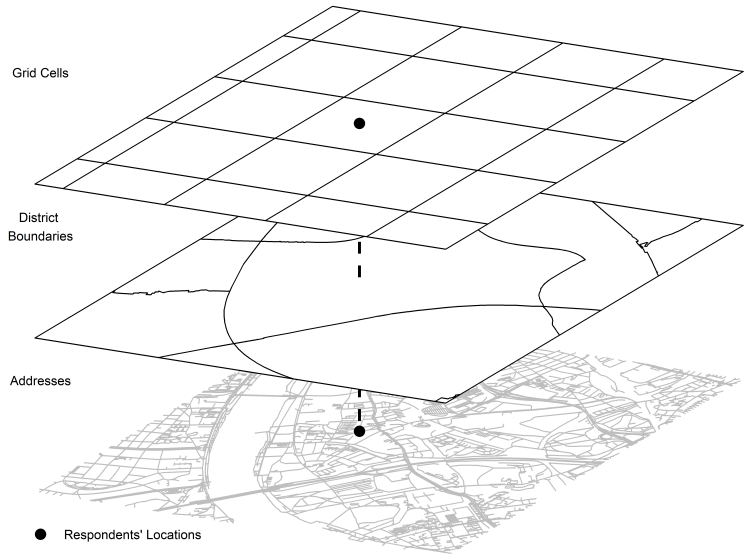
Most common understanding: Geographic Information Systems (GIS) as specific software to process geospatial data for

- Visualization
- Analysis



Screenshot of the Open Source GIS QGIS

Data Specifics



Sources: OpenStreetMap / GEOFABRIK (2018) and City of Cologne (2014)

Formats

- Vector data (points, lines, polygons)
- Raster data (grids)

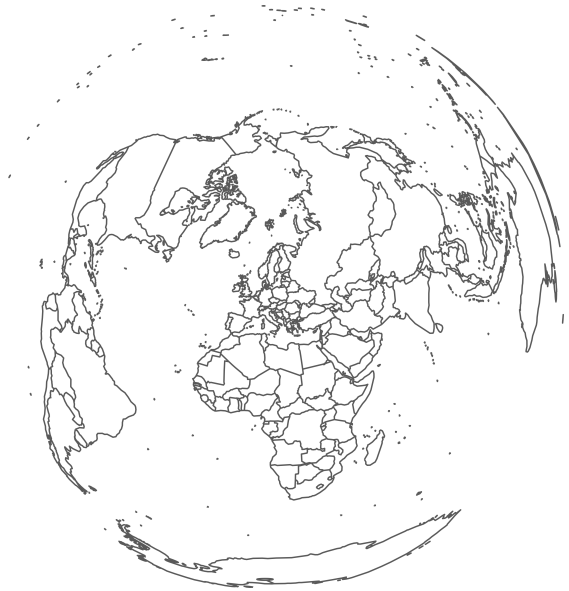
Coordinate reference systems (CRS)

- Allow the projection on earth's surface
- Differ in precision for specific purposes

Layers Must Match!

Projected CRS (EPSG:3857)

Geographic CRS (EPSG:3035)



Source: Statistical Office of the European Union Eurostat (2018) / Jünger, 2019

Types of CRS

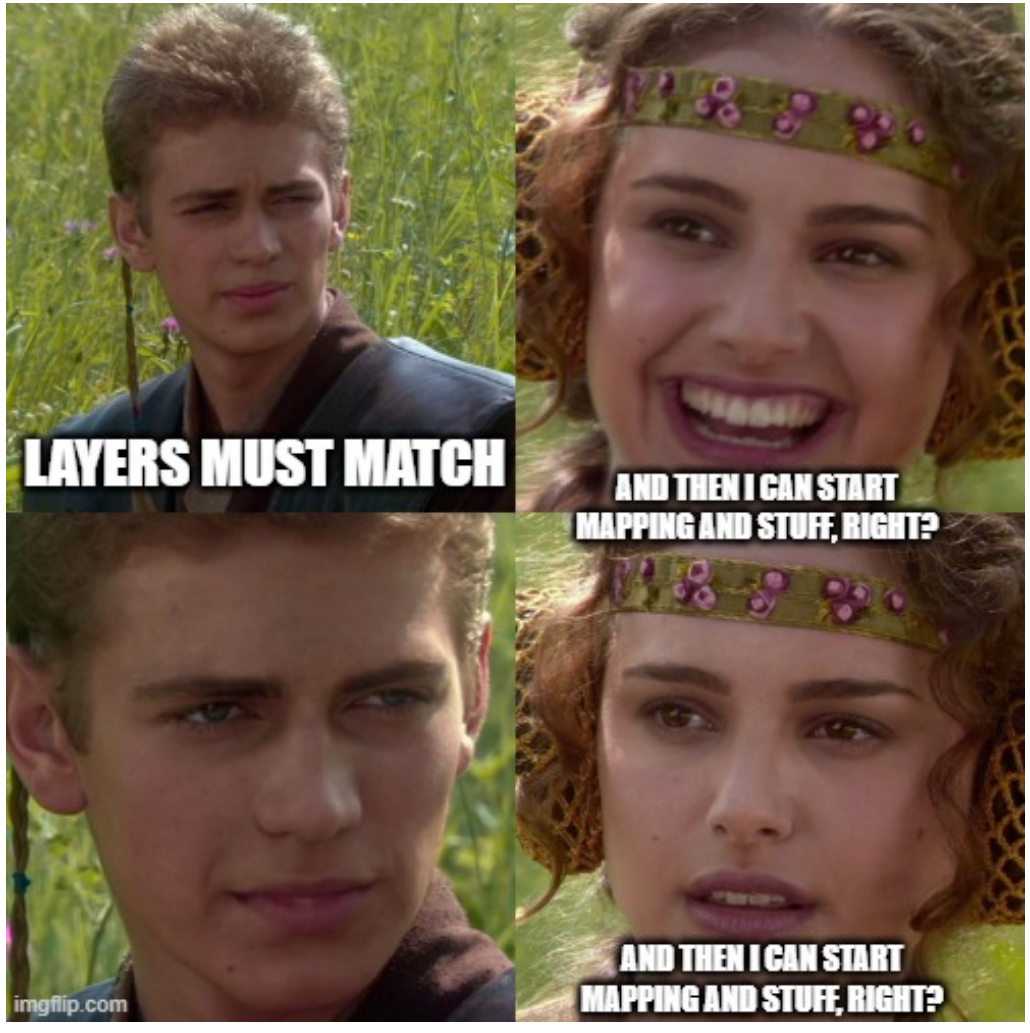
Projected CRS

- projection of geometries on a flat surface (planar)
- distance between two points is a straight line

Geographic CRS

- (unprojected) description of specific points on a sphere
- distance between two points is a bent line

(There are also geocentric CRS requiring a z-coordinate...)


LAYERS MUST MATCH
AND THEN I CAN START MAPPING AND STUFF, RIGHT?
AND THEN I CAN START MAPPING AND STUFF, RIGHT?
imgflip.com

CRS Definitions

Old Standard: PROJ.4 Strings

```
+proj=laea +lat_0=52 +lon_0=10 +x_0=4321000 +y_0=3210000 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs
```

Source: https://epsg.io/3035

New Kid in Town: WKT ("Well Known Text")

```
PROJCS["ETRS89 / LAEA Europe",
    GEOGCS["ETRS89",
        DATUM["European_Terrestrial_Reference_System_1989",
            SPHEROID["GRS 1980",6378137,298.257222101,
                AUTHORITY["EPSG","7019"]],
            TOWGS84[0,0,0,0,0,0,0],
            AUTHORITY["EPSG","6258"]],
        PRIMEM["Greenwich",0,
            AUTHORITY["EPSG","8901"]],
        UNIT["degree",0.0174532925199433,
            AUTHORITY["EPSG","9122"]],
        AUTHORITY["EPSG","4258"]],
    PROJECTION["Lambert_Azimuthal_Equal_Area"],
    PARAMETER["latitude_of_center",52],
    PARAMETER["longitude_of_center",10],
    PARAMETER["false_easting",4321000],
    PARAMETER["false_northing",3210000],
    UNIT["metre",1,
        AUTHORITY["EPSG","9001"]],
    AUTHORITY["EPSG","3035"]]
```

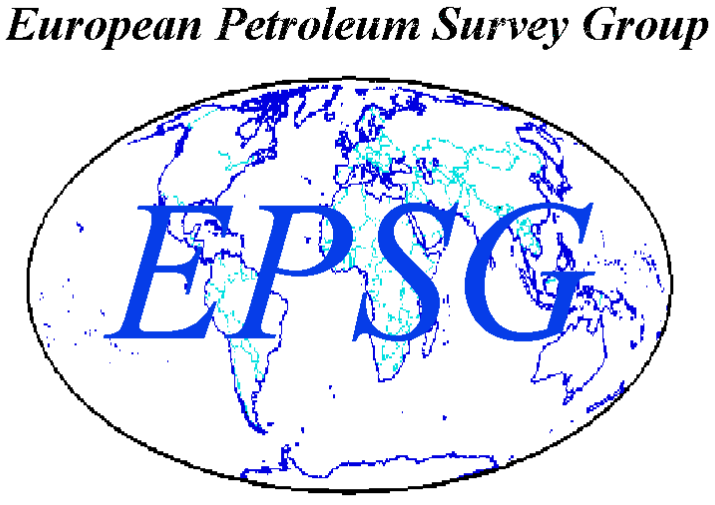
EPSG Codes

Eventually, it's not as challenging to work with CRS in `R` as it may seem

- we don't have to use PROJ.4 or WKT strings directly.

Most of the times it's enough to use so-called EPSG Codes ("European Petroleum Survey Group Geodesy")

- Small digit sequence

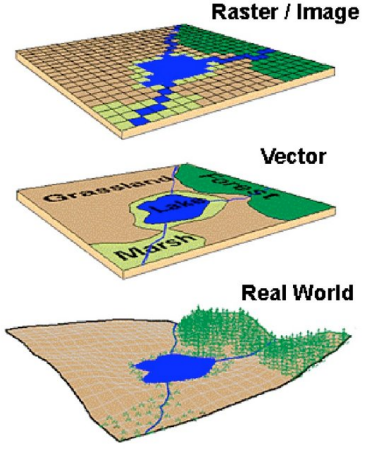


Source

More Details on Geospatial Data

Let's learn about geospatial data as we learn about specific formats:

- vector data
- raster data



Source

Be assured: `R` can serve as a full-blown Geographic Information System (GIS) for all these different data types.

R Packages for Geospatial Data

There have been packages for geospatial data in R already for a long time.

- `sp` for vector data
- `raster` for raster data
  - the successor: `terra`

Cutting-edge for vector data and raster data (cubes)

- `sf`
- `stars`

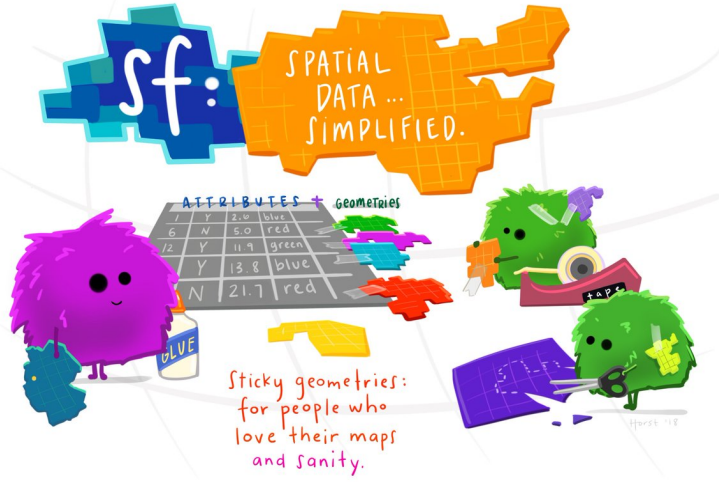


Illustration by Allison Horst

Packages in This Course

We will use plenty of different packages during the course, but only a few are our main drivers (e.g., the `sf` package). Here's the list of packages

- `dplyr`
- `sf`
- `stars`
- `tmap`
- `tmaptools`

Note: Some additional packages will be installed as dependencies.

What You'll See During the Course: Piping In R

Usually, in R we apply functions as follows:

```
f(x)
```

In the logic of pipes this function is written as:

```
x %>% f(.)
```

We can use pipes on more than one function:

```
x %>%
  f_1() %>%
  f_2() %>%
  f_3()
```

More details: https://r4ds.had.co.nz/pipes.html

Exercise 1: R Warm-Up

Exercise

Solution

stefan.juenger@gesis.org
@StefanJuenger
StefanJuenger
https://stefanjuenger.github.io

cessda.eu

@CESSDA_Data