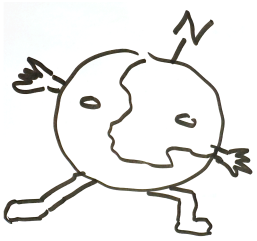
DCN Planetears
Geospatial Data Curation:
an introduction



Module 1: GIS Introduction Instructor Notes

[This handout accompanies a slidedeck with the file name “1_GIS_Introduction”]
2024-02-08 1000

Lesson Plan: GIS Introduction (65 min)

Objectives:

- Describe a geographic information system (GIS)
- Identify common tool sets for working with GIS data
- Describe characteristics of vector GIS data
- Describe characteristics of raster GIS data
- Conceptualize structure of a GIS project
- Understand the three principal components of a Coordinate Reference System

Slides Notes: see slides (35 min)

Activities (10 min)

- Using the data in, 1-4_Excersise_Dataset create a GIS project with one raster and one vector layer
    - Overview of Layers and Browser panes
    - Look at Information Panel: find projection, pixel dimension, other metadata
    - Locate export layer panel: explore options and CRS selector

Check Understanding (5 min)

- Share out GIS project

Quiz (10 min)

Wrap up (5 min)

Notes for Slides (approximately 35 minutes total):

1. **Welcome** to the Data Curation Network's training for Geospatial Data Curation: an introduction.
2. **The Curriculum Development Team** pictured here collaborated from October 202 to March 2024 to create this data curator's intro to geospatial data curation.
3. **Module List:** This is Module 1: GIS Introduction. You should have already been through Module 0: Environment Setup and you should have working software on you machine.
    a. Later modules are: Module 2: Ethics and GIS Data;
    b. Module 3: Common GIS Data Types;
    c. Module 4: GIS Metadata; and,
    d. Module 5: Transformations
4. **Suggested Primers:** The Data Curation Network has created a number of curation Primers that you will also find useful as you curate geospatial data. These are:
    a. Geodatabase Primer https://github.com/DataCurationNetwork/data-primers/blob/master/Geodatabase%20Data%20Curation%20Primer/Geodata-Primer.md
    b. GeoJSON Primer https://github.com/DataCurationNetwork/data-primers/blob/master/GeoJSON%20Data%20Curation%20Primer/GeoJSON-data-curation-primer.md
    c. GeoTIFF Primer https://github.com/DataCurationNetwork/data-primers/tree/master/GeoTIFF%20Data%20Curation%20Primer
    d. netCDF Primer https://deepblue.lib.umich.edu/handle/2027.42/145724

5. **Welcome (back) (30 seconds)**
   We now have our workspace on the computer organized and ready to go (often the hardest part). Let's begin with GIS Basics.

6. **Introduction (2 minutes)**
   We have several learning objectives in this module. Learners should be able to:
    1. Describe a geographic information system (GIS)
    2. Identify common tool sets for working with GIS data
    3. Describe characteristics of vector GIS data
    4. Describe characteristics of raster GIS data
    5. Understand the three principal components of a Coordinate Reference System
    6. Conceptualize structure of a GIS project

   Note that:
    a. We can reduce these objectives to three questions.
        i. What is a GIS?
        ii. What are GIS data?
        iii. What is a GIS Project?

b. We cannot teach GIS in one or two days. Instead we intend to share some working knowledge that will get students started with GIS data curation and perhaps inspire future self- or structured-learning opportunities.
c. Share the GIS Cheat Sheet and note the connections between the learning objectives and the cheatsheet (front side).

Note also that for this entire curriculum we use the terms “geospatial data” and “GIS data” as synonymous, even though there is a great deal of debate within the geospatial data community as to whether they are actually synonyms.

7. **Defining GIS (2 minutes)**
   We can define GIS as:

   A collection of tools, data, hardware, and people used to locate and describe features on the Earth’s surface.

   Humans have been performing this activity well before the advent of computers and digital information systems. The idea of layered data is something we do in our minds or with tools such as paper, plastic, or computers. Key to note in the diagram the two distinct data structures: discrete objects like streets, parcels, and customer locations versus continuous grids or images like elevation or land use.
   a. Often the purpose of GIS is to make a map, a model, or perform an analysis
   b. Maps are models **not** representations

8. **Defining GIS in the Literature (1 minute)**
   There is much jargon and terminology in the GIS world. A good place to start is with the term GIS itself - note that it has several meanings within different user groups
   a. As a system
   b. As a science
   c. As development of tools

   To help you understand all the terms and jargon, these modules are accompanied by a Glossary of GIS Terms

   Often people think of ArcGIS (the ESRI toolset) when they hear GIS, but there is so much more …

9. **Useful Abstractions (1 ½ minutes)**
   This well-known triangle abstraction of how humans understand the world is simple by nature. It reminds us that maps are not representations of the world, but instead models of what we perceive or models of what we want the future to be.
   a. This abstraction can help us locate the use of GIS at all levels of the construction of knowledge -- better said the academic or scientific process.
   b. There are no discrete lines between the layers.
   c. The abstraction on the right can help us locate how the first three (perhaps four) layers of the knowledge creation process can map to physical data files stored on some sort of digital device.

10. **Elements of GIS Assemblages (1 minute)**

When we think of GIS it is useful to separate the tools, data, and languages used to organize the system of geographic information. With this categorization we can then better understand how data is related to software packages and scripting languages.

a. Sometimes these relationships are enforced in proprietary systems where the data will only work with specialized software packages and scripts (an analogy would be SPSS files and associated scripts in the statistics world).
b. In more open systems, or open source systems, the relationships are more fluid, but the distinctions between data, tools, and scripts still exist.
c. Note that common GIS data models include tables, trees, and grids

11. Desktop GIS Assemblages: ESRI (1 ½ minutes)

The proprietary ESRI GIS toolset is the industry standard in North America. Note there are several toolsets and ESRI sells sets of licenses for different collections of tools. Some academic institutions purchase these toolsets through an enterprise collection of licenses, but many institutions cannot afford the steep price and turn to other solutions. As we move forward through this curriculum, we will see these tools, data models, and language referenced, but we will instead use the open-source QGIS for this workshop.

Here are some of the tools that ESRI offers

a. ArcGIS Pro (the standard desktop mapping software)
b. ArcMap Desktop (legacy marked for deprecation, many people still use this)
c. ArcCatalog (a tool for data management)
d. ArcGIS Online (both free and subscription based)
e. ArcGIS Enterprise Server

12: Skippable Slide: Download/Login Instructions for ArcGIS Pro (30 seconds - 5+ minutes depending on if downloads need to be done)

These are the instructions for downloading the program if the participants have not been already prompted to do so. If everyone has already downloaded the program without a problem feel free to skip this slide. If there are still some who need to download the program we suggest navigating to the Esri website and show where the download options are. You may also need to demonstrate how participants can receive a login from your institution.

Then open ArcPro on your computer in order to perform some live demonstrations of tasks.

13. Desktop GIS Assemblages: QGIS (2 ½ minutes)

QGIS, or Quantum GIS, is rapidly becoming a global industry standard for many geospatial data practitioners, especially those that cannot afford the ESRI licenses (disadvantaged communities or underfunded government agencies) or those that are dedicated to open source solutions (as is much of Europe). Note that QGIS is free open-source software (FOSS) mostly written in Python (based on other open source libraries OGR/GDAL).

a. Note that the shp, tif, and sql data structures are shared with ESRI tools as well as the well established and open source python scripting language.
b. Even National Geographic is becoming a QGIS shop …
c. FOSS for GIS - no license fees - and freedom to modify and use as fits your purpose - teaching moment for open source software, versioning, and git if desired ….. https://github.com/qgis/QGIS

d. GDAL - Geospatial Data Abstraction Library
e. OGR - OGR Simple Features Library (prior OpenGIS Simple Features Library)

14: Skippable Slide: Download Instructions for QGIS (30 seconds - 5+ minutes depending on if downloads need to be done)
These are the instructions for downloading the program if the participants have not been already prompted to do so. If everyone has already downloaded the program without a problem feel free to skip this slide. If there are still some who need to download the program we suggest navigating to the QGIS website and show where the download options are.

Then open QGIS on your computer in order to perform some live demonstrations of tasks.

15. Code-Based GIS Assemblages (1 minute)

The world of GIS has gone through rapid changes/evolution in the past several decades. Python is a long standing language used for data crunching, including GIS data. R is a little more recent. Both are experiencing a surge in use for processing and analyzing all kinds of geographic data within the academy, government agencies, and the private sector. This is particularly true for high performance computing (HPC) environments.

a. These tools share certain aspects of data structures from both ESRI and QGIS.
b. FOSS for GIS ….
c. A return to the initial style of command-line computing form the 1960s and 1970s

16. Geospatial Data (2 minutes)

Ok, time to shift gears and dig a little further into the two data structures mentioned earlier: vector and raster. We are also going to continue to familiarize ourselves with some key terminology based on the relationships between file formats and data structures, in this case vector data for discrete features (parcels, streets, customers, and so on) and raster data for continuous phenomena (elevation or land use as examples).

a. Vector and Raster are key terms in the world of GIS
b. Emphasize that vector and raster have different file formats
c. Emphasize the connection between continuous observations and grids (raster) and discrete observations and features (vector)

17. Vector Data Model: Pints, Lines, and Polygons (2 minutes)

The vector data model is built on a relational data model. There are two basic elements for the model that are joined by a feature identifier (FID): locational data stored as x,y coordinate pairs, and attribute data attached (related) to the locational data. Note that collections of points can form line or polygon features which then have the attached attribute data. Put in simple English this means that several different kinds of data (coordinate pairs, collections of coordinates, and attributes) are stored in different tables (like excel spreadsheets) and linked (related) by a feature ID (FID).

a. Key terms to remember: feature, attribute
b. Good for displaying information at different scales
c. Too few points (coordinate pairs) leads to inaccuracy at small scales (very zoomed in)
d. Too many points leads to computational and visualization problems at large scales (very zoomed out).

18. Vector Data Model: File Formats (1 minute)

Just some common file formats and their extensions to look for. We will see this information again with more detail in the data types module, but start to recognize some of these common extensions.

a. Shapefiles: .shp is the main file extension
    i. each shapefile is actually a collection of files that include: dbf, sbn, shx, sbx, prj,, xml (and more - be careful!!) - relational model
    ii. Industry standard for vector data
b. Geojson is a text based format used for information exchange between websites or applications - not necessarily good for archiving.
c. Geodatabase is actually a folder with files inside

19. Raster Data Model: Grids and Images (2 minutes)

The raster data model in GIS is most commonly associated with satellite images or some other color based representation of the surface of the earth. Another common raster file is a digital elevation model or DEM. The use of “raster” instead of “image” in GIS terminology lies in the fact that we can also store non-color data such as decimal or integer numbers in the rows and columns of the data. As examples: elevation, temperature, population density, and so on. Often the terms grid and raster are interchangeable in the world of GIS.

a. Cells are based on center points with a specific cell size.
b. Cells can only contain one value. For example the measured elevation in a DEM is stored in each cell as the unique value.
c. For multiple values, we must have multiple “bands” or layers in the raster data. For example color images are stored with three bands with values for intensity of red, green, and blue.
d. Rasters can be categorical (like the example on the slide). For categorical data, we must have a look up table that associates numerical values with meaning (a data dictionary).

20. Raster Data Model: Resolution (1 minute)

Unlike the vector data model, raster data has a very specific scale that cannot be easily changed. Simply put, it is hard to change the size of the grid cell, known as re-sampling, without introducing errors of some sort. With that said it is easier to make the cell size larger without too much loss of data integrity, but making the cell size smaller can never increase the data resolution. We cannot add more information between the original grid cells without interpolation. Interpolation of rasters in GIS is powerful, but must be treated with caution as we are basically asking the computer to create data where there is none based on some algorithm that a human created.

a. Key terms: scale, resolution, interpolation resampling
b. When curating raster data pay careful attention to any re-sampling or interpolation that the researcher may have done.

21. Raster Data Model: File Formats (1 minute)

Just some common file formats for rasters. Again, we will revisit these in the common data types module.

a. Much like the vector data geodatabase, sometimes raster formats are actually folders with a collection of data (ArcInfo GRID)
b. Emphasize that for all GIS data, there may be more than one file present with the same filename but different extensions. The entire collection of files must be

present to open the dataset.

22. GIS Projects (1 ½ minutes)

A Geographic Information System is not one file; a simple GIS is a project file that links the supporting data and other files together. Hence the project file does not contain the actual data. Instead it stores the location of the data on the file system and any descriptive or visual information needed to display the data layers on a cohesive map.

a. Common file formats for GIS project files are:
   i. .aprx - ESRI ArcGIS Pro
   ii. .mxd - ArcMap Desktop (legacy)
   iii. .qgs .qgz - QGIS
b. Often a GIS will be stored in a project folder that has
   i. The project file as described above
   ii. All supporting data files or databases needed
   iii. Any descriptive information about the project
c. Good data management is a key; where to put your project file, where to put your data files (separate or together with project), where to put the documentation.
d. Often when curating GIS data, the project IS NOT included.
e. This file system architecture is similar for all GIS assemblages

23. Map Projections and Coordinate Reference Systems (1 minute)

Map projections, or how we translate locations on the surface of the earth to a planar x/y coordinate system are an integral part of GIS data visualization and analysis. While most GIS software packages have good tools for managing the transformations between coordinate systems, a high level understanding of these transformations is a key to creating and sharing reusable GIS data packages.

a. The peeling of the orange is a common metaphor for the problem of flattening the Earth's surface. To flatten the orange peel, you must tear and distort the round surface (you can bring oranges and make this a fun break …).
b. There are way too many projections and coordinate reference systems to understand them all … we only need to focus on a few in North America.
c. The same is true for any other place of study, although projections and CRS may change as the area of interest changes (i.e. South America vs. North America).

24. Coordinate Reference Systems (CRS) Basics (2 minutes)

How to flatten the Earth and put it on paper… There are generally two ways to do this. Both approaches rely on the *datum*, or assumed center of the earth, as the center point of the coordinate reference system. All CRS must have a datum. First, a *geographic coordinate system* simply uses the spherical coordinates of latitude and longitude as the planar x and y coordinates. This is known as Platte Carre plotting. The second method uses a *projected coordinate system* that involves a mathematical transformation of the latitude and longitude into specific systems of x and y planar coordinates.

a. Key terms to remember: datum, geographic coordinate reference system, projected coordinate reference system
b. All projections distort either shape, area, distance, or direction. There are many different map projections that minimize these distortions for specific areas on the Earth's surface.

c. Note that to make accurate measurements in linear or areal units, we must first transform latitude and longitude into x and y coordinates with a distance unit (feet, meters, and so on).

25. Datum (1 minute)

We do not know where the center of the earth is … it depends on how the shape of the earth is measured: gravitational fields, distance, angular units. For GIS, we choose a center and work from there. Depending on the center we choose there will be differential measurement/location errors across the surface of the earth because the shape of the earth cannot be defined mathematically, only approximated.

a. Geometric geodesy is the study of the shape of the Earth's surface.
b. Aside: it is a misnomer that we use Datum (Latin for "that which is given") for the center of the Earth. A better name would be that which is chosen. This can be extended to data as well, data is never given to us (ok, there are some exceptions) … we design methods to gather data …

26. Geographic Coordinate Systems (2 ½ minutes)

Geographic Coordinate Reference Systems are what most people think of when they consider location on the Earth's surface - latitude and longitude. Less well known is that there is more than one coordinate reference system for latitude and longitude. The variations depend on which datum is used which in turn is used to name the CRS. The two most common geographic CRS are the WGS 84 (World Geodetic System 84 - good for most global data) and the NAD 83 (North American Datum - good for data in North America).

a. EPSG - European Petroleum Survey Group
b. EPSG have created a set of codes for most common projections used globally, they have become a standard way to identify projections. Note that there are other identifiers for projects as well, but almost everyone uses the EPSG standard now.
c. NAD 83 is actually based on the tectonic plate upon which North America sits and the benchmarks in the ground actually move … this is a problem as it introduces error that changes across time as the tectonic plate moves.
d. The National Geodetic Survey is working to introduce a new set of datums/frameworks sometime in 2024 or 2025 that are based on satellite observations and not benchmarks on the ground - it will be good to pay attention to these changes.
e. Errors between WGS84 and NAD83 vary from 2.2 to about 4.5 feet across the USA: 2.2 SE Florida, 4.5 NW Washington state (see map).

27. Projected Coordinate Systems (2 minutes)

Most serious geographic research uses projected coordinate systems which allow for accurate measurement of distance and area across the study site. These project CRS are specific to the area of interest. Again there are many projected CRS that have specific purposes for certain scales of mapping and certain locations on the globe. Instead of two common coordinate systems as is the case for geographic CRS, there are two common families for projected coordinate reference systems. The State Plane System (SP) for the United States and the Universal Transverse Mercator (UTM) for

global data.

a. Often researchers who are not GIS experts have limited understanding of projected coordinate systems and often simply work in EPSG:3857 (the CRS for web mapping systems). This is likely because much GIS data on the internet uses this coordinate reference system.
b. A good question to ask a researcher: How did you choose the coordinate reference system for your use case?
c. Often governments will use specific datum and projection combinations for government data collections. Miami-Dade County in Florida uses the NAD 83 datum with the State Plane Florida East Projection. Another example, Peru uses the PSAD 56 (1956 Provisional South America Datum) datum and the UTM Zone 11 South projection.

28. Universal Transverse Mercator (UTM) (1 minute)

The Universal Transverse Mercator system of projections is the most common set of projections used world-wide. Each grid zone is 6 degrees wide and is split into a north and south zone at the equator. Often the UTM set of projections is paired with the WGS 84 datum (but not always). For example Florida is WGS 84 UTM Zone 17 North.

a. Fun exercise - find your UTM zone ….
b. Note: yes, Mercator as in the inventor of the Mercator projection in the 16th century. It was the first global projection to preserve direction and hence facilitated navigation across large expanses of water. We still use the math from this 16th century innovation today …

29. State Plane Coordinate System (SPCS) (1 ½ minutes)

In the USA the state plane set of coordinate reference systems is standard in much work that involves surveying or measuring. All government agencies use this system, especially local government offices such as municipalities and counties. For most work this family of projections is based on the NAD 83 datum.

a. Often if a researcher is working at a local scale and is not using the state plane standard, a good question to ask is why did you choose not to use this standard.
b. Note, sometime in 2024 or 2025 the National Geodesy Center may be changing all of this with a new set of satellite based reference systems. It is likely that the State Place family of projections will remain, just with something other than the NAD 83 datum (see note for Geographic Coordinate Systems slide).

30. Review and Skills Checking: (2 minutes)

This module has six objectives. At this point, you should be able to:

1. Describe a geographic information system (GIS)
2. Identify common tool sets for working with GIS data
3. Describe characteristics of vector GIS data
4. Describe characteristics of raster GIS data
5. Understand the three principal components of a Coordinate Reference System
6. Conceptualize structure of a GIS project

Next, we will now have an Activity, and a Quiz.

Skills Check:

31. Activities (10 min)

- Using the data in Dataset_No_ReadMe (this is a subfolder of the 1-4_Excersise_Dataset,folder) create a GIS project with one raster and one vector layer
  - Review of Layers and Browser panes
  - Look at Information Panel: find projection, pixel dimension, other metadata
  - Locate export layer panel: explore options and CRS selector

Knowledge Check

31. Quiz 1

1. Describe, in your own words, what is a GIS?
   a. Answer: [Something along the lines of: A collection of tools, data, hardware, and people used to locate and describe features on the Earth's surface. Slide 7]
2. Describe, in your own words, what are the principal components that make up a GIS?
   a. Answer: [Something along the lines of: A tool set, physical and logical data models, and scripting or programming languages. Slide 10]
3. What are two common GIS software assemblages?
   a. Answer: [Should include at least 2 of these: QGIS, ESRI ArcGIS, Python, and/or R. Slides 11 to 15]
4. What are two common GIS data models?
   a. Answer: [Should be: Vector data model, Raster data model. Slides 17 to 21.]
5. Yes or no. Are data layers stored in the GIS project file?
   a. Answer: [Answer should be "No," plus some explanation of how a GIS project file works. Slide 22]

32. Quiz 2

6. Yes or no. All geospatial data must have a datum defined?
   a. Answer: [Something along the lines of "Yes," plus some explanation of a datum. Slides 24 and 25]
7. What are the names and EPSG codes for two common datums used in North America?
   a. Answer: [World Geodetic System 84 (WGS84) and North American Datum of 1983 (NAD83). Slides 25 and 26]
8. Yes or no. All geospatial data must have a coordinate reference system (CRS) defined?
   a. Answer: [Yes]
9. Name two common projected coordinate reference systems used in North America?
   a. Answer: [Universal Transverse Mercator (UTM) Slide 28, and State Plane Coordinate System (SPCS) Slide 29.]