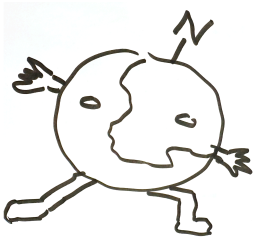
DCN Planetears
Geospatial Data Curation:
an introduction



Module 3: Common GIS Data Types Instructor Notes

[This handout accompanies a slidedeck with the file name “3_Common_GIS_Data_Types”]
2024-02-08 1000

Lesson Plan: Common GIS Data Types (approximately 120 minutes [2 hours]). This module has many short activities for skill building, thus the module length. Participants are expected to have a GIS tool open during this module, and opening and exploring GIS files.

Objectives: At the end of the module, learners will be able to:

1. Differentiate between raster data, vector data, and geodatabases
2. Identify three common file formats for vector data
3. Identify two common raster file formats
4. Identify four common tools to open GIS data
5. Gain an understand complex GIS file structures

Lecture Sections (70 minutes)
Speakers Notes for each slide begin on page 4 below

Activities (38 to 40 minutes)
There are activities of various length on the following slides:

- **16: Slide Title: Adding Vector Data in QGIS: Activity (3 minutes)**
- **19: Slide Title: Vector Activity (10 minutes)**
- **25: Slide Title: Adding Raster Data in QGIS: Activity (5 minutes)**
- **28: Slide Title: File Type Activity (5 minutes)**
- **39: Slide Title: Common GIS File Types Activity (15 minutes)**

Check Understanding (10 minutes)

1. **Quiz**

    Quiz Questions

    1. Describe, in your own words, the characteristics of raster data.

Include at least 1 difference between raster and vector data.
    a. Answer:
2. Describe, in your own words, the characteristics of vector data. Include at least 1 difference between vector and raster data (different from the one you used above).
    a. Answer:
3. Describe, in your own words, the characteristics of a GIS database.
    a. Answer:
4. Name 2 common raster GIS file types.
    a. Answer:
5. Name 3 common vector GIS file types.
    a. Answer:
6. Name 2 common GIS database file types.
    a. Answer:
7. Name 2 GIS Project file types.
    a. Answer:
8. Name 2 common GIS software tools
    a. Answer:
9. Name the primary thing to watch out for when transforming complex file formats like shape files.
    a. Answer:

Quiz Answer Key:

1. Describe, in your own words, the characteristics of raster data. Include at least 1 difference between raster and vector data.
    a. Possible Answer: Raster is a spatial model that defines space as an array of equally sized cells, like pixels in a digital photo. Cells contain attribute values and location coordinates. Groups of cells that have the value, such as color, represent the same type of geographic feature. For example, blue cells are all water features. One difference between raster data and vector data is that raster coordinates are contained in the ordering of the matrix or array, where vector data store the geo coordinates explicitly.
2. Describe, in your own words, the characteristics of vector data. Include at least 1 difference between vector and raster data (different from the one you used above).
    a. Possible Answer: Vector is a coordinate-based data model that represents geographic features as points, lines, and polygons. Each point is represented as a coordinate pair,

such as latitude and longitude. Lines and polygon features are represented with lists of vertices, or points. One difference between vector and raster is that vector graphics scale smoothly, because mathematical equations are used to store and relate data. In contrast, raster does not scale smoothly because it stores information in equally sized cells, like pixels in a digital picture.

3. Describe, in your own words, the characteristics of a GIS database.
    a. Possible Answer: GIS databases are typically associated with a particular software tool, to allow updates and queries. The GIS database includes data about the spatial locations and shapes of geographic features, such as points, lines, areas, pixels, and others. GIS databases can contain both vector and raster data.
4. Name 2 common raster GIS file types.
    a. Possible Answer: Any of: .GeoTIFF; .XML; .ASC; .IMG
5. Name 3 common vector GIS file types.
    a. Possible Answer: Any of: .SHP; .CSV; .GEOJSON; .GML; .KML; .GPX; .OSM
6. Name 2 common GIS database file types.
    a. Possible Answer: Any of: .GDB; .GPKG; .MDTILES
7. Name 2 GIS Project file types.
    a. Answer: Any of: .QGS; .QGZ; .APRX; .MXD; .MXT; .WOR
8. Name 2 common GIS software tools
    a. Possible Answer: Any of: QGIS; ArcGIS; Open Street Map; Text editors
9. Name the primary thing to watch out for when transforming complex file formats like shape files.
    a. Possible Answer: Besure to copy all files, because if you don’t the shape file or other database will not open or operate properly.

Slide Speaker Notes

1. Slide Title: DCN Planetears: Geospatial Data Curation: an introduction: Module: Common GIS Data Types (30 seconds)

Welcome to the next module in the “Geospatial Data Curation: an introduction” curriculum, Common GIS Data Types.
[Next slide]

2. Slide Title: Common GIS Data Types (30 seconds)

In this module we introduce the Common GIS Data Types you may encounter will curating GIS data. We will also offer some tools for and hints on viewing and reviewing GIS data files.

[Next slide]

3. Slide Title: Module Objective: Common GIS Data Types (1 minute)

This module has five objectives. At the end of the module, learners should be able to:

1. Differentiate between raster data, vector data, and geodatabases;
2. Identify one common file formats for vector data;
3. Identify one common raster file formats;
4. Gain an understanding of complex GIS file structures.

The module has Lecture, Activity, and Quiz components to help reinforce new information.

Let us start with a list of common GIS data types.

[Next slide]

4. Slide Title: Common GIS Data Structures (1 minute)

In this module, you will get a brief introduction to four types of GIS data file types. These are:

1. Vector;
2. Raster;
3. Databases; and,
4. GIS Projects.

For each data file type you will learn:

1. How to **Recognize** that file type by file extension;
2. Useful tools to **Open** files with those extensions; and,
3. How to **Assess** the files by asking a number of questions about the files.

Let us get started with Vector Data.
[Next slide]

5. Slide TItle: All Geospatial Data Types: Assess (2 minutes)
Assessing a GIS data file can be part of the “Check” and/or the “Understand” steps of the CURATED workflow, depending on your specific workflow. We will think of Assessing files as part of the “Understand” step, so you will notice that the “U” is underlined in the CURATED graphic in the title box.

We recommend that you ask the following four (4) questions for all Geospatial Data files which have been submitted to you:

1. Are the filenames descriptive enough to ascertain what they display?
    a. Best practices can include the following
        i. No spaces in filenames
        ii. Some descriptive information. For example: a date in the ISO format, a place name, and a subject keyword (20240207_NewYork_BirdSpecies.shp)
2. Do files open in GIS or other software properly?
    a. Does the data display in the correct location
    b. Are there errors in the attribute table headers
3. Are the relationships between the data layers clear?
4. Is the documentation robust and complete enough for re-use?

As we go through different data structures, we will list some specific things to focus on for each structure. So first up, \let’s look at vector data structures a little more.

6. Slide Title: Vector and Raster Data Structures (2 minutes)
Ok, time to review the two data structures (from the introduction): 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)

7. Slide Title: Vector Data Model: Points, 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).

8. Slide Title: Vector Data: Recognize 01 (1 minute)

Recognizing a GIS file type is part of the "Check" step of the CURATED workflow. So you will notice that the "C" is underlined in the CURATED graphic in the title box.

You will recall that vector data represents geographic features as points, lines, and polygons. Point features are represented as a coordinate pair, while line and polygon features are represented as an ordered list of vertices. Attributes are associated with each vector feature. (This definition is edited from the ESRI GIS Dictionary at https://support.esri.com/en-us/gis-dictionary/vector)

In this image we are looking at an example attribute table for point data. Where each point (represented by the symbols) has a unique ID, is given a name, a type, and general location information. We are not seeing the associated coordinate information for the geometries … that is usually not shown as a part of the table (CSV is the exception).
[Next slide]

9. Slide Title: Vector Data: Recognize 02 (1 minute)
When working with vector data there is generally only one geometry type (point, line, or polygon) per data file or feature class. Geometries don't mix in a dataset, but they can be layered in a project. Basically what we are saying is that there are individual layers

for points, lines, and polygons that we stack on top of each other to build our representation of the physical world.

a. OpenStreetMap is an exception to this rule where all geometries are in one dataset or data file.

10: Slide Title: Vector Data Common File Types (2 minutes)
These are some common file types that you will see for vector data structures.

a. Note these are vector only formats (except geodatabase). Later we will se formats that accommodate both vector and raster
b. See the cheat sheet as a handy reference
c. The geodatabase can also hold raster data
d. Leave time for questions

[Next slide]

Table description:
On the slide is a table of Commons Vector File Types. The table has 1 header row; 5 data rows; and 6 columns. The header row contains the following column variables: Extension; Description; QGIS; ArcGIS Pro; R, Python; and, Text Editor.
Row 1 contains the following values: .shp .dbf .shx .sbx .sbn .prj .xml .cpg .qmd; shapefile; check; check; check; null.
Row 2 contains the following values: .json .geojson; geojson; check; check; check; check.
Row 3 contains the following values: .kml .kmz; keyhole markup (google); check; check; check; check (kml).
Row 4 contains the following values: .csv; comma separated values; check; check; check; check.
Row 5 contains the following values: .gdb; Geodatabase (vector); check; check; check; null.

11: Slide Title: Tabular Data: Recognize (2 minutes)
One of the most common formats for distributing geospatial information is just a spreadsheet (.CSV) . If there are columns that contain geospatial information, it can be used to display the data within GIS software. Columns with latitude and longitude are the easiest to display and we will look at how to do that in just a moment.

Knowledge check: what structure is this? Vector or raster?

There are a few other ways to represent location in a spreadsheet. They require some additional work before they can be used by a GIS software. Addresses can be compared against a database of locations within the software using a process called “geocoding.” Also if you have a spreadsheet that doesn’t have spatial information but has a column that can be matched with a dataset that does, they can be joined together.

In the interest of time, we are not going to get into geocoding or joining datasets today, but just wanted to mention that it is a possibility.

What we are going to look at, is how to open spreadsheets that have columns for latitude and longitude.

[Next slide]

12: Slide Title: Tabular Data: Open (5 minutes)
Let's stop for a moment and look at a CSV file (table of values with lat/lon coordinates).

This slide provides steps for how open a spreadsheet with lat/long coordinates in QGIS:

1. Go to Layer → Add Layer → Add Delimited Text Layer
2. Select the AMWO_nests_P.csv file and choose which fields to use for the longitude values (X) and the latitude values (Y)
3. Check that the correct CRS is selected
    a. We are guessing WGS84, EPSG 4326, we will come back to problems with this later
4. Review the spreadsheet headers and data slice in the preview and then click "Add"
5. (If needed) Add a basemap by going to Web → QuickMapServices → and choosing an option (e.g. OSM Standard)
6. Open the attribute table and check column headers and values.
    a. You can also use the "information" tool and click on a point on the map..

NOTE: often we skip step 3!! This can lead to errors of over a meter across the continental United States and even greater errors in other parts of the world.

Next we will look at Assessing Tabular files.
[Next slide]

13: Slide Title: Tabular Data: Assess (2 minutes)
Assessing a GIS file can be part of the "Check" and/or the "Understand" steps of the CURATED workflow, depending on your specific workflow. We will think of Assessing files as part of the "Understand" step, so you will notice that the "U" is underlined in the CURATED graphic in the title box.

We recommend that you ask the following three (3) questions about the Tabular files which have been submitted to you:

1. Anything that you would do when curating non-geospatial tabular data is also appropriate here. So when you inspect the spreadsheet, does anything seem strange?
    a. Column headers
    b. Data types

    c. Truncated headers or truncated data

2. When opened in GIS software, does map location information match the textual description of the dataset? (e.g. Do points show up within the expected extent?) If the documentation says that the study took place in Mongolia, do the points display in Mongolia? Are some of the points showing up somewhere else? (probably errors or missing values in specific rows) Are all of the points showing up somewhere else? (probably an issue with projection)
    a. Sometimes lat/lon are reversed (researcher error)
    b. Sometimes it is not latitude and longitude, but instead projected x/y coordinates (what CRS?)
3. Is the projection / coordinate reference system documented somewhere? Since CSV or other spreadsheet file formats don't have a specific place to store extra spatial information within the file, it's important for that to be written down.
4. Other ideas from the students??

Next, we will look at Shapefiles.
[Next slide]

14: Slide Title: Vector Data: Shapefiles (1 minute)
One of the most common vector File formats out there is the Esri Shapefile. The "shapefile" is, in fact, a set of related files with the same filename and different extensions. A shapefile can have anywhere between three and nine files. The primary file extensions are .SHP, .DBF, .SHX. These three files hold the geometry and location of the data (.SHP), the attributes for each location (.DBF), and an index tying the two together (.SHX). These three files must be present in order for the file to be read correctly in GIS software. There are a number of other extensions, however, that may also be present, containing information about things like the projection (.PRJ), metadata (.XML), and encoding (.CPG).

Next, we talk about Opening Vector files.
[Next slide]

15: Slide Title: Vector Data: Open (5 minutes)
Opening a GIS file can be part of the "Check" and/or the "Understand" steps of the CURATED workflow, depending on your specific workflow. We will think of Opening files as part of the "Understand" step, so you will notice that the "U" is underlined in the CURATED graphic in the title box.

On this slide, there are directions for how to open a shapefile in QGIS:
1. Go to Layer → Add Layer → Add Vector
2. Navigate to the shapefile and select the .shp file. Click "Add."
3. Note, you may need to unzip the file first …

Next we will look at Assessing Vector files.
[Next slide]

16: Slide Title: Adding Vector Data in QGIS: Activity (3 minutes)
Time to add our first set of data. We are starting with vector data and will be using the file type Shapefile. These are the steps we are going to follow and I will demonstrate.

In the 0_Environement_Setup_Datasets folder there is another folder labeled 0_study_area_cover_type. In this folder are multiple file types all by the same name. Right now we are not going to go into all the types of files in this folder; for now just know that we want to use the SHP file type but these other files are important friends that tell the program how to read the SHP file.

Demonstrate adding the vector file 0_study_area_cover_type.shp to a blank project.

Time to add the first set of data. We are going to add vector data in the form of the commonly used Shapefile.

1. Click on the “Open Data Source Manager” icon
2. Select “Vector” from the side menu
3. Click on the three dots next to the Source box
4. Locate where the data layers are saved and select the SHP file extension
    a. Select the file Teaching_study_area_cover_type.shp
    b. In a Shapefile there are multiple file extensions that are necessary for the program to correctly display the data, more details about Shapefiles is given in the Module Common GIS Data Types.
5. Click the “Add” button

[Next slide]

17: Slide Title: Adding Vector Data in QGIS (Example Video) (2 minutes)
If you are comfortable demoing the steps from the previous slide you can skip this slide. Or if you are unable to demonstrate the steps you can play this slide as a demonstration.

Note: This video is not showing the example dataset listed at the start of this module. The purpose of this video is to demonstrate the basic steps necessary to perform the task.

This video illustrates the steps from the previous slide:
1. Click on the “Open Data Source Manager” icon;
2. Select “Vector” from the side menu;
3. Click on the three dots next to the Source box;
4.Locate where the data layers are saved and select the SHP file extension; and,
5. Click the “Add” button.
There is no audio track in the video. Instead the 5 instructions are overlaid on the video in text.

[Next slide]

18: Slide Title: Vector Data: Assess (2 minutes)
Assessing a GIS file can be part of the “Check” and/or the “Understand” steps of the CURATED workflow, depending on your specific workflow. We will think of Assessing files as part of the “Understand” step, so you will notice that the “U” is underlined in the CURATED graphic in the title box.

We recommend that you ask the following six (6) questions about the Vector files which have been submitted to you:
Questions to Ask about Vector Files:

1. (For shapefiles) Are the mandatory files present (e.g. shp, shx, dbf)
2. Do files open in GIS or other software properly?
3. When opened in GIS software, does map location information match the textual description of the dataset?
4. When you inspect the attribute table, does anything seem strange?
5. Other ideas from students??

This list of questions is borrowed from the DCN GeoJSON Data Curation Primer. Dixson, Nadia; Milliken, Genevieve; Mukunda, Keshav; Murray, Reina; Starry, Rachel. (2019). GeoJSON Data Curation Primer, available at http://hdl.handle.net/11299/210208

The answers to the questions above will impact the next steps of the CURATED process. Do you need to Request more information from the data creator or submitter? WIll you need to Augment the metadata, or is it complete enough? Which files will you need to Transform into preservation-friendly formats? Where would you Evaluate these files to fall on the FAIR spectrum?

[Next slide]

19: Slide Title: Vector Activity (10 minutes)
Download and unzip (if needed) the provided example GIS datasets.
Link to example dataset: 1-4_Excersise_Dataset

1. Look at the files in the subfolder labeled study_area_cover_type, what tool or tools can you use to open this data? – Answer ArcGIS and QGIS
2. Is there a specific file extension you would load into the tool to view the data? – Load the SHP file but need the SHX and DBF extensions at minimum

Alright, now let us turn towards Raster files.
[Next slide]

20: Slide Title: Raster Data: Recognize (1 minute)
Recognizing a GIS file type is part of the “Check” step of the CURATED workflow. So you will notice that the “C” is underlined in the CURATED graphic in the title box.

Another important and common way to represent geospatial information is with raster data. Raster data represents space as an array of equally sized cells arranged in rows and columns. Each cell contains an attribute value and location coordinates. Unlike a vector structure, which stores coordinates explicitly, raster coordinates are contained in the ordering of the matrix. (This definition is edited from the ESRI GIS Dictionary at https://support.esri.com/en-us/gis-dictionary/search?q=raster). To represent space with raster all you need is the starting location for the grid, the shape of the grid (the number of columns and size of each grid cell), and then a long string of numbers representing the values in the cells

It is an efficient way to store information, but also higher resolution raster data can have very large file sizes.

[Next slide]

21: Slide Title: 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).

[Next slide]

22: Slide Title: Raster Data: Common File Types (2 minutes)
These are some common file types that you will see for raster data structures.

a. Note these are raster only formats (except geodatabase).
b. Geodatabase can hold both raster and vector (but only raster opens in ArcGIS Pro)
c. See the cheat sheet as a handy reference
d. Leave time for questions

[Table description:
On the slide is a table of Commons Raster File Types. The table has 1 header row; 4 data rows; and 5 columns. The header row contains the following column variables: Extension; Description; QGIS; ArcGIS Pro; and, R, Python.
Row 1 contains the following values: .tiff .tif; geoTIFF; check; check; check.
Row 2 contains the following values: .img; ERDAS Image; check; check; check.
Row 3 contains the following values: .adf; ESRI Grid (legacy, in a folder)
; check; check; check.
Row 4 contains the following values: .gdb; Geodatabase (raster); null; check; null.]

23: Slide Title: Raster Data: Recognize (1 minute)
One of the most common raster file formats is GeoTIFF, with file extension .TIF or .TIFF. This format is similar to other varieties of .TIF images, but with the important addition that it has information about where the data is located in the world. Sometimes this geospatial information is embedded within the primary file - in the header of the TIFF. Sometimes the information is stored in a separate file known as a "world file" (.TFW) There are a few other extensions that may also be present, storing supplemental content such as a lower resolution version of the data for faster display (.OVR) and metadata (.XML).

Next we talk about Opening Raster files.
[Next slide]

24: Slide Title: Raster Data: Open (5 minutes)
Opening a GIS file can be part of the "Check" and/or the "Understand" steps of the CURATED workflow, depending on your specific workflow. We will think of Opening files as part of the "Understand" step, so you will notice that the "U" is underlined in the CURATED graphic in the title box.

This slide provides steps for how to open a GeoTIFF in QGIS:

1. Go to Layer → Add Layer → Add Raster. Select the .TIF or .TIFF. Click "Add."
2. Open the Properties. Look at the Information and Symbology tabs to see whether there are multiple bands or to edit how the raster displays

(If a raster has multiple bands, it means that there are multiple values associated with each grid cell, sometimes representing different parts of the electromagnetic spectrum) You can open GeoTIFFs in other non-geospatial image viewing software as well - there will just be no mechanism for assessing the presence or accuracy of the spatial location. Depending on the software, it may not display multiple bands if present.

Next we will look at Assessing Raster files.
[Next slide]

25: Slide Title: Adding Raster Data in QGIS: Activity (5 minutes)
Let's add in a different data type. We are going to add raster data, this type of data is read more as a grid of values. Raster data comes in many file types for this exercise we will be using a .TIFF file. More information about different raster data types is the module Common GIS Data Types.

Demonstrate adding the raster file 0_warblerProductivity.tif and 0_woodcockProductivity.tif to a blank project.

1. Click on the "Open Data Source Manager" icon
2. Select "Raster" type from the side menu
3. Click on the three dots next to the Source box
4. Locate where the data layers are saved and select the .tif file type
    a. Select the files *warblerProductivity.tif* and *woodcockProductivity.tif*
5. Click the "Add" button

[Next slide]

26: Slide Title: Adding Raster Data in QGIS: Example Video (3 minutes)
Slide Title: Adding Raster Data in QGIS

If you are comfortable demoing the steps from the previous slide you can skip this slide. Or if you are unable to demonstrate the steps you can play this slide as a demonstration.

Note: This video is not showing the example dataset listed at the start of this module. The purpose of this video is to demonstrate the basic steps necessary to perform the task.

This video illustrates the steps from the previous slide:
1. Click on the "Open Data Source Manager" icon;
2. Select "Raster" from the side menu;
3. Click on the three dots next to the Source box;
4.Locate where the data layers are saved and select the SHP file extension; and,
5. Click the "Add" button.
There is no audio track in the video. Instead the 5 instructions are overlaid on the video in text.

[Next slide]

27: Slide Title: Raster Data: Assess (2 minutes)
Assessing a GIS file can be part of the "Check" and/or the "Understand" steps of the CURATED workflow, depending on your specific workflow. We will think of Assessing files as part of the "Understand" step, so you will notice that the "U" is underlined in the CURATED graphic in the title box.

We recommend that you ask the following three (3) questions about the Raster files which have been submitted to you:

1. Are measurement units, temporal information, and spatial information recorded?
2. Does data contain multiple layers or bands?
3. Are the files georeferenced? (i.e. does the image “snap” to the right spot on the map?)
4. Ideas from students??

This list of questions is borrowed from the DCN GeoTiff Data Curation Primer. Kearney, Courtney; Ruhs, Nick; Sedlins, Mara; Tien, Tracy; Trelogan, Jessica; and Watts, John. (2020). GeoTiff Data Curation Primer. Available at: http://hdl.handle.net/11299/216574

The answers to the questions above will impact the next steps of the CURATED process. Do you need to Request more information from the data creator or submitter? WIll you need to Augment the metadata, or is it complete enough? Which files will you need to Transform into preservation-friendly formats? Where would you Evaluate these files to fall on the FAIR spectrum?

[Next slide]

28: Slide Title: File Type Activity (5 minutes)
Download and unzip (if needed) the provided example GIS datasets.
“Dataset_No_ReadMe.zip” can be found in the folder: 1-4 Exercise_Dataset

1. Identify which files are a vector type and which are a raster type.

Next let us turn to Geo Databases.
[Next slide]
Still in the file folder, identify the file types you find there by Type (Vector, Raster, etc.) and File Extension.

29: Slide Title: Database: Recognize (1 minute)
Recognizing a GIS file type is part of the “Check” step of the CURATED workflow. So you will notice that the “C” is underlined in the CURATED graphic in the title box.

A GIS database contains data about the spatial locations and shapes of geographic features recorded as points, lines, areas, pixels, grid cells, or TINs, as well as their attributes. It can contain multiple tabular, vector, and raster data layers.” (This definition is edited from the ESRI GIS Dictionary at: https://support.esri.com/en-us/gis-dictionary/search?q=database)

Common GIS Database File Types include:

1. ESRI File Geodatabase .GDB (The default file format in Esri ArcGIS products)
2. OGC GeoPackage .GPKG (The default file format in QGIS)

Next we talk about opening GIS Database files.
[Next slide]

30: Slide Title: Databases: Open in QGIS (2 minutes)
Opening a GIS file can be part of the “Check” and/or the “Understand” steps of the CURATED workflow, depending on your specific workflow. We will think of Opening files as part of the “Understand” step, so you will notice that the “U” is underlined in the CURATED graphic in the title box.

We will look at how to open an Esri File Geodatabase (They are one of the most common kinds of geodatabase and somewhat complex to work with in QGIS). First it's important to note that if you look inside a folder with a .GDB extension, it is hard to make sense of the contents. There can be dozens or even hundreds of files, all with software-generated and non-descript file-names. These files are not particularly comprehensible to human eyes, but can be interpreted when opened through a GIS software. When opening the example file geodatabase in QGIS, there are three data layers represented by that large number of files. Note: If there is raster data stored in a File Geodatabase it can only be accessed through ArcGIS.

Next we will look at Opening Databases in ArcGIS..
[Next slide]

31: Slide Title: Databases: Open in ArcGIS (2 minutes)
Opening a GIS file can be part of the “Check” and/or the “Understand” steps of the CURATED workflow, depending on your specific workflow. We will think of Opening files as part of the “Understand” step, so you will notice that the “U” is underlined in the CURATED graphic in the title box.

We will look at how to open an Esri File Geodatabase (They are one of the most common kinds of geodatabase and somewhat complex to work with in QGIS). First it's important to note that if you look inside a folder with a .GDB extension, it is hard to make sense of the contents. There can be dozens or even hundreds of files, all with software-generated and non-descript file-names. These files are not particularly comprehensible to human eyes, but can be interpreted when opened through a GIS software. When opening the example file geodatabase in ArcGIS Pro we can see the actual data layers in human readable form.

Next we will look at Assessing Databases.
[Next slide]

32: Slide Title: Databases: Assess (2 minutes)
Assessing a GIS file can be part of the “Check” and/or the “Understand” steps of the CURATED workflow, depending on your specific workflow. We will think of Assessing files as part of the “Understand” step, so you will notice that the “U” is underlined in the CURATED graphic in the title box.

We recommend that you ask the following three (3) questions about the Database files which have been submitted to you:

1. Are there any files stored in the geodatabase that are not in the typical geodatabase format? (i.e. do not begin with the letter a followed by a series of numbers and letters)
2. How many files are contained within the geodatabase? Are any of them repetitive?
3. Is there enough data provided to understand how the data files within the geodatabase were created and what they are intended to display?
4. Ideas from students?

This list of questions is borrowed from the DCN GeoDatabase (.gdb) Data Curation Primer. Created by: Battista, Andrew; Brittnacher, Tom; Garrett, Zenobie; Moore, Jennifer; Pirmann, Carrie (Data Curation Network, 2019). Available at: https://hdl.handle.net/11299/202823

The answers to the questions above will impact the next steps of the CURATED process. Do you need to Request more information from the data creator or submitter? WIll you need to Augment the metadata, or is it complete enough? Which files will you need to Transform into preservation-friendly formats? Where would you Evaluate these files to fall on the FAIR spectrum?

Next we will look at GIS Project files.
[Next slide]

33: Slide Title: GIS Projects: Recognize (1 minute)
Recognizing a GIS file type is part of the “Check” step of the CURATED workflow. So you will notice that the “C” is underlined in the CURATED graphic in the title box.

GIS project files are used in GIS applications. Generally, they all hierarchically store

layers and then display them in a layout. They retain symbology, queries, labeling, and other properties for building maps. They may also contain tools, models, or workflows and non-geospatial files. (This definition is edited from the GIS Geography, and their page, “The Ultimate List of GIS Formats and Geospatial File Extensions” section on “GIS Software Project File Formats,” available at https://gisgeography.com/gis-formats/)

Common GIS Project File Formats include:

1. QGIS Project Files: .QGS (2.X), .QGZ (3.X)
2. ArcGIS Pro Project Files: .APRX; .PPKX
3. Legacy ArcGIS Map: .MXD

Next we talk about opening GIS project files.
[Next slide]

34: Slide Title: Reviewing Common GIS Data Types: Visual Aid (2 minutes)
The image is a visual review of what we have talked about so far. On the left we have icons for Tabular data, Vector points, and a Raster grid.
In the middle we see that the GIS data files can be supplemented by textual documents, maps, tools and scripts. All of these can be stored on a database.
On the right side of the image we see an icon for a Project, which looks like a map in a bag, reminding us that all of these interconnected GIS files should be packaged up together to make them useful.

This diagram can help us understand how the project file and the data layers are related. Important to note is that the project does not actually include the data, but instead is linked to the data. When a researcher copies and moves a GIS project from their work machine to another machine sometimes these links can be broken.

[Next slide]

35: Slide Title: GIS Projects: Open (1 minute)
Opening a GIS file can be part of the “Check” and/or the “Understand” steps of the CURATED workflow, depending on your specific workflow. We will think of Opening files as part of the “Understand” step, so you will notice that the “U” is underlined in the CURATED graphic in the title box.

Project files are software and version specific. They are not very interoperable or durable over the long-term. They can be a good way to see how a researcher has organized their work and visualized the data. From a curation standpoint however, it's important to consider whether the data and research could be understood without the project file as future researchers may not have the right software to access it.

Next we will look at Assessing GIS Project files.
[Next slide]

36: Slide Title: GIS Projects: Assess (2 minutes)
Assessing a GIS file can be part of the “Check” and/or the “Understand” steps of the CURATED workflow, depending on your specific workflow. We will think of Assessing files as part of the “Understand” step, so you will notice that the “U” is underlined in the CURATED graphic in the title box.

Also, it may be worth asking the researcher why the project file is included? Often they are not part of a data package.

We recommend that you ask the following five(5) questions about GIS Project files which have been submitted to you:

1. Does the project open correctly (are the links relative)?
2. Are there any files stored in the project folder that are not in typical GIS formats?
3. How many data layers are contained within the project? Are any of them repetitive?
4. Is there enough documentation provided to understand how the data files within the project were created and what they are intended to display?
5. Could you make sense of the data (ie. how the files are related) without the project file?

Next, we will review what we have learned about GIS data structures and then discuss how to approach unfamiliar file formats.
[Next slide]

37: Slide Title: How to Approach Unfamiliar GIS File Formats (3 minutes)
Today we have tried to cover some of the formats that you will most likely see. But since spatial analysis is relevant to so many disciplines, there are many different software that can create geospatial data. The speed at which technology has changed and continues to change also contributes to the number of GIS file formats out there. There are lots of legacy file formats and constant development of new ways to store information. As a result, there are hundreds of file formats that can contain geospatial data and it's not feasible for one person to have encountered them all.

If you don’t recognize the file format that you are curating:

- It’s okay! This is a common experience for GIS data curators.
- Look it up in the Library of Congress digital formats list (or just on Google). Take a moment to explore this page with the class.

- You also might attempt to open the file with QGIS. QGIS is kind of magic when it comes to open random GIS file formats. If you are able to get something open, look in Properties to see if there is additional metadata or context.
- If that doesn't work (or if you are pressed for time and don't feel like sleuthing) you can always ask the researcher about what software was used to create and view the files.

[For an example to practice with see: ALOS-2 SAR Trinidad]

[Next slide]

38: Slide Title: Common GIS File Types Activity (15 Minutes)
This activity assumes that you have already downloaded and have access to either or both major GIS tools, QGIS or ArcGIS.
Further, this activity assumes that you have some access to other software tools, such as an image or photo tool, a spreadsheet tool, and a basic text editor.

As you work through the steps, you will want to record what you see in a notebook or in a text file.
Activity:

1. Download and unzip (if needed) the provided example GIS datasets with an appropriate tool.
2. Based on the file extensions you find there, identify which tool you will use to open the dataset.
    a. Answer:
3. Still in the file folder, identify the file types you find there by Type (Vector, Raster, etc.) and File Extension.
    a. Answers:
4. Now open "Working_01" with the tool you identified above and make yourself familiar with how the data displays and the menu layout of the tool.
5. Now make a copy of your "Original" dataset and label it "Working_02".
6. Open Working_02 with software tools that are NOT indicated for that data type.
    a. For example, many Vector and Raster files can be opened with image processing tools, such as PhotoShop, Paint, etc.
        i. Try this, and note the differences in how the data displays differently and/or similarly to the appropriate GIS tool.
    b. Now try various text editors, spreadsheet tools, etc.

i. Note how each tool impacts the data display.
ii. This could have an impact on how you transform and preserve the data!

c. Record your observations:

Once you have completed this activity, you may go on to the quiz on the next 2 slides.
[Next slide]

39: Slide Title: Common GIS File Quiz: 01 (10 minutes)

Now that you have had a chance to learn about various GIS file types, and have had experience opening them up, let us test what you have learned.

There are eight (8) questions to this quiz, spread over 2 slides, to make them easier to read. Record your answers to each of the questions in a text file or a notebook. Do your best to answer first from your memory, before going back to review the material on the slides.

Quiz Questions 1 to 5 (of 8)

1. Describe, in your own words, the characteristics of raster data. Include at least 1 difference between raster and vector data.
   a. Answer:
2. Describe, in your own words, the characteristics of vector data. Include at least 1 difference between vector and raster data (different from the one you used above).
   a. Answer:
3. Describe, in your own words, the characteristics of a GIS database.
   a. Answer:
4. Name a common raster GIS file type.
   a. Answer:
5. Name a common vector GIS file type.
   a. Answer:

Now go on to the next slide to answer the rest of the questions.
[Next slide]

40: Slide Title: Common GIS File Quiz: 02

Welcome to the second part of the quiz.

Quiz Questions 6 to 8 (of 8)

6. Name 2 common GIS database file types.

    a. Answer:
7. Name 2 GIS Project file types.
    a. Answer:
8. Name 2 common GIS software tools
    a. Answer:

Once you have finished the quiz, go on the review slide.
[Next slide]

41: Slide Title: Common GIS Data Types Review (3 minutes)
Today we have learned about four common types of GIS data, their file extensions, and practices for handling them during the CURATED workflow. These file types are:

1. Vector: A coordinate-based data model. Scales well.
    a. .SHP; .GEOJSON; GML; .GPX; .OSM; .CSV
2. Raster: A spatial data model that defines space as an array of equally sized cells arranged in rows and columns.
    a. GeoTIFF: .XML; .ASC; .IMG
3. Databases: Includes data about the spatial locations and shapes of geographic features.
    a. .GDB; .GPKG; .MBTILES
4. GIS Projects: All hierarchically store layers and then display them in a layout.
    a. .QGS; .QGZ; .APRX; .MXD; .MXT; .WOR

We also learned about some tools we can use, including:

- QGIS
- ArcGIS
- Open Street Map
- Text editors
- Many Other tools

[End of Module: Time for questions]