DCN Planetears

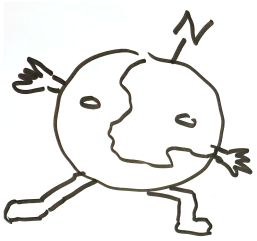
Geospatial Data Curation:

an introduction

Module 4: GIS Metadata Instructor Notes

[This handout accompanies a slidedeck with the file name 4_GIS_Metadata"]

2024-02-16 1000



Lesson Plan: GIS Metadata

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

1. Recognize essential Geospatial metadata elements
2. Be able to evaluate geospatial metadata and documentation
3. Prioritize reproducibility and usability over standards
4. Locate where metadata is recorded in either QGIS or ArcGIS
5. Name two GIS metadata standards

Slide notes: see slides

Activities:

- Locate metadata in GIS files [Slide13 speakers notes]
- Create request for missing metadata from example file [Slide 13 speakers notes]
- Create README.txt for a sample dataset [Slide 25]
- Evaluate the README against the FAIR principles [Slide 25]

Check Understanding

Notes for Slides (approximately 40 to 55 minutes total, depending on how much time is devoted to the final exercise)

1. **Module: GIS Metadata (30 seconds)**
    a. Welcome to the next module in the "Geospatial Data Curation: an introduction" curriculum, GIS Metadata.
2. **Module Objectives: GIS Metadata (1 minute)**

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

1. Locate where metadata is recorded in either QGIS or ArcGIS;

2. Recognize essential Geospatial metadata elements;
3. Be able to evaluate geospatial metadata and documentation;
4. Name two GIS metadata standards; and,
5. Prioritize reproducibility and usability over standards.

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

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

3. GIS Metadata (30 seconds)

To meet the module objectives we will follow this outline in this module

1. Warnings
2. Locating GIS Metadata in Software Tools
3. Recognizing GIS Metadata Elements
4. Evaluating GIS Metadata Documentation
5. Overview of GIS Metadata Standards
6. Reproducibility, Usability, and Standards

[Next slide]

4. Warning: GIS Metadata is … Complicated (2 minutes)

We want to point out a number of factors that make GIS metadata complicated.

1. There are several complex standards to choose from (e.g ISO 19115, ISO 19139, FGDC). It takes substantial effort and time to create this style of metadata. The standards are xml-based and not human-readable. Optimally you shouldn't be looking at the metadata outside of a software program - it's not designed for that. Literally hundreds of fields, usually the only people who create this metadata are paid government employees following strict protocols. NOTE: in research this can be very valuable to re-use, and less valuable to create …. **Check** to see if any exists from pre-existing sources (i.e. the dataset is derived from a government resource)
2. For many file formats (especially databases and projects) the metadata is embedded within the file and can't be reviewed without access to specialized software.
3. The GIS tools themselves don't necessarily support the creation of metadata well.
4. There is little interoperability between geospatial software programs for metadata, so no guarantees that the metadata will be visible to future users of

the data. This is especially true if they are using different software programs than were used to create the original files.

With these warnings in mind, let us learn how to locate metadata in our GIS software tools.

5. **GIS Metadata Standards (30 seconds)**

A little humor, this is certainly the case for GIS metadata (and still continues) Image credit: XKCD https://xkcd.com/927.

6. **A Brief History of GIS Metadata Standards (2 minutes)**

A brief history of GIS Metadata.

a. The M variant is a local state standard for Minnesota (they were not alone in creating a local standard).
b. Currently US gov recommend the ISO series, but there is a good likelihood that this recommendation will change
c. Watch for DCAT and geoDCAT - data.gov is already recommending dcat v2.
d. NOTE The US Government's DCAT-US metadata schema is very specifically for describing datasets within US government data inventories, so that data.gov can ingest them. The DCAT-US version 3 will be final soon. It includes fields very specific to US Federal procurement, etc. It is not meant to replace any other metadata standard. It is only for high-level description.

7. **Warning: GIS Metadata is … Simple (1 minute)**

Use what you know, for example elements that are common across all data formats,

and focus on a few key metadata elements that make GIS metadata more reusable.

Especially citation specific metadata such as author, ORCID, keywords, abstract, so on.

With these key elements in mind, let us learn how to locate metadata in our GIS

software tools. Worse case scenario we ask the researcher to provide these elements

a. Most of what we need is Dublin Core based (or perhaps DataCite)
b. Most repositories have the needed elements for geospatial datasets

8. **Metadata Exceptions for GIS Data (2 minutes)**

There are some important metadata elements that are GIS specific

a. coordinate reference system (datum, coordinate system, projection (and units))
b. process steps (provenance),
c. data types and structures (raster/vector, kinds of attributes),
d. recommended scale of use …
e. Leave some time for questions

With these key elements in mind, let us learn how to locate metadata in our GIS software tools. Worse case scenario we ask the researcher to provide these elements

9. Locating GIS Metadata in Software Tools (2 minutes)

Locating Metadata within or for 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 Locating Metadata as part of the “Check” step, so you will notice that the “C” is underlined in the CURATED graphic in the title box.

In this part of the module, we will talk about locating the metadata in GIS software tools, such as QGIS and ArcGIS. We will also talk about times of external or stand-alone metadata file types you might encounter or choose to create.

First we will discuss embedded metadata. We previously warned you that the interoperability and accessibility of embedded metadata isn’t great in GIS software. But with that caveat, where are some places that you might look for it? Let us first look at QGIS.

10. Demonstration: Finding Layer Metadata in QGIS (1 minute)

To find metadata for a layer in QGIS, right click on the layer and select “Properties.” In the “Information” section of the dialog that opens, you will find metadata about the layer. There will likely be some technical details such as coordinate reference system and number of features. But unless the file was created within QGIS, you may not find very much information here.

Apart from the notice the options for Identification Fields (Descriptive and Discovery), Extent (location - Discovery), Access (administrative), Fields (data dict and technical info), Contracts (Admin), History (process steps and provenance).

Check that everything makes sense: questions to ask ….

Note for Instructor: It is suggested that a live demo be done showing how to look at the information and source information for a raster and shapefile. Can use the data that is in the 1-4_Excersuse_Dataset folder

11. Demonstration: Finding Metadata in ArcGIS (1 minute)

To View Metadata in ArcGIS, open the Catalog Pane. Metadata can be chosen in the Details box on the right.

You will notice this example does not have any added.
Note for Instructor: If showing ArcGIS Pro it is suggested that a live demo be done showing how to look at the information and source information for a raster and shapefile. Can use the data that is in the 1-4_Excersuse_Dataset folder

12. Comparing Metadata Views in ArcGIS Pro and QGIS (1 minute)

One of the biggest problems with geospatial metadata is that there is very little interoperability between programs. Let's say a researcher has taken the time to fill in detailed metadata about a vector data layer in ArcGIS Pro. Here is what that metadata would look like in the program it was made (ie. ArcGIS Pro). Here is what that same layer looks like through QGIS. Nothing If possible, open files in the software that the researcher used to generate them -- but also know that it is pretty uncommon for folks to take the time to fill in standards-based metadata…

13. External Sidecar and Stand-Alone Metadata: Demonstration: .shp .xml (1 minute)

Another place to look … Metadata for many GIS formats is stored in a file with the extension ".xml." You can open the .xml in a text editor (or xml editor if you have one). It's not especially human-readable but you will be able to see if the researcher put a bunch of time into adding detail or if it only has minimal information.

a. Perhaps **check** to make sure no old metadata from a source layer …
b. Also, **Check** for pre-existing metadata in derived files, there may be legacy information that is either important to retain or no longer makes sense.
c. Show this example?

You can export metadata as a free-standing XML file… But at that point, if you intend for the metadata to not be read through software, you might as well make it in a human-readable format - like a spreadsheet or a readme. There are situations where someone might need to create standards metadata (because of the data portal that they are expected to share it through or as a condition of a grant that they are receiving) [also honor the work if someone has done it]. If that is not the case or there is no documentation yet - opt for human-readable.

14. Check and Understand (5 minutes)

Have the students open files and look for the metadata in the 1-4_Excercise_Dataset.

a. As they look through the metadata, have them pay attention to certain details, especially the coordinate reference system, the kinds of data, and the kinds of metadata.
b. This is a good moment to check through all layers and make some notes. If something does not make sense, put it on a list of things to request from the researcher

15. GIS Metadata is Hard (1 minute)

Going back to the conversation about standards … GIS metadata is perhaps one of the most difficult to pin down. Instead of trying to understand these complex standards, let's focus on some basic things that a researcher needs to provide about their geospatial data; key metadata elements you should **check** for or **request** for spatial data.

a. Our discussion focuses on a few things that are unique to spatial data (projection/scale of use).
b. And then also some things that are not completely unique but are issues that frequently come up.

16. Essential Documentation : All GIS Data (1 minute)

Drawing on the conversation about what a researcher needs to re-use geospatial data, these three elements are required.

1. Geographic Coordinate System (CRS - coordinate system and/or projection)
2. Point location, bounding coordinates, or gazetteer name
3. Lineage/process (derived/collected?)

They not only help discovery and the determination of fitness for purpose, but without the CRS the researcher who receives the data will not be able to open it correctly.
The next 3 slides go a little more in depth on each of these.
[Next slide]

17. Essential Documentation: Coordinate Reference System (CRS) (2 minutes)

This is a review from the Introduction to GIS module.

1. For all geographic data the CRS must include the datum and the geographic coordinate system.
2. For all projected data the projection must be named.
3. And the best practice is to include the EPSG code (knowledge check: does anyone remember what EPSG stands for?).

Notes:

a. The EPS code references either two (datum and geographic CRS) or three (additional projection) elements of the CRS
b. If it doesn't "snap to" the correct location when you try to open the file, the information is probably missing.
c. In some cases a researcher can provide a justification of why the specific CRS was chosen
d. In some cases a researcher (or data curator) can document what transformations were used if any

[Next slide]

18. Essential Documentation: Bounding Box or Gazetteer Name (2 minutes)

There are three general ways to describe location in metadata:

1. as a bounding box,
2. as a place name, and
3. as a lat/lon point.

These three approaches are shown here using the element names from the DataCite v4.0 schema.

1. geoLocationBox
    a. westBoundLongitude, eastBoundLongitude, northBoundLongitude, southBoundLongitude (WGS84 lat/lon decimal degrees)
2. geoLocationPlace
    a. A text description, better if referenced to a gazetteer
3. geoLocationPoint
    a. pointLongitude, pointLatitude (WGS84 lat/lon decimal degrees)

Note: This location description is becoming standard in most general purpose repositories (and can also be applied to textual documents and literature)
[Next slide]

19. Essential Documentation: Lineage and Provenance (2 minutes)

Often a researcher will get a dataset, modify the dataset for use in their project (change the coordinate reference system, crop to a specific area, simplify line work, add some label information, and so on). These changes should be documented (tool used, order of operations, justification - this may be in the methods section of the paper as well and can simply be repeated in the metadata for the dataset).

Note that this can be tedious work, but the more that is included, the more re-usable the data will be. Just imagine getting a data set where you do not know how it was created … would you use it in your upcoming publication?

Optional: Talk through examples on slide:
Data source citations: U.S. Census Bureau (2022). Estimated 5-year ACS Total Population at Census Tract Level for Miami-Dade County. The United States Census Bureau.
General description of processing steps: Loaded Florida TIGER line census tract files into QGIS, reprojected to NAD83 Florida state plane east (EPSG 2236), and clipped to Miami-Dade county.
Specific parameters, models, and code: Joined csv data for total population from 2022 5 year ACS tables using a concatenation of the state_fp and county_fp in the csv matched to the geoid in the census tract geometry layer. There were 707 of 709 records matched.

[Next Slide]

20. Essential GIS Metadata Vector and Raster (1 minute)

Essential GIS Metadata vector/raster (2 minute)

For Vector files, it is essential to have:

1. A Data dictionary, and,
2. Recommended scale (or scale of digitization)

For Raster files, it is essential to record the:

1. Pixel size
2. Band information
3. Look up tables (if categorized)

These different metadata elements reveal a little more nuance into the differences between Vector and Raster data. Often the listed descriptive elements will already be provided by the researcher in some format or another; sometimes even in the filenames or titles (pixel dimension for example). The key is to check for them, put each element in a good understandable location (thinking technical, admin, descriptive, and discovery metadata), and then request the parts that may be missing.

It is essential to know these keywords and seek them in the existing metadata in the layers or the metadata provided by the researcher.

We will next focus on each of these points a bit more.

[Next slide]

21. Essential Vector Documentation: Data Dictionary (1 minute)
Slide title: Essential Vector Documentation: Data Dictionary (2 minutes)

As with any type of data submission, it is critical for there to be documentation defining variables.
Some of the important common necessary definitions include:

- Names
- Description
- Units of measure
- Clarification of the difference between Null cells and cells with a “0” (zero) in them

Discussion: Are there other variables that could be added to this list?

Note: Something that comes up pretty frequently for geospatial data is you may get pushback about requesting further documentation for data layers or variables that were not created by the researcher. GIS projects often re-use and combine datasets from a number of different organizations and agencies (for example - demographic information published by the Census or infrastructure data maintained by a municipal government). The original data creators are not always thorough when writing their metadata. Follow-on or re-use researchers likely will not want to take the time to provide “missing” definitions not provided by the original data creators. Further, later researchers will not want to take the time to document “extra” variables that they did not use for their

project.
Possible recommendations: 1) They could link out to existing documentation for the data (Make sure that it is published in a stable place). 2) Remove “extra” fields if they cannot be defined.
[Next slide]

22. Essential Raster Documentation: Band and Pixel Information (1 minute)

Slide Title: Essential Raster Documentation: Band and Pixel Information

For later researchers to have confidence they are displaying Raster images correctly, it is vital that Band and Pixel metadata is recorded. Here are a couple of questions that you can ask the raster file metadata:

- How many bands are in the images and what do they represent?
    - 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.
        - For color images with 3 bands, values are stored for the intensity of red, green, and blue (or RGB Colors)
    - For example three bands: Band 4 , Band 5, and (Band5 - Band4)/(Band5 - Band4) from Landsat 8
- What is the pixel size in CRS units?
    - For example a pixel size 30 meters and a pixel size .3 meters provide vastly different image resolution!
- Need a review? See Module Common GIS Data Types

[Next Slide]

23. Evaluate GIS Metadata for FAIRness: Accessible (1.5 minute)

Slide Title: Evaluate GIS Metadata for FAIRness: Accessible

As you will recall, the “E” step of the CURATE(D) workflow is “Evaluate for FAIRness”
And you will remember that the “A” in the FAIR principles is that metadata and data should be “Accessible.” https://www.go-fair.org/fair-principles/
So when evaluating the GIS data and metadata you are curating for “FAIRness” and Accessibility, you can ask these questions:

Does the README make the data more Accessible?:

- Are the abstract and keywords appropriate for data?
- Is there a contact creator identified?
- Is there a bounding box or gazetteer name
- Are the license and use restrictions appropriate?

Is the data Accessible?

- When you open this data in a new GIS project is the correct coordinate reference system used?
- Is the data in an appropriate format?

If you cannot answer “Yes” to each question, is there more that you can do? Maybe, maybe not. Document the actions you took, where you feel the shortcomings are, and move on.

[Next slide]

24. Evaluate GIS Metadata for FAIRness: Reusable (1.5 minute)

Slide Title: Evaluate GIS Metadata for FAIRness: Reusable

The “R” in the FAIR principles is that metadata and data should be “Reusable.”
https://www.go-fair.org/fair-principles/
So when evaluating the GIS data and metadata you are curating for “FAIRness” and Reusability, you can ask these questions:

Does the README make the data more Reusable?

- Can this data be used by a colleague of yours without speaking to you?
- Are attribute values/units/meaning well explained?
- Are relationships between layers well explained?
    - Derivations/Interpolations
    - Key column and/or spatial joins
- If derived data is included, are the process and tools well documented (version, platform, proprietary)?

Is the Data Reusable?
When you open this data in a new GIS project is the correct coordinate reference system used?

If you cannot answer “Yes” to each question, is there more that you can do? Maybe, maybe not. Document the actions you took, where you feel the shortcomings are, and move on.

[Next slide]

25. Put it in Practice: Exercise (5 to 20 minutes)

Slide Title: Put it in Practice: Exercise

Look at at the files in in https://bit.ly/GIS_1-4_Dataset and

1. Find and evaluate the metadata for each layer
2. Make a list of items to either check with or request from the researcher
3. Prepare a README.txt that focuses on reproducibility and useability
4. Evaluate the README.txt against the FAIR principals