DCN Planetears

Geospatial Data Curation:
an introduction
Module: GIS Metadata

Module Objectives: GIS Metadata

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
5. Prioritize reproducibility and usability over standards

The module has lecture and an activity

GIS Metadata

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

CURATED

Warning: GIS Metadata is ... Complicated

1. There are several complex metadata standards (e.g ISO 19115, ISO 19139, FGDC).
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.

GIS Metadata Standards

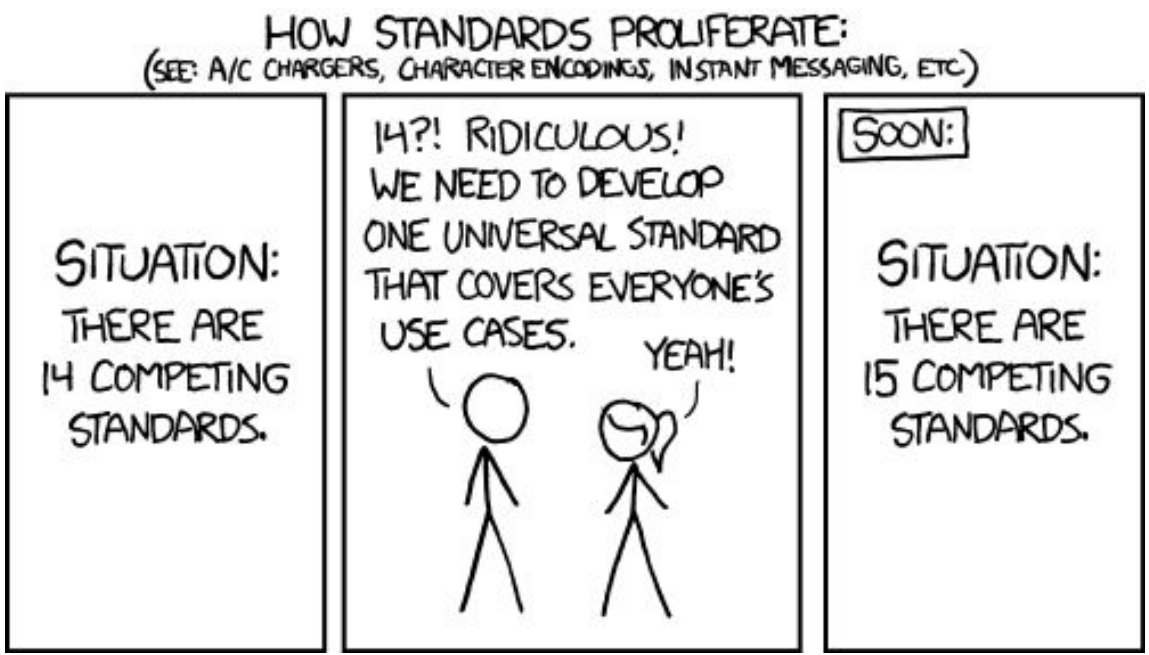


Image credit: XKCD https://xkcd.com/927.

GIS Metadata Standards

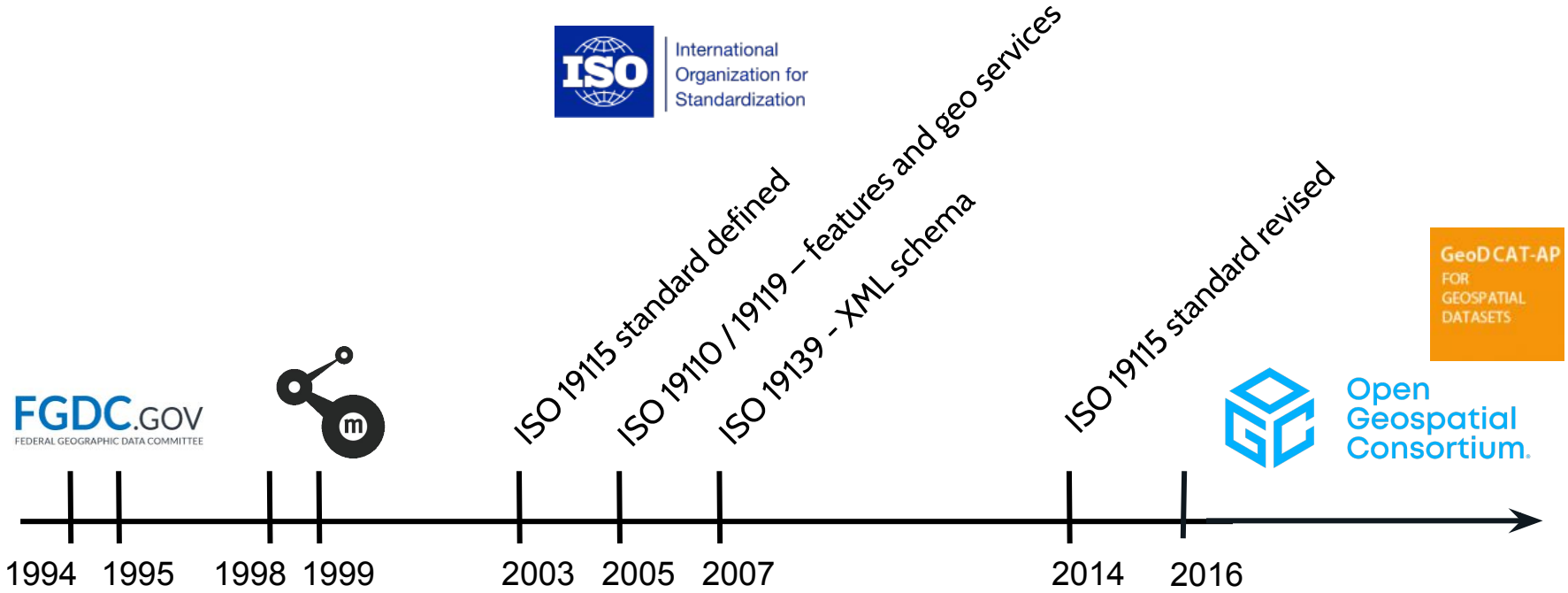


Image credit: Melinda Kernik

CURATED

Warning: GIS Metadata is ... Simple

1. **Technical** * : coordinate reference system, process steps, data types and structures
2. **Admin** : point of contact, license/copyright, last updated
3. **Descriptive** : data dictionary, abstract, purpose
4. **Discovery** : keywords, location, title, time period, recommended scale of use

* perhaps a little complicated

Metadata Exceptions for GIS Data

1. **Technical** * : **coordinate reference system**, **process steps**, **data types and structures**
2. **Admin** : point of contact, license/copyright, last updated
3. **Descriptive** : data dictionary, abstract, purpose
4. **Discovery** : keywords, location, title, time period, **recommended scale of use**

* perhaps a little complicated

Locating GIS Metadata in Software Tools

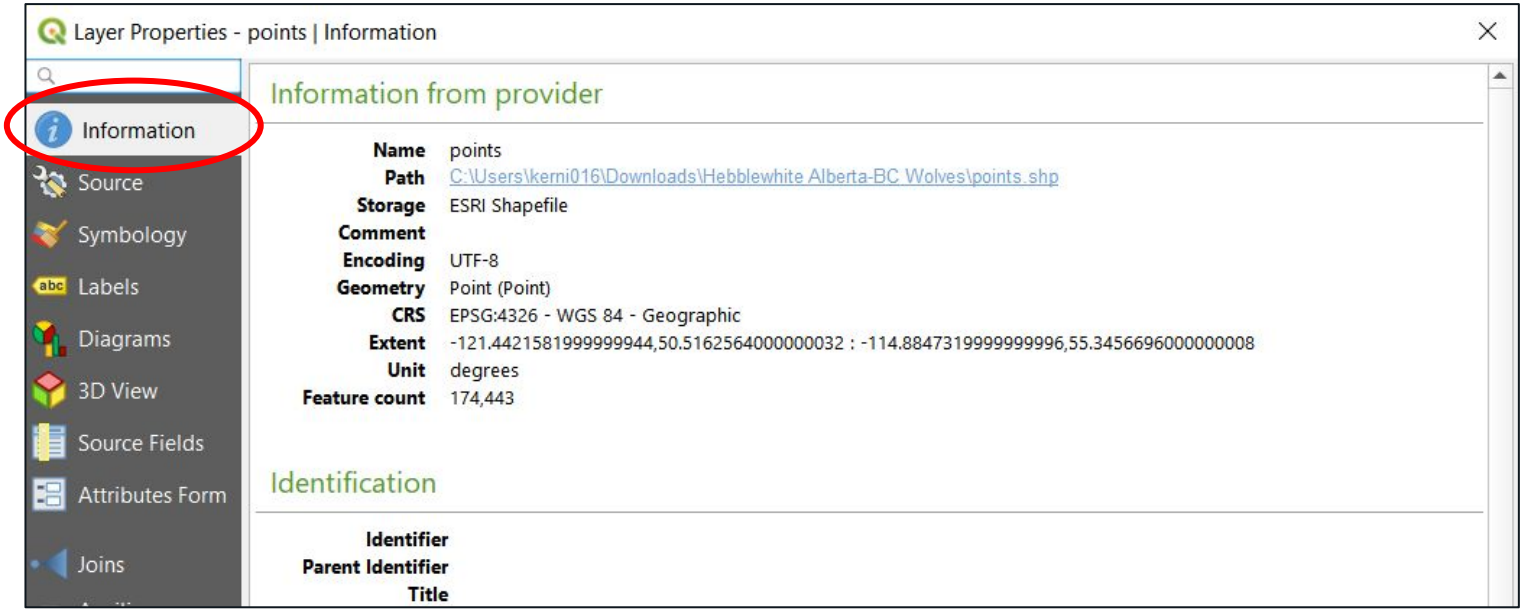
C U R A T E D

Possible (common) Metadata Locations:

1. Embedded QGIS Metadata
2. Embedded ArcGIS Metadata (not today)

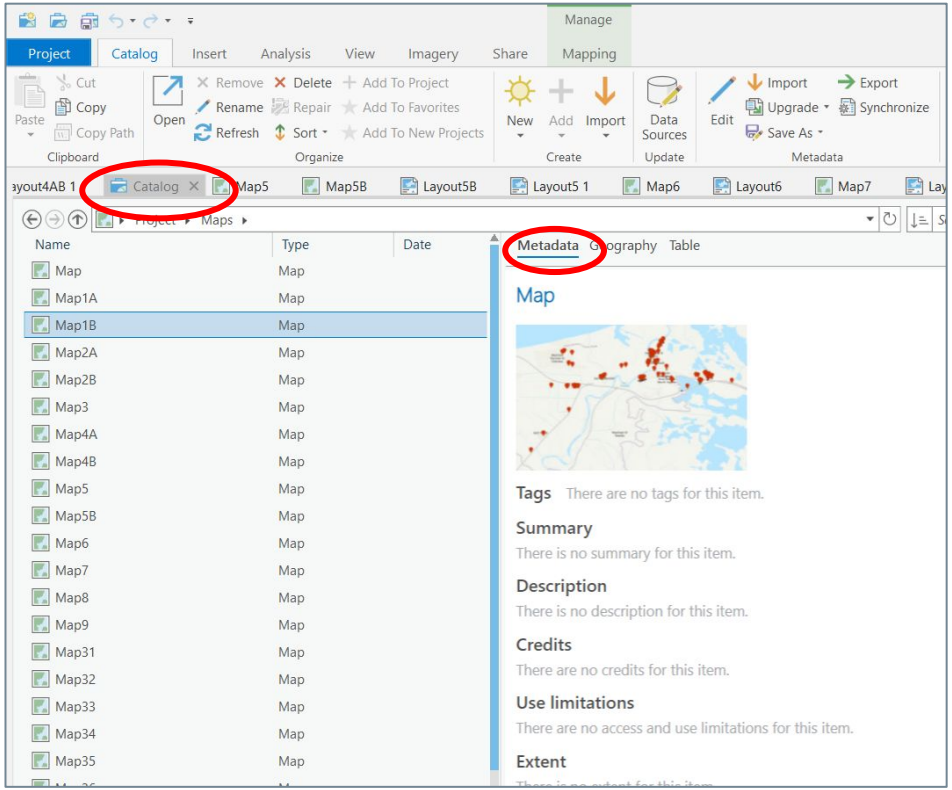
3. External sidecar or stand-alone metadata files
   a. As XML
   b. As PRJ (shapefile)
   c. As WLD (geoTIFF)
   d. As a README

Demonstration: Finding Layer Metadata in QGIS



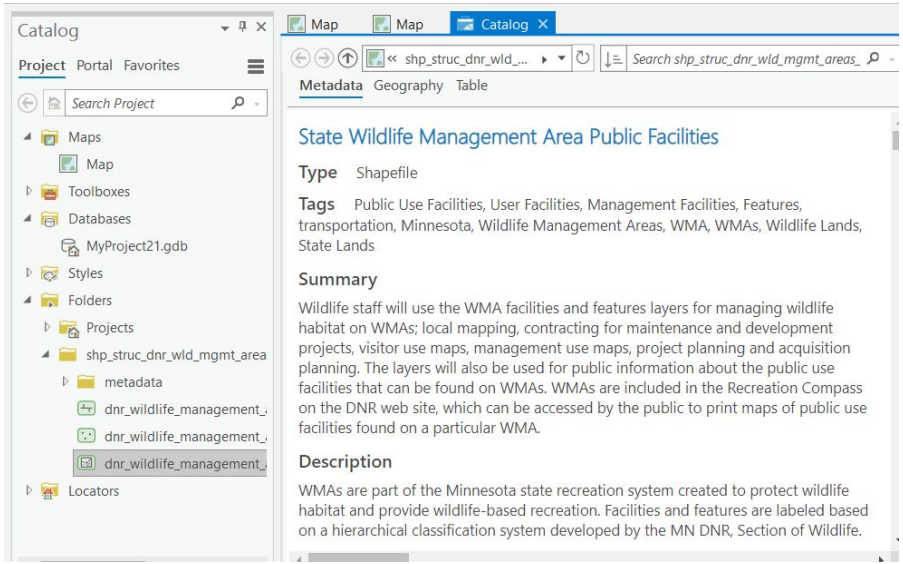
CURATED

Demonstration: Finding Metadata in ArcGIS

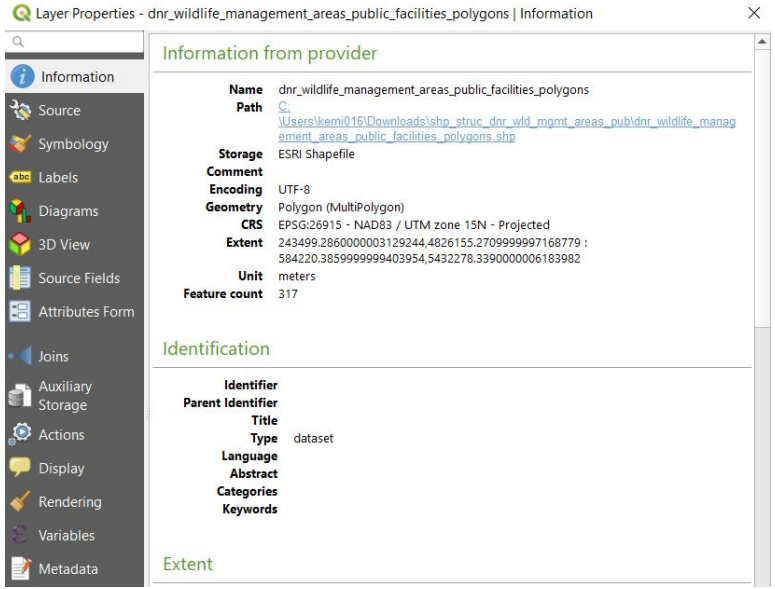


Comparing Metadata Views in ArcGIS Pro and QGIS

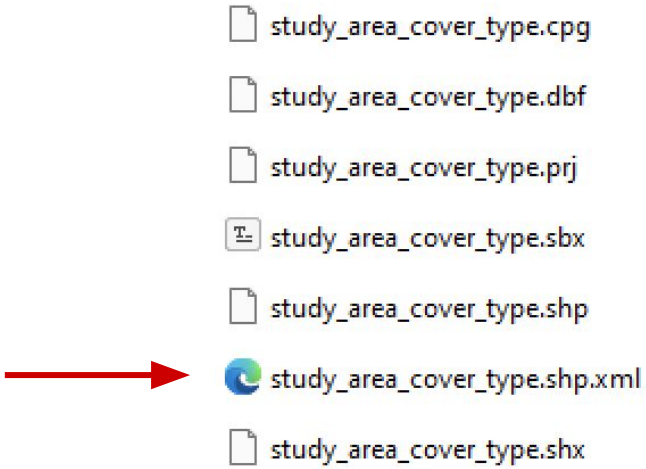
View of the metadata in ArcGIS Pro



View of the metadata in QGIS



External Sidecar and Stand-Alone Metada
Demonstration: .shp.xml



C U R A T E D

Questions to Consider

C

U

R

A

T

E

D

Check and Understand:

1. Does the coordinate reference system make sense for the geographical location (i.e. not web mercator)?
2. Are all the data layers in the same projection?
3. Does the Descriptive and Administrative metadata match what the researcher provided?
4. Is there pre-existing metadata in layers derived from institutional sources?
5. Is anything in this embedded metadata useful for a README file?

If anything does not make sense, **request** clarifications from the researcher

GIS Metadata is Hard

C
U
R
A
T
E
D

At least at first glance. The recommended standards have many elements, but are often considered to be too complex. Best practices suggest that you work with a defined subset that resembles Dublin Core and includes what is necessary for reusability of the GIS data.

Essential Documentation: All GIS Data

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

C U R A T E D

Essential Documentation: Coordinate Reference System (CRS)

The **datum** and the **geographic coordinate system**:

WGS84 (global) or **NAD83** (north american)

The **projection** (if exists)

UTM Zone 16N or **Florida State Plane East**

The **EPSG code** should be included after the text description

NAD83(2011) / UTM zone 17N : EPSG 6346

Essential Documentation: Bounding Box or Gazetteer Name

geoLocation (DataCitev4.0 example)

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)

Essential Documentation: Lineage and Provenance

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.

Increased Confidence and Transparency >> More reproducible

Essential GIS Metadata Vector and Raster

CURATED

Apart from Dublin Core …

- Coordinate reference system
- Bounding coordinates/Gazetteer name(s)
- Process steps (provenance)

Vector

- Data dictionary
- Recommended scale (or scale of digitization)

Raster

- Pixel size
- Band information
- Look up tables (if categorized)

Essential Vector Documentation: Data Dictionary

- What the Attributes or Variables represent (ie. Variable Definitions)
  - Names
  - Description
  - Units of measure
  - Clarification of the difference between Null cells and cells with a “0” (zero) in them
  - Definitions for Coded variables. For example: "GRIDCODE" values are as follows: 0 = non-nesting area, 1 = upland shrubland, 2 = grassland, 3 = wetland shrubland, 4 = mature forest, 5 = forested wetland.
- Common pushback:
  - Defining variables derived from pre-existing datasets
  - Defining variables included but not actively used for a publication

Essential Raster Documentation: Band and Pixel Information

- 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

Evaluate GIS Metadata for FAIRness: Accessible

C
U
R
A
T
E
D

Does the README make the data FAIR?

- 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 FAIR?

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

Evaluate GIS Metadata for FAIRness: Reusable

C
U
R
A
T
E
D

Does the README make the data FAIR?

- 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 FAIR?

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

Put it in Practice: Exercise

CURATED

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

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