GIScience

Data Uncertainty

Image from NASA:
https://commons.wikimedia.org/wiki/File:Earth_Western_Hemisphere_transparent_background.png#filelinks

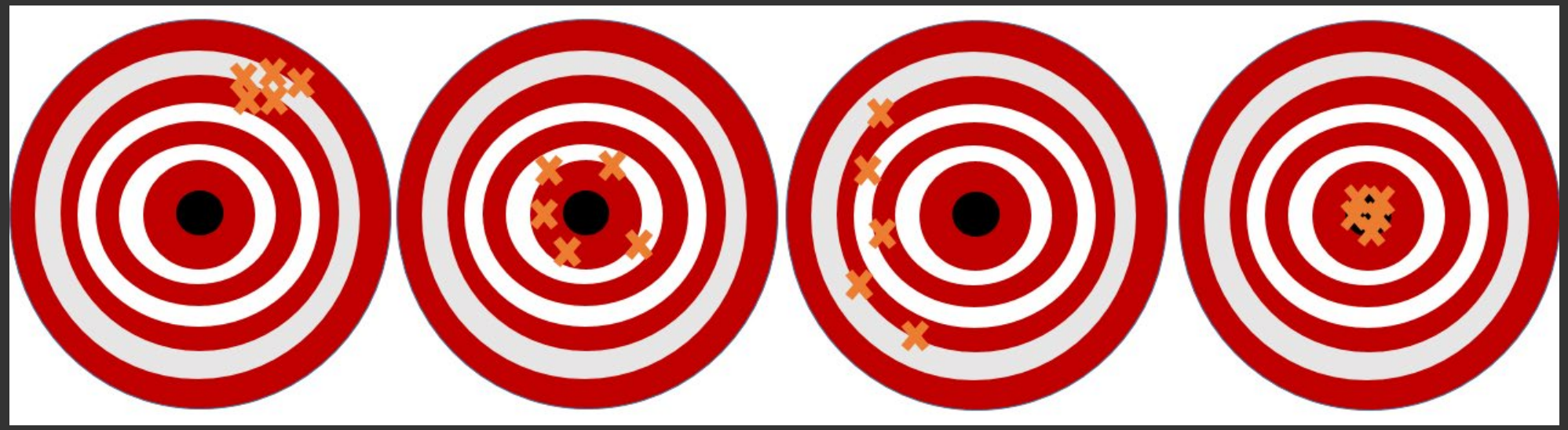
WV VIEW

AmericaView℠
Empowering Earth Observation Education
AmericaView.org - Est. 2003

Data Uncertainty

Accuracy vs. Precision

- Accuracy = agreement with truth
- Precision = agreement with other measurements



Uncertainty

- The difference between the real world and its representation on a map

- Many types of uncertainty: positional, errors, scale and generalization, fuzziness and definitions, sampling of data, timeliness

Positional Uncertainty

- Horizontal (X, Y)
- Vertical (Z)
- Can evaluate using Root Mean Square Error (RMSE)

Example from Jackie Strager

$$RMSE_x = \sqrt{\Sigma(x_{predicted} - x_{actual})^2/n}$$

$$RMSE_y = \sqrt{\Sigma(y_{predicted} - y_{actual})^2/n}$$

$$RMSE = \sqrt{(RMSE_x^2 + RMSE_y^2)}$$

Errors

- Inherent = present in source dataset
- Operational = results from data operations and manipulations
- Typos and simple mistakes, human bias, misinterpretation, misclassification, etc.

Geography 350/550:

Introduction to GIScience

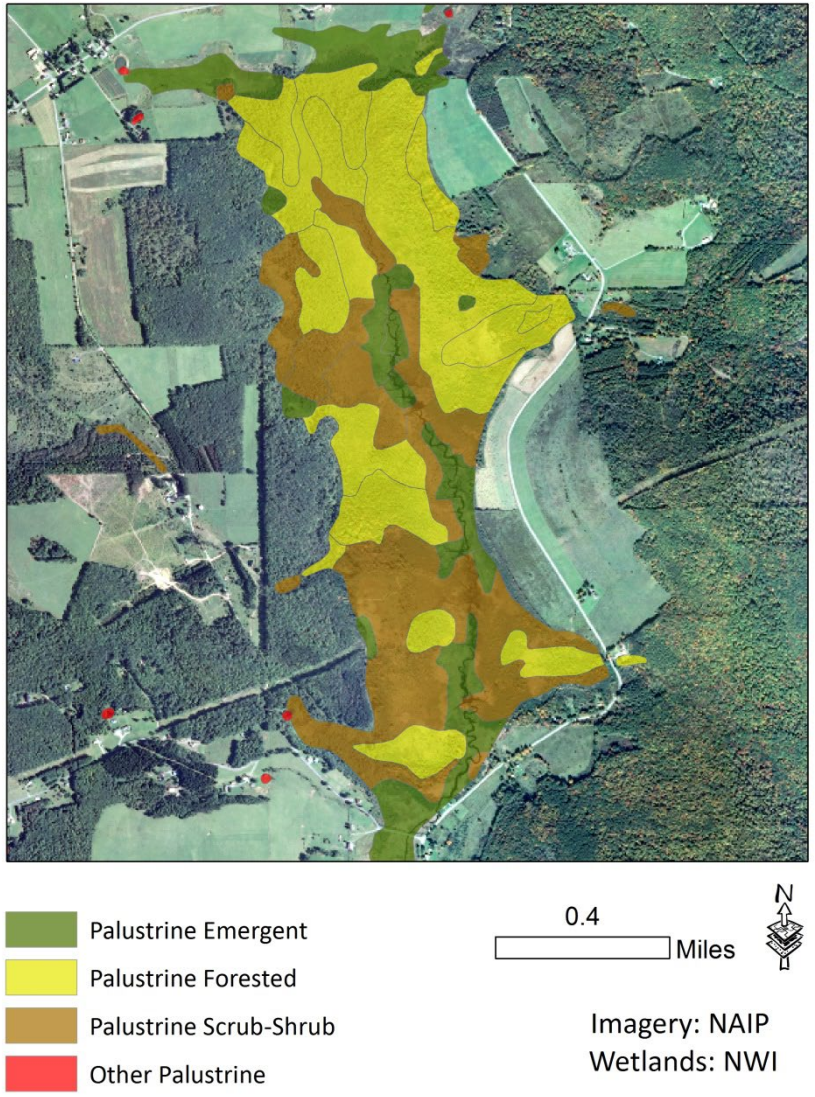
Lab Manal

Scale and Generalization

Red = 1:24,000 scale county boundary
Orange = 1:100,000 scale county boundary

- All maps and geospatial data are generalizations
- Detail and presentation change with scale

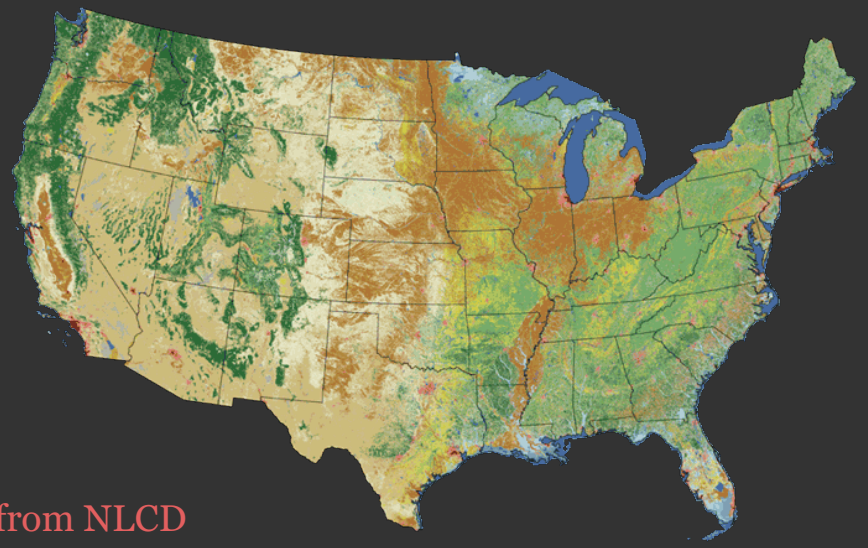
Fuzziness



- Not all features have distinct boundaries
- Not all categories have clear definitions

Fuzziness

- Not all features have distinct boundaries
- Not all categories have clear definitions



Data from NLCD

Class\ Value	Classification Description
Water	
11	**Open Water**- areas of open water, generally with less than 25% cover of vegetation or soil.
12	**Perennial Ice/Snow**- areas characterized by a perennial cover of ice and/or snow, generally greater than 25% of total cover.
Developed	
21	**Developed, Open Space**- areas with a mixture of some constructed materials, but mostly vegetation in the form of lawn grasses. Impervious surfaces account for less than 20% of total cover. These areas most commonly include large-lot single-family housing units, parks, golf courses, and vegetation planted in developed settings for recreation, erosion control, or aesthetic purposes.
22	**Developed, Low Intensity**- areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 20% to 49% percent of total cover. These areas most commonly include single-family housing units.
23	**Developed, Medium Intensity** -areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 50% to 79% of the total cover. These areas most commonly include single-family housing units.
24	**Developed High Intensity**-highly developed areas where people reside or work in high numbers. Examples include apartment complexes, row houses and commercial/industrial. Impervious surfaces account for 80% to 100% of the total cover.
Barren	
31	**Barren Land (Rock/Sand/Clay)** - areas of bedrock, desert pavement, scarps, talus, slides, volcanic material, glacial debris, sand dunes, strip mines, gravel pits and other accumulations of earthen material. Generally, vegetation accounts for less than 15% of total cover.
Forest	
41	**Deciduous Forest**- areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. More than 75% of the tree species shed foliage simultaneously in response to seasonal change.
42	**Evergreen Forest**- areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. More than 75% of the tree species maintain their leaves all year. Canopy is never without green foliage.
43	**Mixed Forest**- areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. Neither deciduous nor evergreen species are greater than 75% of total tree cover.
Shrubland	
51	**Dwarf Scrub**- Alaska only areas dominated by shrubs less than 20 centimeters tall with shrub canopy typically greater than 20% of total vegetation. This type is often co-associated with grasses, sedges, herbs, and non-vascular vegetation.
52	**Shrub/Scrub**- areas dominated by shrubs; less than 5 meters tall with shrub canopy typically greater than 20% of total vegetation. This class includes true shrubs, young trees in an early successional stage or trees stunted from environmental conditions.
Herbaceous	
71	**Grassland/Herbaceous**- areas dominated by gramanoid or herbaceous vegetation, generally greater than 80% of total vegetation. These areas are not subject to intensive management such as tilling, but can be utilized for grazing.
72	**Sedge/Herbaceous**- Alaska only areas dominated by sedges and forbs, generally greater than 80% of total vegetation. This type can occur with significant other grasses or other grass like plants, and includes sedge tundra, and sedge tussock tundra.
73	**Lichens**- Alaska only areas dominated by fruticose or foliose lichens generally greater than 80% of total vegetation.
74	**Moss**- Alaska only areas dominated by mosses, generally greater than 80% of total vegetation.
Planted/Cultivated	
81	**Pasture/Hay**-areas of grasses, legumes, or grass-legume mixtures planted for livestock grazing or the production of seed or hay crops, typically on a perennial cycle. Pasture/hay vegetation accounts for greater than 20% of total vegetation.
82	**Cultivated Crops** -areas used for the production of annual crops, such as corn, soybeans, vegetables, tobacco, and cotton, and also perennial woody crops such as orchards and vineyards. Crop vegetation accounts for greater than 20% of total vegetation. This class also includes all land being actively tilled.
Wetlands	
90	**Woody Wetlands**- areas where forest or shrubland vegetation accounts for greater than 20% of vegetative cover and the soil or substrate is periodically saturated with or covered with water.
95	**Emergent Herbaceous Wetlands**- Areas where perennial herbaceous vegetation accounts for greater than 80% of vegetative cover and the soil or substrate is periodically saturated with or covered with water.

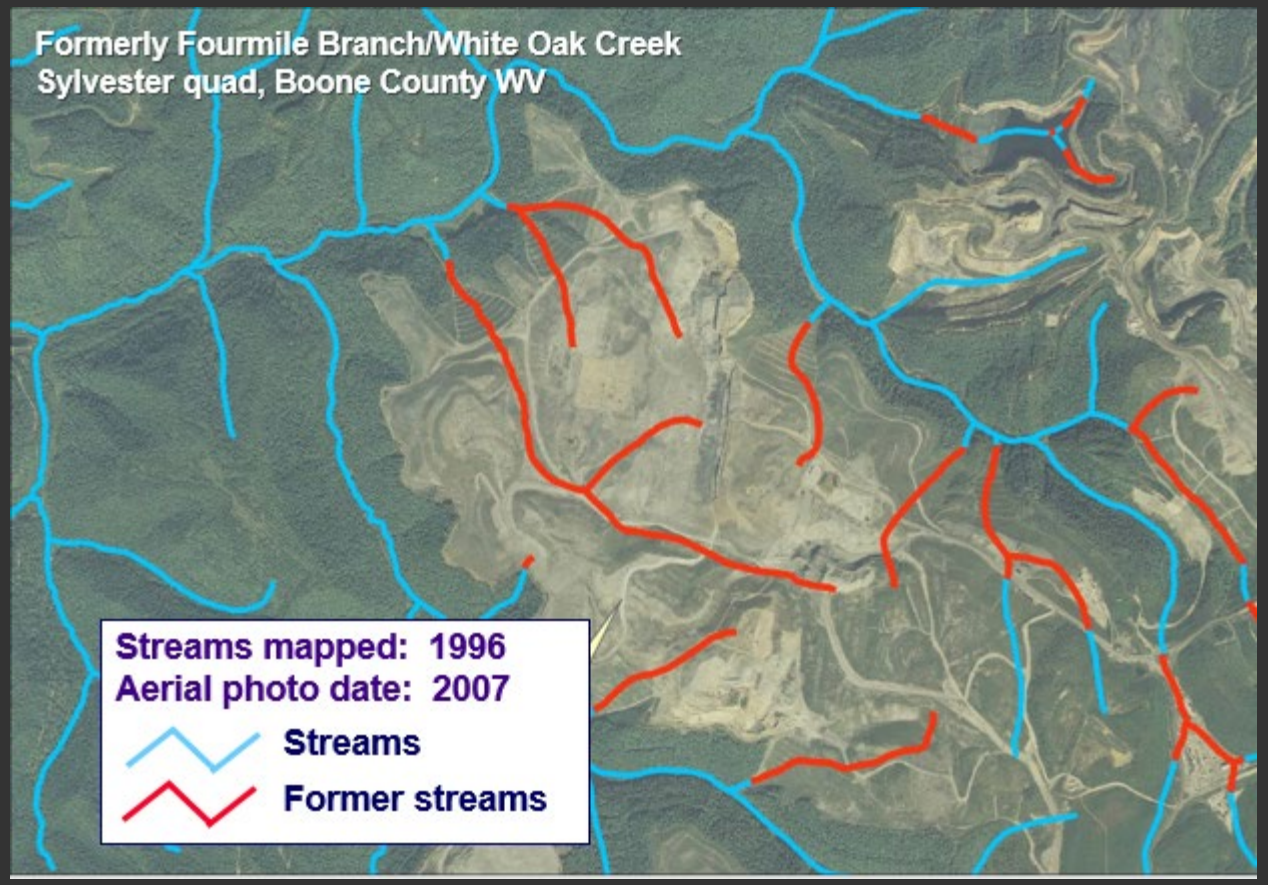
Assessing a Classification

- Commonly use an error matrix
- Explores confusion between classes
- How were the classes confused?
- Should be created from randomized validation data

		Reference Data							
		Forested	*Pasture/ Grass*	*Barren*	*Cropland*	*Developed*	*Water*	Total	User's Accuracy
Classified Data	*Forested*	91	8	1	11	0	2	113	81%
	Pasture/ Grass	4	81	0	15	0	0	100	81%
	Barren	0	0	87	2	10	2	101	86%
	Cropland	3	11	0	69	0	0	83	83%
	Developed	0	0	10	0	73	0	83	88%
	Water	2	0	2	3	17	96	120	80%
	Total	100	100	100	100	100	100		
	Producer's Accuracy	91%	81%	87%	69%	73%	96%		

Temporal Uncertainty

- Things change
- How important is timeliness?
- How often should layers be updates?



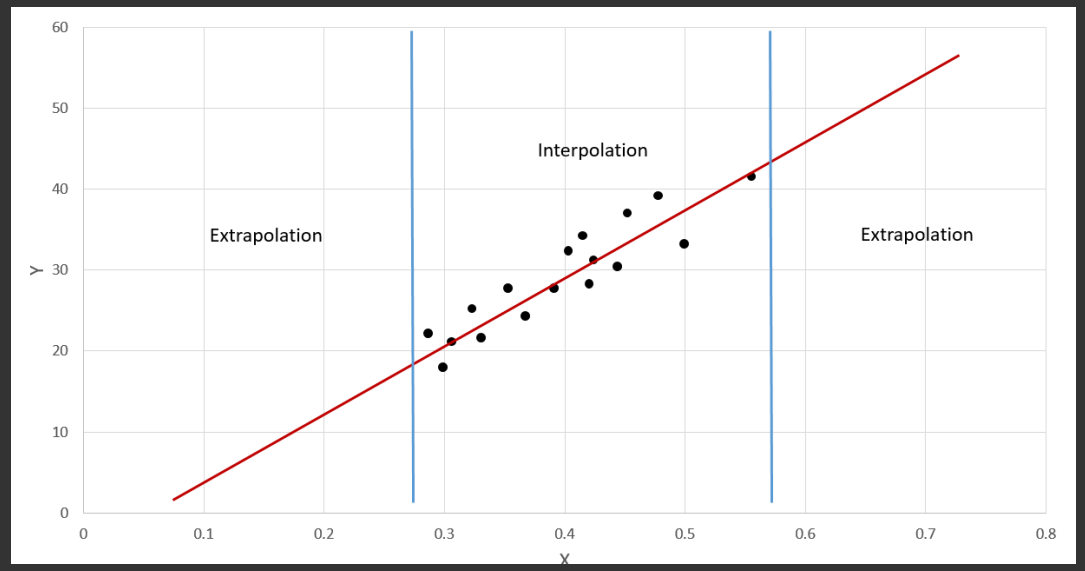
Example by Jackie Strager

Sampling Uncertainty

We can't measure everything everywhere

However, sometimes we need data at all locations

This requires interpolation, estimation, or guessing



Minimizing Uncertainty

1. Implement quality control
2. Use standardize procedures
3. Define scale
4. Define minimal mapping unit (MMU)
5. Double-check data
6. Create metadata
7. Document methods
8. Clearly define definitions
9. Measure uncertainty

Assessing Accuracy/Uncertainty

- Positional Accuracy = RMSE
- Elevation/Vertical = RMSE
- Classification Accuracy = error matrix
- Modeling Accuracy = many methods, will discuss later
- Must compare measurements/ predictions to data of a higher accuracy using statistical methods

$$RMSE_x = \sqrt{\Sigma(x_{predicted} - x_{actual})^2/n}$$

$$RMSE_y = \sqrt{\Sigma(y_{predicted} - y_{actual})^2/n}$$

$$RMSE = \sqrt{(RMSE_x^2 + RMSE_y^2)}$$

		Reference Data							
		Forested	*Pasture/ Grass*	*Barren*	*Cropland*	*Developed*	*Water*	Total	User's Accuracy
Classified Data	*Forested*	91	8	1	11	0	2	113	81%
	Pasture/ Grass	4	81	0	15	0	0	100	81%
	Barren	0	0	87	2	10	2	101	86%
	Cropland	3	11	0	69	0	0	83	83%
	Developed	0	0	10	0	73	0	83	88%
	Water	2	0	2	3	17	96	120	80%
	Total	100	100	100	100	100	100		
	Producer's Accuracy	91%	81%	87%	69%	73%	96%		

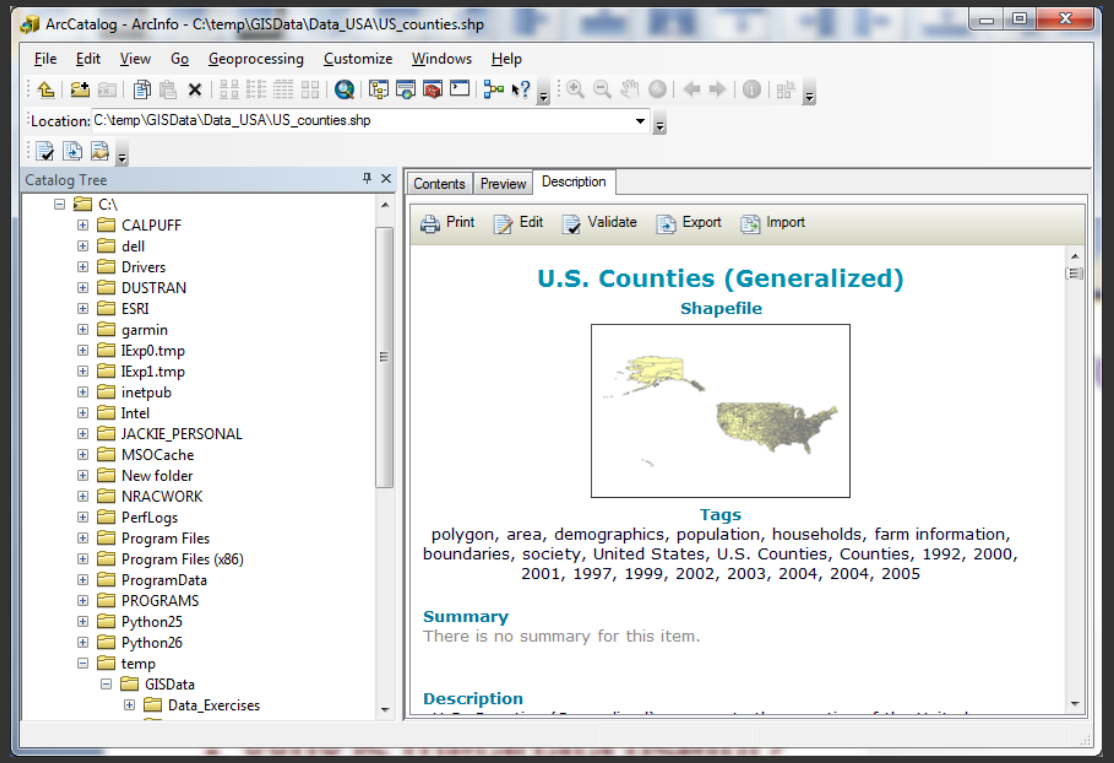
Metadata

Documenting your data: Metadata

- Data about your data
- A description of the basic characteristics of a data resource
- Describes content, source, lineage, methods, developer, coordinate system, extent, structure, spatial accuracy, attributes, and responsible organization
- Is data suitable for an intended use?
- Accurate enough?
- Cover area of interest?
- Provide necessary information?

Why is metadata important?

- Makes your data more interpretable
- Find out more about a dataset
- Minimize assumptions and guesses
- Contact dataset originator for more information
- Ensures long-term viability of your data



US: Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM)

1. Identification, describing the dataset
2. Data quality
3. Spatial data organization
4. Spatial reference coordinate system
5. Entity and attribute
6. Distribution and options for obtaining the data set
7. Currency of metadata and responsible party
8. Citation
9. Time period information, used with other sections to provide temporal information
10. Contact organization or person

US: Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM)

FGDC.GOV
FEDERAL GEOGRAPHIC DATA COMMITTEE

HOME | WHO WE ARE | WHAT WE DO | INITIATIVES | ORGANIZATION | RESOURCES

GEOSPATIAL METADATA STANDARDS AND GUIDELINES

Home / Geospatial Metadata / Geospatial Metadata Standards and Guidelines

When metadata records are formatted to a common standard, it facilitates the location and readability of the metadata by both humans and machines. NSDI Stakeholders have long utilized the Content Standard for Digital Geospatial Metadata (CSDGM). However, Circular A-119 Revised directs agencies to use voluntary consensus standards in lieu of government-unique standards. Furthermore, Circular A-16 promotes the use of international standards to advance the building of the Global Spatial Data Infrastructure (GSDI). Several ISO metadata standards are now endorsed by the FGDC and federal agencies and NSDI Stakeholders are encouraged to make the transition to ISO metadata.

Geospatial Metadata Standards

- International Standards Organization (ISO) 191xx series of metadata standards
- FGDC-authored Content Standard for Digital Geospatial Metadata (CSDGM)

Geospatial Metadata Guidelines

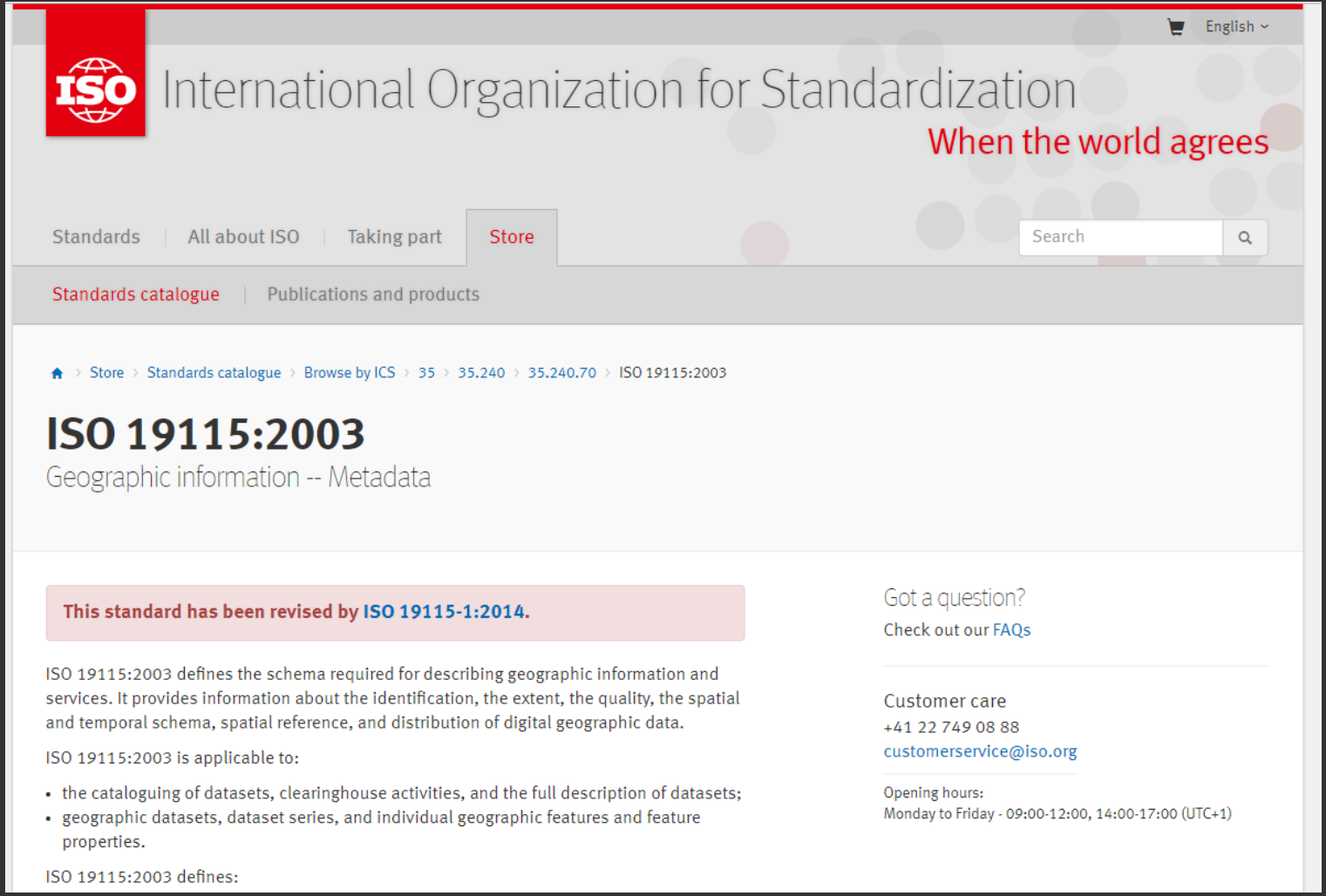
- Selecting a Geospatial Metadata Standard
- Benefits of ISO
- GeoPlatform/Data.gov Metadata Publication Guidelines (pdf)
- National Geospatial Data Asset (NGDA) Metadata Guidelines (pdf)
- Project Open Data (POD) Metadata Schema

https://www.fgdc.gov/metadata/geospatial-metadata-standards

International: ISO 19115 International Standards for Metadata

- Two elements: mandatory and conditional
- Mandatory = data title, dataset reference date, dataset language, dataset topic category, abstract, metadata point of contact, and metadata date stamp
- Conditional = dataset responsible party, geographic location by coordinates, dataset character set, spatial resolution, distribution format, spatial representation type, reference system, lineage statement, online resource, metadata file identifier, metadata standard name, metadata standard version, metadata language, and metadata character list

International: ISO 19115 International Standards for Metadata



https://www.iso.org/standard/26020.html

XML: eXtensible Markup Language

- Metadata can be written using XML
- A markup language like HTML
- Information wrapped in tags
- Can view in a web browser

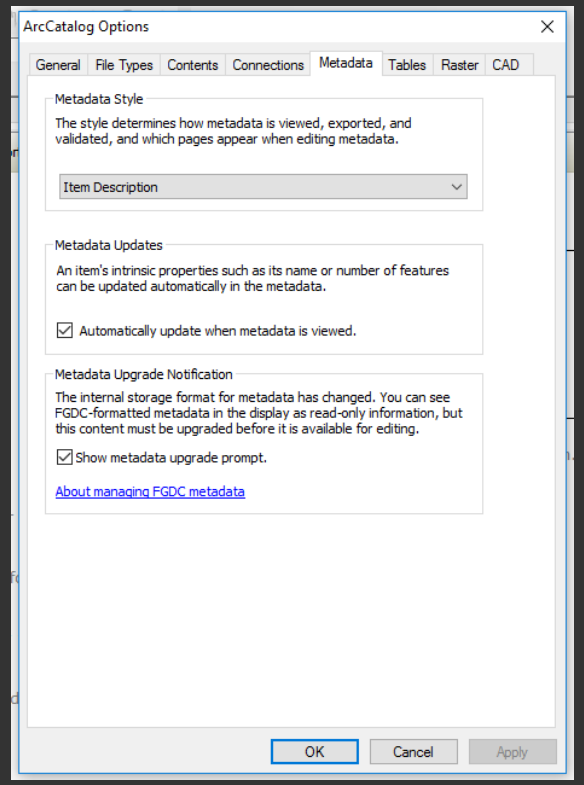
wvgis.wvu.edu/lidar/data/USGS-FEMA_2012_Berkeley-Morgan-Jefferson/BerkeleyMorganFrederickClarke_ClassifiedLAS.xml

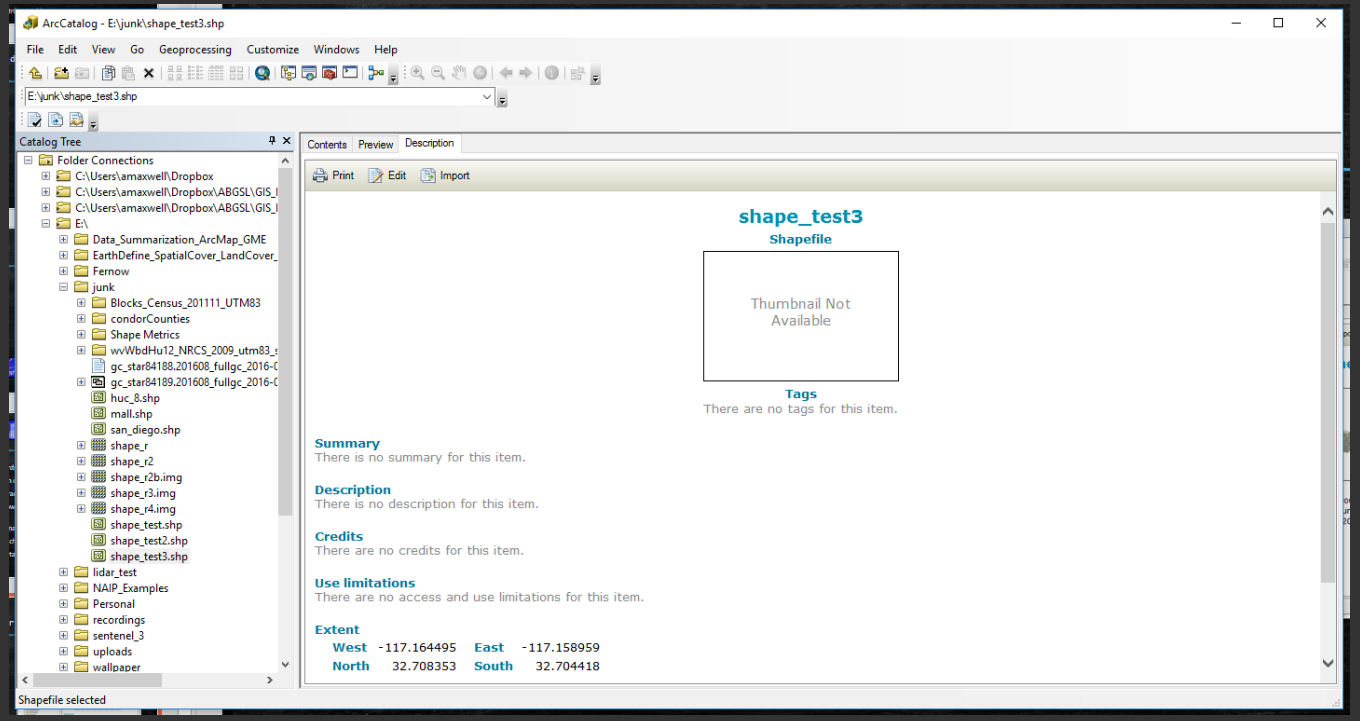
This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
▼<metadata>
  ▼<idinfo>
    ▼<citation>
      ▼<citeinfo>
        ▼<origin>
           Berkeley Clarke Frederick and Morgan County - FEMA Region 3 FY12 LiDAR
        </origin>
        <pubdate>201209</pubdate>
        <pubtime>Unknown</pubtime>
        <title>Classified LAS</title>
        <geoform>map</geoform>
      </citeinfo>
    </citation>
    ▼<descript>
      ▼<abstract>
         Dewberry collected LiDAR for ~3,942 square miles in various Virginia, West Virigia, and Maryland Counties. The acquisition was performed by Geodigital. This metadata covers the LiDAR produced for the Loudoun County project area. The nominal pulse spacing
         for this project is 1.6 ft (0.5 meters). This project was collected with a sensor which collects intensity values for each discrete pulse extracted from the waveform. GPS Week Time, Intensity, Flightline and echo number attributes were provided for each
         LiDAR point. Dewberry used proprietary procedures to classify the LAS according to contract specifications: 1-Unclassified, 2-Ground, 7-Noise, 9-Water, 10-Ignored Ground due to breakline proximity, and 11-Withheld. Dewberry produced 3D breaklines and
         combined these with the final LiDAR data to produce seamless hydro flattened DEMs for the 1154 tiles (1500 m x 1500 m) that cover the this deliverable.
      </abstract>
      ▼<purpose>
         The purpose of this LiDAR data was to produce high accuracy 3D elevation products, including tiled LiDAR in LAS 1.2 format, 3D breaklines, and a 1 m cell size hydro flattened Digital Elevation Models (DEM).
      </purpose>
      ▼<supplinf>
         A complete description of this dataset is available in the Final Project Report.
      </supplinf>
    </descript>
    ▼<timeperd>
```

Creating Metadata

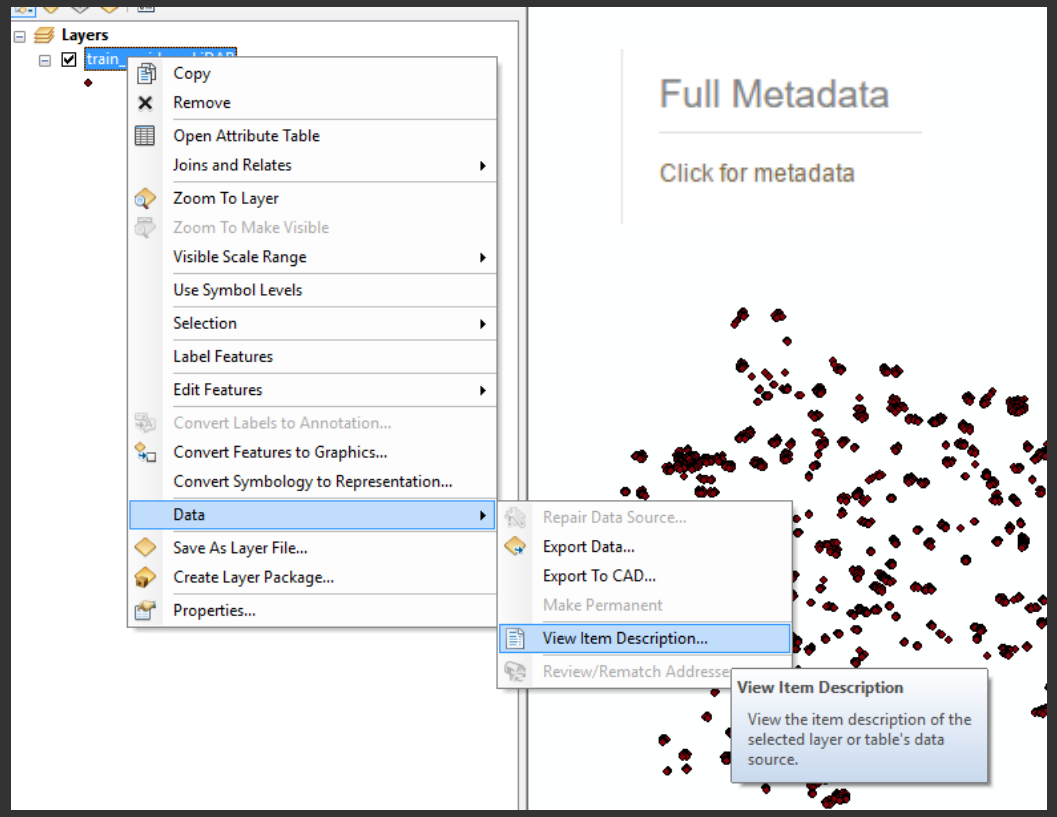
- Metadata can be created in ArcCatalog or the Catalog View in ArcGIS Pro using a variety of standards

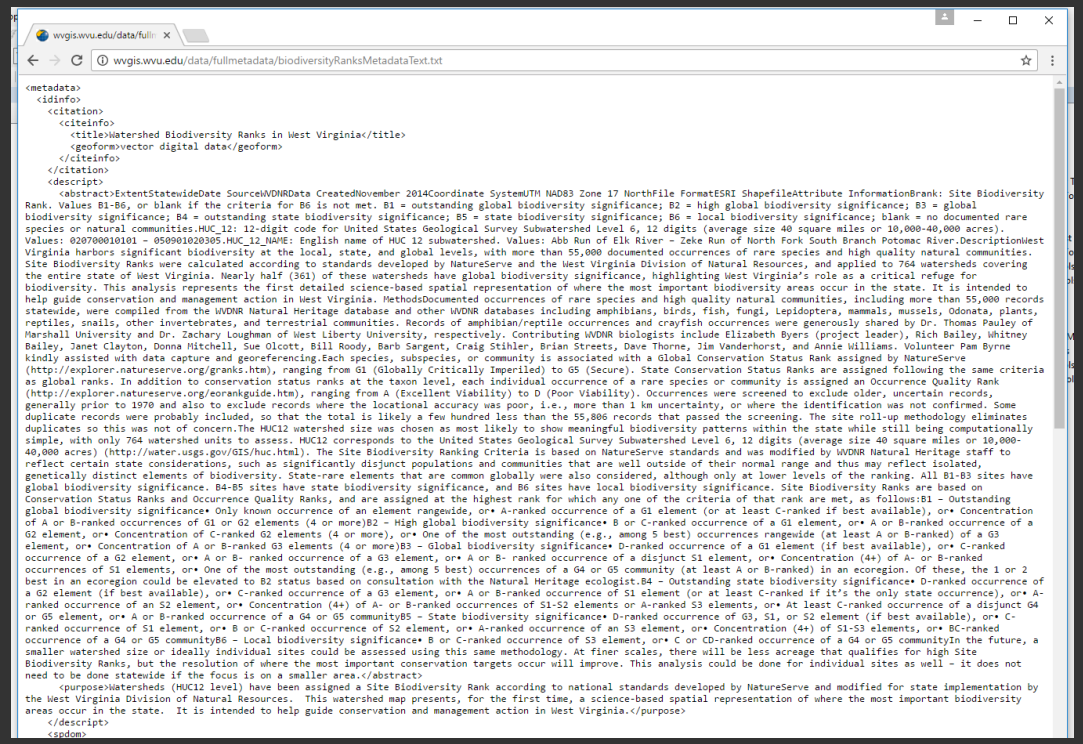




Viewing Metadata

- Metadata can be created in ArcCatalog or the Catalog View in ArcGIS Pro using a variety of standards





This is the end of this lecture module.

Please return to the West Virginia View Webpage for additional content.