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

Geospatial Problem Solving

Data Uncertainty

WV VIEW

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

When working with data, it is important to understand limitations, errors, and uncertainty.

This module will focus on uncertainty in our spatial data.

We will also discuss how data characteristics are documented using metadata.

Module Topics

1. Uncertainty vs. precision
2. Positional uncertainty
3. Inherent and operational errors
4. Scale and generalization
5. Fuzziness
6. Temporal uncertainty
7. Sampling uncertainty
8. Metadata standards and creation

2

These are the topics covered in this module.

Data Uncertainty

3

Accuracy vs. Precision

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

4

Before we move on, I want to provide a review of the difference between accuracy and precision. Since we often use these terms interchangeably colloquially, this can be a bit confusing.

Accuracy represents agreement with truth or the correct measure. Here, we are concerned with whether or not the measurements agree with the correct value on average.

In contrast, precision had to do with whether data agree with each other or how repeatable measurements are. This has nothing to do with agreement with the correct value.

The difference here is often explained with a target where the bull's eye represents the correct answer.

The example on the left represents an archer that is not accurate but precise. So, he or she is missing the bull's eye, but each arrow hits in nearly the same place. So, he or she is consistent even if the bull's eye is missed.

The next example shows a pattern that is less precise but accurate. There is more spread in the pattern, or the archer is less consistent, but the average position would be near the bull's eye.

The third example is neither accurate nor precise since the archer is not consistent and the average location would not be near the bull's eye.

The last pattern is both precise and accurate.

As a second example, imagine you are collecting GPS measurements at a location. If you get roughly the same coordinates each time you take a measurement, then the measurements are precise but not necessarily accurate, as there could be some systematic error or calibration issue causing the measurements on average to not be close to the correct coordinates.

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

5

Uncertainty is the difference between the real world and its representation on a map or within your datasets.

We will now investigate some types or causes of uncertainty in our maps and spatial data.

Positional Uncertainty

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

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

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

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

Example from Jackie Strager

Positional uncertainty relates to differences in the position of your data relative to the correct location.

You can consider uncertainty in 2D space as x and y coordinates in a geographic or projected space. You can also consider the uncertainty of elevation measurements. Combining these, you can consider uncertainty in 3D space.

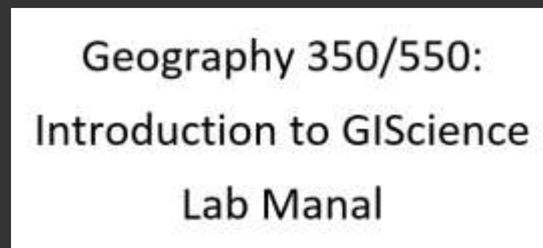
As an example, the provided image shows some uncertainty in a vector line road dataset in comparison to the orthophotography.

Positional error is generally assessed using root mean square error or RMSE, similar to how georeferencing results are assessed.

All spatial data have some positional uncertainty; however, this is only a concern if there is too much uncertainty to support your analysis or mapping needs.

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

7

Errors include typos, simple mistakes, human bias, misinterpretation, and misclassification.

Specifically, inherent errors are present in the source data and are subsequently incorporated into your data product. For example, if you are copying data from a table into your database, errors in the table will be incorporated into your new data.

Operational errors result from data operations and manipulations. This could result from incorrect analyses, misinterpretations of features when digitizing, or misclassifying map objects.

Some errors exist in all datasets. Again, this is really only a concern when the errors could impact your map or analysis.

As a side note, the provided example shows a typo I made on one of my first lab manuals.

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

8

All maps and spatial data layers are generalizations of reality. We can't map every blade of grass, atom, or quark.

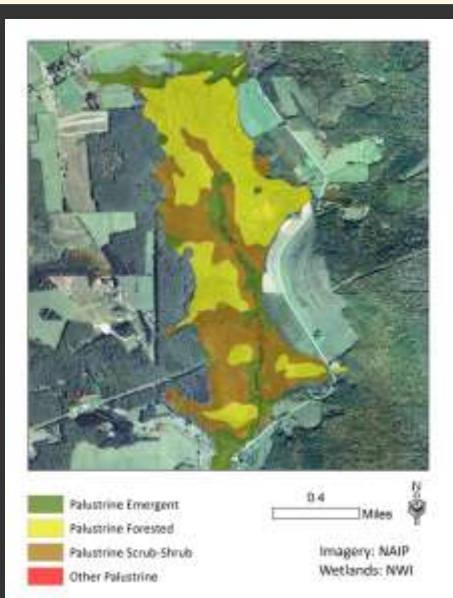
The act of generalization produces uncertainty since generalizations produce differences between the real-world object and its representation on the map.

As an example, the red line represents a county boundary digitized at 1:24,000 scale while the orange line represents the same boundary but at 1:100,000 scale. Uncertainty exists in both datasets, and each boundary is not perfect.

Your concern should be whether the level of generalization is an issue for your specific analysis or mapping tasks.

You may also prefer to generalize the data further. For example, when designing a map you may find that you would like to provide a more generalized representation of the landscape that is more cartographically pleasing.

Fuzziness



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

9

Fuzziness arises from ambiguity in boundaries or definitions.

Some features have definitive boundaries and definitions. For example, roads and sidewalks generally are fairly discrete features with well defined definitions.

In contrast, the boundaries between wetlands and uplands or between soil types or groups are often gradational. So, if you draw a feature to represent a wetland it will automatically have some inherent uncertainty since you are representing a gradational or “gray” boundary as a hard line.

In regards to definitions, not all map features can be clearly defined. For example, the difference between urban areas and suburban areas may be difficult to clearly define. It may also be difficult to fully define a certain type of soil or forest.

Fuzziness arises from the complexity of the real world. The best we can do is create maps and features that describe the landscape the best that we can given this complexity and inherent limitations in how we map and model the landscape.

Fuzziness

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

Data from NLCD

10

As an example of fuzziness, the National Land Cover Database or NLCD separates the country into a set of defined classes. However, a certain pixel or location may not fit well into one of these classes. In other words, the class definitions may be difficult to apply to all landscapes without some ambiguity. A single pixel or cell could also be mixed or contain multiple land cover types.

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%		

11

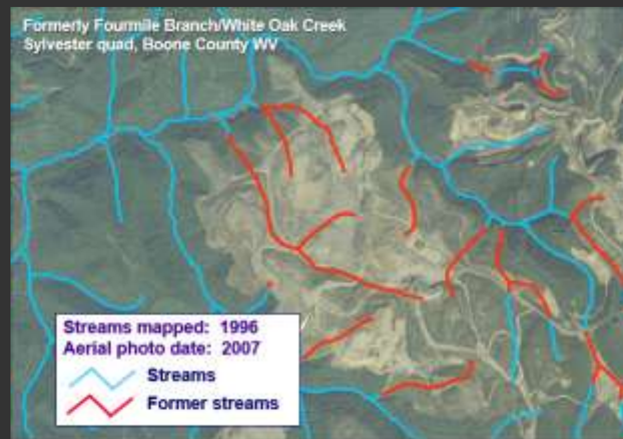
Classification accuracies are often assessed using the confusion matrix in which the classification on the map is compared to the correct classification at that location based on high quality validation data. Using this method, it is possible to get a sense of the accuracy of the classification and map along with the sources of confusion between the classes.

We will not discuss confusion matrices in detail here. This topic is generally covered in a remote sensing course.

This method is simply being used here as an example of how we can assess uncertainty in classification map products.

Temporal Uncertainty

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



Example by Jackie Strager

12

Temporal uncertainty arises from data being outdated.

For example, in the provided example the stream network has changed between 1996 and 2007 as a result of surface mining.

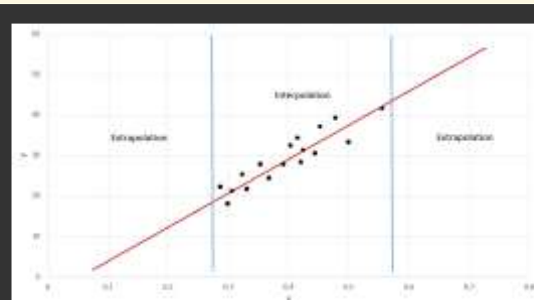
For certain tasks, having access to updated data may not be important. However, timeliness may be necessary for certain tasks. Further, some features may change fairly often, such as development or demographics, whereas other features may change slowly, such as bedrock geology.

Sampling Uncertainty

We can't measure everything everywhere

However, sometimes we need data at all locations

This requires interpolation, estimation, or guessing



13

We can't measure everything everywhere.

For example, it is common to collect measurements at point locations, such as collecting climate or weather data at weather stations.

If you would like to estimate values at locations that were not sampled, you will need to apply interpolation or extrapolation as a means to obtain a continuous surface from your discrete measurements. Since these locations were not actually measured but estimated from nearby sample locations, there will be some error or uncertainty in the product.

Extrapolation is generally considered to be more uncertain than interpolation. As an example, if you are trying to predict a variable y using a variable x with a line or linear relationship, estimates within the bounds of your sample data will likely be less uncertain in comparison to estimates outside of the bounds of your data. Maybe outside of the bounds the relationship is no longer linear, begins to increase exponentially, or levels off. Without data points or samples, this would be hard to determine.

In a spatial context, extrapolation would be making an estimate outside of the geographic extent of your data, as represented by the yellow star. In contrast, interpolation would be making an estimate within the geographic bounds of your data, as represented by the orange star.

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

14

As GIS professionals and analysts, there are many methods available to minimize the uncertainty in our maps and data.

First, we can use quality control procedures, such as spot checking the results or having the results reviewed by multiple analysts.

If standard procedures or protocols are well defined for a specific task, we should follow these protocols. For example, standard protocols could be followed to georeference an image.

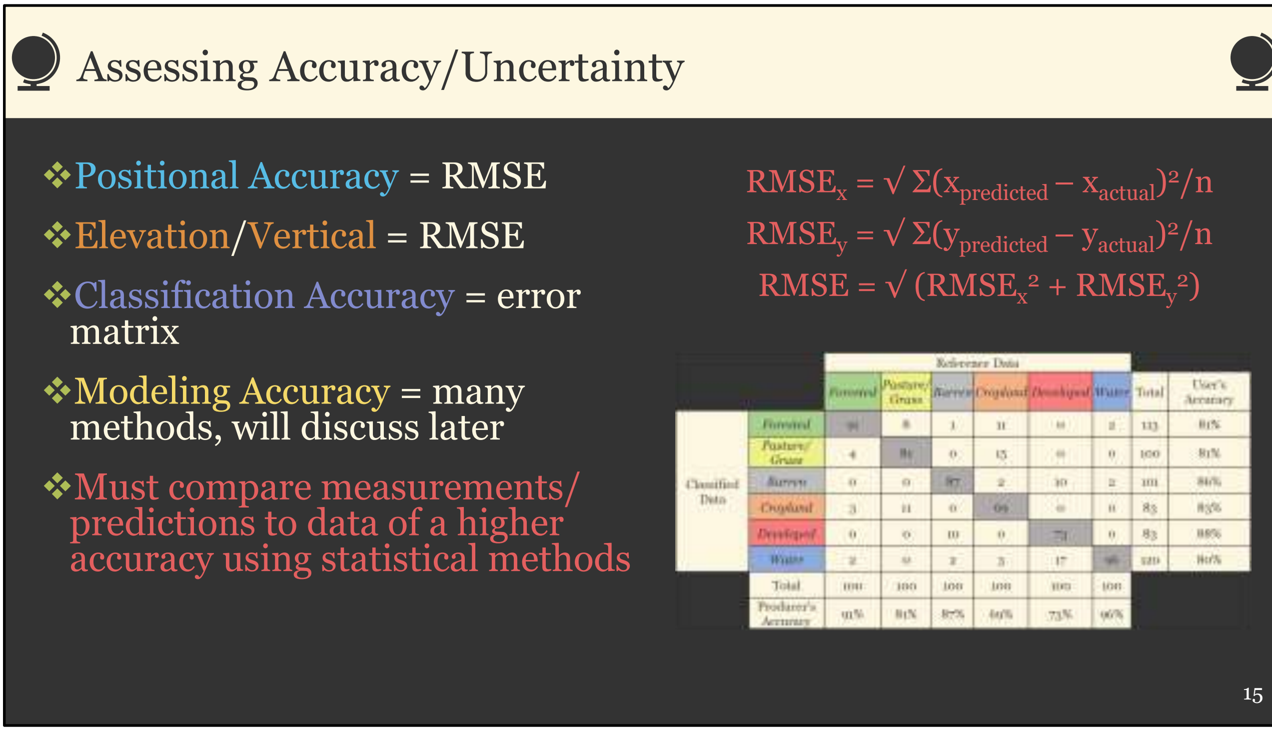
We should clearly define the scale or scale range at which the data were created and the scale or scale range at which they should be used. Data collected at 1:250,000 scale will often look fairly inaccurate when viewed at 1:10,000 scale. Further, these data may not be appropriate for an analysis being performed at 1:10,000 scale. Relating to this, it is important to define a minimal mapping unit, or MMU, for an analysis. This can help inform the appropriate use and scale of analysis to which the data can be applied.

You should double-check results and products. It is a good idea to have others review our work.

Metadata and method documentation should be provided that clearly define the data and how it was produced. Remember to create README files.

Definitions should be clearly defined when possible.

If appropriate and feasible, we should attempt to evaluate and measure the uncertainty or accuracy of our data.



As mentioned above, there are some methods available to measure or assess accuracy and uncertainty.

For example, positions, elevations, and georeferencing results can be assessed with the RMSE statistic. This measure can also be used to assess the prediction of continuous variables.

Classification accuracy can be assessed using a confusion matrix.

In a later module, we will discuss spatial modeling methods and will investigate other assessment methods.

Metadata

16

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?

17

Metadata are data about your data. They commonly include a description of the basic characteristics of the data, content, source, lineage, methods, developer, coordinate system, extent, structure, spatial accuracy, attributes, and responsible organization.

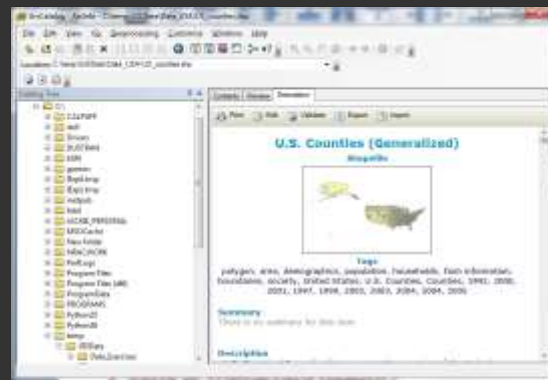
Metadata can help the end user determine if the data are suitable for and accurate enough for an intended use or task and whether or not they cover the area and/or timeframe of interest and provide the desired information.

It is generally considered good practice to create metadata.

Some projects may require the creation of metadata, such as those funded by the federal government.

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



18

Metadata are important because they make your data more interpretable by the user.

They allow users to find out more about the data and minimize assumptions and guesses about the data characteristics and quality.

Most importantly, in my opinion, is that they ensure the long-term viability of your data. Someone in the distant future may want to use your data as part of the historic record. Without metadata, the data may not be useable or interpretable.

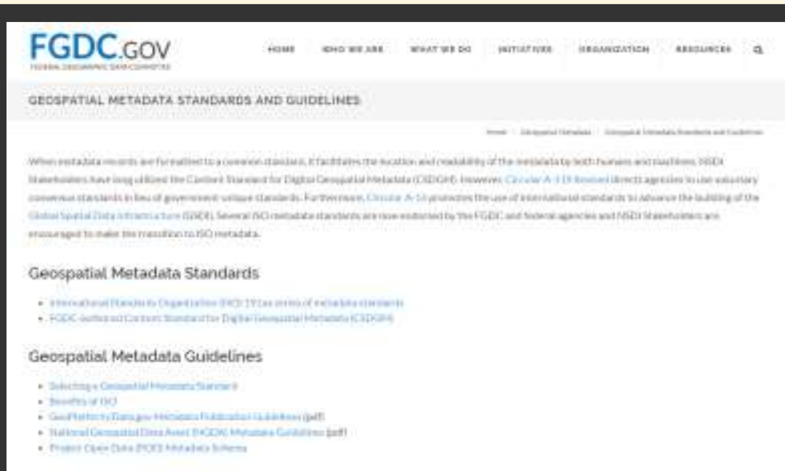
US: Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM)

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

19

The Federal Geographic Data Committee of the US federal government has established standards for geospatial metadata as the Content Standard for Digital Geospatial Metadata. This slide provides a list of the attributes that are required or commonly included.

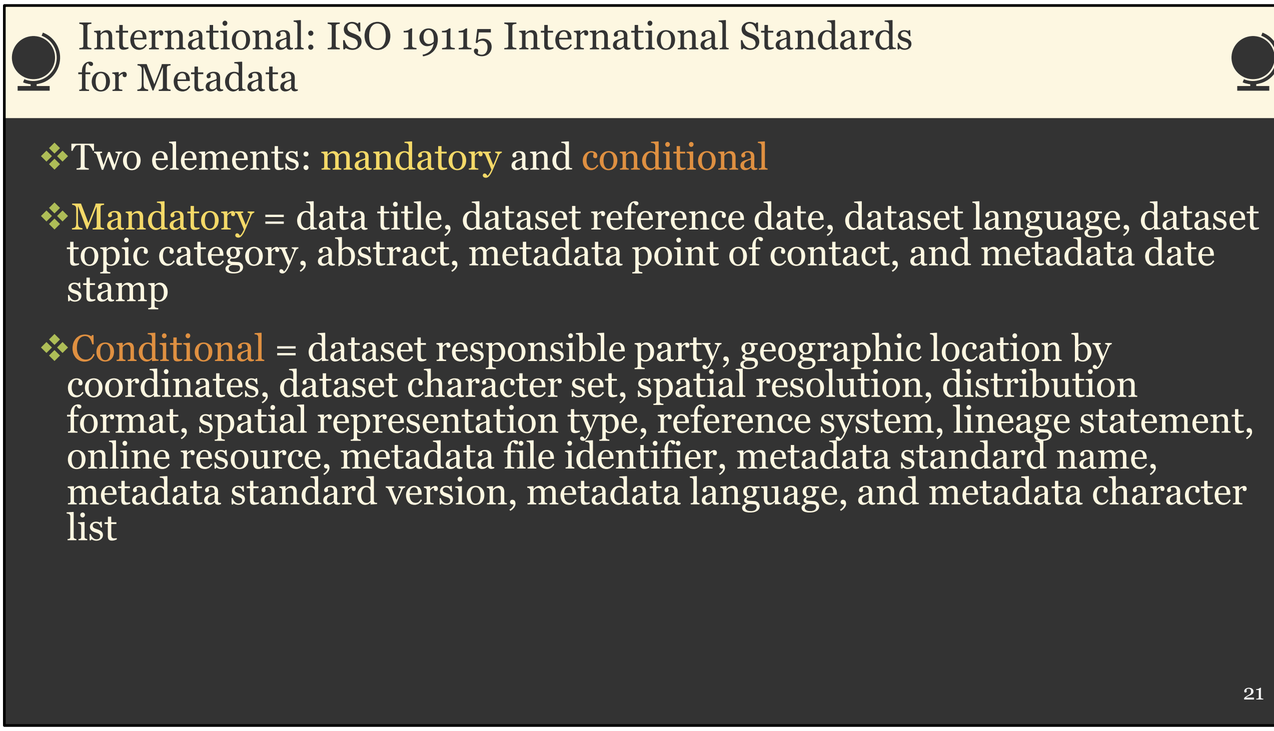
US: Federal Geographic Data Committee (FGDC) Content Standards for Digital Geospatial Metadata (CSDGM)



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

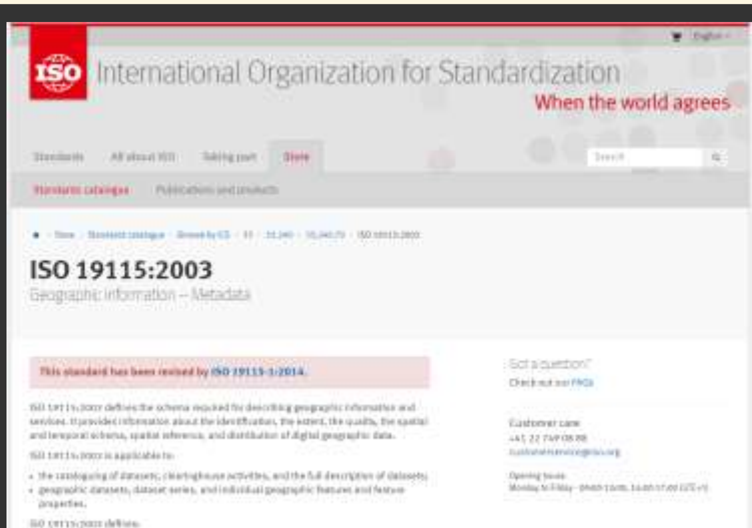
20

Here is a link to the FGDC website. As standards may change, this is a good site to reference.



International geospatial data standards have been established by the ISO 19115 International Standards for Metadata. This standard defines some mandatory and conditional elements as highlighted here.

International: ISO 19115 International Standards for Metadata



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

22

Here is a link to the ISO 19115 standards. As standards may change, this is a good site to reference.

XML: eXtensible Markup Language

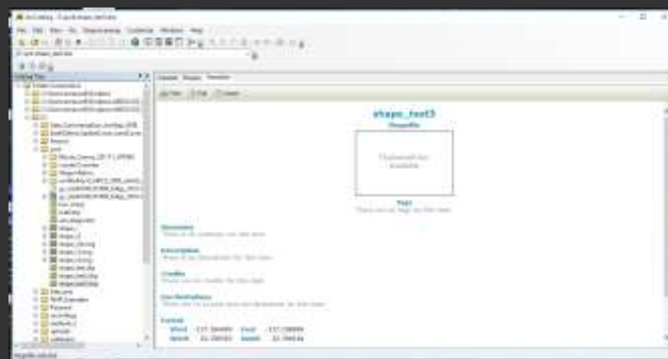
- Metadata can be written using XML
- A markup language like HTML
- Information wrapped in tags
- Can view in a web browser

23

Different data formats can be used to store metadata. One option is eXtensible Markup Language or XML. This format is similar to HTML as it is a tagged text format. XML files can be opened in any web browser or plain text editor.

Creating Metadata

- Metadata can be created in ArcGIS Pro using a variety of standards

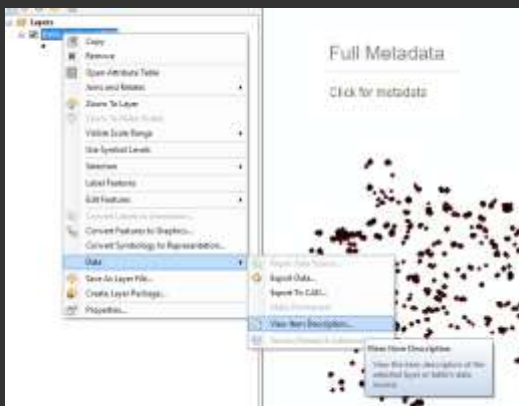


24

Using ESRI software, metadata can be created in ArcGIS Pro. You can define what standard you are using and the software will guide you through entering required and optional elements.

Viewing Metadata

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



25

Metadata can generally be viewed in ArcGIS Pro or using a web browser.

This is the end of this lecture module.

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

Thanks! Hope you found this useful.