Geodatabase Development Considerations: Schemas, Attributes, and Metadata

Data Quality and Schemas

Data Quality

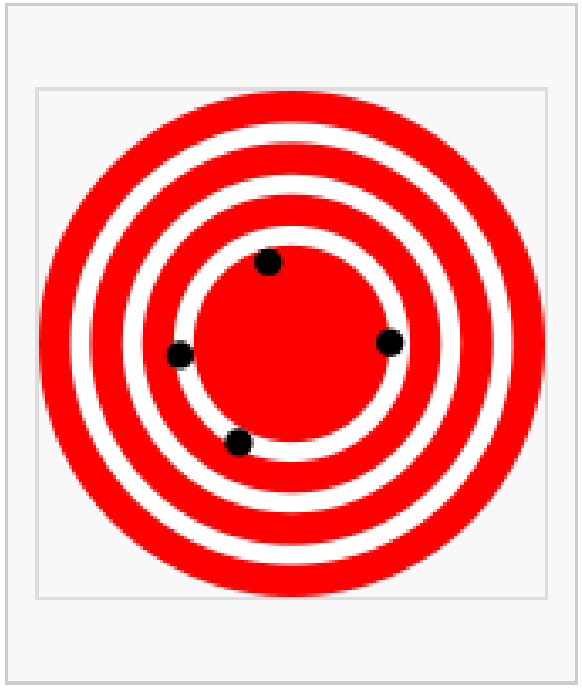
- When you create or acquire spatial data, you need to **verify its quality** before it can be used operationally.
- Two types of data quality assessment:
  - **Accuracy**: Ensuring your data correctly represent the specified location on the Earth's surface.
  - **Integrity**: Can encompass accuracy above but also includes data management and maintaining consistency during processing.

Accuracy and Precision

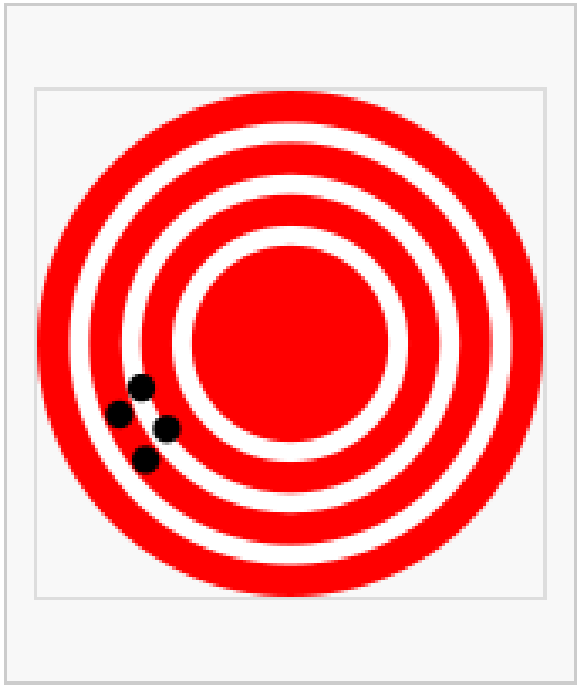
- Measures for assessing spatial data quality:
  - **Accuracy:** Agreement between a data point and its true location (reference value).
  - **Precision:** Number of significant figures stated in a value or the certainty with which a measurement is repeatable at a given location.
- **Margin of error** helps to quantify spatial error.
  - Determined by quality of digitization or field collection instruments.

Accuracy and Precision

Higher accuracy
Lower precision



Lower accuracy
Higher precision



Integrity

- **Data integrity** is a loosely defined concept focused on maintaining quality over the entire lifecycle of data.
  - From production to dissemination.
  - Ensuring data are not **corrupted** during processing.
  - Can also include protecting against inadvertent **disclosure**.
- Includes the design and maintenance of databases.
- ArcGIS includes a number of useful data integrity tools which we will explore in depth.
  - These tools help you manage both your attribute data and spatial data.

Data Schemas

- Depending on the context, a schema can be **conceptual** or **physical architecture**.
- A conceptual schema is much like a blueprint.
- **Database schemas** define the specific roles of each database object, including tables, fields, and relationships.
- Ideally, you should have a data management schema in place for your organization.
  - Helps you to visualize your datasets and their relationships.
  - Improves data quality by preventing repetition and desynchronization of attributes.

Developing a Data Schema

- In ArcGIS, your data schema will involve:
  - Identifying all of your geospatial datasets.
  - Listing and defining their attributes.
  - Establishing rules for the behavior of your geospatial and attribute data.
  - Determining the relationships between geospatial datasets and attribute tables.
- This schema also acts as your **data dictionary**.

Group Exercise

- Let's develop a hypothetical data schema for a census geodatabase.

- We need to list:
  - Feature classes (and feature datasets).
  - The attributes for each feature class.
  - Their appropriate behavior.
  - Their relationships.

Managing Attribute Quality in ArcGIS

Attribute Management

- The File Geodatabase includes a number of useful tools for managing your data attributes.
- These tools are meant to improve editing efficiency and also maintain data quality.
- For example, you can restrict a field to only a few specific values, such as “yes” or “no”.
- These **attribute management tools** are distinct from the **spatial data management tools**, which we will discuss later.
- We access these tools with **ArcCatalog**.

Subtypes

- **Subtypes:** Rules for categorizing distinct features in the same feature class.
- **Example:** Roads.
  - Normally subject to a national transportation classification system, so only a few possible values.
  - E.g., “Trunk Road”, “Primary Road”, “Secondary Road”, “Residential Street”.
- With subtypes, we can **automatically classify roads as we edit**, saving time and improving data quality.

Domains

- **Domains:** Constraints which limit attribute values to a numerical range or a list of possibilities.
- **Example 1:** Population value in a province.
  - A province cannot have a negative population value.
  - If we also know that no province has a population greater than 10,000,000, we could set our domain range to 0-10,000,000.
- **Example 2:** Enumeration area status.
  - During census operations, a field worker could indicate whether enumeration is complete in an EA.
  - We could set a simple yes/no domain value with no other possibilities.
- We will go over subtypes and domains in detail during the exercises.

Metadata

Metadata

- Metadata is "**data about data**".
  - Traditional example of metadata: a library catalog.
- Stores information which describes a file's purpose, contents, methodology, and proper use.
- **Critical for every dataset**, whether geospatial or not.
- Unfortunately metadata are often lacking.

Reasons for Using Metadata

- **Helps data users** understand how to use a specific dataset properly.
- Provides a **structured** and **consistent format** useful for cataloging and organizing.
- Acts as **institutional memory** for data managers who may not revisit a particular dataset for a long time.
- Improves **transparency** and the ability to **discover** new data sources.

Metadata Components

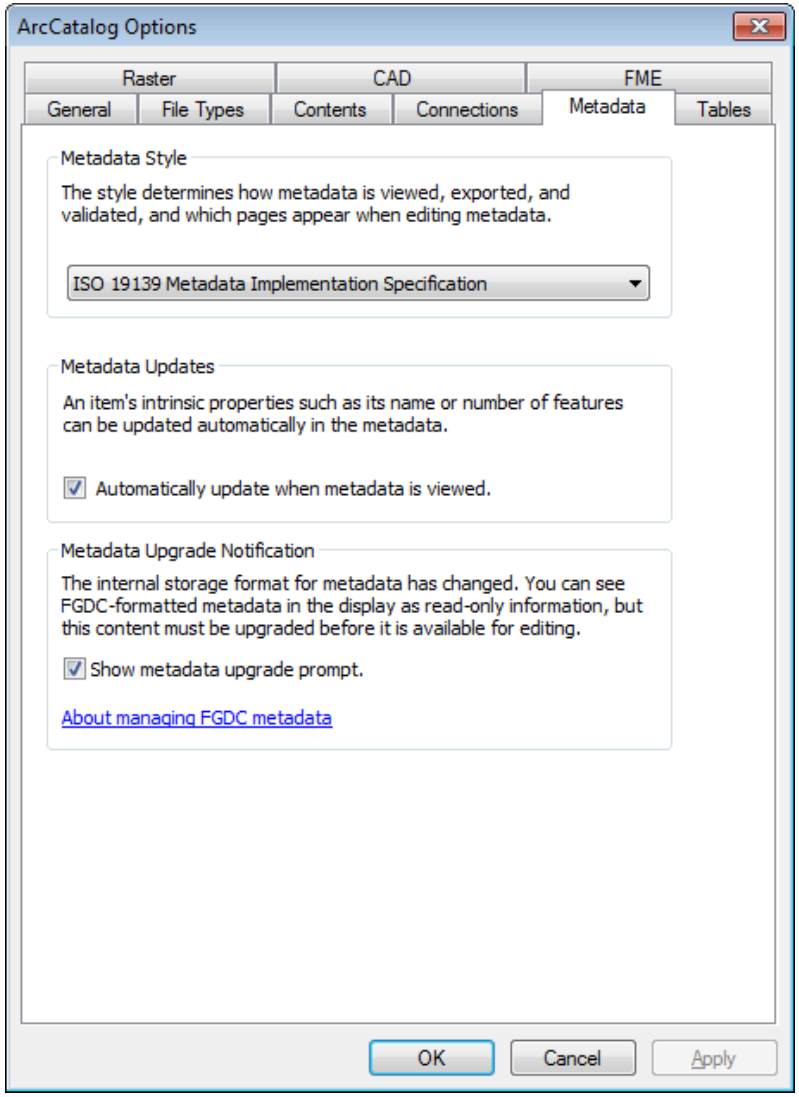
- **All metadata** commonly includes:
  - Technical description such as file format, field names and meanings, methodology, and instructions for proper use.
  - Source/authorship of data.
  - Contact information for questions.
- **Geospatial metadata** can also include:
  - Spatial error, if known.
  - Spatial extent.
  - Coordinate system used.

Metadata Standards

- Several **international organizations** establish standards for metadata.
- The most common metadata standard is maintained by the International Organization for Standardization: ISO 19115 and **ISO 19139**.
  - The latter is a specification for **XML**, which is also a standard file format we discussed previously.

Metadata in ArcGIS

- Metadata in ArcGIS is managed in **ArcCatalog**.
- By default, can store metadata in several formats, including ISO 19139.



Example of Raw XML Metadata

```
- <metadata>
  - <idinfo>
    - <citation>
      - <citeinfo>
        - <origin>
            U.S. Department of Commerce, U.S. Census Bureau, Geography Division
          </origin>
          <pubdate>2013</pubdate>
        - <title>
            TIGER/Line Shapefile, 2013, nation, U.S., Current county and Equivlaent National Shapefile
          </title>
          <edition>2013</edition>
          <geoform>vector digital data</geoform>
        - <onlink>
            http://www2.census.gov/geo/tiger/TIGER2013/COUNTY/tl_2013_us_county.zip
          </onlink>
        </citeinfo>
      </citation>
    - <descript>
      - <abstract>
          The TIGER/Line shapefiles and related database files (.dbf) are an extract of selected geographic and cartographic information
          from the U.S. Census Bureau's Master Address File / Topologically Integrated Geographic Encoding and Referencing
          (MAF/TIGER) Database (MTDB). The MTDB represents a seamless national file with no overlaps or gaps between parts,
          however, each TIGER/Line shapefile is designed to stand alone as an independent data set, or they can be combined to cover
```