Creation of an EA Geodatabase

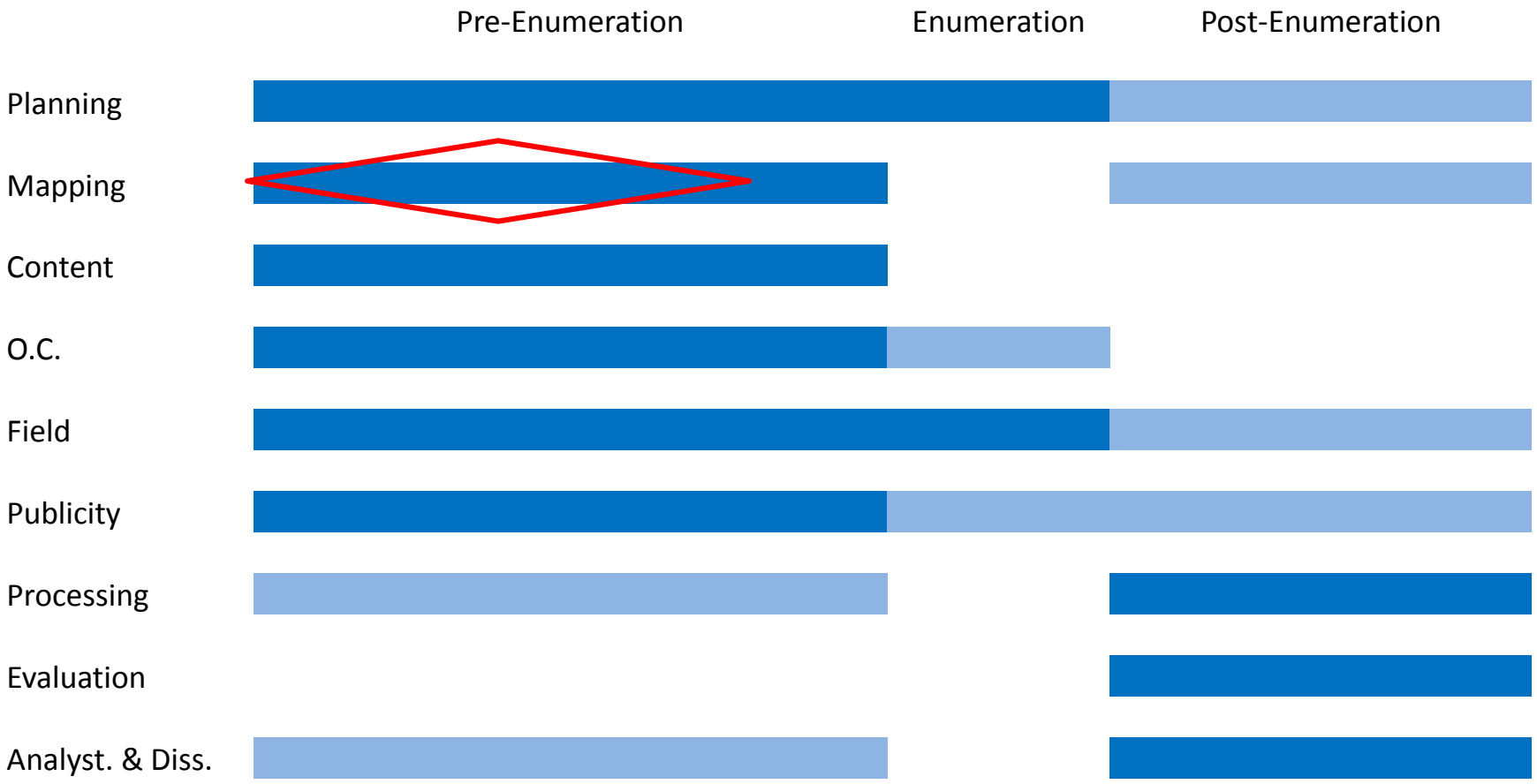
AITRS Workshop on Integration of Geospatial and Statistical Data

February 16 – 19, 2015

Topics

- Digitization
- Background on databases and geodatabases
- Geodatabases in the esri ecosystem
- Metadata
- Conclusion

When do things happen?



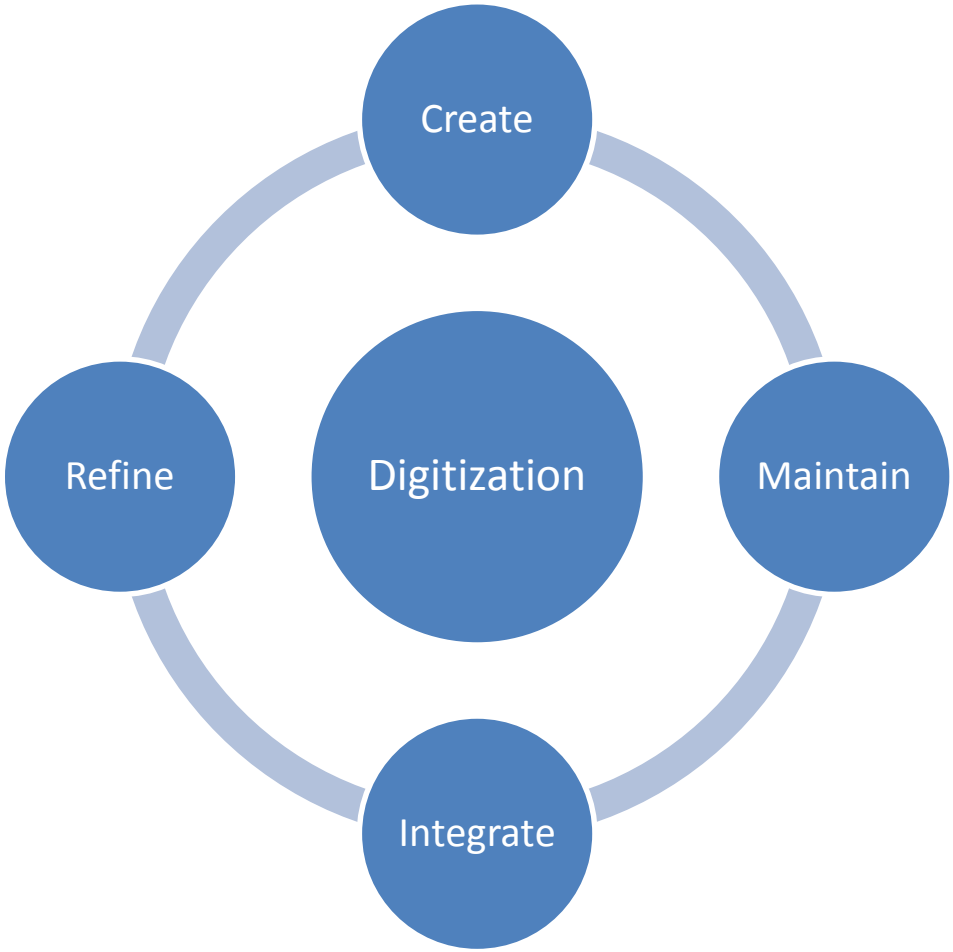
Digitization

- Digitization is not exciting, but is the most important step towards modernized mapping for an NSO.
- Relatively long amount of time required, little payoff until complete.
- Little to no automation possible, difficult to get rest of statistical office “excited” about digitization.

Digitization and Digital Data Maintenance

- May have to champion digitization of administrative and statistical geography.
- Explain the many subsequent geographic activities dependent on digitization.
- Emphasize digitization as an investment.

Digitization Cycle



Digitization Cycle

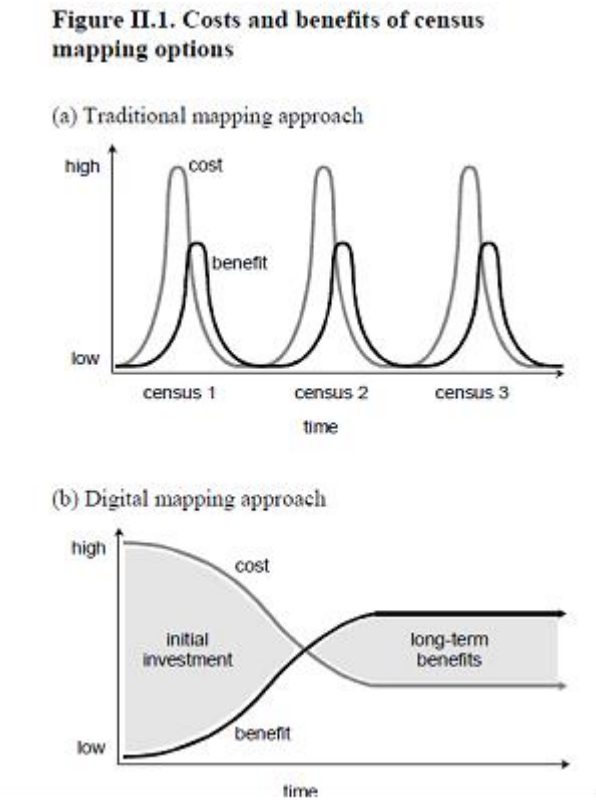


Image source: United Nations. Statistical Division. 2008. *Principles and recommendations for population and housing censuses*

- In the longer term a commitment to digitization will reduce costs.
- Reduces the amount work repeated each census cycle.
- Digitization skills form an important base for more advanced GIS.

Digitization Cycle

Create

Make an entirely new digital layer based on scanned sources or related digital layers.

- This is the most time consuming activity.
- Requires well-organized input data and a plan on how to use those data.

Digitization Cycle

Maintain

Make significant changes to line segments as part of an ongoing program to keep geography up-to-date.

- Ongoing changes necessary due to changes in geographical boundaries or population growth.
- Requires planning the frequency of updates and determination of authoritative data sources for updates to census boundary database.

Digitization Cycle

Integrate

Enforce correspondence between administrative and statistical geography.

- The hierarchy from nation down to enumeration block should be nested.
- Nested geography should correspond topologically. May require work to retroactively correct errors.

Digitization Cycle

Refine

Improve boundary sharing with other physical features. Create new generalization levels.

- Ongoing improvements to the appearance of cartographic products.
- Automated tools exist for generalization but cause topological errors that will require re-digitization of some features and clean-up.

Digitizers

Can range from geographers with academic credentials to draftsmen to other functional specialists brought in from different parts of organization.

- The quality of training and a comprehensive plan will determine success of digitization project.
- Training must teach participants technical skills (button-pushing) as well as problem solving.

Geographic Data Connectivity

- Anything that has a location can be displayed and may be useful when digitizing.
- However, this does not mean that every dataset will be useful when digitizing.
- Consider:
  - Relative position of features
  - Interpretability of shape and size at scale

Base or Precursor Datasets

- Which precursor datasets are necessary or useful when digitizing collection geography?
  - Streets
  - Drainage
  - Parcels
  - Village extents
  - Incorporated or developed areas
  - Landcover
  - Elevation

Data Sources

Source:	Global Landcover Facility	Shuttle Radar Topography Mission (SRTM)	UN Second Level Administrative Boundaries Working Group
Who?	Remote sensing center. Access to land cover products and imagery for local to global areas affiliated with the University of Maryland.	NASA elevation data on a near-global scale to generate the most complete high-resolution digital topographic database of Earth.	Part of the larger Geographic Information Working Group at the United Nations.
What?	An archive of medium and low resolution imagery that may be useful in rural areas. Landcover can also be used to roughly delineate urban.	Provides a digital elevation model (DEM) that can be used to identify topographic features suitable for collection geography digitization.	Can provide boundaries for neighboring countries for use in cartographic products.

Working with Other Agencies

- Establishing relationships with other agencies is often key to a successful digitization project.
- Determine needs early and set out to meet them.
- Early contact with an agency will reduce pressure.

Background on Databases

File and Folder System

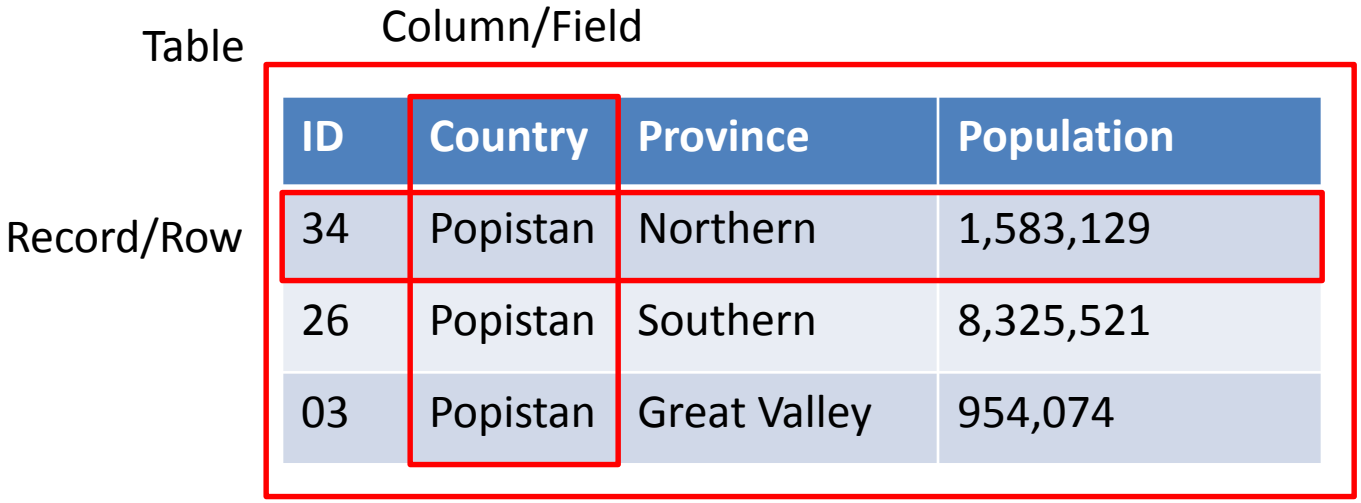
- A **storage system** which uses the **default file and folder structure** found in operating systems.
- Uses the **non-DB formats** we mentioned previously (shapefiles, text/Excel files).
- Data stored on individual computers or shared over a local network.

Database (DB)

- A **storage system** designed to manage large datasets efficiently.
- Users can **query** and **manipulate** data using **joins**, **relates**, and a **Structured Query Language (SQL)**.
- A database can exist on your computer, on a private network (such as your office), or on a server connected to the Internet.

Database Fundamentals

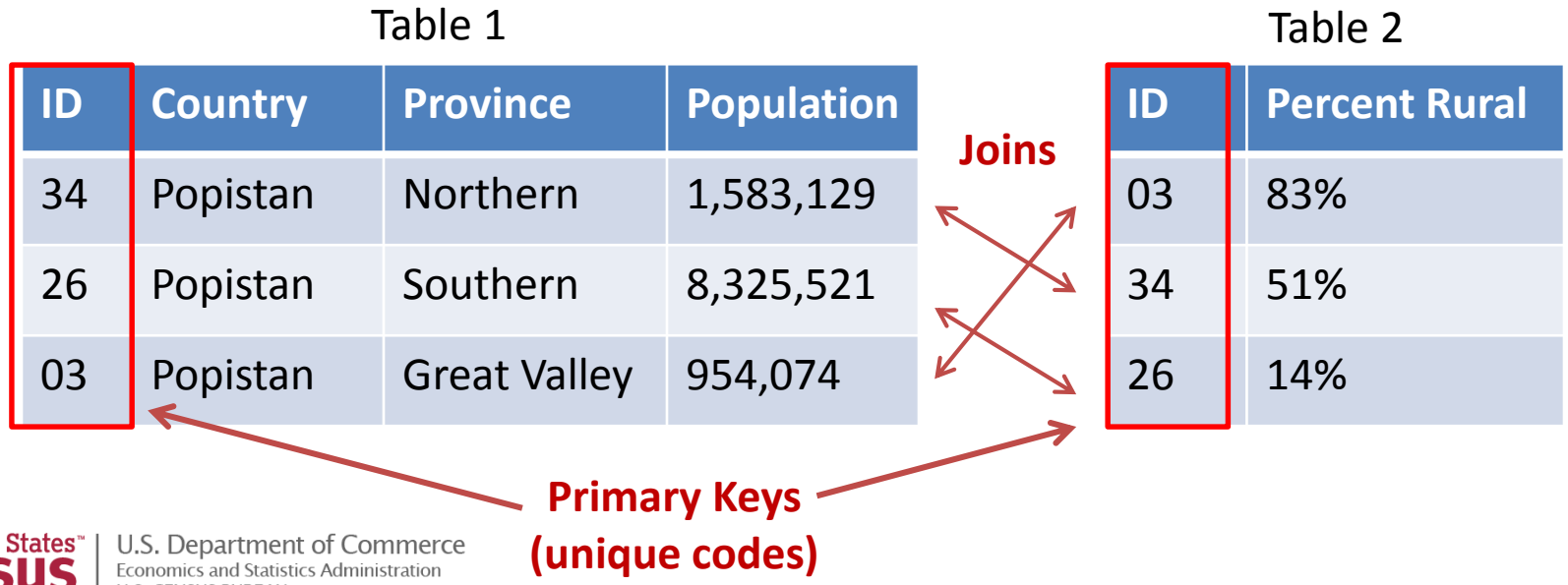
- Databases are a collection of **tables**.
- Each table contains **columns (fields)** and **rows (records)**.



ID	Country	Province	Population
34	Popistan	Northern	1,583,129
26	Popistan	Southern	8,325,521
03	Popistan	Great Valley	954,074

Database Fundamentals

- Tables can be **joined** to each other using a unique identifier or code (a **primary key**).
- It is good practice to assign a primary key to every unit of census geography (including administrative and statistical).
  - Use an **alphanumeric code**.



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 synchronization 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**.

Database Management System (DBMS)

- Software designed to efficiently **administer** one or more database(s).
- In action, users rarely distinguish between a "DBMS" and "database".
- Examples you may be familiar with include **Microsoft Access** and **ArcGIS**.
  - Note: A Microsoft Excel spreadsheet is not a database, though it does share many features.
- You may also be familiar with more advanced DBMS software, such as **PostgreSQL**, **MySQL**, and **Microsoft SQL Server**.

Single User Databases

- Single user means **only one user at a time** can access the database.
  - There may be minimal support for multiple users.
- Useful for managing data for small projects with few participants.
- These databases are generally stored **locally** (i.e., on your computer) or on a **shared drive** for minimal collaboration.
- Both **Microsoft Access** and **ArcGIS** can create single user databases.

Multi-User Databases

- Nearly always a **DBMS** such as PostgreSQL, Microsoft SQL Server, or MySQL.
- Designed to handle **multiple users** retrieving from and updating the **same database simultaneously**.
- Often just called an **enterprise database**.
  - "Enterprise" refers to an office or organization.

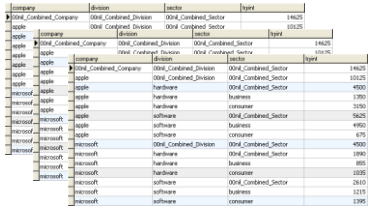
Multi-User Databases

- Managing the edits of multiple users at once is called **versioning** or **deconflicting**.
  - Versioning stores a record of every user's transaction.
  - Thus, each database edit is reversible.
- The databases are **stored remotely** and accessed over a private network or the Internet.
- **Data processing** can either occur on the network or on your local computer.

Diagram of a Multi-User Database

Database Administrator (DBA)

Data Tables



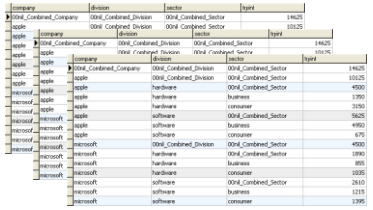
User 1

User 2

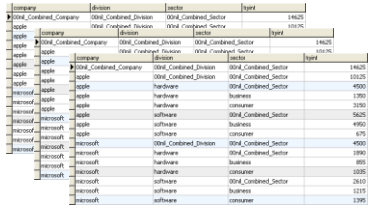
User 3

Database Management System (DBMS)

Database 1



Database 2



Database 3

Servers hosting DBMS and DBs

Spatial Databases

- Also called geospatial databases, geographic databases, or geodatabases.
- Possess all of the same features of other databases, plus the ability to **store location data**.

Database Advantages

- The file/folder system is easier for **quick projects**.
- However, storing geographic data in a database provides numerous **advantages** over the file/folder system:
  - Data are stored more efficiently.
  - Can separate geographic features and attribute data.
  - Larger datasets are easier to access and manage.
  - Specific data can be retrieved using queries.
  - The quality of geographic features can be managed more effectively with **topology**.

Realities

- Databases require extra **knowledge** of administration, querying, managing joins/relates, etc.
- A multi-user enterprise database, while advantageous, requires **expensive hardware**, persistent network **connectivity**, and **highly skilled IT support**.
- May not be feasible for all organizations.

Other Options for geo-enabled RDBMs

Options have increased dramatically, here are some of the major players:

Proprietary

- Oracle Spatial
- Microsoft SQL (post 2008)

Open Source

- PostGIS/Postgre
- SpatialLite

Geodatabases in ArcGIS

Geodatabases in ArcGIS

- As we discussed previously:
  - Geographic data can be stored in many different **formats**.
  - Data can also be stored using standalone **files and folders** or a **database**.
- ArcGIS includes functionality to create **geodatabases** for storing, editing, and managing your data.
- These geodatabases include a number of useful tools which we will explore in detail.
  - Note: the terms "**geodatabase**", "**spatial database**", and "**database**" are used interchangeably.

Fundamental Concepts

- The preferred format for storing data in the ArcGIS environment is the **File Geodatabase**.
  - Another format, the **Personal Geodatabase**, is obsolete.
- You can store all types of data in the file geodatabase, including **vector**, **raster**, and **non-spatial tables**.

Important Caveats

- The file geodatabase is designed as a **single user database**.
  - Very limited multi-user support.
- The file geodatabase is a **proprietary format**.
  - Not easily **compatible** with other GIS software.

File Geodatabase vs. Shapefile

Feature	Shapefile	File Geodatabase
File size limits	2GB ✗	Unlimited (TBs) ✓
Storage efficiency	Less efficient ✗	More efficient ✓
Performance	Slower ✗	Faster ✓
Raster support	No ✗	Yes ✓
Enforcing consistency within and between files	No ✗	Yes (topology, schemas, projections) ✓
Compatibility	Most GIS software ✓	ArcGIS Only ✗
Portability	"Zip and ship" ✓	"Zip and ship" ✓
Multi-User	No ✗	Limited ⚠

Structure of File Geodatabases

- Several file types can exist in the file geodatabase:
  - **Feature dataset:** A geospatial "container" which stores projection information and topology for vector data.
  - **Feature class:** Geospatial point, line, or polygon (vector) data.
  - **Non-spatial table:** A set of attributes which are commonly linked to a feature class.
  - **Raster:** Sits independently within a file geodatabase and cannot be stored in a feature dataset.
  - **Topology file:** Stores the rules which enforce data integrity within the database.
  - **Relationship file:** Creates a join between multiple feature classes and/or non-spatial tables.
  - **Others:** Raster mosaic/catalog, schematic dataset, toolbox, parcel fabric, annotation, network, terrain.
- We will work with feature datasets, feature classes, and tables.

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**
- 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.
  - Can set a simple yes/no domain value with no other possibilities.

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

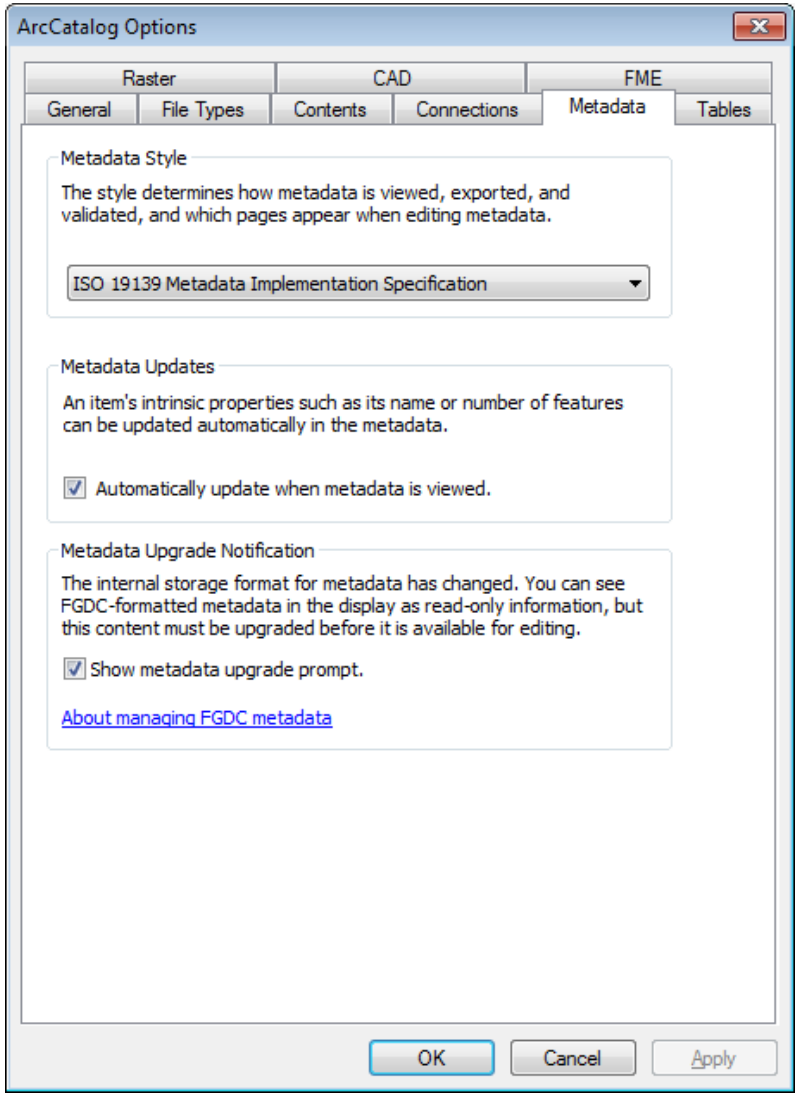
- **All metadata** commonly includes:
  - Technical description such as file format, field names and meanings, methodology, and instructions for proper use.
  - Source/authorship of data.
  - Contact information for questions.
- **Geospatial metadata** can also include:
  - Spatial error, if known.
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

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
```

Good Practices: In General

- Keep spatial data and non-critical attributes **separate**.
  - Exception: attributes critical to the geographic definition of the features in the dataset (e.g., place names, identification codes).
- Use feature datasets to store related geography.
  - Can be structured many different ways.
  - E.g., province-by-province, grouped by type of feature.

Good Practices: Data Preservation

- Maintain **separate** databases for **production** files and **working** files.
  - E.g., one database of data approved for field staff to use and another for data being edited by head office staff.
  - Data from the working DB feeds into the production DB.
  - If you want to experiment, **export** your feature data to a **scratch database** or a **shapefile**.
  - **Never edit production data directly!**
- Produce daily or weekly **backups** of your database and store in a safe place.
  - E.g., a removable hard drive locked in a storage room.

Conclusion

- The creation of a spatially and topologically integrated geodatabase is a significant undertaking
- Digitization is an important task but design and maintenance of database equally as important
- Metadata is critical to working with other organizations and making your data useful for the global community – but is often overlooked