Training course: GIS Data Preparation and Management
GIS Data Preparation and Management

Presentation outline:

- Data quality – the critical element in GIS
- Data acquisition and inspection
- Good data management practice:
  - Data storage, documentation, preparation and cleaning, dissemination, archival
- Choosing spatial data format
The Power of Data – the critical element in GIS

The combination of GIS software with modern computing capabilities holds enormous potential for analysing and understanding the world around us...

**BUT** it all depends on the data!

Remember... ‘G.I.G.O.’ – “Garbage in, garbage out!”

- ‘Dirty Data’ – significant volumes of data are discarded on initial inspection, because they are in some way incomplete or inconsistent
  - Globally, on average, companies estimate 26% of their data to be ‘dirty’
  - Human error is considered as the dominant cause in over 60% of cases
  - Key factors include poor internal communications and protocols, lack of training, inexperience with data collection

*Source: Experian Data Quality research*
Datasets may exist... **BUT** are they suitable for use?

Information may be:

- Out of date
- Incomplete
- Spatially incorrect
- Factually incorrect
- Generated at a scale that is not appropriate to your study

It is important to recognise that a GIS is not a “miracle machine.” You should **critically assess any data** that is being considered for use.
Acquiring GIS data: inspect, transform, integrate

Primary data (e.g. survey)

Tabular data (Database tables, CSVs etc.)

Non-digital data (paper maps)

Inspection
Transformation
Integration

Consistent, standardised GIS data for analysis and visualisation
Data quality: prepare and ‘clean’ input data

Before undertaking any spatial analysis, it is critical to make sure that your data is “clean”

Clean means:

- Data is fully attributed
- Attributes are consistent
- No obvious spatial referencing errors
- Metadata exists, recording all known relevant information on the data

<table>
<thead>
<tr>
<th>wardname</th>
<th>wardcode</th>
<th>Ignaname</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tsamiya Babba</td>
<td>KN1708</td>
<td>NULL</td>
</tr>
<tr>
<td>NULL</td>
<td>KN1708</td>
<td>NULL</td>
</tr>
<tr>
<td>Naibawa</td>
<td>KN2510</td>
<td>NULL</td>
</tr>
<tr>
<td>Zogarawa</td>
<td>NULL</td>
<td>Dawakin Kudu</td>
</tr>
<tr>
<td>Yaraya</td>
<td>KN0914</td>
<td>Dawakin Kudu</td>
</tr>
</tbody>
</table>

- 30/04/171
- 30/4/1971
- 30thApril71
- 30:4:71
- ?th April 17
- Apr-71-30

Missing values or misspelt words

Inconsistent values make data queries impossible

Point locations noticeably outside study area

Poor metadata causes misunderstanding
Good practice in GIS data management

Why establish **standard procedures** within your teams to regulate data acquisition, QA, documentation, storage and archival?

- Increase awareness and use of relevant datasets
- A method of catching and eliminating data errors, as early as possible
- Establish an audit trail of how and when data are used in a project
- It is the **key to working effectively** in GIS, across teams and wider partners

"Data is a precious thing and will last longer than the systems themselves"

Tim Berners-Lee

"Data that sit unused are no different from data that were never collected in the first place"

Doug Fisher
The GRID3 GIS data cycle

Standardise your data management procedures

- Acquire data
  - Save data
    - Save a copy of original data
    - Consider folder structure **
- Metadata
  - Start or update metadata record
- QA data
  - Clean or reformat data, as necessary
- Update Metadata
- 'Save as..' intermediate data file
- 'Save as..' final data file
- Geoprocessing and analysis
- QA results
- Update Metadata
- Data mapping and visualisation
- Archive project and data **

- File and folder management
- Data documentation
- Working on the data

- Individuals and teams should follow **consistent procedures at each stage** of the cycle
- Management should guidance, documentation and training to all data users
- **Think about the long-term when creating folder structures and archiving projects – to ensure these resources are visible and accessible to colleagues and future users**
Good practice: Saving and storing your data

GIS project folder structure:

- Not one-size-fits-all! Customise the structure according to your team/Project requirements
- Once the structure and approach is agreed, it should be adopted and maintained by all team members
- Always consider the long-term when planning folder structures and data protocols – ensure visibility, access and reuse of the data over time

Example of a typical folder structure (used in GRID3 projects)
Good practice: Saving and storing your data

GIS file naming considerations:

- Avoid spaces, periods, hyphens, parentheses, brackets and other special characters, e.g. $, %, @, etc.
- Use acronyms sparingly
- Avoid using reserved database keywords*
- Filenames should be **concise** and **informative**
- If separating words, use underscore (one_two), or 'camelCase'

GRID3 file naming conventions

**Database reserved keywords**

* If using the following spatial databases, you should avoid particular words in your filenames which relate to specific functions in the database: [GeoPackage (SQLite Database)](https://geopackage.org) | [ESRI Geodatabase](https://desktop.arcgis.com/en/arcmap/latest/newproject/provide-naming-guidelines.html)

Note: the restriction also applies to the naming of column headers!
Good practice: Documenting your data (Metadata)

What is the value of metadata?

▪ Vital information about data and how they were collected
  ▪ Method for reporting known limitations of data, i.e.
    ▪ Data currency (when data was generated)
    ▪ Accuracy
    ▪ Completeness
    ▪ Error

▪ Data provenance
  ▪ Provides an audit trail of collection, reformatting & analysis processes applied to the data

Metadata provides a basis for sound decision making!

An example metadata record

| Date created   | 7 February 2024 |
| Date updated   | 7 March 2024   |
| Created by     | Ms A. Learner, Junior data scientist, GRID3 |
| Details        | Health facility data for 8 LGAs in Kano state, Nigeria. Data collected between 6 January and 10 March 2024. |
| Edits made     | Health facility categories updated 2.1 |
| Version        | 2.1           |
| Data source    | Collected by LGA survey teams during 2024 measles campaign |
| CRS            | WGS 1984 UTM Zone 30 |
| Terms of use   | For open external use |
| Known errors   | Data expected from 10 LGAs; received only 8 GPS and ODK forms were used during data collection. Metadata tab created by A.Learner on 31 March 2024 |
| Additional notes |  |
Good practice: Data cleaning for GIS

Non-spatial data can be ‘cleaned’ using a range of software applications

- An example of a common problem – source data organised by column (often exported from a content management system)
- GIS import requires items to be organised by row
- Data must be transposed (in Excel or equivalent)
Good practice: Data cleaning for GIS

Considerations for cleaning non-spatial data:

- The following are not supported in GIS: merged cells, titles, captions
- Field headers should contain no more than 10 characters and no unusual characters (e.g. &, %, £, etc.)
- Investigate duplicated or missing rows
- Remove blank or redundant rows/columns
- Cell values:
  - What is the intended data type of each column? Text? Numeric? Integer? Date?
  - Are the cell characters consistent with the data type?
    - 1000 or 1000?
    - 13th Feb 24 or 13/02/2024?
- Remove trailing- and double-spaces
- Consistent capitalisation?
Good practice: Data cleaning for GIS

Considerations for cleaning spatial data:

- Coordinate system/map projection – is your GIS project set to the same coordinate system as used by the data capture device?
- Do your point locations fall within expected administrative boundaries, or settlement extents?
- Missing attributes? Can you use the location of a feature to fill in missing information?
- Search for duplicate locations using geoprocessing tools
Choosing spatial data format

What are most commonly used formats?

geopackage.org
esri.com
Choosing spatial data format: Shapefile (SHP)

Advantages:

- Universally recognised
- Simple structure, easily shared, good for newcomers

Disadvantages:

- Doesn’t handle large data volumes very well – 2GB limit!
- Cumbersome file management:
  - It’s not a single file, but a collection of components files
  - Metadata must be stored in a separate file (.txt, .xls, etc.)
- One shapefile holds just one geometry type – point, line or polygon
- Limited for international/multilingual data (i.e. non-ASCII character sets)
Choosing spatial data format: **GeoPackage (GPKG)**

**Advantages:**
- Everything is contained in a **single file**, containing multiple spatial datasets
- Suitable for large-scale projects and can hold **massive data volumes**
  - Efficient and quick loading, rendering, planning and zooming
- GeoPackage **supports raster and vector** data seamlessly, plus tile data
- GeoPackage provides **full metadata integration**
- Broad compatibility (ArcGIS, GDAL, QGIS, R, Python)
- **Handles international and multilingual data** (Unicode character encoding)

**Disadvantages:**
- More involved, a steeper learning curve for new users
- Potential compatibility issues with old software

---

*The structure of a GeoPackage; Single file containing multiple spatial datasets*
Choosing spatial data format: **File Geodatabase (FGDB)**

*Note: ESRI File Geodatabase was developed for use in ArcGIS software applications*

Considerations for the QGIS user:

- If you are working solely in QGIS, you should adopt GeoPackage
- However, some teams contain both QGIS users and ArcGIS users!
- Recent QGIS installations come with the `openfilegdb` driver, enabling a level of access and use of FGDB; **note the following:**

**YOU CAN:**

- Read and write to an **existing FGDB**
- Export data from QGIS and create a **new FGDB** to hold the data

**YOU CANNOT:**

- Export your data as a new layer into an existing GDB!
GIS Data Preparation and Management

Wrap up and summary:

- Data is central to working effectively in GIS
- **Investigate** data thoroughly and **critically assess** its suitability and limitations for a given project
- Implement or follow **agreed protocols** within your teams, at all stages of the data cycle
- Consider alternative spatial data formats
- Think about the **long-term** in your data management strategy!
GIS Data Preparation and Management

Now post your questions and comments in the course discussion forum!