# GIS DATA PREPARATION AND MANAGEMENT

Practical Task 1: Initial data setup

Task description

This is the first of four practical tasks that collectively constitute a "data preparation and cleaning" workflow. The objective of this task is to understand best practices for the initial setup of your data to facilitate effective review and allow for any necessary corrections.
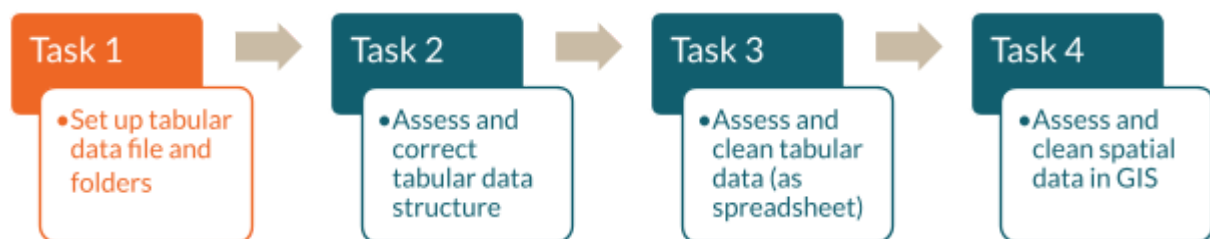
Once you have downloaded, or received, a raw data file, the first step is to a) save a copy of the data in a readily accessible, fixed directory on your computer and b) adapt a consistent and standardized naming convention. This will help other users (and yourself in the future) understand the version of each data file. It is also very important to check/create metadata to describe the data.

Data rules covered in this task:

▪ Always **save a copy** of the raw data before making any edits

▪ Any **updates or changes** to a data file should be recorded in the **metadata**

▪ If a data file comes to you with **no metadata**, you should **create** metadata with as much detail as possible

Part 1: Initial file management

For this course, you have been provided with an existing folder structure, which was determined and developed by your data team to ensure everyone working with the data knows where to find the information they need.

1_SourceData

This folder serves as the point of entry for all data and information related to the project. Data files are placed here before any processing has been done. NO edits should be made to files in this folder.

2_RawData

Before making any edits to the data file, make a **copy** of the original file from the 1_SourceData folder and save it in this 2_RawData folder.

3_IntermediateData

Any edited versions to the data files from the 2_RawData folder should be saved in this 3_IntermediateData folder.

4_FinalData

Once the spreadsheet is ready to be used elsewhere, it can be saved in this folder with a version number.

5_Outputs

This folder is for files created from the spreadsheets, such as maps, graphs, and statistical reports.
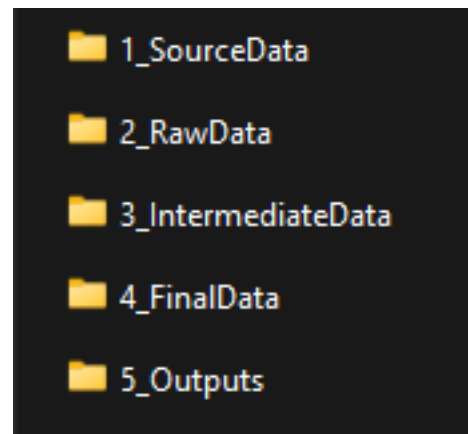
File management example

Step 1: Open the following spreadsheet:

…\GRID3DataMng\Data\11_SourceData\NGA-HFs Kano LGAs sample.xlsx

This spreadsheet contains information on health facilities in eight local government areas (LGAs) in Kano State, Nigeria. The data have been collated from various sources and require cleaning before the data points can be mapped in QGIS.

Note: The data used for this course have been amended for training purposes.

Step 2: Immediately before proceeding, save a copy of the original file in the *2_RawData* folder. It is very important that <u>NO</u> edits are made to the original file.

- ▪ **File > Save As**
  - o Navigate to the course data folder on your computer

- ▪ **GRID3DataMng > Data > 2_RawData**
  - o Save a copy of the data as an Excel file with the following name (change the date to reflect today's date):

    Kano_HealthFacilities_31Mar2024_raw.xlsx

File naming tips
Keep file names concise and informative. Key information such as location (Kano), data type (health facilities), date, and version should always be included. Adopting a naming convention within your organization or team will promote effective data sharing and management.

When preparing data for GIS use, some geoprocessing tools or platforms will not work if the file names do not follow the correct format. Do not use spaces, hyphens, parentheses, brackets, or other special characters (such as $, %, and #). Using underscores (_) instead of spaces solves this problem. This rule can also apply to the entire file path; folder names should not contain spaces or special characters, either.

Examples of INCORRECT file names

HealthFacilities_31032024
- ✗ Geographic extent of data not specified
- ✗ The date format (DDMMYYYY) may not be clear to all data users

NGA_Kano_HF data
- ✗ Not clear what "HF" means
- ✗ Space used between "HF" and "data"
- ✗ No date indicated

Kano.Helth_Facilities-31March2024
- ✗ Spelling mistake ("Health" not "Helth")
- ✗ Full stop (.) used may cause issues for geoprocessing tools in GIS
- ✗ Hyphens(-) used may also cause problems in some programs

Example of CORRECT file name

Kano_HealthFacilities_31Mar2024_raw
- ✓ Correct spelling
- ✓ Location and date indicated in the correct format
- ✓ No use of special characters

Part 2: Creating or updating metadata

Metadata are data about data. It is crucial to maintain clear, informative, and up-to-date metadata about all the data you work with. As well as providing crucial information about how and when the data was collected, metadata help identify known limitations of the data by recording known errors and accuracy issues. Complete and up-to-date metadata provide an audit trail of the data collection, processing, and analysis journey.

The raw data file currently has no recorded metadata, so you need to **create** the information from scratch. Create a new tab and call it *Metadata*.
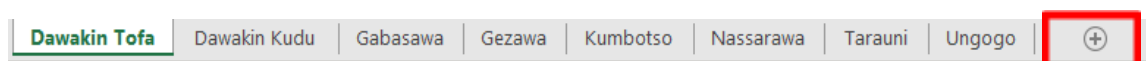
Create the following metadata headers in column A (you do not need to include the text in brackets):

- **Date created** (the date when this file was created)
- **Created by** (include name, job title, organization, and contact details)
- **Details** (information on the data included in the file)
- **Date updated** (date of any data updates or edits)
- **Edits made** (details of any edits made to the data)
- **Version** (include a version of the data file)
- **Data source** (indicates where the data come from; include details of primary data collection or a link to where the data were downloaded)
- **CRS** (indicates whether the data already have a known coordinate reference system)
- **Terms of use** (any licensing or data sharing restrictions)
- **Known errors** (include any known errors in the data)
- **Additional notes** (add any other information about the data that may be useful for others to know)

Fill out the metadata tab as thoroughly as possible using your existing knowledge of the data and the information on the note below, which was provided by a colleague.
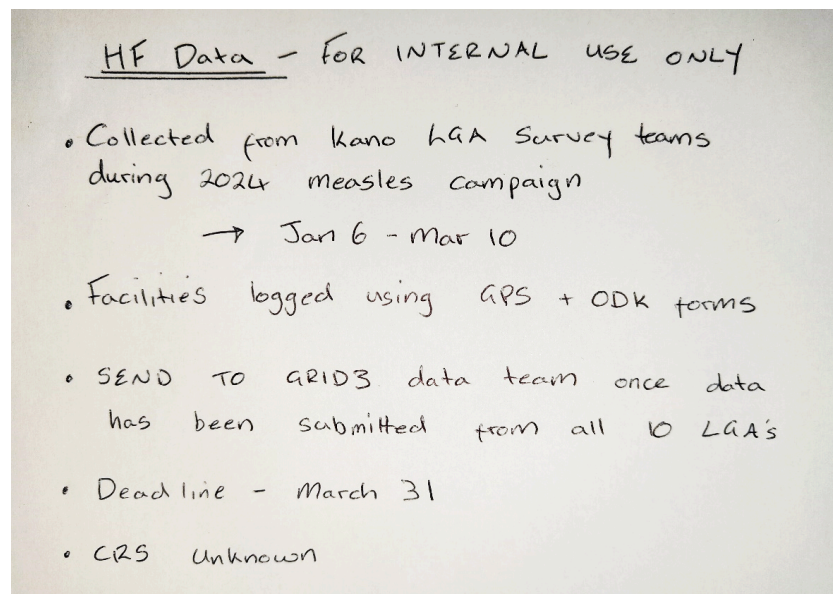
Ensure you have included today's date and your name (and email address, if applicable). So far, no edits have been made to the data; those fields can be left blank (or use N/A).

The metadata tab should now look similar to the table below:

Date created	31 March 2024
Created by	A. Learner, Junior data scientist, GRID3
Details	Health facility data for 8 LGAs in Kano State, Nigeria (Dawakin Tofa, Dawakin Kuda, Gabasawa, Gezawa, Kumbotso, Nassarawa, Tarauni, Ungogo). Data collected between 6 January and 10 March 2024.
Date updated	N/A
Edits made	N/A
Version	1
Data source	Collected by LGA survey teams during 2024 measles campaign
CRS	Unknown
Terms of use	For internal use only
Known errors	Data expected from 10 LGAs but only received for 8
Additional notes	GPS and ODK forms were used during data collection. Metadata tab created by A. Learner on 31 March 2024

It is also important to add a data dictionary to your metadata tab so that other users can easily understand the data fields. Add a list of data field headers and their descriptions to your metadata tab below the information you have already entered.

Data dictionary	
uniq_id	Identification number of each health facility
timestamp	Date of data collection
latitude	Latitude of health facility
longitude	Longitude of health facility
wardname	Name of ward the health facility is in
wardcode	Code of ward the health facility is in
lganame	Name of local government area (LGA) the health facility is in
lgacode	Code of local government area (LGA) the health facility is in
statename	Name of state the health facility is in
statecode	Code of state the health facility is in
updated_on	Date that the data were last updated
func_stats	Status of the functional statistics
category	Health facility category
ownership	Health facility ownership (public or private)
type	Health facility type
source	Organization in charge of collecting the data
prmry_name	Health facility name

Now that you have processed the raw data file and created its metadata, you are ready to move on to Task 2 (in which you will start to look at the contents of the data file). Ensure you have saved your file in the *2_RawData* folder before moving on to the next task!

Learning checklist

By following the above directions correctly, you will have demonstrated good practice in data preparation and management for GIS by having:

- Saved a raw data file in a readily accessible, fixed directory with a consistent naming convention
- Navigated a pre-defined data management folder structure
- Created detailed metadata via previous knowledge of the data

In the next task, you will create an intermediate-level data file after addressing any anomalies with the table structure that would generate errors when importing the data into GIS.