GRID3

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

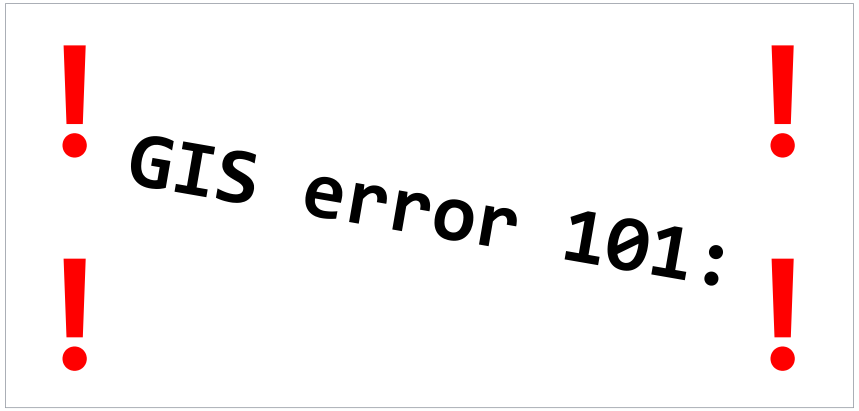
The Power of Data – the critical element in GIS

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

The Power of Data – the critical element in GIS

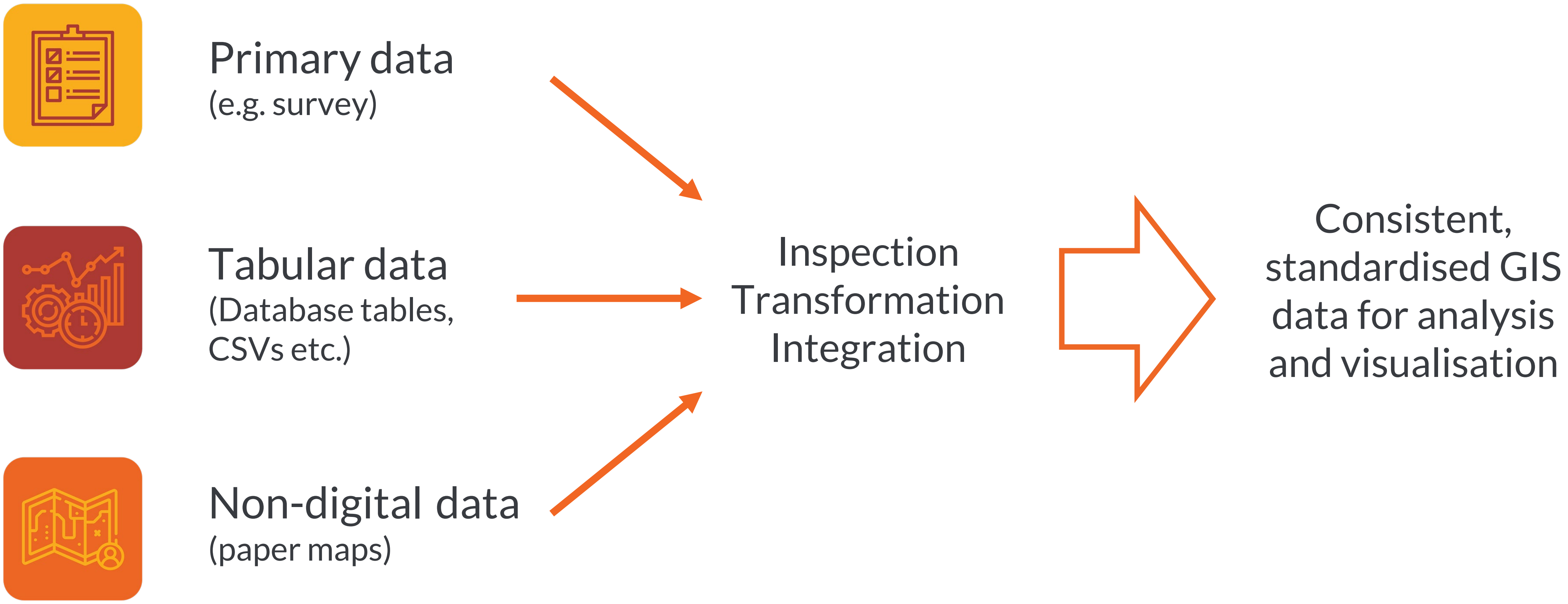
Datasets may exist… **BUT** are they suitable for use?

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- Out of date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

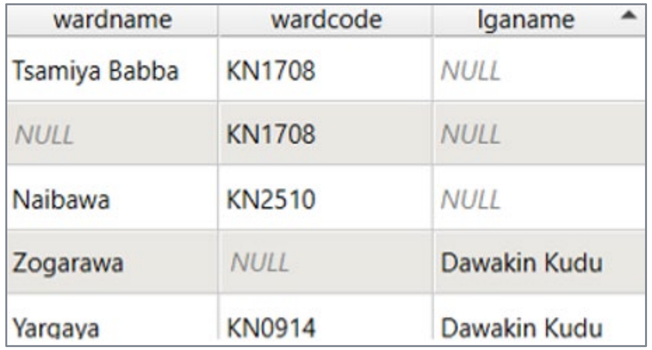
Consistent, standardised GIS data for analysis and visualisation

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



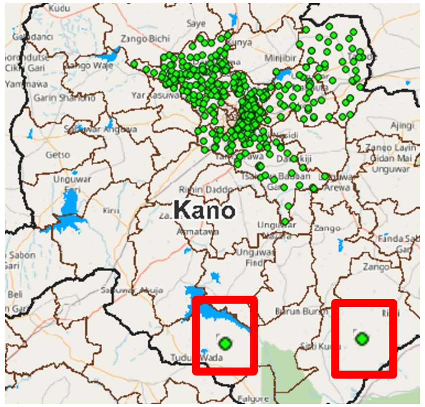
wardname	wardcode	lganame
Tsamiya Babba	KN1708	*NULL*
NULL	KN1708	*NULL*
Naibawa	KN2510	*NULL*
Zogarawa	*NULL*	Dawakin Kudu
Yargaya	KN0914	Dawakin Kudu

Missing values or misspelt words

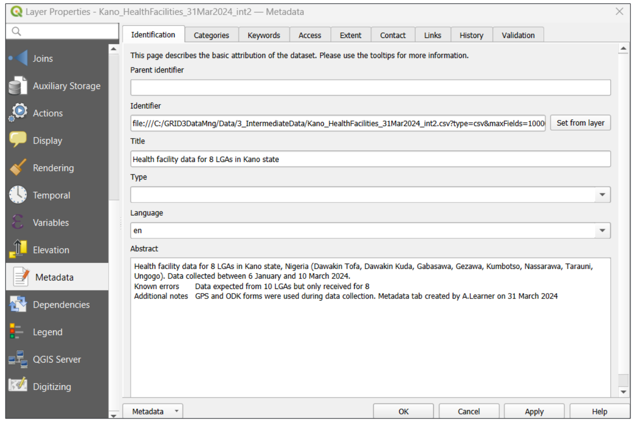
30/04/171
30/4/1971
30thApril71
30:4:71
?th April 17
Apr-71-30

100.123
100,123
100.123%

Inconsistent values make data queries impossible



Point locations noticeably outside study area



This page describes the basic attribution of the dataset. Please use the tooltips for more information.

Parent identifier

Identifier: file:///C:/GRID3DataMng/Data/3_IntermediateData/Kano_HealthFacilities_31Mar2024_int2.csv?type=csv&maxFields=1000

Title: Health facility data for 8 LGAs in Kano state

Type

Language: en

Abstract:
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.
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

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

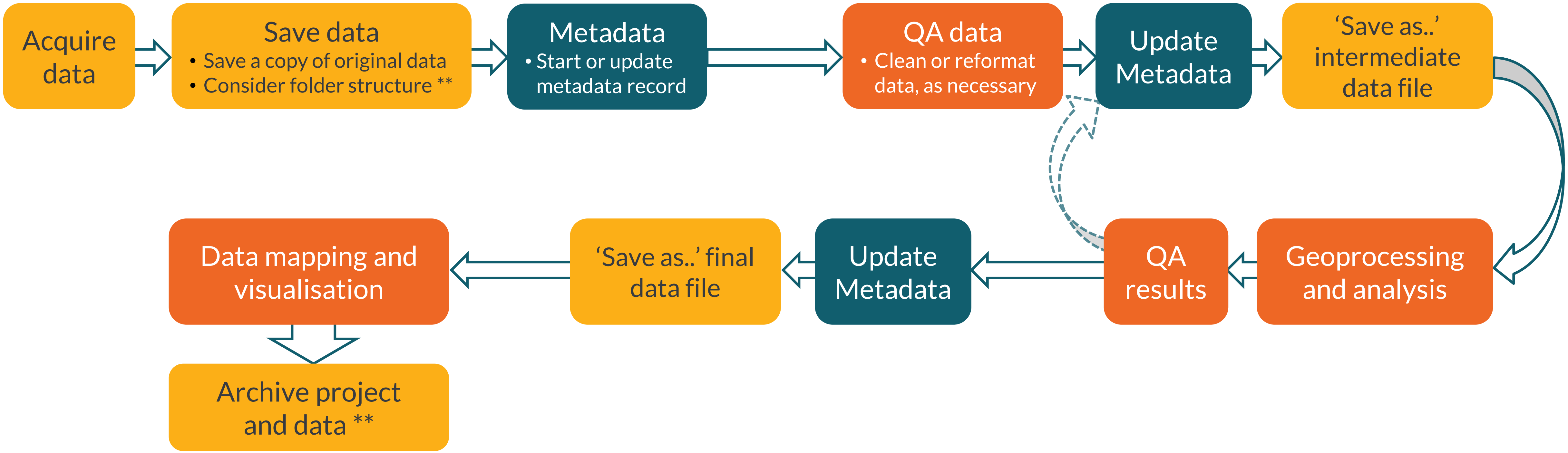
The GRID3 GIS data cycle

Standardise your data management procedures

🟨 File and folder management

🟦 Data documentation

🟧 Working on the data



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

Geoprocessing and analysis

QA results

Update Metadata

'Save as..' final data file

Data mapping and visualisation

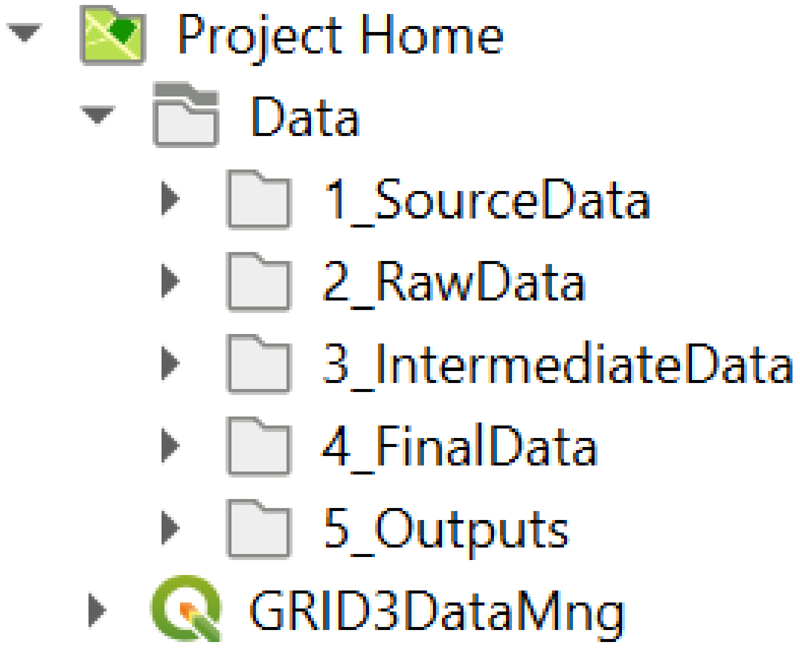
Archive project and data **

- 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



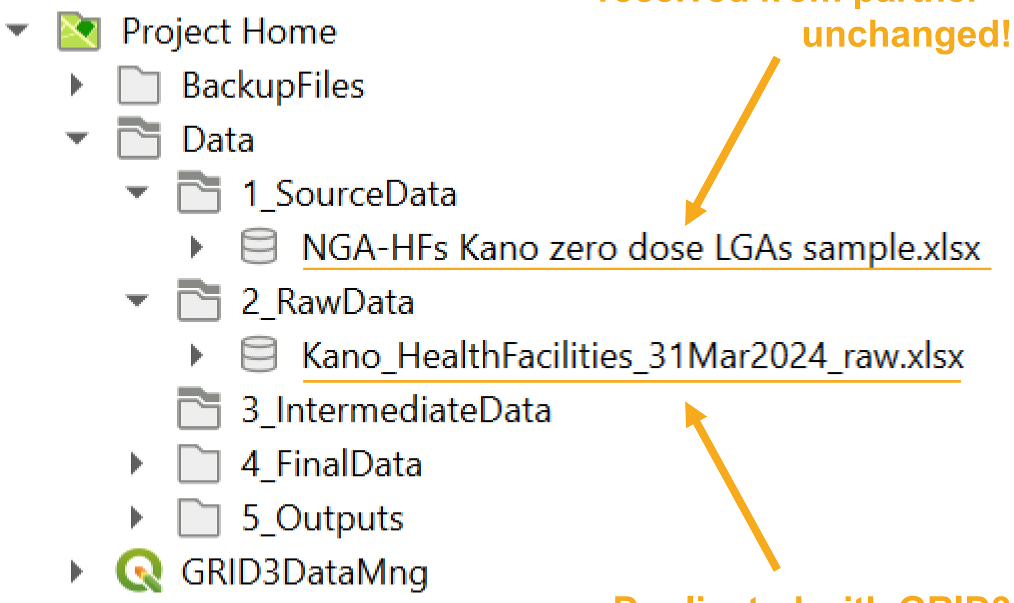
```
Project Home
  Data
    1_SourceData
    2_RawData
    3_IntermediateData
    4_FinalData
    5_Outputs
  GRID3DataMng
```

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'



```
Project Home
  BackupFiles
  Data
    1_SourceData
      NGA-HFs Kano zero dose LGAs sample.xlsx
    2_RawData
      Kano_HealthFacilities_31Mar2024_raw.xlsx
    3_IntermediateData
    4_FinalData
    5_Outputs
  GRID3DataMng
```

Source data file received from partner – unchanged!

Duplicated with GRID3 filename

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) | ESRI Geodatabase

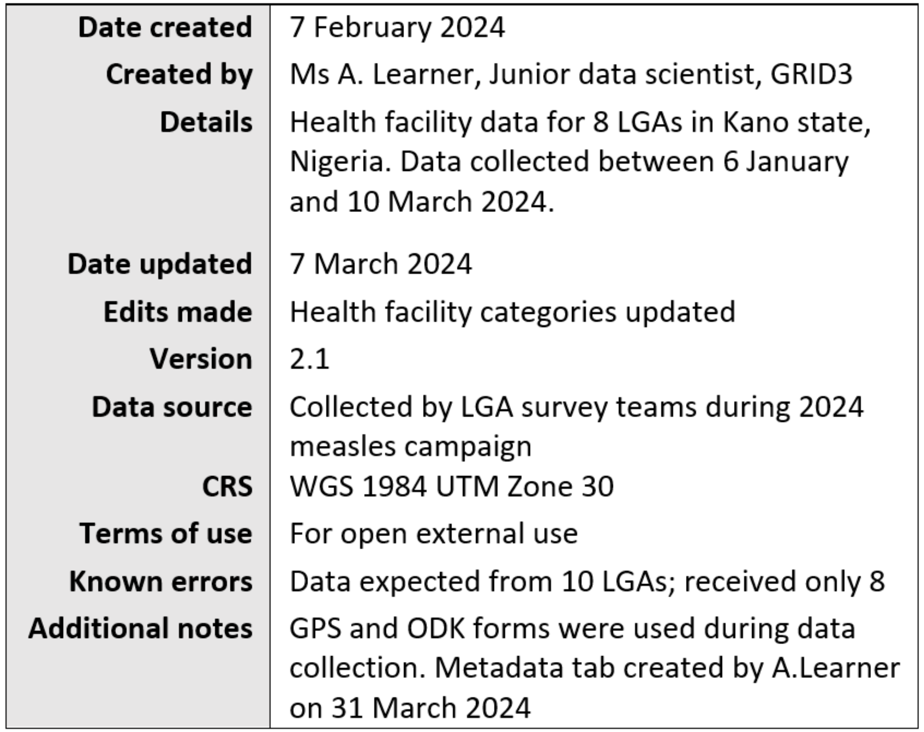
Note: the restriction also applies to the naming of column headers!

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!



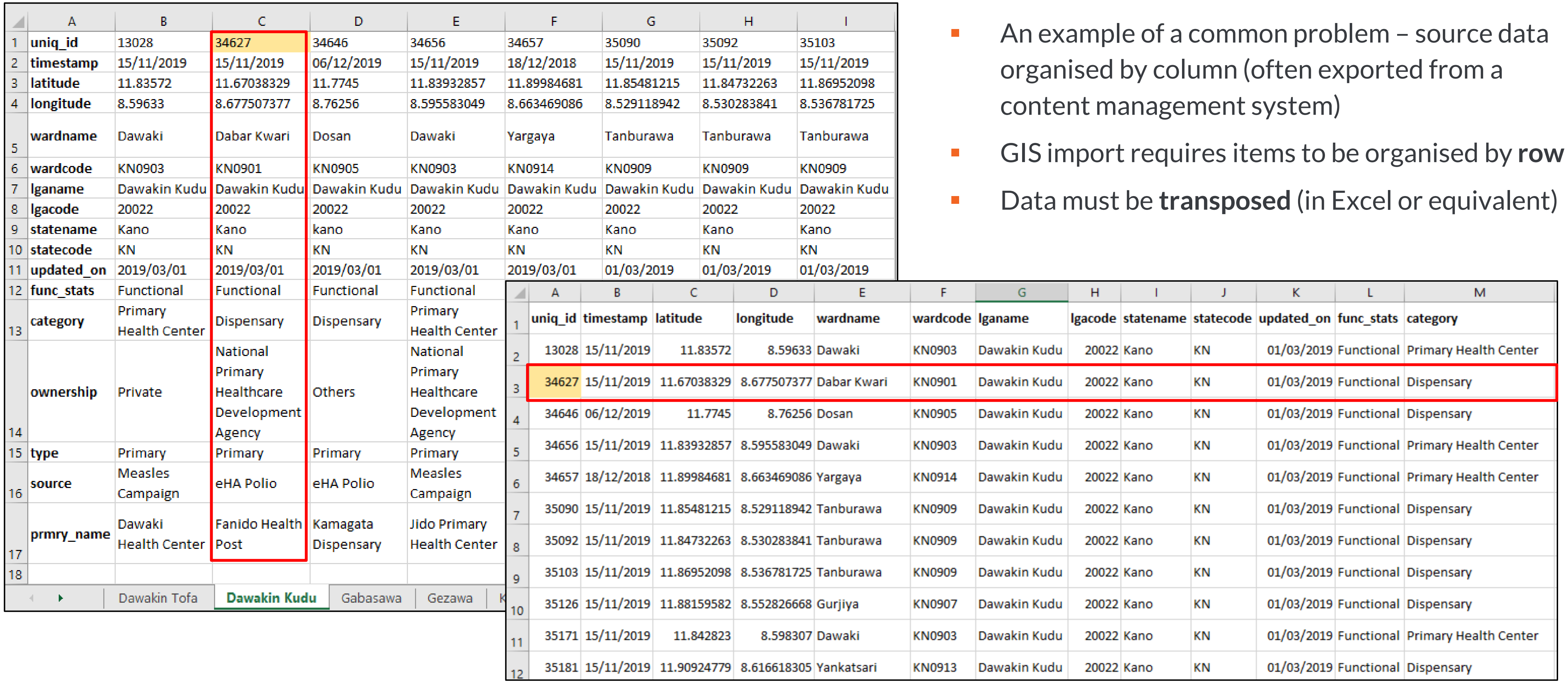
Date created	7 February 2024
Created by	Ms A. Learner, Junior data scientist, GRID3
Details	Health facility data for 8 LGAs in Kano state, Nigeria. Data collected between 6 January and 10 March 2024.
Date updated	7 March 2024
Edits made	Health facility categories updated
Version	2.1
Data source	Collected 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
Additional notes	GPS and ODK forms were used during data collection. Metadata tab created by A.Learner on 31 March 2024

An example metadata record

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)



	A	B	C	D	E	F	G	H	I
1	**uniq_id**	13028	34627	34646	34656	34657	35090	35092	35103
2	**timestamp**	15/11/2019	15/11/2019	06/12/2019	15/11/2019	18/12/2018	15/11/2019	15/11/2019	15/11/2019
3	**latitude**	11.83572	11.67038329	11.7745	11.83932857	11.89984681	11.85481215	11.84732263	11.86952098
4	**longitude**	8.59633	8.677507377	8.76256	8.595583049	8.663469086	8.529118942	8.530283841	8.536781725
5	**wardname**	Dawaki	Dabar Kwari	Dosan	Dawaki	Yargaya	Tanburawa	Tanburawa	Tanburawa
6	**wardcode**	KN0903	KN0901	KN0905	KN0903	KN0914	KN0909	KN0909	KN0909
7	**lganame**	Dawakin Kudu	Dawakin Kudu	Dawakin Kudu	Dawakin Kudu	Dawakin Kudu	Dawakin Kudu	Dawakin Kudu	Dawakin Kudu
8	**lgacode**	20022	20022	20022	20022	20022	20022	20022	20022
9	**statename**	Kano	Kano	kano	Kano	Kano	Kano	Kano	Kano
10	**statecode**	KN	KN	KN	KN	KN	KN	KN	KN
11	**updated_on**	2019/03/01	2019/03/01	2019/03/01	2019/03/01	2019/03/01	01/03/2019	01/03/2019	01/03/2019
12	**func_stats**	Functional	Functional	Functional	Functional				
13	**category**	Primary Health Center	Dispensary	Dispensary	Primary Health Center				
14	**ownership**	Private	National Primary Healthcare Development Agency	Others	National Primary Healthcare Development Agency				
15	**type**	Primary	Primary	Primary	Primary				
16	**source**	Measles Campaign	eHA Polio	eHA Polio	Measles Campaign				
17	**prmry_name**	Dawaki Health Center	Fanido Health Post	Kamagata Dispensary	Jido Primary Health Center				
18									

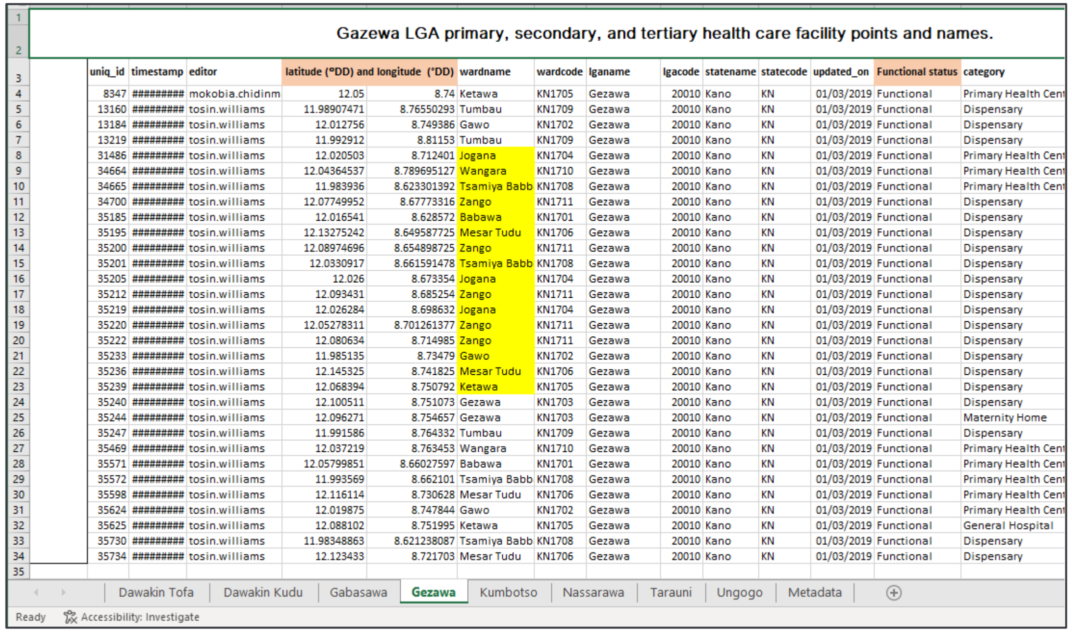
Dawakin Tofa | **Dawakin Kudu** | Gabasawa | Gezawa | K

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**uniq_id**	**timestamp**	**latitude**	**longitude**	**wardname**	**wardcode**	**lganame**	**lgacode**	**statename**	**statecode**	**updated_on**	**func_stats**	**category**
2	13028	15/11/2019	11.83572	8.59633	Dawaki	KN0903	Dawakin Kudu	20022	Kano	KN	01/03/2019	Functional	Primary Health Center
3	34627	15/11/2019	11.67038329	8.677507377	Dabar Kwari	KN0901	Dawakin Kudu	20022	Kano	KN	01/03/2019	Functional	Dispensary
4	34646	06/12/2019	11.7745	8.76256	Dosan	KN0905	Dawakin Kudu	20022	Kano	KN	01/03/2019	Functional	Dispensary
5	34656	15/11/2019	11.83932857	8.595583049	Dawaki	KN0903	Dawakin Kudu	20022	Kano	KN	01/03/2019	Functional	Primary Health Center
6	34657	18/12/2018	11.89984681	8.663469086	Yargaya	KN0914	Dawakin Kudu	20022	Kano	KN	01/03/2019	Functional	Primary Health Center
7	35090	15/11/2019	11.85481215	8.529118942	Tanburawa	KN0909	Dawakin Kudu	20022	Kano	KN	01/03/2019	Functional	Dispensary
8	35092	15/11/2019	11.84732263	8.530283841	Tanburawa	KN0909	Dawakin Kudu	20022	Kano	KN	01/03/2019	Functional	Dispensary
9	35103	15/11/2019	11.86952098	8.536781725	Tanburawa	KN0909	Dawakin Kudu	20022	Kano	KN	01/03/2019	Functional	Dispensary
10	35126	15/11/2019	11.88159582	8.552826668	Gurjiya	KN0907	Dawakin Kudu	20022	Kano	KN	01/03/2019	Functional	Dispensary
11	35171	15/11/2019	11.842823	8.598307	Dawaki	KN0903	Dawakin Kudu	20022	Kano	KN	01/03/2019	Functional	Primary Health Center
12	35181	15/11/2019	11.90924779	8.616618305	Yankatsari	KN0913	Dawakin Kudu	20022	Kano	KN	01/03/2019	Functional	Dispensary

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?
    - *1ooo or 1000?*
    - *13th Feb 24 or 13/02/2024?*
  - Remove trailing- and double-spaces
  - Consistent capitalisation?



Gazewa LGA primary, secondary, and tertiary health care facility points and names.

uniq_id	timestamp	editor	latitude (°DD) and longitude (°DD)		wardname	wardcode	lganame	lgacode	statename	statecode	updated_on	Functional status	category
8347	#########	mokobia.chidinm	12.05	8.74	Ketawa	KN1705	Gezawa	20010	Kano	KN	01/03/2019	Functional	Primary Health Cent
13160	#########	tosin.williams	11.98907471	8.76550293	Tumbau	KN1709	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
13184	#########	tosin.williams	12.012756	8.749386	Gawo	KN1702	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
13219	#########	tosin.williams	11.992912	8.81153	Tumbau	KN1709	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
31486	#########	tosin.williams	12.020503	8.712401	Jogana	KN1704	Gezawa	20010	Kano	KN	01/03/2019	Functional	Primary Health Cent
34664	#########	tosin.williams	12.04364537	8.789695127	Wangara	KN1710	Gezawa	20010	Kano	KN	01/03/2019	Functional	Primary Health Cent
34665	#########	tosin.williams	11.983936	8.623301392	Tsamiya Babb	KN1708	Gezawa	20010	Kano	KN	01/03/2019	Functional	Primary Health Cent
34700	#########	tosin.williams	12.07749952	8.67773316	Zango	KN1711	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35185	#########	tosin.williams	12.016541	8.628572	Babawa	KN1701	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35195	#########	tosin.williams	12.13275242	8.649587725	Mesar Tudu	KN1706	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35200	#########	tosin.williams	12.08974696	8.654898725	Zango	KN1711	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35201	#########	tosin.williams	12.0330917	8.661591478	Tsamiya Babb	KN1708	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35205	#########	tosin.williams	12.026	8.673354	Jogana	KN1704	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35212	#########	tosin.williams	12.093431	8.685254	Zango	KN1711	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35219	#########	tosin.williams	12.026284	8.698632	Jogana	KN1704	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35220	#########	tosin.williams	12.05278311	8.701261377	Zango	KN1711	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35222	#########	tosin.williams	12.080634	8.714985	Zango	KN1711	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35233	#########	tosin.williams	11.985135	8.73479	Gawo	KN1702	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35236	#########	tosin.williams	12.145325	8.741825	Mesar Tudu	KN1706	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35239	#########	tosin.williams	12.068394	8.750792	Ketawa	KN1705	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35240	#########	tosin.williams	12.100511	8.751073	Gezawa	KN1703	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35244	#########	tosin.williams	12.096271	8.754657	Gezawa	KN1703	Gezawa	20010	Kano	KN	01/03/2019	Functional	Maternity Home
35247	#########	tosin.williams	11.991586	8.764332	Tumbau	KN1709	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35469	#########	tosin.williams	12.037219	8.763453	Wangara	KN1710	Gezawa	20010	Kano	KN	01/03/2019	Functional	Primary Health Cent
35571	#########	tosin.williams	12.05799851	8.66027597	Babawa	KN1701	Gezawa	20010	Kano	KN	01/03/2019	Functional	Primary Health Cent
35572	#########	tosin.williams	11.993569	8.662101	Tsamiya Babb	KN1708	Gezawa	20010	Kano	KN	01/03/2019	Functional	Primary Health Cent
35598	#########	tosin.williams	12.116114	8.730628	Mesar Tudu	KN1706	Gezawa	20010	Kano	KN	01/03/2019	Functional	Primary Health Cent
35624	#########	tosin.williams	12.019875	8.747844	Gawo	KN1702	Gezawa	20010	Kano	KN	01/03/2019	Functional	Primary Health Cent
35625	#########	tosin.williams	12.088102	8.751995	Ketawa	KN1705	Gezawa	20010	Kano	KN	01/03/2019	Functional	General Hospital
35730	#########	tosin.williams	11.98348863	8.621238087	Tsamiya Babb	KN1708	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary
35734	#########	tosin.williams	12.123433	8.721703	Mesar Tudu	KN1706	Gezawa	20010	Kano	KN	01/03/2019	Functional	Dispensary

Good practice: Data cleaning for GIS

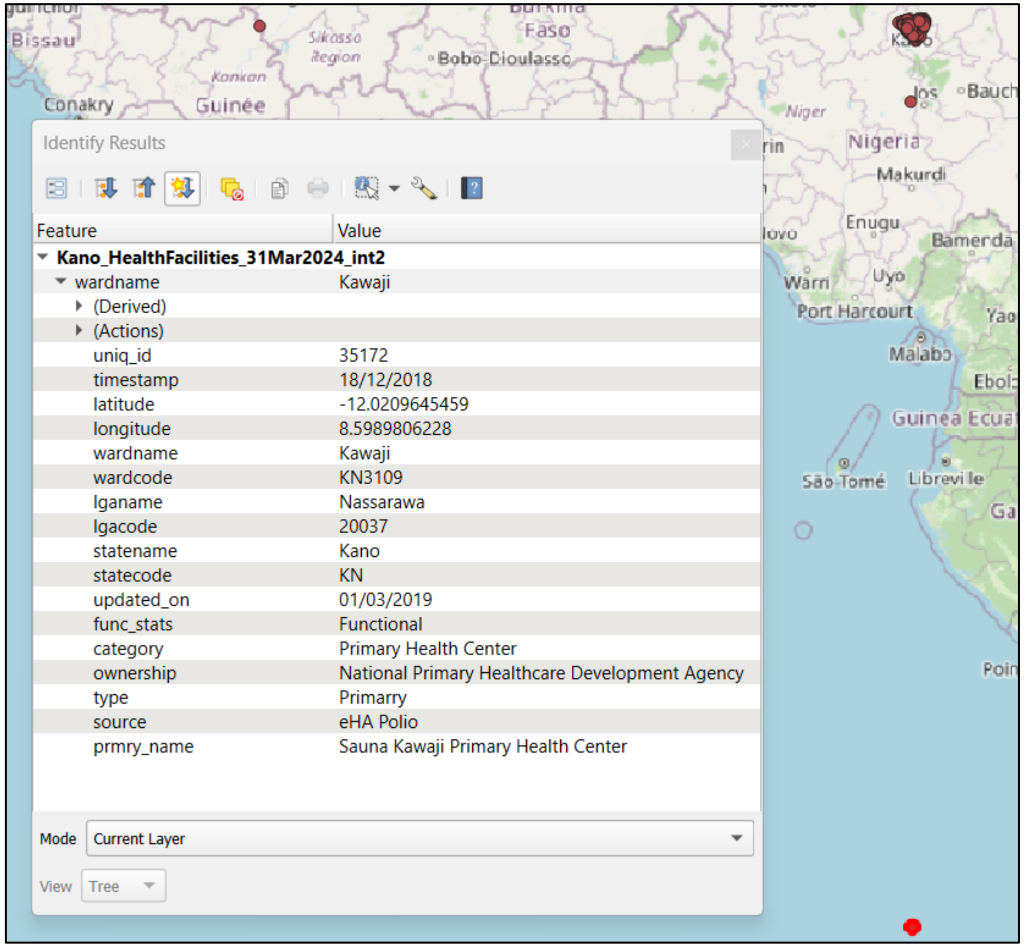
Considerations for cleaning spatial data:

- 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



Identify Results

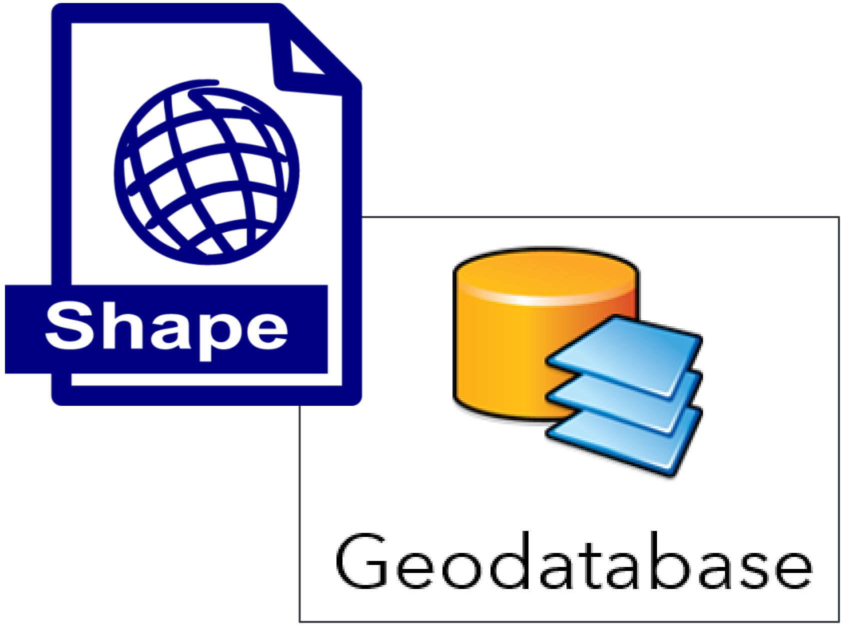
Feature	Value
Kano_HealthFacilities_31Mar2024_int2	
wardname	Kawaji
(Derived)	
(Actions)	
uniq_id	35172
timestamp	18/12/2018
latitude	-12.0209645459
longitude	8.5989806228
wardname	Kawaji
wardcode	KN3109
lganame	Nassarawa
lgacode	20037
statename	Kano
statecode	KN
updated_on	01/03/2019
func_stats	Functional
category	Primary Health Center
ownership	National Primary Healthcare Development Agency
type	Primarry
source	eHA Polio
prmry_name	Sauna Kawaji Primary Health Center

Mode: Current Layer

View: Tree

Choosing spatial data format

What are most commonly used formats?



Open Geospatial Consortium

geopackage.org

esri™

esri.com

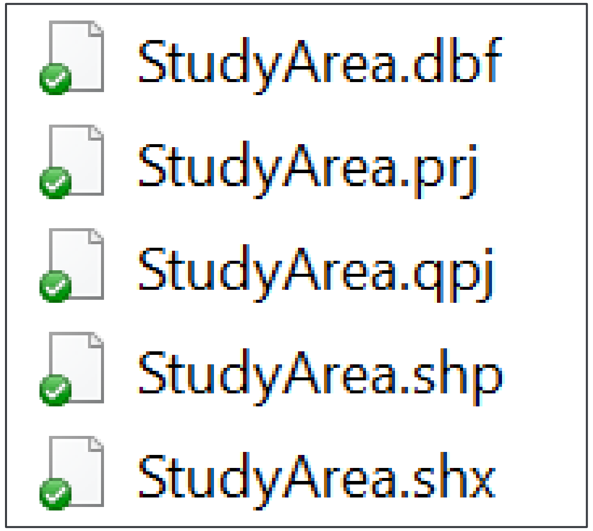
Choosing spatial data format: Shapefile (SHP)

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)



StudyArea.dbf
StudyArea.prj
StudyArea.qpj
StudyArea.shp
StudyArea.shx

The structure of a shapefile; multiple components files

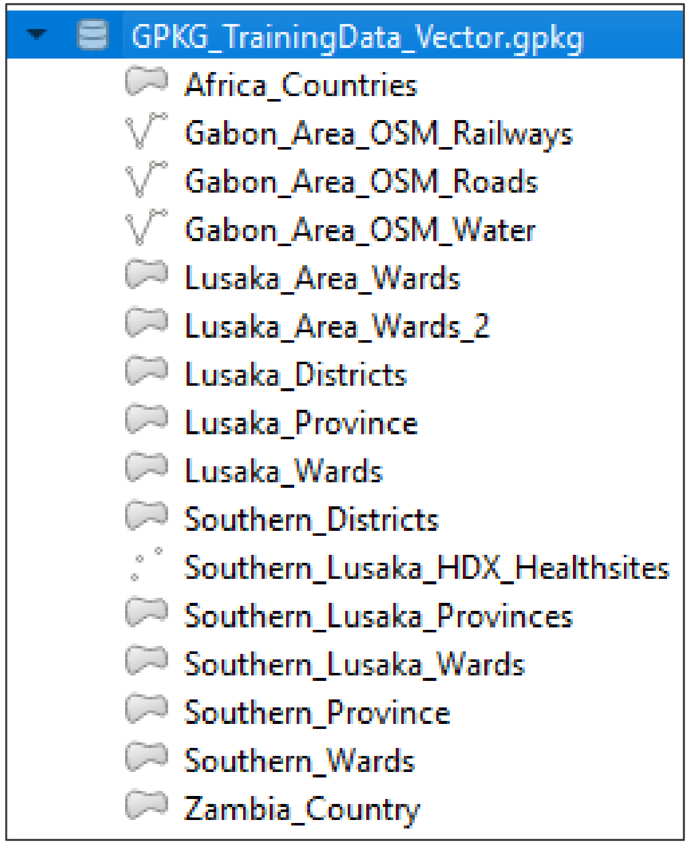
Choosing spatial data format: GeoPackage (GPKG)

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



- GPKG_TrainingData_Vector.gpkg
  - Africa_Countries
  - Gabon_Area_OSM_Railways
  - Gabon_Area_OSM_Roads
  - Gabon_Area_OSM_Water
  - Lusaka_Area_Wards
  - Lusaka_Area_Wards_2
  - Lusaka_Districts
  - Lusaka_Province
  - Lusaka_Wards
  - Southern_Districts
  - Southern_Lusaka_HDX_Healthsites
  - Southern_Lusaka_Provinces
  - Southern_Lusaka_Wards
  - Southern_Province
  - Southern_Wards
  - Zambia_Country

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:

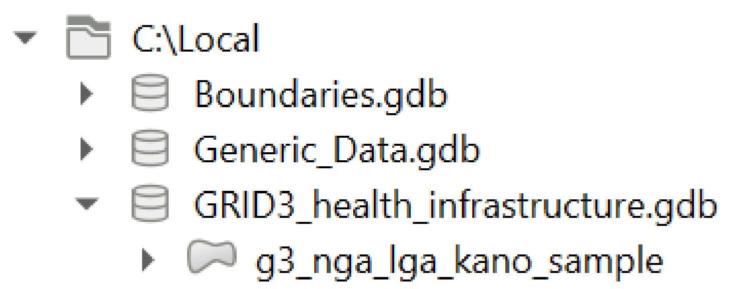
- If you are working solely in QGIS, you should adopt GeoPackage
- However, some teams contain both QGIS users and ArcGIS users!
- Recent QGIS installations come with the *openfilegdb* driver, enabling a level of access and use of FGDB; **note the following:**

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!



QGIS users can explore Geodatabases directly from the Browser panel

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!

GRID3

GIS Data Preparation and Management

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