Guidance for the management and use of geospatial data and technologies in health

Part 2 - Implementing the geospatial data management cycle: 2.3 Compiling existing data and identifying gaps

Version 1.4 (last update: 20.05.2025)

In collaboration and with the support of:

Revision history

Version	Release date	Comment	By
1.0	20 June 2018	Document created	Steeve Ebener, Izay Pantanilla, Chris Erwin G. Mercado, Richard J. Maude
1.1	04 March 2020	Adjustment of the terminologies to align with other volumes of the series	Steeve Ebener
1.2	10 January 2022	Inclusion of the reference to the new HGL guidance documents and improvement of section 5 based on recently published guides	Izay Pantanilla, Steeve Ebener
1.3	19 February 2024	Update the layout, section 4.2.1 and the URLs across the document	Steeve Ebener
1.4	20 May 2025	Addition of one annex describing the process to assess the quality of a list	Steeve Ebener

Authors

Izay Pantanilla[1]
Steeve Ebener[1]
Chris Erwin G. Mercado[2]
Richard J. Maude[2,3,4]

1. Health GeoLab, Manila, Philippines
2. Mahidol-Oxford Tropical Medicine Research Unit (MORU), Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand
3. Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK
4. Harvard TH Chan School of Public Health, Harvard University, Boston, USA

Acknowledgements

Our gratitude goes to the Asian Development Bank (ADB), the World Health Organization (WHO) and the United Nations Population Fund (UNFPA) for the support provided to the Health GeoLab for the development and update of the present document. MORU is funded by Wellcome.

Table of Contents

Purpose and audience

The purpose of the Health GeoLab series of guidance is to inform concerned practitioners about the key elements they need to be aware of when it comes to managing and using geospatial data and technologies in public health and guide them through the processes to be followed in that regard.

The audience for this guidance includes geospatial data managers, technical advisors, and any other practitioners that are directly or indirectly involved in the collection and use of geospatial data and technologies in public health.

Please note that some of the sections in the present guidance require a basic understanding of concepts pertaining to the management and use of geospatial data and technologies.

Abbreviations

ADB	Asian Development Bank
AeHIN	Asia eHealth Information Network
DEM	Digital Elevation Model
GIS	Geographic Information System
HGL	Health GeoLab
HIS	Health Information System
MORU	Mahidol-Oxford Tropical Medicine Research Unit
SDG	Sustainable Development Goal
WHO	World Health Organization

1. Background

The Health GeoLab (HGL) is a regional resource supporting low- and middle-income countries in Asia and the Pacific for them to fully benefit from the power of geography, geospatial data, and technologies to reach the health-related Sustainable Development Goal of healthy lives and well-being for all (SDG 3)[1].

The HGL uses the HIS geo-enabling framework to strengthen in-country capacity. The present document has been developed as part of this approach and with the objective of being used by the largest number of users possible.

This volume is part of a series of guidance started under the umbrella of the AeHIN GIS Lab and now continued by the HGL. The complete series is organized as follows:

- Part 1 - Introduction to the data-information-knowledge-decision continuum and the geospatial data management cycle [1]
- Part 2 - Implementing the geospatial data management cycle:
    - 2.1 Documenting the process and defining the data needs [2]
    - 2.2 Defining the terminology, data specifications, and the ground reference [3]
    - 2.3 Compiling existing data and identifying gaps (the present document)
    - 2.4 Creating geospatial data
        - 2.4.1 Extracting vector format geospatial data from basemaps [4]
        - 2.4.2 Collecting data in the field [5]
    - 2.5 Cleaning, validating, and documenting the data
        - 2.5.1 Documenting the data using a metadata profile [6]
        - 2.5.2 Using advanced Microsoft Excel functions [7]
    - 2.6 Distributing, using, and updating the data
        - 2.6.1 Creating good thematic maps using desktop GIS software [8]
        - 2.6.2 Using thematic maps for decision making [9]
        - 2.6.3 Developing and implementing the appropriate data policy [10]

This guidance is a living document made to evolve based on the inputs received from the users. Please don't hesitate to contact us if you have any suggestions for improvement.

The terms used in the present guidance are defined in the following glossary of terms maintained by the Health GeoLab: https://bit.ly/3ctoHiS

Please also contact us using the same email address should you use this document as part of your activities and would like to have your institution recognized as one of the document's users.

[1] https://www.un.org/sustainabledevelopment/health/

2. Introduction

Once the data needs have been identified [2] and before collecting new data in the field, it is advisable to compile the data already available and see if it is appropriate for the initial purpose and that it complies with the data specifications and ground reference that have been pre-defined [3]. This process prevents duplication of efforts, saves time and money, and allows identification of potential gaps.

The present document's objective is to guide users through the above-mentioned process. While this process is to be applied to both geospatial and statistical data, the present guide focuses mainly on the former.

3. Compiling existing datasets

The compilation process needs to cover the following to lead to a quality dataset:

1. The master list for the geographic features considered in the data model [2]
2. The geospatial data containing the geometry (geographic objects) for the considered geographic features.
3. The statistical data to be attached to these features.
4. Basemaps to serve as ground reference when checking the geospatial data that has been collected.

While master lists should only come from the governmental entity having the official mandate over the considered geographic feature(s), geospatial and statistical data as well as basemaps can themselves come from different sources depending on the needs identified at the beginning of the process and their availability. It is therefore important to consider all these sources as they might be complementary and under different use and redistribution rights constraints.

Table 1 gives the list of the governmental entities generally in charge of the master list and associated geospatial data for the geographic features most often used in public health.

Geographic feature	Master list	Geospatial data	Governmental entity
Health facilities	✓	✓	Ministry of Health
Health districts or other reporting divisions	✓	✓	Ministry of Health
Administrative units and villages	✓	✓	Ministry of Interior, National Statistical Agency, National Mapping Agency
Transportation network	Not necessary	✓	Ministry of Public Works, Ministry of Transportation
Hydrographic network	Not necessary	✓	Ministry of Environment/Agriculture
Climate data (temperature, precipitation, etc.)	Not applicable	✓	Ministry of Meteorology, Meteorological agency
Digital Elevation Model (DEM)	Not applicable	✓	National Mapping Agency
Land cover	Not applicable	✓	National Mapping Agency, Ministry of Environment/Agriculture

Table 1. Governmental entities generally having the mandate over the geographic features mainly used in public health

The other potential sources of local, regional or global geospatial and statistical data can be non-government organizations (NGOs), volunteer/community groups, research groups, universities, and the private sector. When accessible, much of this data can be downloaded directly from the internet, with some requiring registration with the institution that distributes the data.
Basemaps are accessible either through the GIS software you are using or through online web mapping services such as ArcGIS Online or Google Maps.

Whatever the source of the data being compiled, it is always very important to collect the metadata associated with it [6]. If such metadata is not directly attached to the data file itself, this should be collected separately and kept together with it (e.g., in the same folder).

At minimum, the metadata should include the following to be useful:

1. What is the data about?
2. Who created the data?
3. When was the data created/collected/last updated and what is its temporal validity?
4. How was the data created?
5. What are the data specifications (geographic coordinate system/projection system, scale, accuracy, language,...)?
6. Are there any access, use or redistribution restrictions or limitations attached to the data?
7. Who can I contact if I have questions about the data?

4. Organizing the compiled data

As you are compiling the data, it is important to organize it on your computer in such a way that it is easy to find, including by other people. The filing structure should change as little as possible to avoid losing the path to these datasets once they are stored in a project file (.mxd, .qgs) generated from a GIS software. Figure 1 provides an example of a folder organization structure.

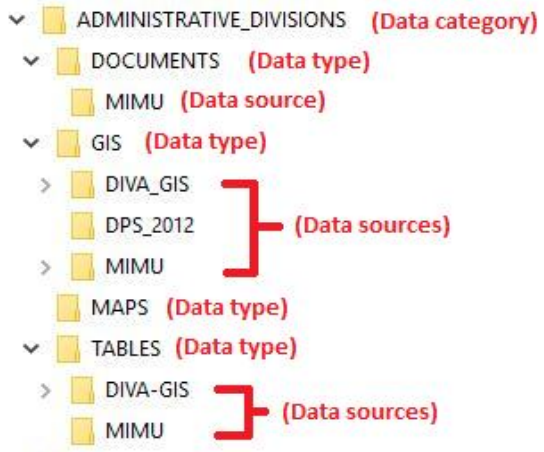


Figure 1. Example of folder organization structure

In the folder organization structure shown in Figure 1, files are organized by:

1. **Data category** corresponding to the different geographic features being collected (health facilities, administrative units, Digital Elevation Model (DEM), etc.)
2. **Data type**. Four main types are generally considered:
    a. DOCUMENTS: for reports, publications and other narrative documents
    b. GIS: for geospatial data saved in a GIS-compatible format (shapefile, GeoJSON, etc.)

c. MAPS: for maps saved in pdf, Microsoft Word format or images (.jpg, .bmp, .png, etc.)
d. TABLES: for any data saved in a tabular format (Microsoft Excel, CSV, etc.)

3. **Data source** with one folder for each source, the corresponding data being saved in each of these folders. Please note that the year of data production is included together with the data source in the folder name when known (e.g., DPS_2012).

5. Assessing the compiled data and documenting the gaps

Once all the available data sets are compiled and organized, it is important to assess them to identify:

- If at least one source is usable (no restriction of use) for each of the data that is needed as defined at the beginning of the process [2]
- Which source(s), present the required level of quality across the six dimensions (Completeness, Uniqueness, Timeliness, Validity, Accuracy, Consistency) and this in alignment with defined data specifications and the ground references (imagery, master lists [3]

While the first part of the assessment is straightforward and consists of making a list of data that it has not been possible to find, or that could not be used, the second part requires a more in-depth analysis.

When it comes to master lists, Annex 1 contains the questions that the assessment should aim at answering as well as the process to be followed to answer them. This process can also be used to assess the quality of non-master list against the master list.

The implementation support guide for the development of a national georeferenced community health worker master list hosted in a registry [11] does itself provide the questions and a description of the process specifically for this type of master list.

Annex 2 contains the questions that the assessment should look at answering for geospatial data (vector or raster format) when it comes to the first 5 data quality dimensions. Consistency is itself reached once the benchmarks or conditions fixed for the other dimensions are reached or respected. Please note that the ability to answer these questions very much depends on the availability of data specifications and ground references (master lists and satellite images) being defined and identified prior to the data compilation exercise [3]. Without these, only a limited number of data quality dimensions can be assessed, making the choice of the source with the highest quality more difficult.

While the sources with the best “score” at the end of the assessment is most likely to be the most appropriate for the project, there is no perfect rule in this regard and the amount of work to fill the remaining gaps will have to be considered when taking this decision. For example:

- Temporal discrepancies (criteria 1 and 2 in Annex 1): Such discrepancies between datasets are a common issue that can have a big impact on the results and might require a lot of work to be addressed. Comparing the location of health facilities as observed today with the road network as it was in place 10 years ago is one example.
- The lack of documentation (metadata) might not only lead to technical issues such as the impossibility to correctly project a geospatial dataset or using a dataset that was not meant to be shared in the first place (limited use and/or redistribution rights). Using such datasets

might end up being more problematic than using other sources of lower quality but for which this information is known.
- The gaps in authoritative data from the government might be too big compared to other sources for them to be considered.

When it comes to statistical data, the questions to be covered are as follows:

- Validity:
  - Is the dataset accompanied by a data dictionary and metadata containing all the information necessary to use it properly?
  - Is the dataset available in a format that allows its use or could it be converted in such a format?
- Timeliness:
  - What is the temporal validity of the statistical data?
  - Does it match the temporal period of validity set in the data specifications?
- Completeness:
  - Are statistics available for each of the records captured in the corresponding master list?
  - Is a value available for each of the data elements included in the data dictionary?
- Uniqueness:
  - Does the dataset contain record duplicates?
- Consistency:
  - Are any inconsistencies observed between records for data elements based on a classification table (example: health facility type, ownership) or associated master list (example: name of the administrative units in which they are located)?

It is possible that, at the end of the assessment, none of the data that has been compiled presents a quality sufficient to justify their use. The following options should be considered in this case:

- Look for additional sources that might have been missed during the first round of data compilation.
- Identify if combining different sources of data together could help cover the gap(s).

If none of the above is possible then remaining data gaps should be documented and properly mentioned not only in the metadata profile but also on any maps that would be created using this data.

The next natural step in the process, when possible, would consist of extracting data from other sources [4] or collecting data in the field [5].

References

[1] Ebener S. (2016): Guidance for the management and use of geospatial data and technologies in health. Part 1 - Introduction to the data-information-knowledge-decision continuum and the geospatial data management chain. Health GeoLab document: https://www.healthgeolab.net/DOCUMENTS/Guide_HGL_Part1.pdf [Accessed 19 February 2024]

[2] Ebener S. (2016): Guidance for the management and use of geospatial data and technologies in health. Part 2 - Implementing the geospatial data management cycle: 2.1 Documenting the process and defining the data needs. Health GeoLab document: https://www.healthgeolab.net/DOCUMENTS/Guide_HGL_Part2_1.pdf [Accessed 19 February 2024]

[3] Ebener S. (2016): Guidance for the management and use of geospatial data and technologies in health. Part 2 - Implementing the geospatial data management cycle: 2.2 Defining the terminology, data specifications, and the ground reference. Health GeoLab document: https://www.healthgeolab.net/DOCUMENTS/Guide_HGL_Part2_2.pdf [Accessed 19 February 2024]

[4] Ebener S. (2021): Guidance for the management and use of geospatial data and technologies in health. Part 2 - Implementing the geospatial data management cycle: 2.4 Creating geospatial data - 2.4.1 Extracting vector format geospatial data from basemaps. Health GeoLab document: https://healthgeolab.net/DOCUMENTS/Guide_HGL_Part2_4_1.pdf [Accessed 19 February 2024]

[5] Ebener S., Maude R.J., Gault P. (2018): Guidance for the management and use of geospatial data and technologies in health. Part 2 - Implementing the geospatial data management cycle: 2.4 Creating geospatial data - 2.4.2 Collecting data in the field. Health GeoLab document: https://www.healthgeolab.net/DOCUMENTS/Guide_HGL_Part2_4_2.pdf [Accessed 19 February 2024]

[6] Ebener S., Pantanilla I., Mercado C.E., Maude R. (2018): Guidance for the management and use of geospatial data and technologies in health. Part 2 - Implementing the geospatial data management cycle: 2.5 Cleaning, validating, and documenting the data - 2.5.1 Documenting the data using a metadata profile. Health GeoLab document: https://www.healthgeolab.net/DOCUMENTS/Guide_HGL_Part2_5_1.pdf [Accessed 19 February 2024]

[7] Ebener S., Pantanilla I. (2019): Guidance for the management and use of geospatial data and technologies in health. Part 2 - Implementing the geospatial data management cycle: 2.5 Cleaning, validating, and documenting the data - 2.5.2 Using advanced Microsoft Excel functions. Health GeoLab document: https://healthgeolab.net/DOCUMENTS/Guide_HGL_Part2_5_2.pdf [Accessed 19 February 2024]

[8] Pantanilla I., Ebener S., Maude R. (2018): Guidance for the management and use of geospatial data and technologies in health. Part 2 - Implementing the geospatial data management cycle: 2.6 Distributing, using, and updating the data - 2.6.1 Creating good thematic maps using

desktop GIS software. Health GeoLab document: https://www.healthgeolab.net/DOCUMENTS/Guide_HGL_Part2_6_1.pdf [Accessed 19 February 2024]

[9] Ebener S. (under preparation): Guidance for the management and use of geospatial data and technologies in health. Part 2 - Implementing the geospatial data management cycle: 2.6 Distributing, using, and updating the data - 2.6.2 Using thematic maps for decision making. Health GeoLab document.

[10] Ebener S. (under preparation): Guidance for the management and use of geospatial data and technologies in health. Part 2 - Implementing the geospatial data management cycle: 2.6 Distributing, using, and updating the data - 2.6.3 Developing and implementing the appropriate data policy. Health GeoLab document.

[11] CHAI, CHIC, HGL, LivingGoods, The Global Fund, UNICEF (2021): Implementation support guide: Development of a National Georeferenced Community Health Worker Master List Hosted in a Registry: https://www.unicef.org/media/113081/file/National-Georeferenced-Community-Health-Worker-Master-List-Hosted-in-a-Registry-2021.pdf [Accessed 19 February 2024]

Annex 1 – Process to assess the quality of a list

The present annex describes the process through which the quality of a master, or non-master list against the master list, can be assessed across the six dimensions of data quality and this by answering the questions listed in Table A.1.

Table A.1 - Questions to be answered during the quality assessment of a list, including master lists (adapted from Table 7 in [11])

Quality dimension	Question to be answered	Method to answer the question	Resulting information/measurement
Timeliness	When was the list last updated?	Access to metadata and/or interview data source	Date (DD.MM.YYYY)) when the list was last updated
	Were all the data elements updated or only some of them?		List of data element updated during the last update
Validity	Are the data dictionary, metadata and classification tables provided with the list and complete?	Access to the data dictionary, metadata and classification tables for the list	Absence or incompleteness of the data dictionary, metadata and classification tables
	Are the values captured according to the format and standards captured in the data dictionary and classification tables?	Manual or pseudo automatic identification of records not matching the defined format/standards	Number and percentage of records not matching the defined format/standards for each data element
Consistency	Are inconsistencies observed between records for given data elements?	Manual or pseudo automatic identification of inconsistencies	Percentage of records presenting inconsistencies with the rest of the list
	When applicable, are there inconsistencies with other master lists?		Percentage of records presenting inconsistencies with the other master lists
Uniqueness	Does the list contain duplicate records? If yes, how many?	Manual or pseudo automatic identification of duplicates	Number of duplicates identified in the list
Completeness	Does the list contain all the geographic features currently active in the country?	Access to metadata and/or interview data source	List of geographic features not currently included in the list
	Does the list contain all the data elements included in the data dictionary of the master list?	Visual analysis of the list using the data dictionary of the master list as reference	List of data elements from the master list not currently included in the list
	Is the value for each data element available for all records in the list?	Manual or pseudo automatic identification of empty records	Percentage of missing values for each data element
Accuracy	Does the information captured in the list correspond to the reality?	Access to SOP used for data collection, random check, comparison between sources	Percentage of records checks for which the value does not match the reality
		Specific check for geographic coordinates	Percentage of records presenting unprecise and/or inaccurate geographic coordinates

A specific MS Excel template has been created to facilitate the conduct of the different analysis as well as capture the answer to each of the questions reported in Table A.1. This template can be downloaded from here: https://tinyurl.com/yypy8cm7. In this template, data quality dimensions are listed according to the order in which they should be assessed.

A separated copy of the template should be complete for each separated list even if these lists contain information about the same geographic feature.

The fake master list of health facilities for Atlantis accessed in September 2022 is used as an example on how to conduct the assessment and complete this template. The data dictionary, classification tables and metadata for this master list are included in Sub-Annex 1.1.

Before conducting the assessment, it is recommended to create a separated folder on your computer, folder in which you will place:

1. The different lists to be assessed together with their associated data dictionary, metadata and classification tables. These files should remain untouched during the whole exercise and this to allow to come back to the original list if needed.
2. The downloaded template.

Other files will be placed in that same folder during the assessment.

The next thing to be done is to capture the generic information about the first considered list in the *Summary* worksheet of the MS Excel template. For this:

1. Open the MS Excel template
2. Save the file under a name that will allow you to easily recognize to which list it refers to (example: *Assessment_HFML_Atlantis_01012023*)
3. Complete the Information about the list in the *Information about the list* section of the *Summary* worksheet. Here is an example of the result using the NHFR of the Philippines:

	A	B
1	**Information about the list**	
2		
3	Name of the list:	*Health facility master list of Atlantis*
4	Source (organization)	*Ministry of Health of Atlantis*
5	Source (URL)	https://moh.atlantis.gov/hfml
6	Publication/release date	*02-Sept-22*
7	Temporal validity	*Unknown*
8	Recognized as the master list (Yes/No/to be confirmed)?	*Yes*
9	Coverage (National/subnational)	*National*
10	Number of records included in the list	*1256*
11	Format	*MS Excel (.XLS)*

4. Indicate the date at which the assessment is being conducted on line 13 of the Summary worksheet. Example:

13	Assessment date:	*01-Jan-23*

Once this done, the process described in the next sections can be applied to answer the questions reported in Table A.1 for each data quality dimension for that list.

The same process would then be repeated for any other list included in the assessment.

Timeliness

The questions included in Table A.1. for timeliness might be answered by looking at the metadata associated to the list if available (example in Sub-Annex 1.1) or by contacting the entity that created it.

Two scenarios can occur:

1. The information is available, in which case, it can be entered in the *Summary* worksheet of the template. Examples:

Quality dimension	Question to be answered	Answer
Timeliness	When was the list last updated?	*January 2022*
	Were all the data elements updated or only some of them?	*Yes*

Quality dimension	Question to be answered	Answer
Timeliness	When was the list last updated?	*January 2022*
	Were all the data elements updated or only some of them?	*Only the hospitals contact information has been updated*

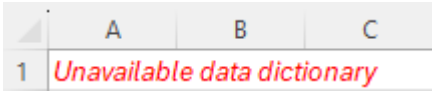
2. The information is unknown. In this case, unknown would be specified in the *Summary* worksheet. Example:

Quality dimension	Question to be answered	Answer
Timeliness	When was the list last updated?	*Unknown*
	Were all the data elements updated or only some of them?	*Unknown*

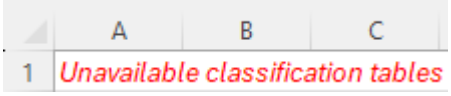
Validity

The first question for this data quality dimension (Are the data dictionary, classification tables and metadata provided with the list and complete?) is easily being answered by checking the availability of this information (examples in Sub-Annex 1.1) and then capturing the following in the list assessment MS Excel file:

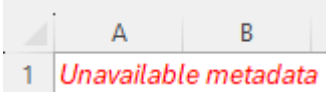
1. Data dictionary:
   a. If available, copy and paste it in the *Data dictionary* worksheet
   b. If not available, indicate *Unavailable data dictionary* in the *Data dictionary* worksheet:



2. Classification tables:
   a. If available, copy and paste these tables in the *Classification tables* worksheet
   b. If not available, indicate *Unavailable classification tables* in the *Classification tables* worksheet:



3. Metadata:
   a. If available, copy and paste the list metadata the *Metadata* worksheet
   b. If not available, indicate *Unavailable metadata* in the *metadata* worksheet:



4. Include a summary of the situation regarding the availability and completeness of the data dictionary, classification tables and metadata in the *Summary* worksheet. Examples:

Validity	Are the data dictionary, classification tables and metadata provided with the list and complete?	The data dictionary, classification tables and metadata are not available

Validity	Are the data dictionary, classification tables and metadata provided with the list and complete?	The data dictionary, classification tables and metadata are avaiable but the classification table for the position of health facility head is missing

Validity	Are the data dictionary, classification tables and metadata provided with the list and complete?	The data dictionary, classification tables and metadata are avaiable and provide all the necessary information to use the list

The second question for this data quality component (Are the values captured according to the format and standards from the data dictionary and classification tables?) concerns the following data elements and corresponding checks to be performed:

1. Unique identifier: Identify any records for which the unique identifier does not respect the coding scheme described in the data dictionary or any other documentation associated with the list.
2. Data elements which values come from a classification table (e.g. health facility type or ownership): Identify any records for which the value captured in the list does not correspond to one of the options included in the corresponding classification table.
3. Geographic coordinates: When the list contains geographic coordinates, identify those that are not captured according to the format of the coordinate system documented in the data dictionary/metadata and/or switched (latitude captured as longitude and vice versa).

The following process is to be implemented to perform and document each of these checks:

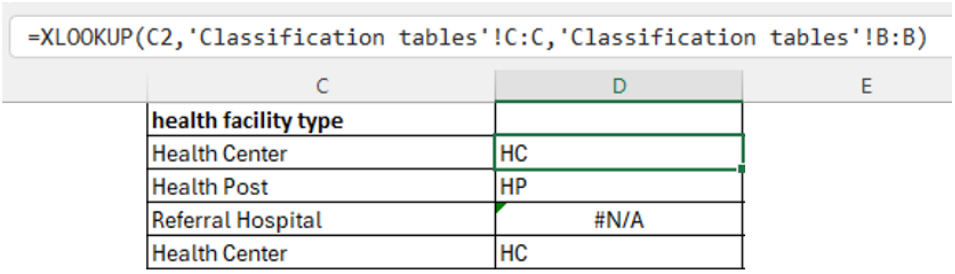
1. Copy and paste the content of the list being assessed in the *Validity* worksheet of the list assessment MS Excel file
2. If a unique identifier is attributed to the geographic objects in the list:
   a. Sort by alphabetical order the content of list according to the column containing the unique identifier
   b. Scroll down the sorted column to identify records for which the structure of the unique identifier is different from the defined coding scheme and color the cells in question. Example:

	A	B
1	**Unique id**	**health facility name**
2	HF506667	Thunukkai health center
3	92627822	Karuwalaga health post
4	HF347219	Analai referral hospital
5	HF573470	Blearla clinic

   Note: while records presenting a unique identifier different from the coding scheme might often find themselves on top or at the bottom of the sorted column this might not always be the case depending on the coding scheme being used
3. For the data elements based on a classification table, if such a table is available, check if the value captured for each record matches one of the options in the corresponding classification table using the following steps:
   a. Add one blank column on the right of each column containing one of the concerned data element. Example:

C	D
health facility type	
Health Center	
Health Post	
Referral Hospital	
Health Center	

b. Use the MS Excel XLOOKUP or VLOOKUP function (see Section 3.9 of Health GeoLab Guidance 2.5.2[2] for more information about these functions) to bring the code, acronym or description from the corresponding classification table in the column on the right of the analyzed data element (column D). Example:



C	D	E
health facility type		
Health Center	HC	
Health Post	HP	
Referral Hospital	#N/A	
Health Center	HC	

c. Highlight the cells containing the data elements for which the value returned by the XLOOKUP/VLOOKUP function is #N/A as these corresponds to values not matching the content of the classification table. Example:

C	D
health facility type	
Health Center	HC
Health Post	HP
Referral Hospital	#N/A
Health Center	HC

4. When the list contains geographic coordinates:
   a. Coordinates not captured according to the defined coordinates system:
      i. Look at the data dictionary and the metadata to identify in which coordinate system the geographic coordinates are supposed to be captured in the list:
         - At this stage in the process, we are only checking if they are captured in decimal degrees (±DD.DDDDD for the latitude and ±DDD.DDDDD for the longitude), in meters (±MMMMMMM.MM for the Northing and ±MMMMMMM.MM for Easting) or in Degrees, minutes and seconds (±DD° MM'SS.SSSS'' for the latitude and ±DDD° MM'SS.SSSS'' for the longitude)
         - If this information is not available in the data dictionary or the metadata, contact the source or look at how the coordinates are captured in the list)
      ii. Sort by increasing order the columns containing the geographic coordinates. The values expressed in decimal degrees will be on top of the list, those in meters at the bottom of it
      iii. Highlight the cells for which the coordinates are expressed in a coordinate system different from the one specified in the data dictionary/metadata. Example of a record for which the coordinates are expressed in meter while they should have been in decimal degrees:

K	L
Lat	**Long**
6.406941	80.331795
6.8568435	80.830729
6.879718	80.814321
76836.819	412512.15

   b. Switched geographic coordinates:
      i. Use the method reported in Section 3.6 of Health GeoLab guidance document 2.5.2[2] to identify these geographic coordinates
      ii. Highlight the cells for which the coordinates are switched. Example:

[2] https://healthgeolab.net/DOCUMENTS/Guide_HGLC_Part2_5_2.pdf

H	I
Lat	Long
6.406941	80.331795
6.8568435	80.830729
80.215206	6.9446736
6.879718	80.814321

5. Summarize the result of the different validity checks that have been performed in the *Summary* worksheet of the list assessment MS Excel file. Examples which could end up being combined into the cell in question depending on the result of this part of the assessment:

Validity	Are the values captured according to the format and standards from the data dictionary and classification tables?	*Yes, the values for all the data element match the format and standards from the data dictionnary and classification tables*
Validity	Are the values captured according to the format and standards from the data dictionary and classification tables?	*No, the health facility type information for 67 health facilities (9.5%) was not matching one of the options included in the corresponding classification table*
Validity	Are the values captured according to the format and standards from the data dictionary and classification tables?	*No, the geographic coordinates for 15 health facilities (2.1%) were not captured according to the coordinate system specified in the data dictionary*
Validity	Are the values captured according to the format and standards from the data dictionary and classification tables?	*No, the geographic coordinates for 7 facilities (1%) were switched*

The issues identified for the data elements capturing values based on a classification table as well as the switched coordinates and, if possible, coordinates captured in a different format, should be corrected in a copy of the original list and such a copy used to implement the rest of the process.

Consistency

Answering the first question for this data quality dimension (Are inconsistencies observed between records for given data elements?) is done by applying the following process:

1. Copy the version of the list being assessed that has been adjusted after checking its validity and paste it in the *Consistency* worksheet of the list assessment MS Excel file
2. For each of the data elements included in the list, except the geographic coordinates:
    a. Sort the list by alphabetical order according to the column containing the data element
    b. Scroll down the column to identify potential inconsistencies in the way the values are being captured (examples for a health facility master list: health facility name structured differently (type before or after the name, no type), Health facility type captured in full or as an acronym, district name captured in upper case or lower case)
    c. Highlight the cells containing the information captured in an inconsistent way. Examples:

B
health facility name
Thunukkai health center
HP Karuwalaga
ANALAI PROVINCIAL HOSPITAL
Blearla

3. Capture the result of this part of the assessment in the *Summary* worksheet of the list assessment MS Excel file. Examples:

Consistency	Are inconsistencies observed between records for given data elements?	*None*

Consistency	Are inconsistencies observed between records for given data elements?	*- Health facility name: 123 health facilities for which the name is captured in an inconsistent way (e.g. type stored as acronym, name without type, full name in upper case* *- Other data elements: None*

Answering the second question for this data quality dimension (When applicable, are there inconsistencies with other master lists?) applies generally to data elements capturing the unique identifier and name of the subnational unit (administrative, health, statistical, postal) in which the geographic feature is located and this across levels (e.g. 1st, 2nd, 3rd subnational level of the administrative structure).

Identifying this kind of inconsistencies requires having access to the master list for the subnational units as observed at the time of conducting this part of the assessment. If this is the case:

1. Sort the copy of the list already in the *Consistency* worksheet according to the columns containing the information about the structure of the subnational units you are assessing and this from the upper to the lower level. Example:

F	G	H	I
PRO_ID	**PRO_NAME**	**DIS_ID**	**DIS_NAME**
TLK01	Andustar	TLK0101	Andunie
TLK01	Andustar	TLK0102	Eldalond
TLK02	Forostare	TLK0201	Ondosto
TLK02	Forostar	TLK0202	Sorontil
TLK03	Hyarrostar	TLK0301	Nindamos
TLK03	Hyarrostar	TLK0308	Romenna
TLK04	Mittalmar	TLK0401	Armenelos

2. Add one blank column on the right of each column containing one of the concerned data element. Example:

F	G	H	I	J	K	L	M
PRO_ID		**PRO_NAME**		**DIS_ID**		**DIS_NAME**	
TLK01		Andustar		TLK0101		Andunie	
TLK01		Andustar		TLK0102		Eldalond	
TLK02		Forostare		TLK0201		Ondosto	
TLK02		Forostar		TLK0202		Sorontil	
TLK03		Hyarrostar		TLK0301		Nindamos	
TLK03		Hyarrostar		TLK0308		Romenna	
TLK08		Mittalmar		TLK0401		Armenelos	

3. Starting from the upper level in the structure of the subnational units, use the MS Excel XLOOKUP or VLOOKUP function (see Section 3.9 of Health GeoLab Guidance 2.5.2[3] for more information about these functions) to bring the official unique name of the subnational units from the corresponding master list in the blank columns that have been added using the unique identifier as the common data element. Example:

```
=XLOOKUP(F2,[Admin_divisions_ML_01012000.xlsx]County!$C:$C,[Admin_divisions_ML_01012000.xlsx]County!$D:$D)
```

F	G	H
PRO_ID		**PRO_NAME**
TLK01	Andustar	Andustar
TLK01	Andustar	Andustar
TLK02	Forostar	Forostare
TLK02	Forostar	Forostar
TLK03	Hyarrostar	Hyarrostar
TLK03	Hyarrostar	Hyarrostar
TLK08	#N/A	Mittalmar

[3] https://healthgeolab.net/DOCUMENTS/Guide_HGLC_Part2_5_2.pdf

4. Highlight the cell(s) containing the unique identifier of the subnational units for which the XLOOKUP/VLOOKUP function returned #N/A as these corresponds to unique identifiers that are not included in the master list. Example:

F	G	H
PRO_ID		**PRO_NAME**
TLK01	Andustar	Andustar
TLK01	Andustar	Andustar
TLK02	Forostar	Forostare
TLK02	Forostar	Forostar
TLK03	Hyarrostar	Hyarrostar
TLK03	Hyarrostar	Hyarrostar
TLK08	#N/A	Mittalmar

5. In the blank column on the right of the subnational unit's name from the assessed list use the IF function to identify potential difference of spelling between that name and the included in the master list. Example:

```
=IF(G2=H2,"",1)
```

F	G	H	I
PRO_ID		**PRO_NAME**	
TLK01	Andustar	Andustar	
TLK01	Andustar	Andustar	
TLK02	Forostar	Forostare	1
TLK02	Forostar	Forostar	
TLK03	Hyarrostar	Hyarrostar	
TLK03	Hyarrostar	Hyarrostar	
TLK08	#N/A	Mittalmar	#N/A

6. Highlight the cell(s) containing the name of the subnational units for which the IF function returned the value 1 as these corresponds to names that are spelt differently in the master list. Example:

G	H	I
	PRO_NAME	
Andustar	Andustar	
Andustar	Andustar	
Forostar	Forostare	1
Forostar	Forostar	
Hyarrostar	Hyarrostar	
Hyarrostar	Hyarrostar	
#N/A	Mittalmar	#N/A

   Note: the comparison for the records presenting difference in unique ID (steps 4 and 5) will have to be repeated once this information has been adjusted in the list being assessed as the unit name from the master list would now appear in column G
7. Repeat steps 3 to 6 for the other levels in the structure of subnational units being assessed
8. Capture the result of this part of the assessment in the *Summary* worksheet of the list assessment MS Excel file. Examples:

Consistency	When applicable, are there inconsistencies with other master lists?	None

Consistency	When applicable, are there inconsistencies with other master lists?	- Province unique ID: the unique ID for 1 of the provinces included in the list do not exist in the master list - Province name: the name for 1 of the provinces included in the list does not macth the one used in the master list - Other level: None

The inconsistencies identified during this part of the assessment should ideally be corrected in a new copy of the list that will contain these adjustments.

Performing these adjustments might not necessarily be straightforward as the list being assessed might contain information that has not been updated for some time. When this is the case attributing the correct unique identifier, or even name, would require to first identify potential changes that have occurred in the structure of the considered subnational units since the last

update of this information in the list being assessed. Historic changes, such as the ones captured in the context of the implementation of the Second Administrative Level Boundaries (SALB) programme[4] for example, can help in this regard.

The copy of the list resulting from these adjustment should then be used to implement the rest of the process.

Uniqueness

Identifying potential duplicates is the next part of the assessment to be conducted as this might have an impact on the total number of records to be considered for the remaining data quality dimensions.

To identify potential duplicates:

1. Copy the content of the adjusted list resulting from the consistency part of the assessment and paste it in the *Uniqueness* worksheet of the MS Excel template
2. Identify the data element(s) that can be used to identify potential duplicates. In most cases, several data elements will have to be used to identify real duplicates (when assessing a list of health facilities for example, the name of the facilities together with their location in the administrative structure and their geographic coordinates could be used).
3. Use one of the two approaches described in Section 3.8 of Health GeoLab's guidance 2.5.2[5] to identify duplicate values for the different data elements selected under step 2. Example
4. Sort the content of the list according to the data elements selected under step 2 and this to have the records presenting duplicate values on top of it. Example (sorting based on the cell color):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	HF_ID	HF_NAME	HF_T	HF_OWN	HF_OWN_G	Reg_Code	Reg_Name	Pro_Code	Pro_Name	Mun_Code	Mun_Name	LAT	LONG	SOUR_COOR
2	HF110002	Guilopan Health Post	Health Post	Government	DOH	PH130000000	Northern Land	PH139900000	Tolkien	PH139914000	Guilopan	14.500430	121.039020	DOH (GPS)
3	HF110023	Guilopan Health Center	Health Center	Government	DOH	PH130000000	Northern Land	PH139900000	Tolkien	PH139914000	Guilopan	14.500430	121.039020	DOH (GPS)
4	HF110013	San Juan Hospital	Hospital	Government	DOH	PH130000000	Northern Land	PH139900000	Tolkien	PH139916000	Tolkien City	14.412830	121.033090	DOH (GPS)
5	HF110001	Tolkien Provincial Hospital	Provincial hospital	Government	DOH	PH130000000	Northern Land	PH139900000	Tolkien	PH139916000	Tolkien City	14.414290	121.044060	DOH (GPS)

5. Identify potential duplicates among the records finding themselves at the top of the list. In the example illustrated under step 4 it seems that the Guilopan health facility might have been captured twice, once when it was a health post and once after being upgraded as a health center.
6. Capture the number of identified duplicates in the Summary worksheet of the list assessment MS Excel file. Examples:

Uniqueness	Does the list contain duplicate records? If yes, how many?	*No*

Uniqueness	Does the list contain duplicate records? If yes, how many?	*Yes, 12 potential duplicates have been identified*

The potential duplicates should ideally be confirmed and, if needed, removed from the list before conducting the rest of the assessment.

[4] https://salb.un.org/en
[5] https://healthgeolab.net/DOCUMENTS/Guide_HGLC_Part2_5_2.pdf

Completeness

Answering the first question for this data quality dimension (Does the list contain all the data elements included in the data dictionary?) requires to either have access to a complete master list or to be in contact with the governmental entity having the curation mandate over that same master list.

Two scenario can occur:

1. The information is available. When this is the case, the answer can be captured in the Summary worksheet. Examples:

Completeness	Does the list contain all the geographic features currently active in the country?	*Yes, the list contains all the active health facilities as of January 1, 2023*

Completeness	Does the list contain all the geographic features currently active in the country?	*No, new health facilities have been opened since the last update in January 2022*

2. The information is unknown. In this case, unknown would be specified in the *Summary* worksheet. Example:

Completeness	Does the list contain all the geographic features currently active in the country?	*Unknown*

The second question (Does the list contain all the data elements included in the data dictionary of the master list?) is being answered by comparing the data elements included in the data dictionary of the master list with those included in the list being assessed.

Different scenario can occur:

1. The data dictionary of the master list has not yet been defined. In this case, the *Summary* worksheet would be completed as per the following example:

Completeness	Does the list contain all the data elements included in the data dictionary of the master list?	*The data dictionary of the master list has not yet been defined. The present list contains the following data elements: health facility name, health facility type, health facility address, province and district in which the health facility is located as well as geographic coordinates in decimal degrees*

2. The data dictionary of the master list has been defined and is accessible (example in Sub-Annex 1.1). In this case:
    a. Create a table containing the correspondence between the data elements included in the master list and those included in the list being assessed (example in Sub-Annex 1.2). When doing this, it is crucial to ensure that the data elements being match do indeed contain the same information. The final table should be included in the *Completeness – data elements* worksheet of list assessment MS Excel file
    b. Based, complete the corresponding cell in the *Summary* worksheet. Example based on the table included in Sub-Annex 1.2):

Completeness	Does the list contain all the data elements included in the data dictionary of the master list?	*The list contains information for 11 of the data elements included in the master list data dictionary (47.8%)*

Answering the third question (Is the value for each data element available for all records in the list?) requires to perform the following in MS Excel:

1. Copy the content of the list resulting from the uniqueness part of the assessment and paste it in the *Completeness - records* worksheet of the MS Excel template

2. Identify the number of empty records for each data element included in the list either by:
    a. Sorting the content of each column by alphabetical order and counting the number of empty cells appearing at the bottom of that same column
    b. Using the MS Excel's COUNTIF function (http://tinyurl.com/3jmadh2b) to identify how many records are empty (Note: for the result to be correct with this method, the empty cells should not even contain a space). The complete formula in this case would look like this for column A with 1,256 records + the header: =COUNTIF(A2:A1257,"")
3. Measure the percentage of records without value for each data element by dividing the number of empty cells obtained in step 2 by the total number of records to obtain a percentage of missing values by data element
4. Capture the resulting percentages in a table like this one in the *Completeness - records* worksheet of the list assessment MS Excel file:

Data element in the assessed list	Percentage of missing values
Health Facility Code	0%
Facility Name	0%
Health Facility Type	0%
Ownership Major Classification	10%
Street Name and #	63%
Province code	0%
Province name	0%
Latitude	5%
Longitude	4%
Phone number	57%
Email address	70%

5. Capture a summary of the results in the Summary worksheet of the list assessment MS Excel file. Examples

Completeness	Is the value for each data element available for all records in the list?	*Yes*

Completeness	Is the value for each data element available for all records in the list?	*No, values are missing for 6 data elements. The percentage of missing values for the other data elements varies between 4 and 70%*

Accuracy

There are two parts to the question for the accuracy data quality dimension (Does the information captured in the list correspond to the reality?):

1. The first part consists in going through the available documentation describing how the content of the list has been generated/updated or, if the necessary resources are available, to perform a random check on a sampled number of records remotely (e.g. contacting someone having the necessary local knowledge) or onsite (field data verification). Another option is to compare the content of the list being assessed with other reliable sources of information, starting with the master list when available. The result of this exercise is to be captured as follows in the *Summary* worksheet:
    a. If no documentation is available, that resources to perform random check were not accessible and that no other reliable source of information currently exists for that geographic feature:

Accuracy	Does the information captured in the list correspond to the reality?	*Unkown for the data elements outside the geographic coordinates due to the lack of documentation, resources to perform random checks and other reliable sources of information*

b. If it has been possible to perform some accuracy check using one or several of the approaches mentioned here above. A summary of the result of this exercise is then to be included in the *Summary* worksheet of the list assessment MS Excel file. Example:

Accuracy	Does the information captured in the list correspond to the reality?	*A random check has been performed over 100 health facilities and allowed to identify inaccuracy in the name of the health facility for 25 records as well as in the name of the districts in which the health facility is located for 15 records*

2. The second part, when applicable, consists in assessing the precision and accuracy of the geographic coordinates included in the list:
    a. Implement the process detailed in Sub-Annex 1.3
    b. Capture a summary of this part of the assessment in the *Summary* worksheet. Example:

Accuracy	Does the information captured in the list correspond to the reality?	The geographic coordinates for: - 135 health facilities present a precision level higher than the meter - 25 health facilities fall outside the province in which they are indicated to be located in the list and outside the district for 34 health facilities - 56 health facilities fall outside a built up area

The final content of the *Summary* worksheet together with the information (data dictionary, classification tables, metadata) and detailed analysis captured in the other worksheets constitute the result of the quality assessment conducted on the list.

Sub-Annex 1.1 – Data dictionary and metadata of the Atlantis health facility master list

Data dictionary

Data element group	Data element label	Data element description
Uniquely identify	HF_ID_N	Official national unique identifier of the health facility (coding scheme structure: HFXXXXXX)
	HF_N_RO	Official complete name of the health facility (Romanized)
	HF_N_LOC	Official complete name of the health facility (Atlantean language)
Classify	HF_T_RO	Health facility type (Romanized)
	HF_T_LO	Health facility type (Atlantean language)
	HF_OWN_T	Type of organization having the ownership or managing authority (government, private, other)
	HF_OWN_O	Full name of the organization owning or managing the health facility (Example: Ministry of Health, Ministry of interior,..)
	STATUS	Health facility status (under construction, open, temporarily closed, etc.)
Locate	HF_ADD	Street number and name
	PRO_C	Official unique identifier of the Province in which the facility is located
	PRO_N_RO	Official name of Province in which the health facility is located (Romanized)
	PRO_N_AT	Official name of the first subnational level administrative unit in which the health facility is located (Atlantean language)
	DIS_C	Official unique identifier of the District in which the facility is located
	DIS_N_RO	Official name of the District in which the facility is located (Romanized)
	DIS_N_AT	Official name of the District in which the facility is located (Atlantean language)
	LAT	Latitude of the health facility in decimal degrees (EPSG 4326)
	LONG	Longitude of the health facility in decimal degrees (EPSG 4326)
	S_COOR	Source and method used to obtain the geographic coordinates of the health facility (including unknown)
	AC_COOR	Qualitative measure of the accuracy level for the geographic coordinate
Contact	HEAD_N	Full name if the health facility head
	HEAD_POS	Position of the Head of Facility head
	LAND_NBR	Health facility landline telephone number
	EMAIL	Email address of the health facility

Classification tables

Health facility type

Health Facility Type Code	Acronym	Health Facility Type	Description
T1	NH	National Hospital	Health facility used to train health personnel and undertake research studies, in addition to providing specialised referral services.
T2	PH	Provincial Hospital	Health facility providing advanced services to which patients are referred to if they cannot be treated at the health center level
T3	HC	Health Center	Health facility delivering primary health care services
T4	HP	Health Post	Health facility located in remote areas and function as the first point of contact with the population in low population density districts

Health facility ownership

Ownership Code	Acronym	Health facility owner	Description
O1	MOH	Ministry of Health	Health facilities managed by the Ministry of Health
O2	MOI	Ministry of Interior	Health facilities managed by the Ministry of Interior
O3	MOD	Ministry of Defense	Health facilities managed by the Ministry of Defense

Metadata

Title:	Health facility master list of Atlantis
Originator:	Ministry of Health of Atlantis
Publication date:	02-sep-2022
Temporal validity	Unknown
Abstract:	This master list has been created and is being maintained by the Department of Planning (DP) as the officially curated master list of public health facilities in Atlantis
Process:	The master list has been established by combining, organizing and cleaning information stored in different tables.
Progress:	Ongoing
Access constraints:	This data is publicly accessible for non-commercial use
Use constraints:	The use of this data is limited to non-commercial use. Users are encouraged to inform Department of Planning if they discover any error or would have information that would allow to complete or update the master list
Acknowledgments	Department of Planning, Ministry of Health of Atlantis
Disclaimer:	This dataset is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the data lies with the user. In no event shall the Ministry of Health of Atlantis be liable for damages arising from its use.
Primary Contact	
Contact Name	Gadeirus Azaes
Organization	Ministry of Health/Department of Planning (DP)
Contact Telephone number:	+820 77436515
Contact Email Address:	gaseirus.azaes@gmail.com

Sub-Annex 1.2 – Example of correspondence between the data elements from the master list data dictionary and the data elements included in the assessed list

Data element in the master list	Data element in the assessed list
HF_ID_N	*Health Facility Code*
HF_N_RO	*Facility Name*
HF_N_LOC	*Not available*
HF_T_RO	*Health Facility Type*
HF_T_LO	*Not available*
HF_OWN_T	*Ownership Major Classification*
HF_OWN_O	*Not available*
STATUS	*Not available*
HF_ADD	*Street Name and #*
PRO_C	*Province code*
PRO_N_RO	*Province name*
PRO_N_AT	*Not available*
DIS_C	*Not available*
DIS_N_RO	*Not available*
DIS_N_AT	*Not available*
LAT	*Latitude*
LONG	*Longitude*
S_COOR	*Not available*
AC_COOR	*Not available*
HEAD_N	*Not available*
HEAD_POS	*Not available*
LAND_NBR	*Phone number*
EMAIL	*Email address*

Sub-Annex 1.3 – Process to assess the precision and accuracy of geographic coordinates

This annex describes the process to follow to assess the precision and accuracy of the geographic coordinates included in the list being assessed.

This process requires to have:

2. Basic skills in the management and use of geospatial data and GIS software (e.g. QGIS or ArcMap)
3. Access to geospatial datasets containing:
    a. The boundaries of the subnational units (administrative, health, statistics, postal) for which the information is being captured in the list being assessed. The content of this geospatial dataset should match the content of the corresponding master list.
    b. Building footprints (e.g. Open Buildings: https://sites.research.google/gr/open-buildings/#open-buildings-download)

Once the above available, and before starting this part of the quality assessment, there is a need to know in which coordinate system the geographic coordinates are being captured in the list:

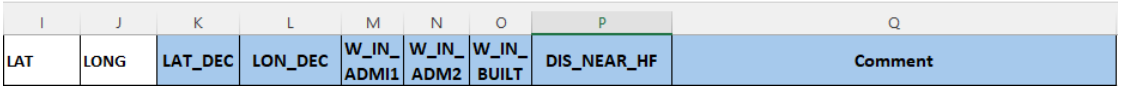
For this, use the metadata, or data collection protocol, associated with the source of the geographic coordinates, if any, to identify the coordinates system. Different scenarios can occur:

1. The information about the coordinate system is not available even after contacting the source. In this case:
    a. If the coordinates are captured in a decimal degree-looking format (±DD.DDDDD for the latitude and ±DDD.DDDDD for the longitude), display these coordinates on top of satellite imagery in a GIS software and zoom to different areas to identify if their location is plausible (falls on top of a building that looks like a health facility):
        - If this is the case, continue the process considering that these coordinates are captured in decimal degrees (WGS 84, EPSG: 4326)
        - If this is not the case, it is better not to conduct the assessment as these coordinates might be captured in degrees, minutes and seconds (±DD° MM'SS.SSSS'' for the latitude and ±DDD° MM'SS.SSSS'' for the longitude)
    b. If the coordinates are captured in a metric-looking format (±MMMMMMM.MM for the Northing and ±MMMMMMM.MM for Easting) then the assessment should not be implemented as there is too much uncertainty regarding the coordinate system in which they are being captured.
2. If the coordinate system is known then the assessment can take place

Once the coordinate system is known:

1. Copy the content of the copy of the list resulting from the uniqueness part of the assessment in the *Accuracy – Coordinates* worksheet of the assessment MS Excel file
2. If the coordinates are projected (not captured in decimal degrees), they will first have to be unprojected (WGS84, EPSG: 4326) using a GIS software before continuing the rest of the process
3. Add the following set of column on the right of those containing the geographic coordinates captured in decimal degrees:
    a. LAT_DEC: To capture the number of digits after the decimal point for the latitude as a measure of precision

b. LON_DEC: To capture the number of digits after the decimal point for the longitude as a measure of precision
c. W_IN_ADM1: To document if the coordinates are falling within the boundaries of the 1st subnational level administrative unit (e.g. province) in which the geographic feature is indicated to be located in the list
d. W_IN_ADM2: To document if the coordinates are falling within the boundaries of the 2nd subnational level administrative unit (e.g. district) in which the geographic feature is indicated to be located in the list
e. W_IN_BUILT: To document if the coordinates are falling within a built up area



I	J	K	L	M	N	O	P	Q
LAT	LONG	LAT_DEC	LON_DEC	W_IN_ ADMI1	W_IN_ ADM2	W_IN_ BUILT	DIS_NEAR_HF	Comment

Notes:

- When applicable, a column could be added to document if the coordinates are falling within the health unit (e.g. health district) in which the geographic feature (e.g. health facility) is indicated to be located in the list
- In countries where Google Street View presents a good coverage and where the geographic features can be easily identified from the street (e.g. board at the entry of health facilities), an additional column can be added to capture the result of this test or even replace the one aiming at checking if the coordinates are falling within a built up area (W_IN_BUILT column).
- Depending on the geographic feature, a column could also be included to capture the distance between the geographic coordinates of the geographic feature and the nearest geographic feature (DIS_NEAR_HF in the above screenshot). Set of coordinates that are too close to each other while this is not the case in the reality (e.g. vaccination points) could be an indication of an issue with the coordinates of one if not both features.
- We generally don't perform any check below the 2nd level of the country's administrative structure unless we are sure about the accuracy of the administrative boundaries GIS layer below that level.

4. Perform the following test on the geographic coordinates of each record and capture the result of each test in the corresponding column(s):
   a. <u>Coordinates precision</u>: follow the process described in Section 3.7 of Health GeoLab guidance 2.5.2[4] to capture the number of digits after the decimal points for both the latitude and longitude in the LAT_DEC and LON_DEC columns (note: conditional formatting can be used to highlight coordinates presenting 5 or more (green), 4 (yellow) and less than 4 digits (red)). Example:

I	J	K	L
LAT	LONG	LAT_DEC	LON_DEC
9.117	80.268	3	3
8.4855	79.929	4	7
9.615398	79.9705	6	6
9.52547	79.69244	5	5

   b. <u>Coordinates falling outside the correct subnational unit:</u>
      i. In a GIS software:
         - Upload the content of the list being assessed and convert it into a geospatial data layer (e.g. shapefile)

- Upload the geospatial dataset containing the boundaries of the subnational unit for which the check is being performed
- Identify if the geographic coordinates of each geographic feature included in the list are falling within the boundaries of the subnational unit (administrative, health, statistics, postal) mentioned in the list being assessed and this across levels. This can be done either visually or by using functions like the *Join data from another layer based on spatial location* tool in ArcMap or the *Join Attributes by Location* tool in QGIS.

ii. Capture the result of this exercise in the corresponding columns of the *Accuracy – Coordinates* worksheet. Example (columns M and N):

I	J	K	L	M	N
LAT	LONG	LAT_DEC	LON_DEC	W_IN_ ADMI1	W_IN_ ADM2
9.66607	79.76224	5	5	No	No
7.013971	80.92688	6	6	No	No
6.90401	80.19911	5	8	No	No

Notes:

- In these columns, a Yes means that the coordinates are falling within the unit in question, a No that they don't
- Conditional formatting can be used to highlight the cells in these column in green for Yes and red for No to better visualize those presenting issues

ii. When the geographic coordinates are not falling within the concerned boundaries, indicate in the *Comment* column:

- The distance between geographic coordinates and the boundary of the unit in which they are supposed to fall measured using the measuring distance tool in the GIS software
- Any additional information that can help understand the issue with the coordinates. Example:

M	N	Q
W_IN_ ADMI1	W_IN_ ADM2	Comment
No	No	In the sea; 441 meters away from the correct District boundary
No	No	1324 meters from the correct Province and District boundaries
No	No	17 meters away from the correct Province and District boundaries

Notes: As we can see from the example illustrated here above, sometimes the issue is not with the coordinates but with the accuracy of the subnational unit boundary layer

c. <u>Coordinates falling outside a builtup area</u>:

i. In a GIS software:

- Upload the geospatial dataset containing the location of the geographic features coming from the list being assessed and created during the previous step
- Upload the geospatial dataset containing the building footprints
- Convert the building footprint into points using the appropriate GIS software tool to generate polygon centroids (e.g., Feature To Point

tool in ArcMap or Centroids tool in QGIS). This will reduce the size of the shape file.

- Use the GIS software tool allowing to calculate the distance (in meters) between each geographic feature and the nearest point from the building footprint geospatial dataset (e.g., Near (Analysis) tool in ArcMap or Distance Matrix tool in QGIS).
- Check the result of the distance calculation: geographic features that are within 50 meters of a point are considered as located in a built-up area.
- For geographic features presenting a distance above 50 meters, use satellite imagery as basemap in the GIS software to identify If the geographic feature finds itself within buildings that are not captured in the building footprint geospatial dataset:
    - If this is the case, the geographic feature is considered as being located within a built-up area
    - If this is not the case, then it is considered as being outside a built-up area

ii. Capture the result of this exercise in the corresponding column of the *Accuracy – Coordinates* worksheet. Example (column O):

I	J	K	L	M	N	O
LAT	LONG	LAT_DEC	LON_DEC	W_IN_ ADMI1	W_IN_ ADM2	W_IN_ BUILT
7.013971	80.92688	6	6	No	No	Yes
6.90401	80.19911	5	8	No	No	Yes
6.406941	80.3318	6	6	No	No	No
9.615398	79.9705	6	6	Yes	Yes	No
9.52547	79.69244	5	5	Yes	Yes	No
7.785583	80.47506	9	5	Yes	Yes	No
6.144636	80.65243	7	7	Yes	Yes	No

iii. When the geographic coordinates are not falling within a built-up area, indicate in the *Comment* column:

- The distance between the geographic coordinates and nearest building measured using the measuring distance tool in the GIS software
- Any additional information that can help understand the issue with the coordinates. Example:

O	Q
W_IN_ BUILT	Comment
Yes	1324 meters from the correct Province and District boundaries
Yes	17 meters away from the correct Province and District boundaries
No	233 meters away from the correct Province and District boundaries; 73 meters away from nearest building
No	185 meters away from nearest building; In an open field
No	190 meters away from nearest building; In a lake
No	200 meters away from nearest building; In a forested area
No	204 meters away from nearest building; In a forested area

Annex 2 – Question to be answered during the assessment for geospatial data

Quality dimension	Applicability		Questions to be answered	Method to answer the question	Resulting information/ measurement
	Vector format	Raster format			
Timeliness	X	X	What is the temporal representativity of the dataset?	Access to metadata and/or interview data source	Date or period of validity matching or not the data specifications
Completeness	X		With master list: Does the geospatial data contain all the geographic objects contained in the master list?	Compare the content of the geospatial data with the content of the master list	% of geographic objects from the master list missing in the geospatial data
			Without master list: Does the geospatial data contain all the features observed on the satellite images used as ground reference?	Visually assess the level of completeness using satellite imagery as ground reference	Estimated % of missing geographic objects
Uniqueness	X		With master list: Does the geospatial data contain duplicates based on the master list?	Compare the content of the geospatial data with the content of the master list	% of identified duplicates
			Without master list: Does the geospatial data contain duplicates that can be identified based on the content of the attribute table and/or geographic location or extent?	Visually check content of attribute table as well as location or geographic extent	% of identified duplicates
Accuracy	X		Is the scale at which the geospatial data has been created matching the one defined in the data specifications?	Access to metadata and/or interview data source, SOP used for data creation	Difference in scale been the geospatial data and the data specifications
			Are the geographic objects in the geospatial data located with the expected positional accuracy defined in the data specifications?	Visually assess the level of accuracy using satellite imagery as ground reference; access to SOP used for data collection, random check, comparison between sources,	Estimated % of geographic objects that are not located with the expected horizontal accuracy
		X	Is the resolution of the geospatial data matching the one defined in the data specifications?	Check the properties of the geospatial data	Difference in resolution between the geospatial data and the data specifications
Validity	X	X	Is the metadata for the geospatial data available ?	Access to metadata and/or interview data source	Availability of metadata
			Are the geographic coordinate system and map projection known?	Access to metadata and/or interview data source	Availability of projection information
			Is the geospatial data available in a format that is compatible with the ones defined in the data specification or can it be converted accordingly	Check the data format and/or interview data source if unknow	Compatibility of format
			Does the geospatial data cover the study area as defined in the data specifications?	Visually assess the coverage of the geospatial data using the satellite images as ground reference	% coverage of the study area
Consistency	X		Are inconsistencies observed between records in the attribute table?	Manual or pseudo automatic identiication of inconsistencies in the attribute table	Description of the inconsistencies that have been observed