HEALTH GEOLAB COLLABORATIVE

Supporting countries to use geospatial data and technologies to achieve SDG 3

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 1.2 (last update: 10.01.2022)

In collaboration and with the support of:

ADB

wellcome trust MORU Tropical Health Network

UNIVERSITY OF OXFORD

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 HGLC guidance documents and improvement of section 5 based on recently published guides	Izay Pantanilla, Steeve Ebener

Authors

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

1. Health GeoLab Collaborative, 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) and the World Health Organization (WHO) for the support provided to the Health GeoLab Collaborative. MORU is funded by Wellcome.

Table of Contents

Purpose and audience

The purpose of the Health GeoLab Collaborative series of guidances 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
HGLC	Health GeoLab Collaborative
HIS	Health Information System
MORU	Mahidol-Oxford Tropical Medicine Research Unit
SDG	Sustainable Development Goal
WHO	World Health Organization

1. Background

The Health GeoLab Collaborative (HGLC) is a collective of institutions and individuals sharing the same vision - for low- and middle-income countries in Asia and the Pacific 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 HGLC uses the 4Ts (Training, Tooling, Testing, and Teaming) approach to strengthen in-country capacity. The present guidance has been developed as part of this approach and with the objective to be used by the largest number of users possible.

This volume is part of a series of guidances started under the umbrella of the AeHIN GIS Lab and now continued by the HGLC. 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 vocabulary, the data set 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 at info@healthgeolab.net 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 Collaborative: 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.

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.

[1] www.healthgeolab.net

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 in order 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 of 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 divisions 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 in order 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.

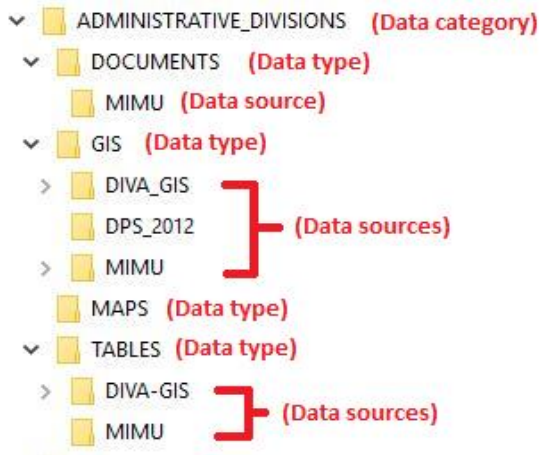

Figure 1. Example of folder organization structure

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 divisions, 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 is compiled and organized, it is important to assess it in order to identify:

- If you have been able to find at least one source for each of the data needed at the beginning of the process [2]
- Which source(s) are of best quality across the dimensions documented in the data specifications (timeliness, uniqueness, completeness, validity, accuracy, and consistency) and the use of the defined ground references [3]

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

When it comes to master lists, Table 7 implementation support guide for the development of a national georeferenced community health worker master list hosted in a registry contains the questions that the assessment should answer [11].

Annex 1 does itself contain 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 potential remaining gaps will have to be taken into account 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 the use of data that was actually not meant to shared in the first place but using data without knowing its use and redistribution rights might end up being more problematic than using other sources of lower quality but for which these information are known.
- The gaps in authoritative data from the government might be too big compared to other sources for them be considered

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

- Timeliness: What is the temporal validity of the statistical data? Does it match the temporal period of validity set in the data specifications?
- Uniqueness: Does the dataset contain duplicates?
- Validity:
    - Is the dataset accompanied by a metadata and does it contain all the information necessary to use, including data dictionary?

- Is the dataset available in a format that allows its use or could be converted in such format?

It is possible that, at the end of the assessment, none of the data that has been compiled presents a quality sufficient enough 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 Collaborative document: https://www.healthgeolab.net/DOCUMENTS/Guide_HGLC_Part1.pdf [Accessed 10 January 2022]

[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 Collaborative document: https://www.healthgeolab.net/DOCUMENTS/Guide_HGLC_Part2_1.pdf [Accessed 10 January 2022]

[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 vocabulary, the data set specifications, and the ground reference. Health GeoLab Collaborative document: https://www.healthgeolab.net/DOCUMENTS/Guide_HGLC_Part2_2.pdf [Accessed 10 January 2022]

[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 Collaborative document: https://healthgeolab.net/DOCUMENTS/Guide_HGLC_Part2_4_1.pdf [Accessed 10 January 2022]

[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 Collaborative document:

https://www.healthgeolab.net/DOCUMENTS/Guide_HGLC_Part2_4_2.pdf [Accessed 10 January 2022]

[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 Collaborative document: https://www.healthgeolab.net/DOCUMENTS/Guide_HGLC_Part2_5_1.pdf [Accessed 10 January 2022]

[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 Collaborative document: https://healthgeolab.net/DOCUMENTS/Guide_HGLC_Part2_5_2.pdf [Accessed 10 January 2022]

[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 Collaborative document: https://www.healthgeolab.net/DOCUMENTS/Guide_HGLC_Part2_6_1.pdf [Accessed 10 January 2022]

[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 Collaborative 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 Collaborative document.

[11] CHAI, CHIC, HGLC, 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 10 January 2022]

Annex 1 – 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 (Table AK): 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 (Table AK): 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 (Table AK): 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 (Table AK): 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