

Moving Towards Global Spatial Data Infrastructure: Improving Data Integration at the Global Level

Steeve Ebener
World Health Organization (WHO)
Switzerland
ebeners@who.int

Carrie Stokes
United States Agency for International Development (USAID)
USA
cstokes@usaid.gov

Kate Lance
International Institute for Geo-Information Science and Earth Observation (ITC)
The Netherlands
lance@itc.nl

Carmelle J. Terborgh
ESRI
USA
cterborgh@esri.com

Abstract

There is general agreement in the international community about the need for establishing a Global Spatial Data Infrastructure (GSDI) to ensure interoperability, enhance return on investment, and improve geospatial data sharing among the geographic information user community. However, to realize such a global SDI, the international community needs to effectively adopt common practices and geospatial data standards.

This paper therefore proposes a framework with suggested practices and standards that would improve data integration, data quality, and sustainability at the global level. If applied by the geographic community around the world, this framework would maximize investments and minimize redundant creation of geospatial information for use in many areas such as public health, disaster response, land use management, urban planning, agriculture, forestry, education, and other important sectors needed for social and economic development.

The opinions expressed in this paper are solely those of the authors and do not represent official policy of their respective organizations.

Keywords: data integration, spatial data infrastructure, standards,

Introduction

The amount of geospatially referenced information available is growing significantly and is reaching larger audiences than ever before through visualization tools such as Google Earth. While users benefit from having multiple data sources, and data redundancy may contribute to better quality, there still is considerable need for improving the integration of data originating from different sources.

Over the past several years, numerous workshops dealing with geospatial information have concluded that common specifications should be adopted and a consistent data framework should be developed (e.g., Lance and Gavin, 2003; Highland Surveyors Licensed Land Surveyors & Geomatic Consultants, 2003; ICA, 2003). Yet while there is a strong consensus for interoperability of geospatial information technology and sharing of data, a significant void still exists for achieving efficient integration of disparate geospatial datasets. In other words, efforts to achieve interoperability between geographic information systems (GIS) has not necessarily led to the interoperability of the geospatial data generated by and for use in the different systems.

The challenge, then, is for producers of geospatial information to ensure that their data are compatible with data from other sources. To do so, this requires:

- applying a set of internationally agreed upon standards regarding scale, projection, and levels of accuracy & precision as specific data collection protocols (e.g. regarding the use of GPS devices); and
- proper documentation about how data was created (assumptions & methods behind both geometric and attribute data)
- proper documentation about when the data was collected

The wide community of geospatial information producers and users needs to build a common understanding and a consensus regarding the processes, protocols, practices, and standards that can be applied to ensure the integration of data originating from different sources. This paper makes a first step in establishing such a common understanding. It is intended to advance the international dialogue about global data interoperability and data integration. Specifically, it proposes a data production process, or “data production chain”, with specific steps to follow for those in the geospatial community dealing with this issue.

The framework

The proposed framework outlines a set of selected or proposed processes, protocols, practices, and standards for use within the data production chain. By data production chain we understand the logical succession of steps that drives to the creation of a specific data set (Figure 1).

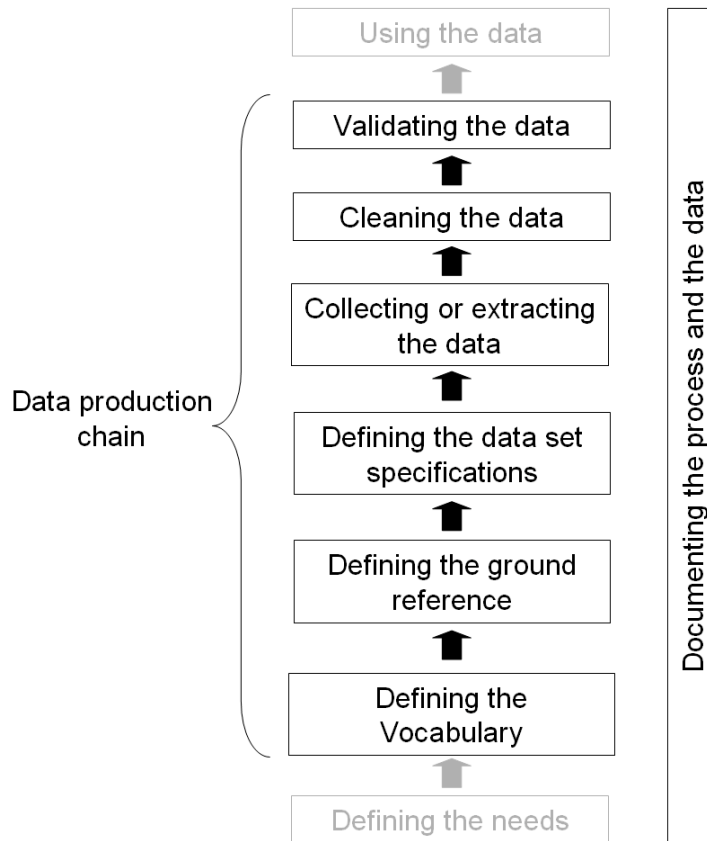


Figure 1 - Proposed data production chain

In order to be operational globally, this framework must remain independent of the level of application (local, national, regional, global) and from any specific technology. If applied, it will allow for horizontal and vertical integration of globally important foundation layers such as roads, hydrology, administrative boundaries, populated places, population distribution, land cover, and elevation, with more detailed level geospatial information covering villages, health facilities, schools, and neighborhoods, over time.

The proposed protocols, process, practices, and standards coming from different sources around the globe are described in the following sections of this paper according to the order of the steps shown in the data production chain (Figure 1). The aspects of data visualization and data sharing, steps that would typically appear in the continuity of the data production chain, are not addressed in this paper as they have already received considerable attention by the geospatial community are being addressed. However, the important issue of data integration has been poorly addressed.

Defining the vocabulary - geographic terminology

With an increasing number of global geospatial data consumers, rather than just a few highly specialized and technical producers, there is an increased need among the growing geospatial information user community to clearly communicate with each other. Among the resources that cover technical terms related to geographic information systems (GIS), online GIS dictionaries or glossaries from three organizations stand out: the Australian and New Zealand Land Information Council (ANZLIC), the Association for Geographic Information (AGI), and ESRI. A review of the definitions that these three respected resources provide for similar terms reveals differences, demonstrating the need to agree on common geospatially-related terminology. To minimize confusion and maximize data integration between GIS project investments, the use of standard terminology when referring to geospatial data and issues is therefore essential. *The SDI Cookbook* underscores the importance of this topic by dedicating a full chapter on the development and management of terminology in the field of Geographic Information (Jones, 2004).

The International Standards Organization (ISO) 19104 standards on Geographic Terminology (currently being developed) may provide an alternative resource. However, this paper suggests the creation of a “wikipedia” for GIS-related terms. Such a solution would provide the advantage of being more easily accessible to the public and could be established by the international community through a consensus-based process.

Defining the Ground Reference

One of the biggest challenges when working with disparate geospatial datasets is that they were likely produced using different methods, extracted from different sources (e.g. paper maps and satellite images) of different quality, or were based on different reference systems. The first challenge for the user is to homogenize these datasets in terms of projection. The second challenge is to compare them. If the data sets are different, the user must find a way to select the one that most closely represents the real situation on the ground. Some of these challenges could be solved if an agreement were reached among global data producers about using a standard ground reference system. Such an agreement is dependent upon the use of a globally consistent ground reference.

Therefore, to ensure a system that

- provides precision for any place on the planet;
- clearly delineates the hemispheres (rather than having variable reference lines dependant upon the part of the world where the data is used, which can lead to confusion when the data is subsequently mapped);
- easily stores data as decimal values (rather than degrees, minutes and seconds);
- and
- easily imports latitude and longitude values in decimal degrees into a GIS, mapping, or visualization application;
-

this paper recommends:

- using the World Geodetic System (WGS) 84 datum;

- using the IAG-GRS80 spheroid;
- using geographic projection; and
- using decimal degrees for projection units.

Obtaining a ground reference at the global level is more complex. This paper proposes using a free global seamless mosaic of Landsat Enhanced Thematic Mapper Plus (ETM+) scenes collected between April 2003 and June 2005. This free data set is for example accessible from the University of Maryland website at <http://glcfapp.umiacs.umd.edu:8080/esdi/index.jsp>.

The positional accuracy of these scenes, reported by the National Aeronautics and Space Administration (NASA) to be below 50 meters for the GeoCover-Ortho 1990 dataset (Dykstra & Storey, 2004) and the high resolution that they offer (between 15 m for the panchromatic band to 60 m for thermal infrared) makes this source reliable as a ground reference for generating geo-referenced datasets to the 1:50,000 scale.

However, if these scenes were to be used as a standard ground reference, the following issues would need to be addressed:

- The quality of the MrSid format tiles is not high enough to identify features such as roads, rivers of small size, or settlements;
- The large size of each scene in TIFF format (about 200 Mb each) restricts countries with limited internet connectivity;
- The scenes are projected in the UTM 1983 system, which is not consistent with the proposed ground reference system proposed in this paper and limits the extraction of data for countries whose territory spans several UTM zones;
- The location of the control points is not publicly available, which reduces the possibility for users to have a confidence level for the area they are working on;
- The Landsat Program will soon be concluding, and the sensor on board the latest mission (Landsat 7) is having a technical problem that reduces the direct utility of the scenes captured by this satellite. Furthermore, the future of the Landsat program is uncertain.

To address these issues, this paper proposes:

- creating an un-projected version of the mosaic, equivalent to a geographic projection
- facilitating access to resources in countries with low internet capability;
- providing public access to the location of the control points and allowing users to provide potential new control points to improve the georeferencing of the mosaic;
- identifying another program that could provide such a fundamental dataset in the future.

Defining the data set specifications - the geospatial data set design

Before starting any geospatial data collection or extraction process it is important to have a clear understanding of the geographic features which will be involved, and the associated attributes that are of interest for the final data use.

Using common data models is one method for ensuring that the dataset is designed according to the specifications set forth by industry standards. The following data models are mentioned as examples which could be looked at:

- Global Map Data Model for base maps at the 1:1M scale,
- Data Models for industries posted on the ESRI web site.

To facilitate data integration at the global level, the different scale(s) or resolution(s) of data must be determined from the beginning of the project. For example, to make sense of the combined results for two raster format data layers, they both need to be at a similar resolution.

Table 1 presents the scale of work, with the equivalent cell size (resolution) in decimal degrees for raster based data, as proposed by FAO (2003). It is based on practices within the United Nations and includes the 1:250,000 scale, which is the minimum scale required by the United Nation’s humanitarian community. The corresponding resolution in meters is provided here only for reference as the proposed framework in this paper recommends working with un-projected datasets.

Scale	Equivalent cell size in degrees	Equivalent cell size in meters
1:250,000	0.0004167 (1.5 arc-sec)	50 m
1:1,000,000	0.001666° (6 arc-sec)	200m
1:5,000,000	0.008333° (30 arc-sec)	1000m
1:10,000,000	0.016666° (1 arc-minute)	2,000m
1:40,000,000	0.083333° (5 arc-minutes)	8,000m

Table 1 - Proposed scale of work and corresponding cell-size (resolution)

Collecting data and extracting data

This paper distinguishes between “data collection” and “data extraction.” Data collection refers to collecting the locations of geographic features in the real world. Today’s technology allows multiple users to collect geographic coordinates of features using hand-held GPS devices. The level of accuracy and precision that should be maintained with GPS devices depends on the ultimate use of the data. For example, for collecting the location of health facilities as part of the MEASURE Health Facility Survey Assessment (Spencer & Ebener, under preparation), 30 meter accuracy is sufficient. In contrast, collecting data about the precise locations of landmines may require a higher level of accuracy for which detailed surveying measurements would be made. To ensure that the GPS data consistency, regardless of its level of detail, a clear and easy-to-use data collection protocol must be developed for the person(s) conducting the data collection activities in the field.

To ensure integration of data collected at different times and by different entities, the data collection process must include a unique identifier for each of the geographical features it

locates. For example, the health facility surveys (mentioned above) plans to integrate GPS coordinate into a “signature” domain. This signature domain would include the following fields:

- Date of the survey
- Health Facility Country Registry Code
- Health Facility Survey ID
- Health Facility Name
- Health facility contact information
 - o Health Facility Postal Address (street number, city, postal code, other)
 - o Main Phone Number
 - o Main Fax number
 - o Main Email address
 - o Name of the Director
 - o Director’s phone number
- Facility’s geographic administrative unit (at least first and second level)
- GPS Coordinates (latitude, longitude waypoint ID) in decimal degrees.

This paper recommends applying a similar practice to the collection of any geographic feature coordinates.

Data extraction refers to the generation of a derived geospatially referenced data product. The data used at the origin of this process can originate from different sources: paper or digital maps, aerial photography, satellite images, etc. Regardless of the origin of the data, however, to ensure final data integration, the data extraction process must keep in mind that:

- the final data must be un-projected and congruent with the global Landsat mosaic,
- a certain level of precision, to be defined, must be respected and rules applied regarding the size of the minimum mapping units according to the scale of the source (see Table 2),
- the scale or resolution of the final product should correspond to one of those proposed in Table 1.

Scale	Minimum mapping unit size
1:250,000	125 m ²
1:1,000,000	4Km ²
1:5,000,000	100Km ²
1:10,000,000	400Km ²
1:40,000,000	6,400Km ²

Table 2 - Minimum mapping unit size to berespected when extracting data

Cleaning the data

The amount of effort and resources needed for cleaning is inversely related to the amount of effort and resources put into the previous steps of the data production chain. This means that the more effort put into clearly defining the geographic vocabulary, the dataset specifications, the data model, the methods for data collection and extraction, etc.,

the less work will have to be done at this step. Less work at this stage of the data production process can translate into significant time and resource savings.

As an example of the importance of this stage, the World Health Organization (WHO) World Health Survey (WHS) was not able to provide interviewers with the appropriate GPS training and data collection protocols in time to conduct the survey. As a result, it has taken almost one year to clean about 130,000 GPS coordinate points, distributed across twenty-seven countries. The cleaning process did, however, necessitate the creation of a GPS data cleaning protocol that is now freely accessible to the public via the WHO WHS web site. The global Landsat mosaic also provides the possibility for specific layers, such as roads or rivers, to be cleaned and corrected when extracted from another source.

Validating the data

To ensure that the information and data produced ultimately corresponds to the reality observed on the ground in-country, it is crucial to identify and engage the appropriate governmental body or non-governmental organization (NGO) that is in charge of validating the final dataset. When possible, it is most advisable to engage the appropriate authority directly in the data creation process as the validation process could therefore take place during the data collection process itself.

For example, in the case of the Second Administrative Level Boundaries (SALB) dataset project, the National Mapping Agency (NMA) of each country serves to authorize final datasets (Ebener et al, 2006). Since its launching in 2001, the SALB project has closely collaborated with the NMAs of 192 United Nations member states to provide the international community with information and data regarding the evolution and spatial representation of their respective administrative boundary layers. As a result, the project has been able to avoid redundant work to maximize time and resources by coding geographic features at the data collection step, rather than at the final validation step. Waiting until the last step of the data production chain risks having to redo the codes if they were not originally based on validated information. In addition to the SALB project, the US Federal Information Processing Standard (FIPS) codes respect this specific order in the process of creating data sets at the global level.

Documenting the process and the data

More than eight years ago, it was recognized that “without proper metadata, an organization cannot inform data sharing partners of the spatial reference and spatial data structure information that will make data integration and consistent coding possible” (Cote, 1998). An integral part of data creation must be data documentation. Lack of up-front documentation affects users’ ability to integrate data later. It often leads to duplication of effort, as users must recreate a dataset that does not have enough information ensuring that it meets their needs. Too frequently, the documentation of the data transformation through the data production chain often takes place long after the data has been produced, when it should be included throughout the process to ensure

production of metadata records that are of value to the data user. Metadata cannot be an afterthought, otherwise, much valuable information about the data transformation can be lost, especially if staff turnover occurs during the data production process.

To be useful to the user, the metadata record attached to the data should therefore capture all the elements defined in the context of the framework presented in this paper. It is therefore crucial that the documentation process start from the beginning of the data production chain (see Figure 1).

Basic metadata content standards have been agreed upon at the international level through ISO. The number of standards currently in use tends to be limited, with the most common being the ISO 19115 and the US Federal Geographic Data Committee (FGDC) metadata standards. However, the *mandatory* fields proposed by these most-widely used metadata standards are not sufficient to capture all of the necessary information recommended by this paper. A complete and useful metadata record should also include information about:

- the quality of the particular data layer;
- the purpose for which the dataset was created;
- the time period over which the dataset was created;
- the accuracy of the dataset; and
- the protocols that were used in the data production process.

In addition, terms used to define some of the fields already recommended in the ISO 19115 standards need clarification (e.g. “start date”, “end date”). Guidance for filling out the free text fields for the abstract, data description and lineage is also needed. The absence of guidelines for these important fields results in metadata content that does not facilitate the work of the user who is comparing two different data sets to select the most appropriate one.

Summary of Recommendations:

Defining the vocabulary - geographic terminology:

- Possibly the future ISO 19104 standards on Geographic Terminology or the development of a "wikipedia" for GIS-related terms

Defining the Ground Reference:

- Use the global Landsat mosaic as ground reference
- Geodetic Control
 - Datum - WGS84
- Spheroid - IAG-GRS80
- Projection System - Geographic
- Projection Units – Decimal degrees

Defining the data set specifications - the geospatial data set design:

- Scale/Resolution (please refer to Table 1)
- Use appropriate data models to assist in data design

Collecting data and extracting data:

- GPS coordinates – Decimal degrees
- Development of signature domain for the unique identification of objects
- Accuracy/Precision (please refer to Table 2)

Cleaning the data:

- Plan in advance of data collection to save resources and time on this step
- Provide training on data collection specifications to data collectors

Validating the data:

- Use local, in-country authorities to validate the data

Documenting the process and the data:

- Document all steps in the data production effort – do not wait until after the data set is completed
- Use a common metadata content standard, such as ISO 19115 or the future ISO19119
- Extend the list of mandatory fields when creating metadata profiles
- More guidance should be provided to users for filling out the free text fields

Discussion and Conclusion

This paper offers an overview of practices and standards that could be used to improve data integration at the global level. By presenting these practices and standards within a data production chain, this paper provides a new perspective to the discussion calling for an integrated approach to development of new data sets and standards. In an era when government bodies, NGO's, international organizations, and the private sector are all engaged in the process of building or maintaining SDI's, it is important to ensure global data integration to facilitate the emergence of a global SDI.

While it may be difficult to reach a global consensus regarding the standards to be used, this paper proposes at least the creation of a data integration framework. Among the different steps proposed, the use of the global Landsat mosaic represents a major opportunity to move forward quickly. By providing an interactive, user-friendly access to this mosaic, Google Earth and other similar data visualization applications already have started the process as more users are in a position to evaluate the accuracy or quality of the georeferenced data they are using. It is therefore time to encourage data producers to create their data according to the framework proposed in this paper, one similar to it. Utilization of data lifecycle planning and implementation standards will improve data quality and data exchange among users, thereby increasing efficiency and cost effectiveness for all.

In the second chapter of the SDI Cookbook (Luzet & Murakami, 2004), the following list of reasons is given to explain why GIS users tend to develop their own data sets when many already exist:

- they may not know available existing data sets that could be appropriately used for their applications; or access to these data sets is difficult;
- they are not used to sharing data sets with other sectors and/or organisations; and
- existing geospatial data sets stored in a certain GIS may not be easily exported to another system.

This paper highlights that another reason could also be added to this list: the lack of data interoperability (spatial and temporal) between the data sets currently available due to non-standard specification and lack of documentation.

In reality, there is only one road on the ground, or one river. There should therefore be no need to have more than one data set for each layer, for a given time period and at a specific scale. While a given feature may have numerous attributes, use of common data models will improve the consistency of content and coding schemes within attribute tables.

Applying the comprehensive framework suggested in this paper will only happen if a redistribution of priorities and funding for the production of geospatial data takes place, and the terms of reference for projects require this level of data interoperability. To encourage it, donors could play an important role in advancing the global SDI by requesting that data produced using their funds follow these specific sets of practices and

standards. Large data producers and consumers, such as the United Nations, could work harder to promote the use of such approaches among the more than 40 agencies and organizations using geographic information.

In addition, national activities should focus attention on improving data integration. Several efforts underway include the Working Group 3 - Cadastre of the Permanent Committee on GIS infrastructure for Asia and the Pacific (PCGIAP) which held a workshop during the 17th UN Regional Cartographic Conference for Asia and the Pacific (UNRCC-AP) to look at data integration of natural and built environmental datasets; and the Mapping Africa for Africa (MAfA) initiative (ICA; 2003). It would also be important at the other aspects related to data integration, such as the institutional, policy or legal ones as identified by Mohammadi et al. (2006).

In return for such investments, the benefits to society would be tremendous. We would not only save a lot of money and reduce the cost of the data, but would also be able to more quickly obtain a complete and reliable view about the distribution of resources, actual problems, and potential challenges that we might have to face in the future at the scale of the planet. *Is this not what a Global SDI should be able to address?*

Acknowledgements

The authors wish to express our thanks to Dr. Abbas Rajabifard for his suggested revisions to a draft version of this paper.

References

ANZLIC Glossary, www.anzlic.org.au/glossary.html

AGI Dictionary,

www.agi.org.uk/POOLED/articles/bf_trainart/view.asp?Q=bf_trainart_156551

Cote, Carmelle J., 1998. Dissertation - *Content Standards for Geospatial Metadata: Impacts on International Spatial Data Sharing*, 1998.

Data model for industries web site,

<http://support.esri.com/index.cfm?fa=downloads.dataModels.matrix>

Dykstra J.D., Storey J.C., 2004, *Landsat 7 definitive ephemeris: An independent verification of Geocover-ortho 1990 positional accuracy*. NASA internal report.

Ebener S., Brookes B., Silva A., Gagliano E., Guigoz Y., 2006, *Lessons learned towards the creation of a global SDI: the example of the SALB project in the Americas*, 9th Global Spatial Data Infrastructure (GSDI) Conference, Santiago de Chile, 3-11 November 2006.

ESRI GIS Dictionary,

<http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.gateway>

FAO, 2003, *A Spatial and Norms Proposal for FAO Interdisciplinary GIS Database*, Technical Report of the Spatial Standards and Norms Task Force, SPATL-PAIA Work Group.

Foote, Kenneth E. and Donald J. Huebner, 1995, *The Geographer's Craft Project*, Department of Geography, The University of Colorado at Boulder.

www.colorado.edu/geography/gcraft/notes/error/error_f.html

Global Map Data Model for base maps,

[http://downloads.esri.com/support/datamodels/Basemap/Global_MapDetail\(Esize\).pdf](http://downloads.esri.com/support/datamodels/Basemap/Global_MapDetail(Esize).pdf)

Highland Surveyors Licensed Land Surveyors & Geomatic Consultants, 2003, *Study for Establishing Kenyan Standards for Spatial Data*.

http://kism.iconnect.co.ke/NSDI/SURVEY_REPORT.pdf

ICA, 2003, *Mapping Africa for Africa: Durban Statement*. Mapping Africa for Africa Workshop, International Cartographic Congress, 14 August 2003, Durban South Africa.

- Jones A., 2004, Geospatial Terminology in: *The SDI Cookbook*, version 2.0, 25 January 2004. Editor: Douglas D. Nebert, Technical Working Group Chair, GSDI.
- Lance, K. and E. Gavin, 2003, *Implementing Standards for Geographic Information in Africa*: Outcomes of workshop held 10 August 2003, Durban, South Africa.
- Luzet C., Murakami H., 2004, Geospatial Data Development: building data for multiple uses. In: *The SDI Cookbook*, version 2.0, 25 January 2004. Editor: Douglas D. Nebert, Technical Working Group Chair, GSDI.
- Mohammadi H., Binns A., Rajabifard A., Williamson I.P., 2006, The development of a framework and associated tools for the integration of multi-sourced spatial datasets, 17th UNRCC-AP Conference, Bangkok, Thailand, September 2006
- Spencer J., Ebener S., under preparation, *Standardization of Geographic Data for Health Facility Surveys*, MEASURE Health Facility Survey Assessment task group.
- WHO World Health Survey, www.who.int/healthinfo/survey/en/

Annex I: *Acronyms used in this paper*

ANZLIC	Australia and New Zealand Land Information Council
ESRI	Environmental Systems Research Institute, Inc.
ETM	Enhanced Thematic Mapper
FAO	Food and Agriculture Organization of the United Nations
FIPS	Federal Information Processing Standards of the United States
GPS	Global Positioning System
GRS	Geodetic Reference System
GSDI	Global Spatial Data Infrastructure
HFA TG	MEASURE Health Facility Survey Assessment Task Group
IHO	International Hydrographic Organization
ICA	International Cartographic Association
ISO	International Organisation for Standardization
ITC	International Institute for Geo-Information Science and Earth Observation
MAfA	Mapping Africa for Africa
NASA	National Aeronautical and Space Administration of the United States
NGA	National Geospatial-Intelligence Agency of the United States
NGO	Non-Governmental Organization
PCGIAP	Permanent Committee on Geographic Information for Asia and the Pacific
SALB	Second Administrative Level Boundaries
SDI	Spatial Data Infrastructure
SRTM	Shuttle Radar Topographic Mission
UNEGN	United Nations Group of Experts on Geographic Names
UNGIWG	United Nations Geographic Information Working Group
UNRCC-AP	United Nations Regional Cartographic Conference for Asia and the Pacific
USAID	United States Agency for International Development
USBGN	United States Board on Geographic Names
UTM	Universal Transverse Mercator
WGS	World Geodetic System
WHO	World Health Organization
WHS	World Health Survey