An Overview to Spatial Data Protocols for HIV/AIDS Activities

Why and How to Include the "Where" in Your Data

MEASURE Evaluation
www.cpc.unc.edu/measure

MEASURE Evaluation is funded by the U.S. Agency for International Development (USAID) through Cooperative Agreement GHA-A-00-08-00003-00 and is implemented by the Carolina Population Center at the University of North Carolina at Chapel Hill, in partnership with Futures Group, ICF Macro, John Snow, Inc., Management Sciences for Health, and Tulane University. The authors' views expressed in this publication do not necessarily reflect the views of USAID or the United States government.

November 2011

MS-11-41A

Acknowledgments

MEASURE Evaluation acknowledges the work of MEASURE Evaluation staff members James Stewart, Becky Wilkes, John Spencer, and Nash Herndon for their contribution to this publication.

Acronyms

ART	antiretroviral therapy
EA	enumeration area
FGDC	Federal Geographic Data Committee
GIS	geographic information systems
GPS	global positioning system
IP	implementing partner
ISO	International Organization for Standardization
KML	keyhole markup language
LGA	local government area
M&E	monitoring and evaluation
NMA	National Mapping Agency
NSDI	National Spatial Data Infrastructure
UNGIWG	United Nations Geographic Information Working Group
UNSALB	United Nations Second Administrative Level Boundaries
USAID	U.S. Agency for International Development
USG	United States government

Table of Contents

Introduction

The foundation of effective decision making is the effective use of data. However, there is a journey that data must make before being used to support decisions. Data must be accurate and, for maximum utility, conform to standards. Data users must be thoughtful about their use of data, making sure to use only appropriate data and to consider fully the limitations of each data set they use.

Data that are of high quality, relevant, and complete can help lead to a better understanding of health interventions and human activity. Often multiple data sets are necessary to paint a picture that is useful to decision makers. In addition to collecting the data, there are issues associated with managing multiple data sets. Decision makers need effective ways to join and synthesize multiple data sets and pick out important data points and patterns.

The synthesis of data requires a common denominator, and there is one common denominator across nearly all human activity: It happens somewhere on Earth. The spatial component of data can be used to great advantage not only for understanding where things are happening, but also for understanding why things are happening. Including the geographic perspective in data will not only make it possible to produce maps that can serve as effective decision-support tools, but will also enable a common link across multiple data sets that can make it easier to join data sets and synthesize information.

Key Message

Solid data form the foundation of effective decision making, and the best way to have a strong foundation is by including the geographic context.

Purpose of this Guide

This guide focuses on data and how to use geography to facilitate linkages among data. It presents an overview of the ways to structure health data to take maximum advantage of existing geographic data and to facilitate future inclusion of the geographic context of data being gathered. Data with a geographic component are particularly well-suited to support monitoring and evaluation (M&E) efforts and evidence-based decision making. The use of geospatial technologies has been hampered by limitations that often exist with the data. This document seeks to address some of these limitations by presenting key concepts involved in the collection and use of spatially referenced health data. The document presents several key concepts:

- role of data in the decision-making process
- value of using data wisely
- importance of data quality and standards
- significance of standard data schemas and identification of data schema best practices

- how geographic context can strengthen data infrastructure
- value of maps as decision support tools
- illustration of the importance of linking data sets and how geography can be used to facilitate that link

This guide is not intended to provide technical guidance on the use of spatial software, such as geographic information systems (GIS). While this guide presents material in terms of the geographic context, the concepts can be applied to non-geographic contexts as well. Regardless of whether data include a geographic component, the concepts of data standards, data quality, and effective data software are important in all settings.

Intended Audience

The concepts presented in this document have wide applicability. Staff and decision makers within national governments, program managers, implementing partner (IP) staff, and U.S. government (USG) staff are all examples of people who might benefit from a stronger data infrastructure and from using a geographic context in support of the decision-making process.

Because this document focuses on data and the ways to make data more useful for decision making, no specific knowledge of software programs, such as those involving GIS is required. However, familiarity with spreadsheet programs or database programs will be helpful.

Advantages to Including the Geographic Component to the Data

The geographic component of an epidemic is an important one, as geography not only influences the spread of the disease, but also its treatment. Geography is an important factor in early attempts to understand an epidemic.[1-2] In the global response to the HIV epidemic, GIS provides an important tool in addressing such issues as areas of high transmission,[3] most-at-risk populations,[4-5] access to services,[6] and understanding the epidemiology of the disease.[7-9]

Challenges to Using Spatial Data for HIV Programs

The issues surrounding HIV are complex. Risk of infection and receptivity to treatment can be affected by confounding factors such as poverty, gender, religion, and societal structure. Finding data sets that help explain the relationships among these factors can be challenging. Finding such data sets that include a spatial component can be even more challenging.

Not all HIV services are provided at a health facility. Some services and programs are community based or mobile, and are not confined to one geographic location. This means that including a geographic identifier for them can be difficult.

Another notable constraint is confidentiality. Because of the need to protect data at an individual's level, spatial context is often omitted from individual patient-level data. There are, in fact, other ways to deal with this issue. For more information on privacy issues and spatial data, see page 20 of this document.

Data Flow and the Role of Data in M&E

Monitoring and evaluation of programs and activities can help managers and decision makers understand better the effectiveness of interventions and health activities. Data are the foundation of effective M&E; therefore, it is vital that relevant data are collected accurately and are timely.

Including the spatial perspective in M&E data provides two advantages to program planners and decision makers. First, the spatial perspective provides a foundation to understand better the geographic context behind activities, through the production of maps and spatial analysis techniques. The second advantage to inclusion of spatial data lies in the fact that geographic context can help facilitate linkages of datasets and keep data from being used in isolation.

Data used for M&E follows a path that often starts with individual patients and can end up in a national or even international destination (figure 1). Consider the following illustration: A patient walks into a facility to receive antiretroviral therapy (ART). In a setting with a functioning, robust M&E system, this visit will be entered into a log at the facility. In addition to helping ensure that the patient stays current with his or her treatment regime, this information

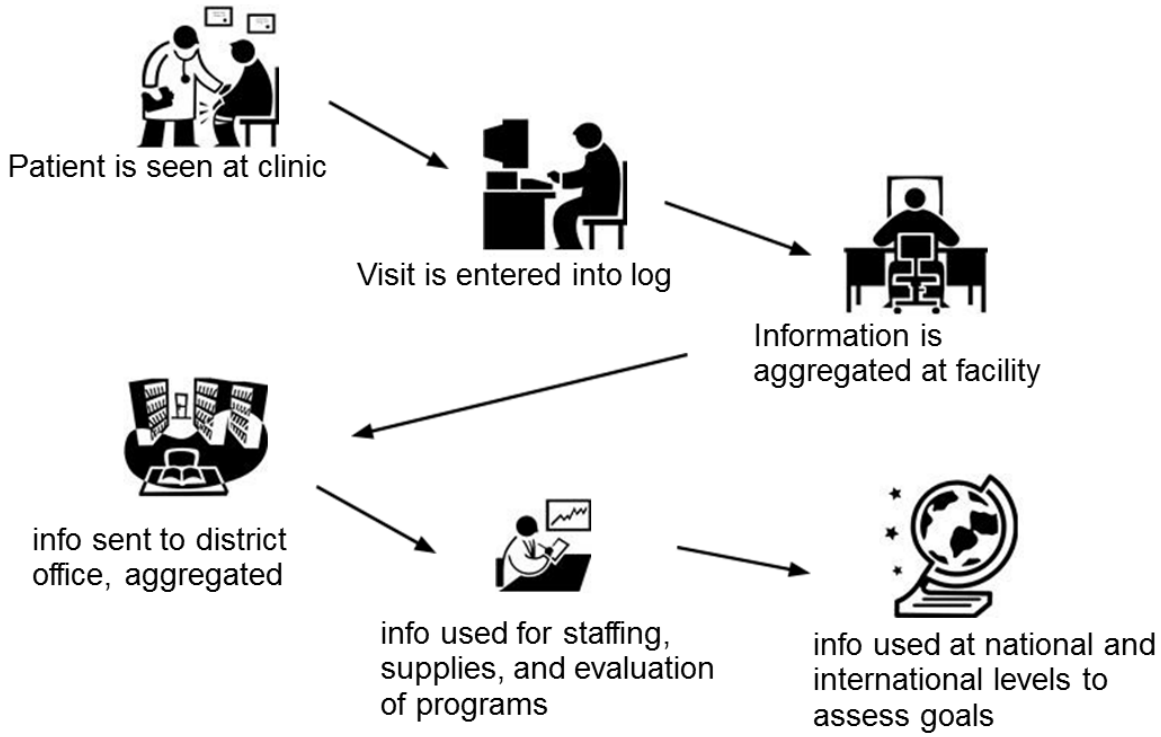


Figure 1. Information used at district, national, and international levels often starts with data collection from individual patients.

can also be aggregated at the facility to help that facility track patient loads, compliance, and other factors. Aggregate data from the facility that includes not only this patient's visit, but all patients' visits for all services, will be sent regularly to a regional or district office. This regional or district office will also receive similar data from all facilities in the region or district. The regional or district officer planners will then aggregate all of the data from the district's facilities to help them plan at the regional or district level. This can help ensure having adequate staffing in facilities to serve the clients' needs or prevent running out of supplies, or even identify areas that may be over- or under-served. The next step in the data's journey might possibly be at the national level, where staff might receive and aggregate data from all districts. This allows national planners to evaluate effectiveness of programs within a national context and can provide updates and feedback to decision makers in the country's legislature or to other national leaders. Sometimes data can then move beyond the national level to the international level, with national governments reporting to donors or multi-national bodies such as the United Nations or World Bank. The data can then be used to evaluate aid programs or help inform decision making at the global level regarding progress to global goals on health and well-being.

At the end of the journey, whether it arrives at the national or inter-national destination, the data flow that started with one patient's visit to one facility now contains information about thousands or even millions of visits across many facilities. At each step of the journey, there are potential risks from problems with storing, manipulating, aggregating, or analyzing the data. Each country is different; and in some countries, the journey may vary slightly from data use in another country. But regardless of the exact path, the opportunity for problems to arise with the data at each step along the way are universal. Including a geographic component to the data can add even more challenges. This document presents strategies that can ensure the integrity of the data.

If procedures to ensure the integrity of the data aren't maintained, then the flow of data can stop, or inaccurate data can be used for M&E. This can result in bad decisions, inefficient programs, wastes of resources and, most critically, poorer outcomes that may lead to increased illness or unnecessary deaths. The process, protocols, and guidance provided in this document can help strengthen the data infrastructure, which can minimize such risks.

The Purpose, Limitations, and Schema of Data

The purpose of having data is to provide a summary of what is going on in the real world. Data can be thought of as representations of information. Information, in turn, leads to knowledge. Data analysis is a process of gathering, modeling, and transforming data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making.

Data provide a snapshot of the real world; data are gathered in, and are representative of, a specific place and time. Without a record of the places and times the data belong to, data lose their usefulness. Place and time components of data sets also provide common linkages among different data sets. This type of data commonality "offers benefits to the developers and the users, as well as spreading the cost of database creation among multiple users. There are also synergies in multiple uses and analysis of a common spatial database by diverse groups, with one group's insights sparking another's, thereby creating value."[10] This chapter presents some key data concepts important to the discussion that follows in later chapters.

Key Message

Data are more useful if formatted in ways that allow easy linking with other data. Data schemas need to foster linkages among data sets.

Evidence-Based Decisions

The findings from analysis of health statistics "are sometimes used both as background for policy development and for the planning and implementation of specific interventions, on the one hand, and as factual material presented in education strategies to inform the public and specific groups, on the other hand."[11] In order for decisions about policy to be completely and appropriately informed, we must have as complete a picture of reality as can be provided by the data. This includes spatial as well as temporal components to the data.

Spatial data allow for key linkages among data sets, and wide diversity of health data, covering a wide range of policy and planning issues. Health data with a geographic component can be integrated with other social and environmental data, which often also have spatial components. These "foundation data" are useful when planning, evaluating, and researching health-related issues. Types of foundation health data include vital statistics, disease surveillance, survey information (e.g., M&E), and health services data.[10]

Health care surveys in particular can "serve to provide essential information to inform program development requirements for specific population subgroups with specific needs. A first step in the process is to identify where there is variation in health and health care in the population."[11] This process can be accomplished by the inclusion of a geographic component to the data.

Limitations of Data

What can data NOT do? All data are merely representations of the real world, and as such cannot provide a complete and perfectly accurate picture of that world. Data collected in a particular instance are merely snapshots in time; the data have no cause or effect per se. And some aspects of reality may be difficult to capture, due to rapid change or difficulties in measurement.

Biases in the gathering of data can further cause data to reflect reality incorrectly. The initial act of choosing which variables to measure can draw the researcher down the road of bias, and necessarily limits the information that will be able to be gleaned from that particular data set in the future. This is known as *selection bias*. Selection bias can also be caused by the use of non-representative sample populations.

Another type of bias is *random error*. This can occur due to errors in actual measurement (equipment errors or user errors) or misunderstandings about questionnaires, errors in coding, etc. This type of error is unpredictable. Yet another type of error is *systematic error*, which can also be caused by miscalibrated equipment — but unlike random error, this can be corrected if the problem becomes apparent.

One type of error that occurs from data misuse (rather than from data gathering) is *ecological fallacy*. Ecological fallacy is defined by Environmental Systems Research Institute (ESRI) as "the assumption that an individual from a specific group or area will exhibit a trait that is predominant in the group as a whole." An example would be assuming an individual had 12 years of education because that was the average number of years of education in the district where they lived. The data user is applying the characteristics of a group to an individual.

Data Use Cycle

Data are of no value unless they are used. Therefore, it is important to think from the beginning about how the data are going to be used, by whom, and what questions the data are going to help answer. Data use inevitably creates a demand for more data, either in the form of updates or expansion, which leads to more data collection. This leads to increased availability of data, which in turn leads to increased utilization of data. Thus, we have a cycle that continues to drive data collection and use (figure 2).

Generally, data collection should support better decision making. Data collection should be thorough and relevant to the questions at hand. Any particular data set's eventual uses may be difficult to predict. Thus, the data must be stored in a format that can be readily updated and easily combined with other data or otherwise modified.

What Is the Data Infrastructure?

The data infrastructure refers to the aggregate collection of all potentially available data. Data infrastructure refers to the systems in place to collect, maintain, and analyze the data, as well as the actual data. Not only is the data infrastructure dynamic, changing as new diseases

and population patterns emerge, but the data require regular updates. The format of the data will ideally be one of maximum flexibility, providing for ease of reporting and linking with other data sets. Specifically, National Spatial Data Infrastructure (NSDI) is a term that refers to spatial data sets, an important component of the entire data infrastructure. Spatial information, such as location of road networks and administrative boundaries, is often changing and requires updates as well. The full data infrastructure includes data from multiple realms — public health, economic, environmental, and demographic — though it is not always necessary to utilize the full data infrastructure. Within the public health sector alone are multiple realms: orphans and vulnerable children, ART and HIV/AIDS, malaria, family planning and reproductive health, etc.

Metadata

Metadata compose an important part of the data infrastructure. Metadata are data about the data. They tell when the data were collected, how data were collected, and who collected them. Metadata give definitions of each variable (sometimes called a *data dictionary*) and provide insights into data organization and original intention. This information is essential both to the people who collect the data and to those who may use the data later. For example, when it comes to data collected via administrative area, it is crucial to have dates recorded, as those boundaries may change, affecting both the collection of future data and the base maps used to show the data.

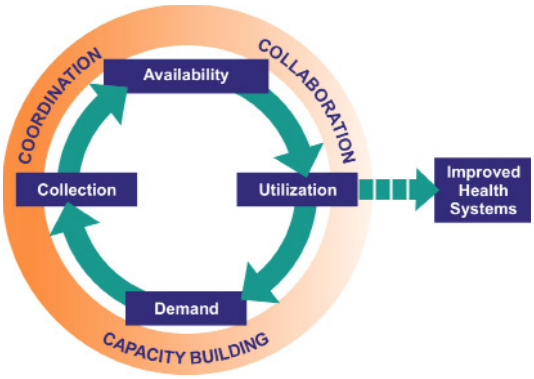


Figure 2. The data-use cycle.

Data Schema

A data schema is a description of a database structure. The term refers to the formal language used to describe data and their arrangement in tables, and the ways in which the data interrelate. It is similar to an outline showing the way data are organized.

Properly organized data can be used to explore useful questions (figure 3), such as: Who is being served by certain programs? Where is the greatest need? Are the programs making a difference? What improvements can be made? Are there relationships and connections between the data that have not been previously explored?

Why the data schema is so important — The data schema is an important consideration when

Data on programs serving orphans is organized according to location:

Number Orphans	Orphans Served	Province Name	Year Program Began
23,300	4,536	Dirada	2005
76,500	5,231	Modiju	2005
32,400	3,627	Helmaga	2004
36,900	6,342	Klit	2006
47,800	5,482	Redishin	2006

Possible questions:

What locations have been served the longest?
What percentage of orphans are being served in each location?
Is there a particular portion of the country which is underserved?

Figure 3. Properly organized data generate useful questions.

deciding to collect or organize data in an orderly fashion into one or more tables. Having an organized data structure can help ensure maximum usage and effectiveness of the data in that this structure fosters an ease of understanding for other users of the data. Data standards can be established that allow comparison of statistics from one situation to another, including across time.

Data schema standards also allow the easy and effective use of one or more sets of data by multiple people or even more than one organization, greatly aiding communication. An established data schema makes it much easier to share data among different users and different computers, which may even be using different tools or software packages.

A good data schema will also allow maximum ways to display and compare data, and to generate statistics and informative graphics. It allows users to sort and sift data and to perform calculations on the data. The best schema allow comparison of data from one data set to another, and joining of data sets based on common denominators, such as place (e.g., country or district) or time (e.g., month or year). Table 1 is an example of a less flexible data schema. It shows a hypothetical case of numbers of children served and is arranged by type of funding organization, and then by province and district.

Table 1. Example of a Table Formatted Using Poor Data Schema to Show Number of Children Served in a Hypothetical Country

Funding organization	*Province*	*District*	*Children served*
Alpha	Alboma	Getar	354
	Delmet	Huma	45
		Jedanga	267
	Spudow	Flennet	12
		Pranan	78
		Spilitina	426
Bravo	Alboma	Getar	34
	Bronip	Huma	69
		Star	62
		Trethel	587
	Spudow	Pranan	324
Charlie	Delmet	Huma	213
		Dumwick	95

Suppose you wanted to compare which districts served the greatest number of children, regardless of funding organization? Cells with no values make this difficult because most data software will consider empty cells as missing data. Table 1 shows that several districts are served by more than one funding organization. Table 2 shows a new data schema, which makes analysis easier. To make this new schema, children served are totaled by district. Note that district Huma includes more than one province (Delmet and Bronip). Thus, if all values for Huma were added, the totals would not be valid for purposes of the comparison. By creating a new data schema that includes unique identifying codes by district, and by formatting a new table based on these unique codes, comparing children served by district can be more easily examined. For further visual clarification, these data can also be displayed on a map. In this schema, items in each column can be sorted or aggregated according to the needs of the user.

Table 2. Example of Hypothetical Country Data Reorganized

District code	*District-province (unique name)*	*Children served*	*Alpha funding (U.S. dollars)*	*Bravo funding (U.S. dollars)*	*Charlie funding (U.S. dollars)*
11	Getar-Alboma	388	$35,400	$3,400	0
12	Huma-Delmet	258	$4,500	0	$21,300
13	Jedanga-Delmet	267	$26,700	0	0
22	Dumwick-Delmet	95	0	0	$9,500
14	Flennet-Spudow	12	$1,200	0	0
15	Pranan-Spudow	412	$7,800	$32,400	0
16	Spilitina-Spudow	426	$42,600	0	0
17	Huma-Bronip	69	0	$6,900	0
18	Star-Bronip	62	0	$6,200	0
19	Trethel-Bronip	587	0	$58,700	0

One Record per Unit

There are many ways data can be stored in a table, but storage of one record per geographic unit offers tremendous flexibility for use and analysis of the data. Storing data with one record per geographic unit, such as by district or province, provides such benefits as:

- allowing for linkages among data sets;
- imposing a standard structure that can help promote data accuracy; and
- allowing for use by database programs (such as Microsoft Access), even if data are originally stored in a common spreadsheet format (such as Microsft Excel); such programs provide a means for searching (querying) and for import to a GIS (mapping program).

One of the most important concepts in structuring data is that of a unique identifier. In order for one value to be easily retrieved and examined from a database, it needs to have a unique way of being singled out. The name of a person, for example, is not a unique identifier (there may be two David Hills living in the same town), but a personal identification number can be unique. A date is not unique unless it includes the year (e.g. December 15 vs. December 15, 2008). A place name is not necessarily unique, either. There may be two occurrences of the same place name in different locations, such as Durham, England and Durham, North Carolina, USA.

In the case of names of organizations or locations, a unique identifying code is also preferable to a text-based name due to possible differences in spelling or the use of special characters (e.g., Mombaso vs. Mombasa vs. Mom-B'asa).

Figure 4 shows maps of a hypothetical country produced from the data tables illustrated above.

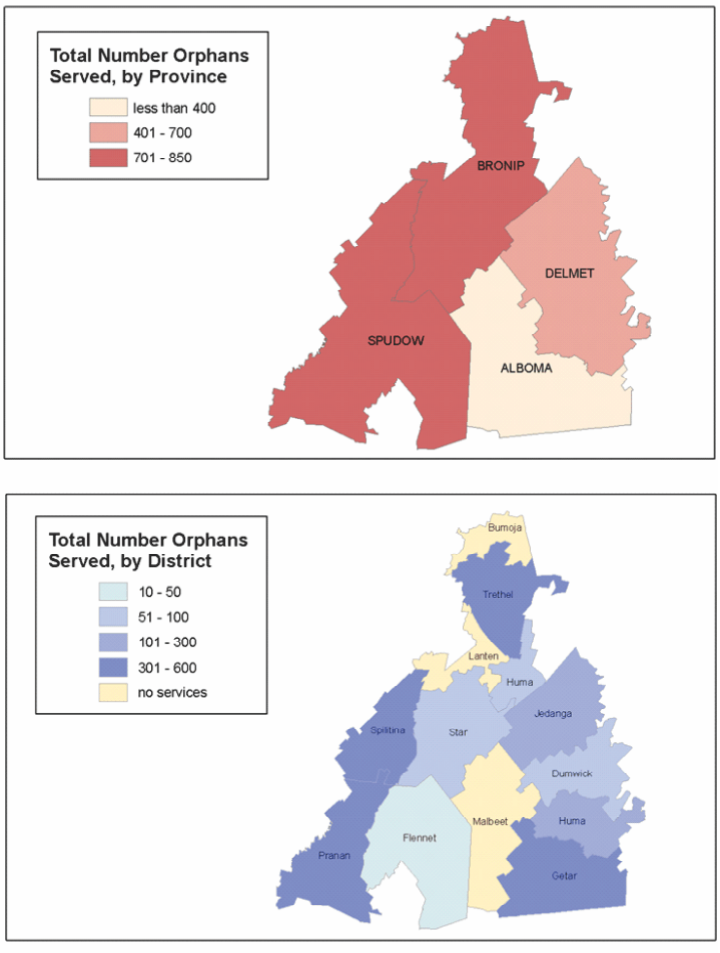


Figure 4. Example of showing data in a hypothetical country.

Data Formats

Data can be reported in a variety of formats.* Part of the task of organizing data into a schema is to identify the different types of data.

Text — Text is useful for names or descriptions of locations or facilities. Text data can also be used to designate certain classifications, such as type of disease.

Numeric — Numerals can be used either for unique identifiers (codes), as discussed above, or to show amounts, such as total population or numbers of children under age 15 attending school. They can also be used to show comparative (ordinal) values, such as 1 for highest, 2 for middle values, and 3 for low values of a particular type of data. If data are to be added (aggregated), averaged, or otherwise mathematically manipulated, it is important that the numeric format be distinguished

* The term *format*, as used here, refers to data formats (i.e., the plan for organization or arrangement of the data), not file formats, a more technical term referring to various types of encoding of computer files.

from text. If a number in a table represents a code (unique identifier) it can actually be stored as text.

Coordinate — Coordinates can be stored as X,Y values (longitude and latitude) using degrees, minutes, and seconds, or using decimal degrees. Coordinates must be tied to a system or grid (origin). Since more than one datum may exist at a particular location, coordinates are not necessarily unique identifiers. But a coordinate location can be the most accurate way of locating a geographic point.

Other formats — Other special formats that can be designated include financial information (i.e., dollars), date and time formats (month, year, etc.), and percentages. These are often stored as numeric data, as different software programs may have a variety of formatting options.

Figure 5 lists some examples of different formats.

text	Central City Children's Hospital	No data
numeric	549	0.478
coordinate	35°55′44″N 79°2′22″W	-79.039444, 35.928889
other	12/15/08	25%

Figure 5. Examples of various formats of data.

Software Considerations

There are many software programs available to facilitate the management and analysis of data. Some software programs are simple to use, yet offer little power for management and analysis. Other programs are very powerful and allow users to query and analyze the data, but require specialized training to use properly. Selecting the most appropriate software involves thinking about the type of data being used, the ways data will be analyzed, and the capacity of staff who will be using the software.

Relational database — There are a large number of ways to store and manipulate data. A relational database, such as Microsoft Access, can help keep data organized and allows easy manipulation. Microsoft Access is a program that requires that a data schema be defined that can store multiple data types and multiple tables. It is very particular about data structures. Data that have been organized this way can easily be imported into a GIS for use in mapping or graphing.

Spreadsheet — Many people also store data in what is called a "flat file," which is an independent data table that can produce certain mathematical calculations or sorting and groupings. An example of this is Microsoft Excel, used for creating individual spreadsheets and tables. A spreadsheet is not as particular about data structures as more complicated software might be. As a result, a spreadsheet can be easier use than other software options, but may be less effective in its creation of organized data structures.

Other electronic formats — A poor choice for storing data is in a word processor or other document management or graphic display program. These do not permit data formatting, aggregating, or sorting. Data stored this way also cannot be easily imported into a GIS program. The data can be displayed nicely and copied or shared, but in its original table format only. No further manipulation of the data is possible. Examples would include Microsoft Word and various images of documents, such as Adobe's Portable Document Format (PDF) or the Joint Photographic Experts' Group format commonly called JPEG. Creation of a data schema for data stored in any of these formats is often possible only by means of complete re-entry into a spreadsheet or, preferably, a more sophisticated database. If data are stored in a Microsoft Word table format, the data can becopied into a Microsoft Excel spreadheet, in chunks, but often it must copied cell-by-cell or even re-typed by hand, a time-consuming and error-prone process.

Summary of Data Collection Considerations

As a general rule, data are more useful if formatted in ways that allow easy linking with other data. Data schema need to foster linkages among data sets. They need to take into account existing data structures and existing data-gathering practices. They also need to encourage the gathering and maintenance of a complete and accurate picture of reality. Such data considerations are extremely important when supporting M&E and evidence-based decision making. In short, data collection should consider:

- errors and biases potential to the data
- data users (current and future)
- metadata and data dictionaries
- data schema (which will influence ways to display and join data, and to perform calculations)
- data formats (important when considering data schema)
- data storage (software and file formats)

The Spatial Context in Data Collection

Key Message

A spatial context to data can be utilized by adding geographic identifiers, using a well-defined geographic hierarchy. GPS deserves particular consideration when capturing geographic information.

As previously stated, one thing that most data have in common is that they refer to human activity taking place somewhere on Earth. When the spatial context is included in each data set, it becomes possible to make linkages across data sets and to integrate effectively newly collected data into a country's data infrastructure. This section presents the key issues that should be considered when building data sets that use spatial information. Specifically, adding a spatial context to data can be accomplished by using geographic identifiers, which can be any data elements that indicate the geographic location of the data. For example, in figure 6, the columns called "District" provide geographic identifiers.

There are a variety of ways to include geographic identifiers in data. Some common geographic identifiers include administrative division names (such as province or district), place names (such as city, village, neighborhood, or barrio), and exact locations (such as a street address within a city or global positioning system [GPS] latitude and longitude coordinates).

Administrative Divisions as Geographic Identifiers

Administrative divisions are areas defined by governments to ease administration of government services or identify distinct geographic, ethnic, or cultural regions in a country. These may change over time due to government action, and the geographic data will have to be updated accordingly. This can be an important consideration when using these divisions as primary geographic identifiers. Administrative units often have a hierarchical relationship, where the divisions are made up of a collection of units from a lower level. For example, in 1999, the Kenya National Bureau of Statistics used a well-defined hierarchy for

Population data from national census bureau

District	Population
North	3253
South	5621
East	8732
West	7715

OVC data from health agency

District	OVC Population
North	812
South	1011
East	2709
West	1411

Data linkage aided by common locations, allowing further computation and examination

District	Population	OVC	Percent OVCs
North	3253	812	25
South	5621	1011	19
East	8732	2709	31
West	7715	1411	18

Figure 6. Example of linking data from common locations.

census data collection (table 3).* The primary administrative division is the nation, which is separated into provinces. Provinces are further divided into districts, districts into divisions, and divisions into locations. Locations are comprised of sub-locations, and below that is the lowest level of the hierarchy, the enumeration area (EA), which usually contains between 51 and 150 households.

Table 3. Kenyan Geographic Hierarchy for 1999 Census

Administrative Unit	*Number of Units*
National	1
Province	8
District	69
Division	497
Location	2,427
Sub-location	6,612
Enumeration area	61,921

Source: Odhiambo, E. "Census Cartography: The Kenyan Experience," presented at the UN Expert Group Meeting on Contemporary Practices in Census Mapping and Use of GIS, May 29-June 1, 2007. New York.

Importance of Hierarchy

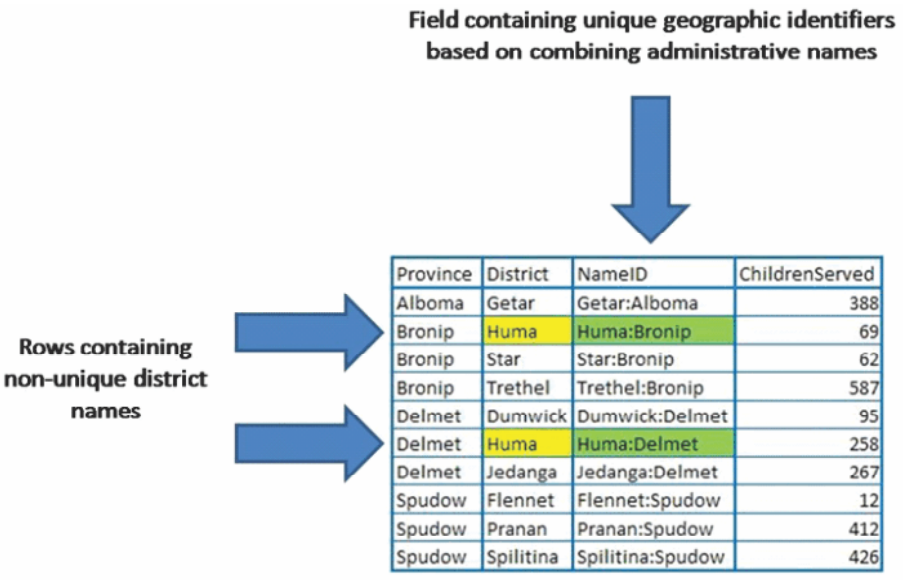
A well-defined geographic hierarchy is critical for aggregating data from a detailed to a higher level for analysis and reporting. This can greatly facilitate the decision-making process, as many decisions rely on a more synoptic view of data. This is logical, given that public health efforts are generally funded and managed by administrative divisions.

Also, a well-defined geographic hierarchy can provide the foundation for the establishment of unique geographic identifiers. Unique geographic identifiers differ from the unique identifiers discussed previously, in that they refer to just the geographic identifiers. Having unique geographic identifiers is important because it is possible that the same names may be used more than once for a particular level of the geographic hierarchy. If the full hierarchy is not included, then it could be difficult to tell which unit a name is referring to. This becomes more common as one travels down the hierarchy to the more detailed level. In Nigeria, for example, there are two local government areas (LGAs) named Bassa, one in the state of Kogi, and one in the state of Plateau (figure 7). For geographic analyses, these non-unique administrative names could cause confusion. The geographic hierarchy can help eliminate this confusion. In Nigeria, for instance, combining the LGA name with the state name would yield a unique geographic identifier. Figure 8 gives another example of how unique geographic identifiers can be generated; by combining province and district names in this example, since two districts have identical names of Huma.

* Since 1999, Kenya has changed its administrative structure. Consequently, the structure in table 3 is no longer current.

Figure 7. In 2002, duplicate local government area names, Bassa, were being used in two states.

Source: Administrative boundaries for Nigiera downloaded October 2008 from http://gisweb.ciat.cgiar.org/povertymapping/.



Province	District	NameID	ChildrenServed
Alboma	Getar	Getar:Alboma	388
Bronip	Huma	Huma:Bronip	69
Bronip	Star	Star:Bronip	62
Bronip	Trethel	Trethel:Bronip	587
Delmet	Dumwick	Dumwick:Delmet	95
Delmet	Huma	Huma:Delmet	258
Delmet	Jedanga	Jedanga:Delmet	267
Spudow	Flennet	Flennet:Spudow	12
Spudow	Pranan	Pranan:Spudow	412
Spudow	Spilitina	Spilitina:Spudow	426

Figure 8. Example of creating unique geographic identifiers by combining administrative names from different levels of the geographic hierarchy. In this case, combining district and province names to produce NameID to resolve the problem of two districts with the same name.

In figure 8, note that the field NameID uses a colon as a delimiter, to make it easier to separate out the values. For data quality purposes, it is important to identify duplicate administrative names and to verify the accuracy of data assigned to them before proceeding.

Numeric codes can also be assigned to units to help create a unique id for an administrative unit. For example, in Kenya, the geographic hierarchy used for the census has been augmented with numeric codes to create unique geographic identifiers (table 4). The use of unique geographic identifiers, which ensures uniform identification of geographic entities throughout the system, provides the added benefit of allowing census data to be merged more easily with data from other sources. It is important to ensure that any system can accommodate future changes to the underlying geography, since boundaries can change. Many countries had developed official geographic identifiers. Whenever possible, these official identifying schemas should be used. For more information, refer to a country's mapping agency, bureau of statistics, or census bureau.

Table 4. Geographic Identifier Scheme for 1999 Kenya Census

Administrative Unit	*Number of Digits**	*Example Unit*	*Example Code*	*Example of Full Code*
Province	1	Central	2	2
District	2	Kiambu	01	201
Division	2	Limuru	01	20101
Location	2	Ngecha	01	2010101
Sub-location	2	Kabuku	01	201010101
Enumeration area	4	EA	0011	2010101010011

* Total number of digits = 13.

Source: Odhiambo E. "Census Cartography: The Kenyan Experience," presented at the United Nations Expert Group Meeting on Contemporary Practices in Census Mapping and Use of Geographical Information Systems, May 29-June 1, 2007. New York.

Beyond administrative units, such as districts, there are other ways that geography can be represented in data. For instance, communities, towns, cities, and other settlement areas are important sources of potentially useful data. These can be identified via name or code. Additionally, buildings or other specific locations can be represented by either an address or a GPS coordinate. The process of converting street addresses into point locations for use in a GIS is called *geocoding*.

GPS Data Collection

GPS (the satellite-based global positioning system) provides users the ability to locate phenomena very accurately. While all GPS receivers will provide coordinates locating a single point, some receivers can be used to collect multiple points to construct paths or polygons. These features make it possible to record roads or streets or create polygons that can represent

service areas, neighborhoods, or other important features. GPS receivers record locations on Earth with a high level of accuracy by receiving signals from satellites in space. Most GPS receivers are simple to operate; however, some care is needed to make sure that the points are as accurate as possible. Before collecting locations with a GPS receiver, it is important to develop a solid data collection strategy. The exact process of data collection will vary based on the type of receiver and the data collection needs. However, there are important considerations in all uses of GPS (see Four Steps for Successful GPS Data Collection, next page).

Once the GPS receiver has locked onto signals from the satellites, it will display the current location as coordinates. There are many coordinate systems the receiver can use, but a latitude and longitude reading is the most common (figure 9). There are several different ways that latitude and longitude can be stored. The best format for use in most mapping programs is decimal degrees. Another common format used by many GPS receivers is decimal minutes. In order to overlay points with other existing map information, it may even be necessary to use a different coordinate system other than latitude/longitude. When using any coordinate system, it is extremely important to note the format, to be consistent in using that format, and to make sure that such details as positive/negative signs and decimal precision are maintained when storing the coordinates in a database.

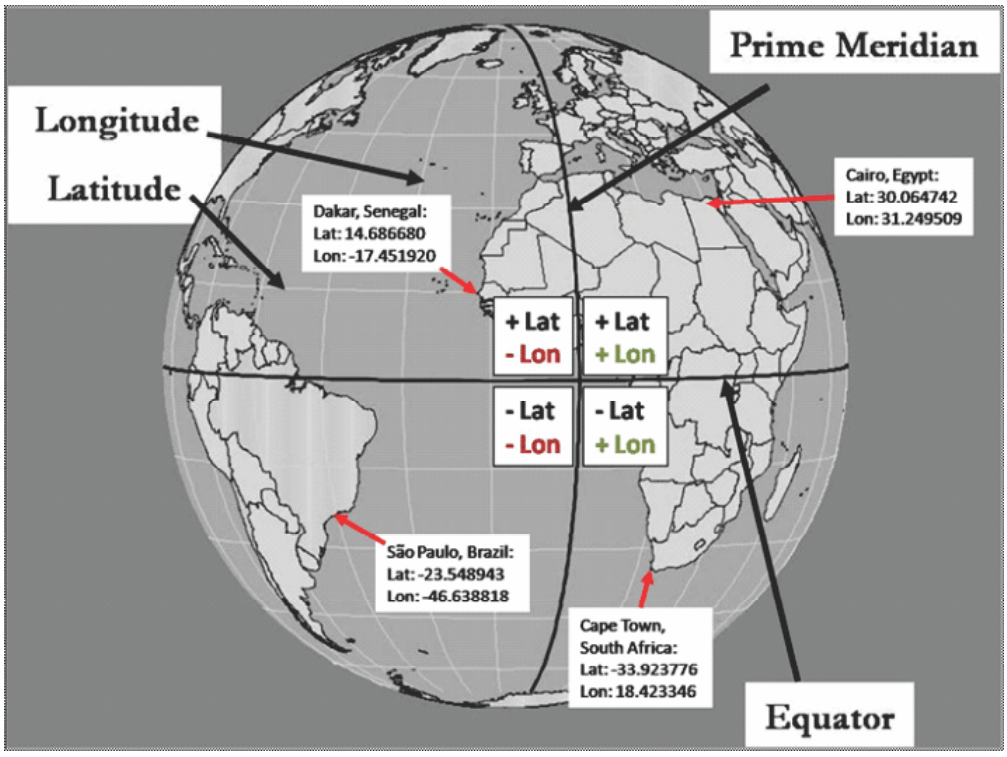


Figure 9. Examples of latitude and longitude readings in different locations around the globe.*

* Note the use of positive and negative readings above and below the equator, and also east and west of the prime meridian.

Four Steps for Successful GPS Data Collection

1. *Plan a data collection strategy* — Know what you want to locate and the rules for collecting the point. For instance, if collecting the location of a facility, do you collect at the front door or at the street?
2. *Plan a data entry strategy* — How will you transfer the data from the GPS receiver to a database? Many GPS receivers will do an electronic transfer of points from the receiver's memory to a computer file, which makes it easy to use your data and ensures fewer transfer errors. However, it can still be important to record manually on paper the coordinates when you collect them, in case the receiver is lost or malfunctions before the points can be downloaded.
3. *Consistent data* — Make sure your GPS data IDs match up with any other data you are collecting. For example, IDs might be assigned to participants in an interview; these IDs will need to be recorded as part of each GPS data point for future point identification of participants and locations. ID assignments can be noted in the metadata.
4. *Set up receiver properly* — Make sure you have enough batteries or cables for the receiver. Allow adequate time and space for the receiver to receive an accurate signal. Additionally, it is important that the settings of the receiver are correct. Collecting data in latitude/longitude with WGS 84 datum is a common setting and is generally considered a good practice. Be consistent in using degrees/minutes/seconds or decimal degrees. Using decimal minutes is a common setting; whatever setting is used can be noted in the metadata.

Putting GPS Data into a Spreadsheet

GPS data can be exported from a GPS receiver in ASCII text format and then imported into a spreadsheet for further review or for importing into a GIS. To export GPS data from a GPS receiver, follow the instructions provided in the corresponding GPS manual. For an example of how to download GPS data using the Garmin GPS 72 receiver, please see *MEASURE Evaluation Global Positioning System Toolkit* (http://www.cpc.unc.edu/measure/publications/ms-07-21/at_download/document). This publication also addresses data quality considerations during the download process, so is a good starting point for beginners or is an excellent refresher for more experienced GPS users. GPS data can also be entered into a spreadsheet manually, but this method is not recommended because of the possibility of data-entry errors.

Regardless of the method of entering GPS data into a spreadsheet, it is important to ensure the following:

- Individual GPS observations should possess a unique identifier. Duplicate identifiers will create problems, and should be corrected before the data are used for display or analysis.
- The spreadsheet field containing the unique identifier should be formatted as either "text" or "number," depending on the naming convention used. If the unique identifier begins with anything other than a number, or begins with leading zeroes, the field should be formatted as "text."
- Latitude and longitude should be entered in the spreadsheet in decimal degrees, as this format is easily understood and actionable using GIS software.

- To store decimal degrees correctly, the spreadsheet fields containing the latitude and longitude information should be formatted as a "number" with the maximum number of decimal places necessary to capture all of the coordinate information provided by the GPS receiver (commonly .000001).
- For positive latitudes (above the equator) and positive longitudes (east of the prime meridian but west of the international date line), it is not necessary to use a plus sign (+) within the spreadsheet cell. A minus sign (-) should be used, however, to identify negative latitudes (below the equator) and negative longitudes (west of the prime meridian but east of the international date line). The sign becomes especially important in areas straddling the equator or prime meridian.

Figure 10 is an example of coordinates extracted from a GPS receiver and imported into a spreadsheet. Note that the unique identifier field (column A, GPSID) is formatted as "text" in

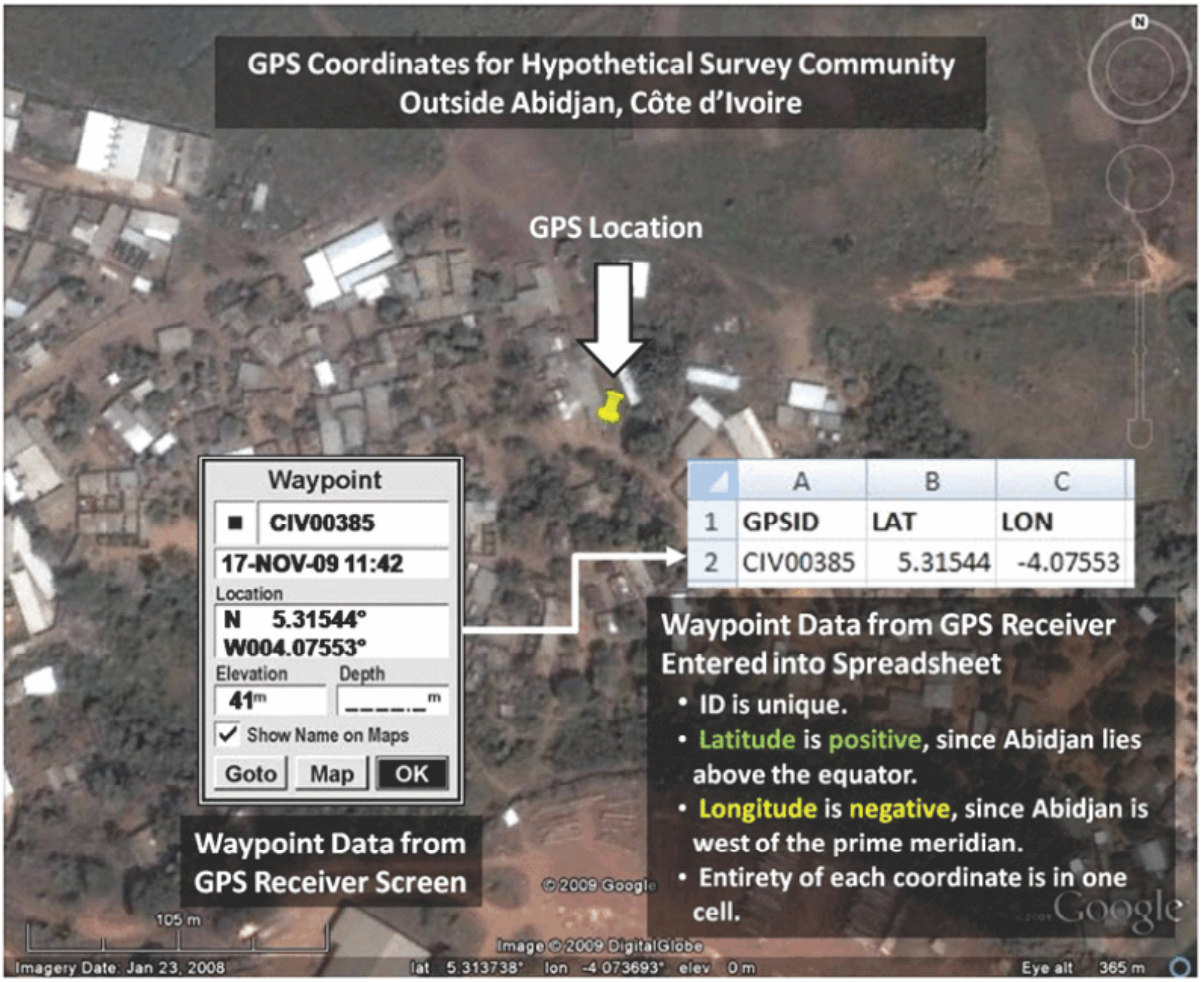


Figure 10. Example of GPS data from a handheld receiver (left), and the same coordinates after being moved into a spreadsheet (right).

Map image source: Google Earth, accessed December 2009.

the spreadheet. The latitude (LAT) and longitude (LON) were returned by the GPS receiver with five decimal places, so the fields LAT and LON in the spreadsheet are formatted as "numbers" with five decimal places. (The default number of decimal places within Microsoft Excel spreadsheets is two places, so it is important to change this within the spreadsheet according to the circumstances, and to save the change. If not, there could be a loss of positional accuracy.)

Another important setting with a GPS is the datum setting. Datum refers to the specific mathematical model of the earth used for mapping. There are many different datums GPS receivers may use, but the default for most receivers is WGS84. If points collected with a GPS receiver do not line up with other features properly when mapped, it could be the result of a datum conflict. In such cases, it may be necessary to use a GIS to convert the points to a consistent datum.

The GPS receiver's manual should be consulted to determine how to make sure the unit collects data in the proper coordinate system and datum. Many receivers are different, but there are some rules that should be followed with any GPS collection:

- *Allow time for the receiver to receive an accurate signal* — When first turned on, GPS receivers need time to receive the signals fully from the GPS satellites (generally a minimum of four visible satellites) in order to calculate an accurate position. The time required for this varies among different receivers, length of time since a receiver was last turned on, position of satellites in the sky, and even atmospheric conditions. Most receivers will indicate when they have received a strong enough signal to calculate an accurate location. Do not record the coordinates until the receiver has indicated it has received enough information to calculate an accurate location (incorrect coordinates may be hundreds of kilometers off).
- *GPS coordinates are not a unique identifier* — It is important to remember that GPS coordinates are not a unique identifier. Because of variations in signal quality, for example, readings taken in exactly the same location at two different times can have different coordinates. This could mean that when a set of coordinates are mapped, such as for a building, that the location might appear to be building A on one occasion and building B the next. Also, it is possible for more than one geographic entity to have the same coordinates, such as two different addresses in the same building, both collected at the front of the building. The resulting non-unique geographic identifiers could present difficulty for subsequent analysis of the data.

Privacy Issues Concerning Point Location Data

Spatial location of an individual can serve as a de facto identifier. For some data, this means inclusion of spatially identifying information requires precautions consistent with preserving the confidentiality of the data. If this is impractical, there are approaches that can be employed that can mask the true location, and thereby preserve confidentiality.

Some examples of masking the true location include:

- shifting point locations randomly within a maximum allowable distance to hide true locations;

- distorting the location or shape of identifiable map features such as roads, rivers, or populated places; and
- generalizing locations by rounding up or down the values of the coordinates.

There are no one-size-fits-all solutions to preserving confidentiality and maintaining the spatial integrity of the data. Each situation will require data collectors to make decisions about the risks and rewards of a particular strategy. For more information, consult the following:

VanWey LK, Rindfuss RR, Gurmann MP, Entwisle B, Balk DL. Confidentiality and spatially explicit data: concerns and challenges. *PNAS.* 2005;102(43):15337-15342.

MEASURE Evaluation GIS Working Group. Overview of issues concerning confidentiality and spatial data [working paper WP-08-106]. Chapel Hill, NC: MEASURE Evaluation; 2006. Available at: http://www.cpc.unc.edu/measure/publications/wp-08-106/at_download/document.

Finding and Using Existing Spatial Data

Key Message

How to format and collect spatial data is presented, including the importance of metadata, accuracy, currency, and source. Specific file formats and software are discussed.

In order to map coordinate or area data that are collected as part of a routine health information system, a minimum of nearby reference information is required. This information is commonly called a base map. The first place to begin the search for supporting base map information is with the national mapping agencies (NMAs), which maintain national-level data sets such as political boundaries, topographic map series, geodetic control networks, and aerial photography or satellite imagery. All of these can be of great use in locating key health-related features. Topographic maps, for example, can help identify such features as populated places, transportation routes, and areas of poor drainage (figure 11). Depending on their scale, they can also help identify facilities such as hospitals and schools. Often the NMAs offer complementary spatial data sets as well, such as land cover or transportation. By virtue of their central role in NSDI efforts, these agencies are uniquely positioned to serve as a first point of contact. The most well-maintained list of NMA contacts available on the Web is provided through the United Nations Geographic Information Working Group (UNGIWG) via the Web site for second administrative level boundaries (www.unsalb.org).

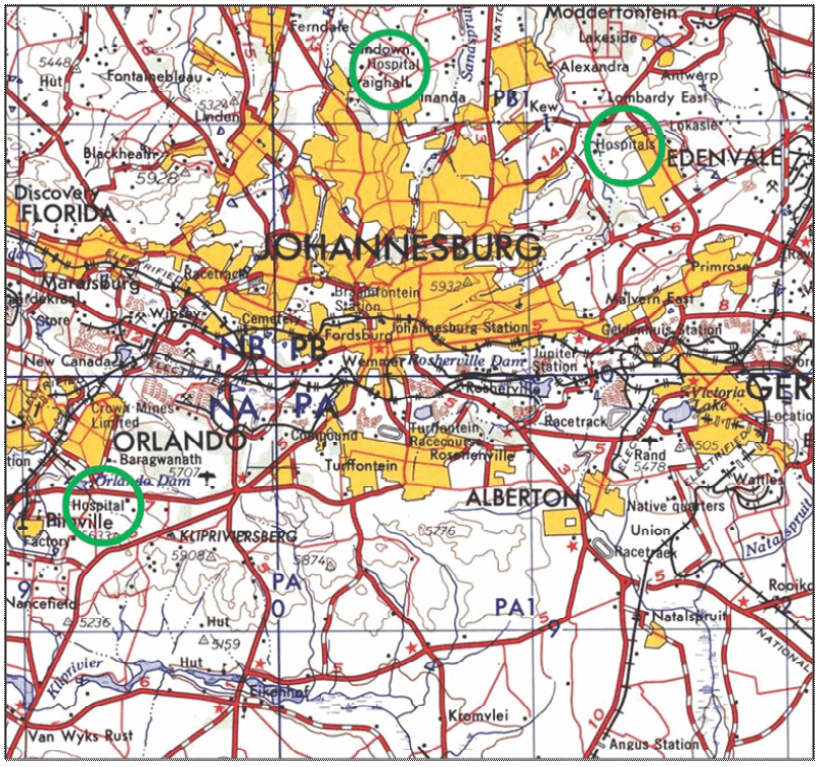


Figure 11. Topographic map of Johannesburg, South Africa, with three hospitals circled.

Map image source: U.S. Army Map Service, 1962. Courtesy of University of Texas Libraries, University of Texas at Austin. Downloaded December 2009.

The NMAs are not the only source of spatial data, however. A country's national statistical office or census bureau, for example, can be a rich source of spatial data, such as boundary files for census enumeration

areas. This is significant, as census enumeration boundaries can often be used in combination with demographic, socioeconomic, and health data collected at a common scale to create maps and other spatial products that help decision makers identify patterns that might not otherwise be detected.

Beyond the NMAs and statistical offices, spatial data are often available from an increasing variety of sources, such as regional GIS centers, academic institutions, local governments, private vendors, public-private partnerships, and Web sites. For more detail on spatial data sources, please see appendix 4.

Before spatial data are used, there are some key attributes of the data that should be considered. These include accuracy, currency, source, coordinate system and datum, file format, and availability of metadata.

Accuracy

The accuracy of spatial data is important, as this determines the suitability of the data for specific uses, as well as the reliability of the data for decision making. Inaccurate data can lead to false statements and conclusions. For example, figure 12 shows two different boundaries for the same informal settlement outside São Paulo, Brazil, in 2002. If health officials were to plan services based on a household count obtained using the border in image A, the level of services would be underestimated. Using the correct boundary, as shown in image B, would allow a more accurate count. Using an inaccurate boundary could have an impact on resource assignments for fieldwork, lead to false assumptions about population density in the settlement, or generally have an adverse effect on tasks associated with providing the best level of services to the population in question.

A large part of accuracy assessment is determining how well the different spatial data themes or layers align with one another. District boundaries, for example, might not align with street files or other data sources. This could have a significant impact on the ability to achieve reliable results from any analysis undertaken.

Likewise, the accuracy of the attribute data, especially the geographic identifiers, is also important. If the geographic identifiers are missing or incorrect, they could have a serious impact on one's ability to use the data, especially for mapping and spatial analysis (table 5).

Currency

The currency of spatial data is of critical importance, as the data can change with the passage of time, and therefore become useless. Revisiting the example of the informal settlement outside São Paulo, Brazil, an out-of-date boundary might look quite different from a current boundary (figure 13).

Out-of-date data should not be trusted to be representative of current conditions. As a result, it is important to know the time period for which the spatial data are relevant.

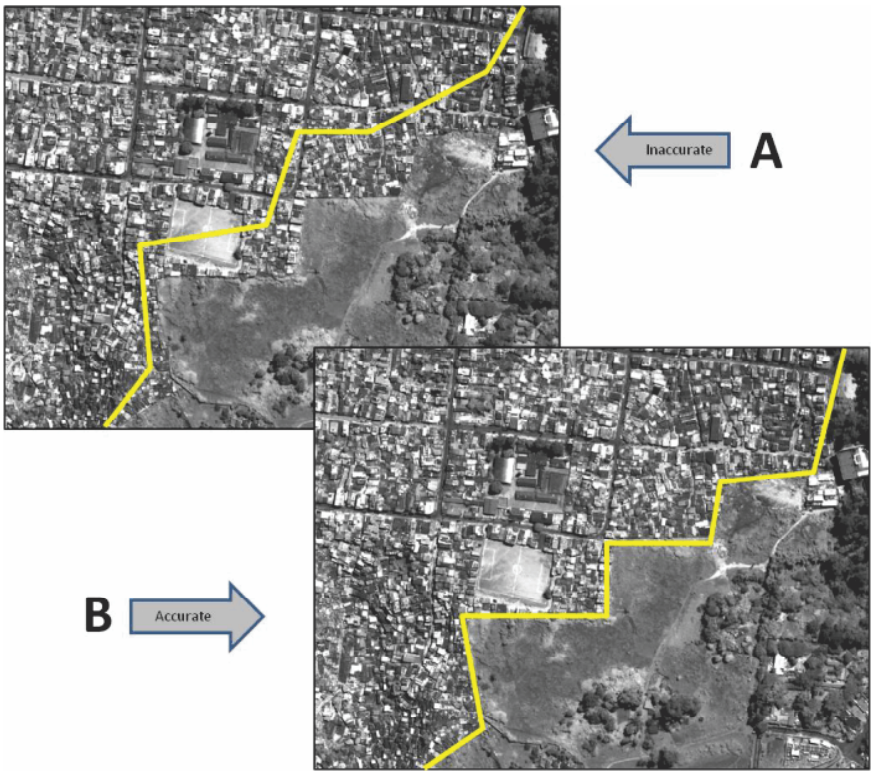


Figure 12. Settlement near São Paulo, Brazil, showing inaccurate boundary (A), and accurate boundary (B).

Map image source: QuickBird panchromatic satellite image from Digital Globe, 2002.

Table 5. Table with Inaccurate (Missing*) Geographic Identifiers

Province	*District*	*Children Served*
Alboma	Getar	388
Bronip	Star	62
Bronip	Trethel	587
Delmet	Dumwick	95
Delmet	Jedanga	267
Spudow	Pranan	412
Spudow	Spilitina	426
Spudow	***Unkown***	12
Unknown	Huma	69
Unknown	Huma	258

* Missing geographic identifiers are marked ***unknown***, in bold italics.

Source

Knowing the source of spatial data is essential, as some sources are much more reliable than others. For administrative boundaries, for example, the United Nations provides administrative boundaries through the Second Administrative Level Boundaries (SALB) project (www.unsalb.org). The boundaries obtained from the UNSALB project may not always be completely current, but they are among the most reliable boundaries available for the time period to which they correspond by virtue of the fact that they have been vetted and approved by the national mapping agencies involved in their creation. Many other data sources have not been subjected to the same rigorous level of quality control.

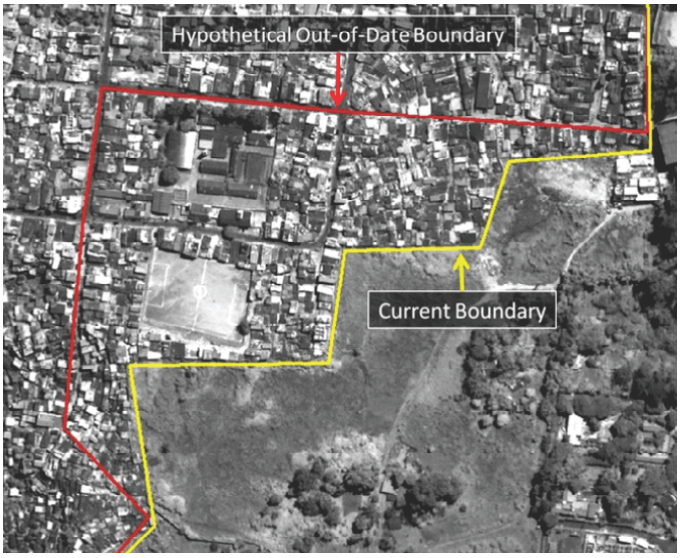


Figure 13. Example of current boundary (yellow) and out-of-date boundary (red).

Map image source: QuickBird panchromatic satellite image from Digital Globe, 2002.

Knowing the source of spatial data should also allow one to contact the originating organization in case there are issues with the data or there is a need for technical support.

Coordinate System and Datum

When spatial data layers are used together within a GIS for analytical purposes (as opposed simply for display, for example) they should share the same coordinate system and datum. This can be accomplished using built-in commands within a GIS such as ArcGIS or QGIS, or via third-party tools, such as the Geographic Calculator from Blue Marble Geographics.

File Format

To use spatial data, it is essential to understand the format in which the data are supplied. Spatial data can be stored in raster or vector formats. Raster is used for imagery, often captured via satellite. A common type of raster format for spatial data is called GeoTIFF, which stores not only pixel-by-pixel values in a grid pattern, but also a georeference for that grid so that it may be tied to a particular point on the earth. Gridded population maps are stored in this format. For more information on the many open source formats for raster data, visit the Geospatial Data Abstraction Library, at www.gdal.org.

Points, lines, and polygons (areas) are types of vector data. Three common formats are *shapefiles*, *KML*, and *plain text* (these formats are examined further on the next page). For health information and population demographics, it is most likely the data will be in vector format.

Shapefile — One of the most common formats for spatial data is the electronic shapefile, which was developed by ESRI (www.esri.com).

A shapefile is actually a collection of at least three files:

- a main file, which contains geometric information for the features of interest on a record-by-record basis;
- index file, which identifies the positional offset of each record in the main file from the beginning of the main file; and
- dBASE file, which contains a table of attribute data for each geometric feature described in the main file.

A shapefile can also have a projection file, which is optional but highly valuable.

To maintain their association, each of the component files for a shapefile must have the same root name. To be recognized as the individual component files, however, they must possess different file name extensions. Following is an example shapefile for health districts in Sénégal:

- main file: Senegal_districts_sanitaires.shp
- index file: Senegal_districts_sanitaires.shx
- dBASE file: Senegal_districts_sanitaires.dbf
- projection file: Senegal_districts_sanitaires.prj

The ESRI suite of software tools offers several ways to create shapefiles. Due to the open nature of the shapefile specification, it is also used by many open source GIS applications, such as QuantumGIS.

KML — Keyhole markup language or KML is a file format used to display geographic data in a Web-based application. KML can be displayed using a variety of "geo-browsers," such as Google Earth, Google Maps, NASA WorldWind, and ESRI's ArcExplorer and ArcGlobe. (These geo-browsers all use raster-based imagery as background information. The KML information is typically displayed on top of this imagery.)

KML files can be e-mailed or posted on a Web server for others to access. As a result of its open nature and broad application to geographical data display via the Web, KML has been strongly supported by Google and adopted as an international standard by the Open Geospatial Consortium (www.opengeospatial. org).

KML files, which have the KML extension (e.g., Senegal_districts_sanitaires.kml), can be converted from other vector-based formats or created using XML or simple text editors, such as the following example from kml-samples.googlecode.com (accessed January 2010):

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://www.opengis.net/kml/2.2">
  <Placemark>
    <name>Simple placemark</name>
    <description>
      Attached to the ground. Intelligently places itself at the
```

```
    height of the underlying terrain.
   </description>
   <Point>
    <coordinates>-
122.0822035425683,37.42228990140251,0</coordinates>
   </Point>
 </Placemark>
</kml>
```

KML files can be compressed into ZIP-format archives (.KMZ extension) for more efficient storage and display. The ArcGIS software package from ESRI, for example, provides a tool to convert a spatial data layer into a KMZ file. The best online sources of information on KML are as follows:

- Google: code.google.com/apis/kml/documentation/
- open Geospatial Consortium: www.opengeospatial.org/standards/kml/

Text — Some spatial data, such as GPS coordinates, can be provided in simple text files. The following, from Google Maps and Google Earth, is an example of venues for the 2010 World Cup soccer games, in simulated GPS data using simple text comma-delimited format:

WaypointID	***Latitude,Longitude***	***Description***
SV0001	-33.903429,18.411171	Green Point Stadium; Cape Town
SV0002	-29.829073,31.030327	Moses Mabhida Stadium; Durban
SV0003	-26.197585,28.060733	Ellis Park Stadium; Johannesburg
SV0004	-26.234987,27.982577	Soccer City Stadium; Johannesburg

The preceding text file contains geographic (unprojected) coordinates in latitude and longitude expressed as decimal degrees. The decimal degrees format for coordinates is easily understood within a GIS, so is highly practical. What is not identified is the datum. To avoid difficulties, it is important to identify and keep track of the datum used for the creation of spatial data. If the datum is known, reprojecting coordinates to another coordinate system and datum in a GIS is relatively straightforward.

Availability of Metadata

Regardless of the format in which spatial data are provided, best practices require that they be accompanied by metadata. Metadata are summary, text-based data used to describe spatial data. They provide information on such things as the scale, currency, source, coordinate system and datum, file format, access constraints, etc. Spatial data obtained without metadata should be considered of suspect quality as the spatial data could be out-of-date or even incorrect.

To be of practical use, the metadata provided with a set of spatial data should, at a minimum, include the following:

- name of the data set
- description (short narrative that includes identification of geographic area covered)
- source (originating organization and contact information)
- dates of data collection/creation

- coordinate system and datum (especially if projection file not provided)
- scale at which data is intended to be used

Table 6 contains an example of metadata provided with a file of administrative boundaries downloaded from the UNGIWG's Second Administrative Level Boundaries (SALB) Web site .

The International Organization for Standardization (ISO) has published an international metadata standard known as ISO 19115. According to the ISO Web site (www.iso.org, accessed December 2009), ISO 19115 "defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data." The ISO 19115 metadata standard also defines the following:

- mandatory and conditional metadata sections, metadata entities, and metadata elements;
- the minimum set of metadata required to serve the full range of metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data);
- optional metadata elements, to allow for a more extensive standard description of geographic data, if required;
- a method for extending metadata to fit specialized needs.

For the United States, the federal metadata standard was developed by the Federal Geographic Data Committee (FGDC), an interagency committee responsible for national geospatial data standards (www.fgdc.gov). The resulting standard is called the Content Standard for Digital Geospatial Metadata (CSDGM), Vers. 2 (FGDC-STD-001-1998). For compatibility with ISO 19115, the FGDC is leading the development of a U.S. profile of the standard.

Table 6. Example of Metadata

Term	Description or Comments
Dataset Title	Second Administrative Level Boundaries of the Democratic Republic of the Congo
Theme Keywords	Democratic Republic of the Congo, administrative boundaries, sub
Dataset Topic Category	Boundaries
Geographic Location	Democratic Republic of the Congo
Dataset Reference Date	20080703 (Dataset Reference Date Type: Publication)
Representativity	Start Date: March 2003 End Date: April 2008
Abstract	This ARC/INFO polygon coverage contains the second level administrative boundary. The coverage is part of the SALB dataset.
Supplemental Information: In order to ensure a close match between different national coverages, and to obtain compatibility with other standard medium resolution data sets, all national boundaries were replaced with the international borders boundary database developed by the UN Task Group on International Borders (International Boundaries Database, Version 1.0). Due to differences in the quality of the documents compiled, the SALB data set is better adapted for thematic mapping than for depicting precise locations or spatial modeling. It is therefore recommended not to use this data at a scale below 1:1 000000. The coding scheme combines the ISO3 alpha code and a numerical code.	
Dataset edition	First edition
Data quality comments	Due to differences in the quality of the documents compiled, the SALB dataset is better adapted for thematic mapping than for depicting precise locations or spatial modeling.
Source map format	Digital map
Source map name	Limites Administratives RGC
Distributor	UNJLC
Lineage: Some units have been merged in the source file and some other have been added by digitization of paper maps. The resulting file has been adjusted to fit the international borders standard.	
Spatial representation type	Vector
Geometric object type	Surface
Map projection	Un-projected (geographic)
Reference system	WGS 84 datum
Geographic box	***X min:*** 12.206632 ***X max:*** 31.305912 ***Y min:*** -13.455676 ***Y max:*** 5.386098
Data exchange format	ArcInfo Interchange format: e.00, ArcView Shapefile: shp
Restrictions: The "Référentiel Géographique Commun, RD Congo" owns the copyright of the original document and the UN of the modifications made on this document.	
Acknowledgments: The source map has been provided by the "Référentiel Géographique Commun" and modified by the UN (2007) in collaboration with the "Institut Géographique du Congo" in the context of the SALB project (2008).	
Validation Institution	AInstitut Géographique du Congo
Online Linkage	http://www.who.int/whosis/database/gis/salb/salb_home.html
Dataset Language	en
Dataset Character Set	usAscii
Metadata Provider	World Health Organization - EIP/KMS/EHL/STK
Metadata Contact	Evidence and Information for Policy Knowledge Management and Sharing World Health Organization 20, AV. Appia 1211 Geneva 27 Switzerland Fax: +41.22.791.43.28
Metadata Date	20080703
Metadata Language	en
Metadata Character Set	usAscii
Metadata Standard	ISO 19115

Table source: SALB Web site, accessed July 2008.

Analyzing Data Using Spatial Tools

Advantages of Analysis When Including Geographic Components of Data

When data have spatial components, they can be grouped and compared in many different ways, some of which have been shown in the data schema section of this guide. Information on setting characteristics becomes much richer when geographic components are included in the data. Not only can such geographically determined variables as urban/rural be extrapolated when the locations of the data are known, but a whole host of other possibilities open up. Other variables that might also be extrapolated may include distance to a water source, distance to an urban area, and distance to transportation routes or to a medical clinic. Using a GIS, these extrapolations can all occur long after the initial data set has been collected.

A geographic component also lends itself to the creation of more meaningful variables for comparison purposes. It becomes possible to incorporate the percentages of other variables on an area-by-area basis; e.g., the number of orphans per population under age 18, by province. The geographic division (administrative unit) is a useful way of grouping and comparing data at several different levels of detail (i.e., country, province, district).

Key Message

Geographic components of data provide an effective way to compare variables and complicated concepts. Insights may be easier to discover, while errors or missing data may become more apparent.

Visual Display of Data

Another advantage to data that contain spatial components is that they allow comparison of several different variables within the same graphic (map). Complicated concepts can be distilled into useful graphics using simple color and symbol changes. For example, the map in figure 14 shows percentages of children attending school and a breakdown of the types of school students are attending by province in a Nigerian study area, using thematic classing and pie charts. It is an interesting comparison of related types of data shown together in one graphic. It is much harder to tell from a series of tables than from a single map that there may actually be a relationship between the percentages of children attending school and the number of school choices available in an area.

By exploring the data in these types of visual ways, a researcher may unearth various patterns or anomalies in the data that warrant further analysis. The top map in figure 14 follows the incidence of disease A in a hypothetical country, with the disease appearing to be high in the districts located along the Current River, in the western portion of the country. Further exploration of the data may reveal a link between variables associated with water quality and variables associated with disease A. The map reveals these sorts of possible relationships between variables that may not have been otherwise apparent.

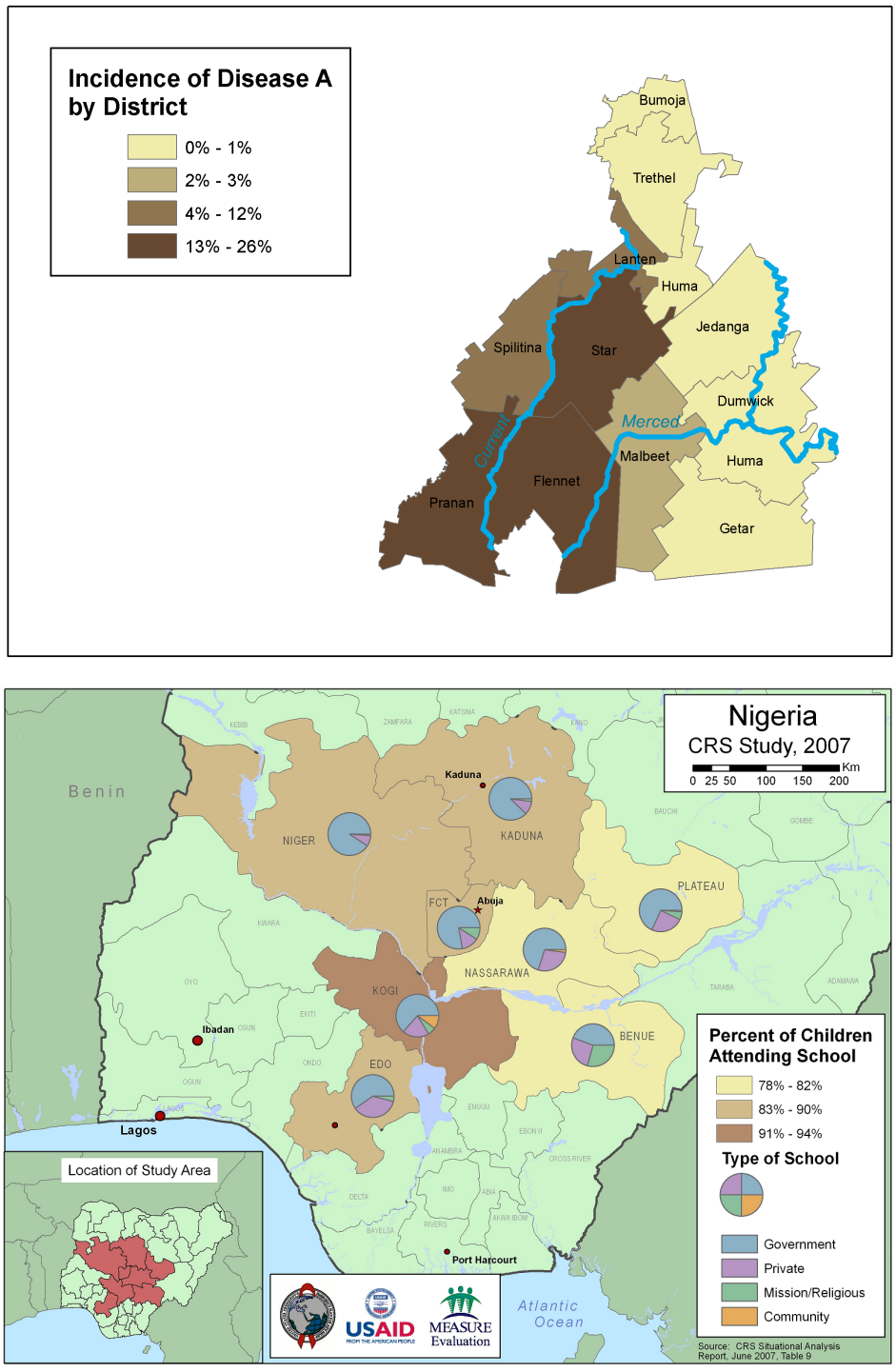


Figure 14. Examples of displaying concepts on the same map show the incidence of a disease in a hypothetical country (top) and percentages of Nigerian children attending school, as well as types of school (bottom). The bottom map uses data gathered by Catholic Relief Services.

Errors or anomalies in the data will also tend to stand out when displayed spatially. In figure 15, data from a spreadsheet are provided spatially on maps, where missing data and incorrect data can be easily seen.

Another beneficial aspect of visual (map) display of data is that these diplays can provide an opportunity for "aha" moments during the data exploration or analysis process. Sometimes unexpected relationships between variables appear. In the map of Nigeria (figure 16), one might expect percent orphans to increase with poverty, but putting the two variables together on a map clearly shows the inverse to be true, at least for the case of these particular locations.

GIS vs. Geo-display

What's the difference between a GIS (geographic information system) and a geo-display? A GIS is generally used to perform analyses on data that are geographical in nature. A program that does only geographic display typically will display a map or a satellite image with simple line or point overlays. Most software that carry the GIS connotation use data in a database and can perform spatial analysis actions, such as overlays and buffers. Such software can measure distances in a gradational fashion and compare data values in a number of tables at once.

Other programs (geo-display) simply take one table of data and class it for display on a map. These programs generally do not perform any kind of

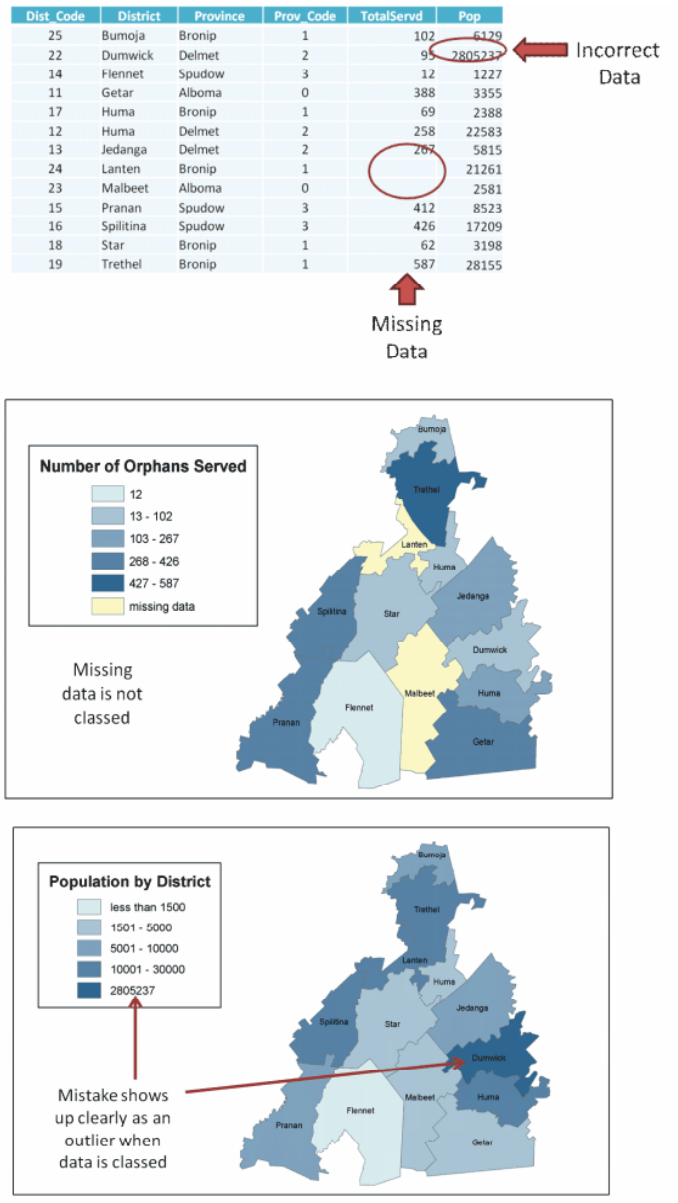


Figure 15. Missing data on number of children served in a hypothetical country (top map) and an incorrect population for a district (bottom).

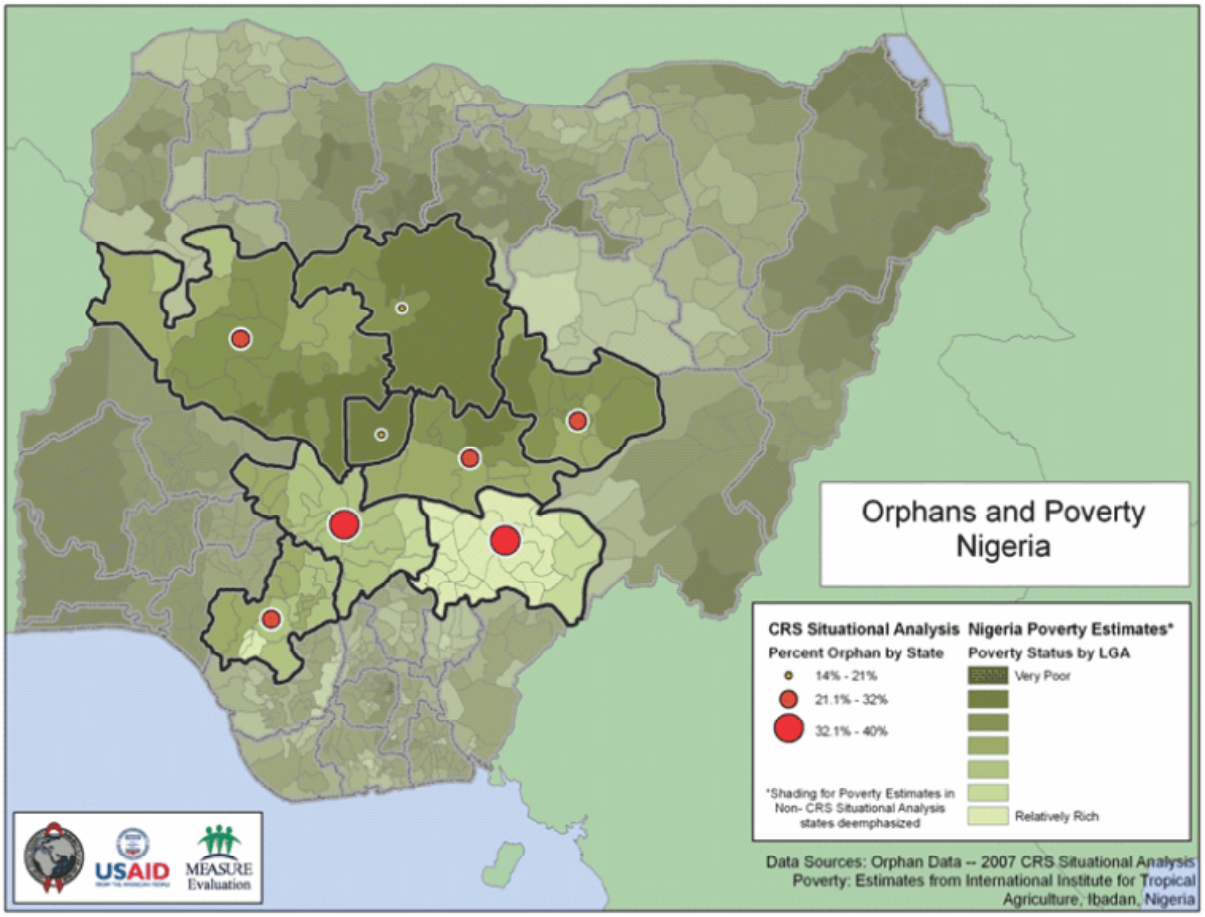


Figure 16. Percentages of Nigerian orphans within a study area are contrasted with levels of poverty.

measurement or special statistics on the data. They may or may not have a graphing capability in addition to their mapping capability — though graphs from the data may need to be created using another program, such as a spreadsheet or database program.

GIS programs typically come with a high learning curve. The better-known programs come with good documentation, but are expensive. There are a number of "open-source" programs that are free or inexpensive, but these can lack good (or up-to-date) documentation and users sometimes create their own modules, or pieces of software, which can be put together in various ways for different types of data analysis. These often require the services of a programmer/analyst to be useful.

Geo-display programs are generally less expensive or free and have a much smaller learning curve. Those that have good documentation can be quite useful in producing simple thematic maps, which, as said earlier, can help to identify problems in the data or possible relationships

in the data that may warrant further analysis. The simple maps produced from these types of programs can be quite useful in making a particular point with policy makers that would be much less obvious or less visually striking and memorable in the more traditional tabular format. A detailed comparison of software options is provided in appendix 1.

Service Maps

Mapping access to health care services is important for national planning purposes. Point locations of service providers can be mapped. Distances can be calculated from these points to nearby major roads and populations to indicate ease of access. In a study by Noor and colleagues, the locations of public health service providers were derived using a combination of GPS coordinates, topographic maps, hand-drawn maps, and Google Earth. Then using a raster (100x100m pixels) population density map and calculating distances from each pixel to the nearest facility (with the help of ESRI ArcGIS software), a classified population map was created. The resulting map showed population living more than than five kilometers from a facility. This was a relatively simple method of mapping access, and did not take into account road networks, physical barriers such as mountains or rivers, or modes of travel; but the map is nonetheless informative and serves to visually indicate areas where population to facility ratios are potentially high.[12]

Another way to show service availability with maps is by the use of buffer analysis. Buffers can be drawn in simple circular or more complicated irregular polygon patterns around each health facility. Then, using a GIS, the buffers can be compared to raster population density maps (figure 17). The population that falls inside of each buffer can then be calculated. The resulting map can help identify where the population may lack access to services.

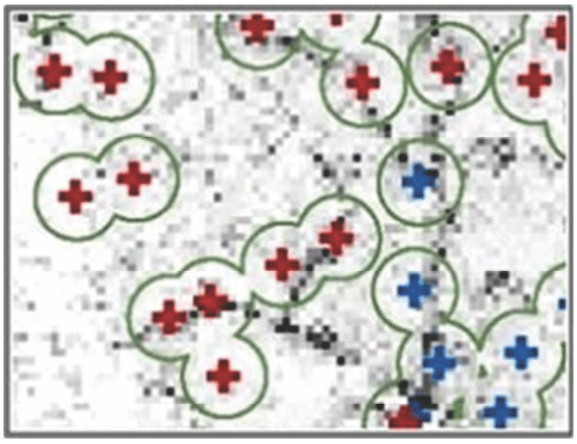


Figure 17. A simple buffer map shows uniform distance measurements from a series of points (red and blue crosses) overlaid with a population density map (darker pixels indicate higher density).

Another more detailed and realistic way to examine service availability is to use kernel density estimation to create an accessibility map (figure 18). This is a way of statistically weighing certain points to discover areas that are more or less influenced by them without regard for administrative boundaries. For more on this method see Spencer and Angeles.[13]

PLACE Studies

The Priorities for Local AIDS Control Efforts (PLACE) method is a tool to identify areas systematically that are likely to have a high incidence of HIV in order to identify specific sites within these areas where AIDS prevention programs should be focused. It can be adapted for use at the city or district level.

At the point location level, once sites of high transmission and infection (ordinarily, those with underlying sexual and drug injection networks) have been identified, they can be mapped on a simple point map. Areas of interest can often be initially identified by review of available demographic, epidemiologic, and contextual data, or through interviews with knowledgeable parties such as transportation providers or community leaders. These sites can then be visited and people socializing there can be interviewed to confirm behavior. Each site's coordinates can be recorded on a GPS as these visits are made, and then downloaded to a mapping program such as ArcGIS or Google Earth. The resulting maps, showing items such as condom use vs. new partner acquisition, can be shown to stakeholders and implementing partners. When combining point maps with other base information, such as city or slum boundaries and roads or railroads, clustering patterns may appear, such as in urban areas, in areas of overcrowding and underemployment, or along transportation routes (figure 19).

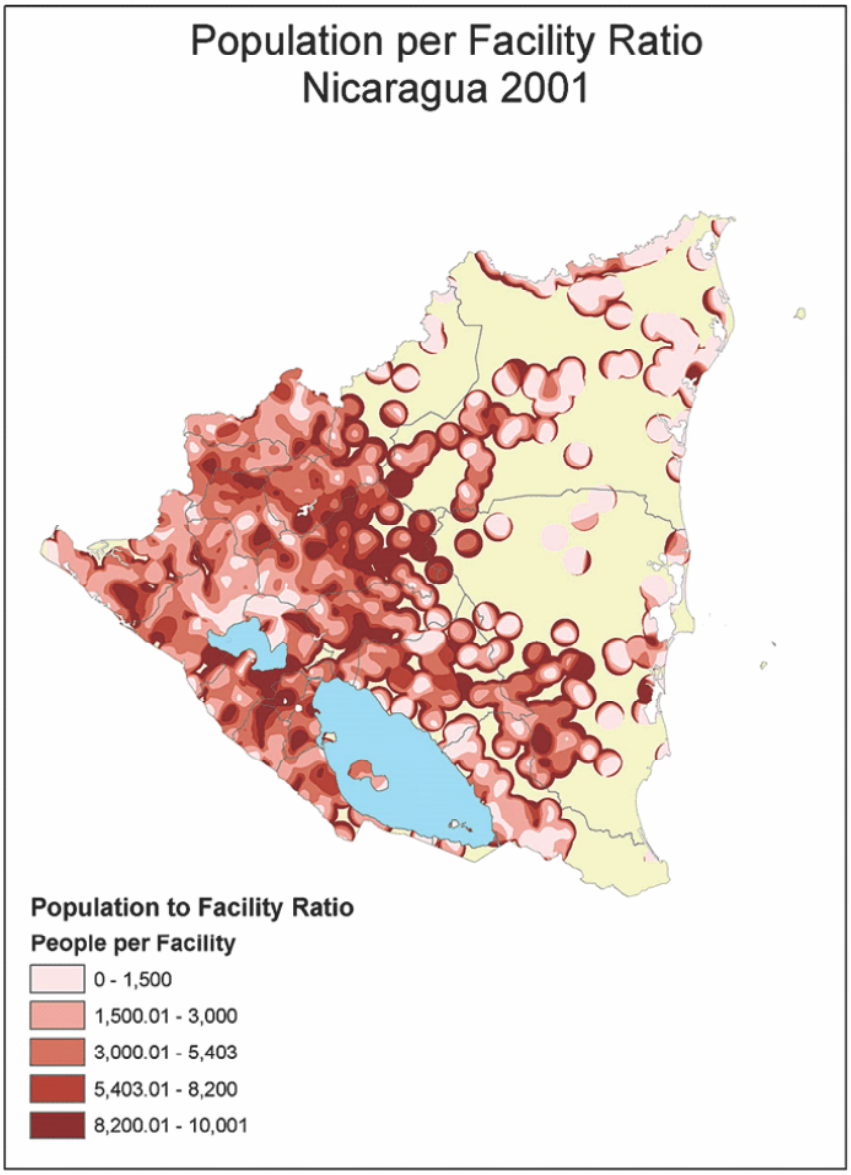


Figure 18. Kernel density estimation is used to indicate underserved communities in Nicaragua.

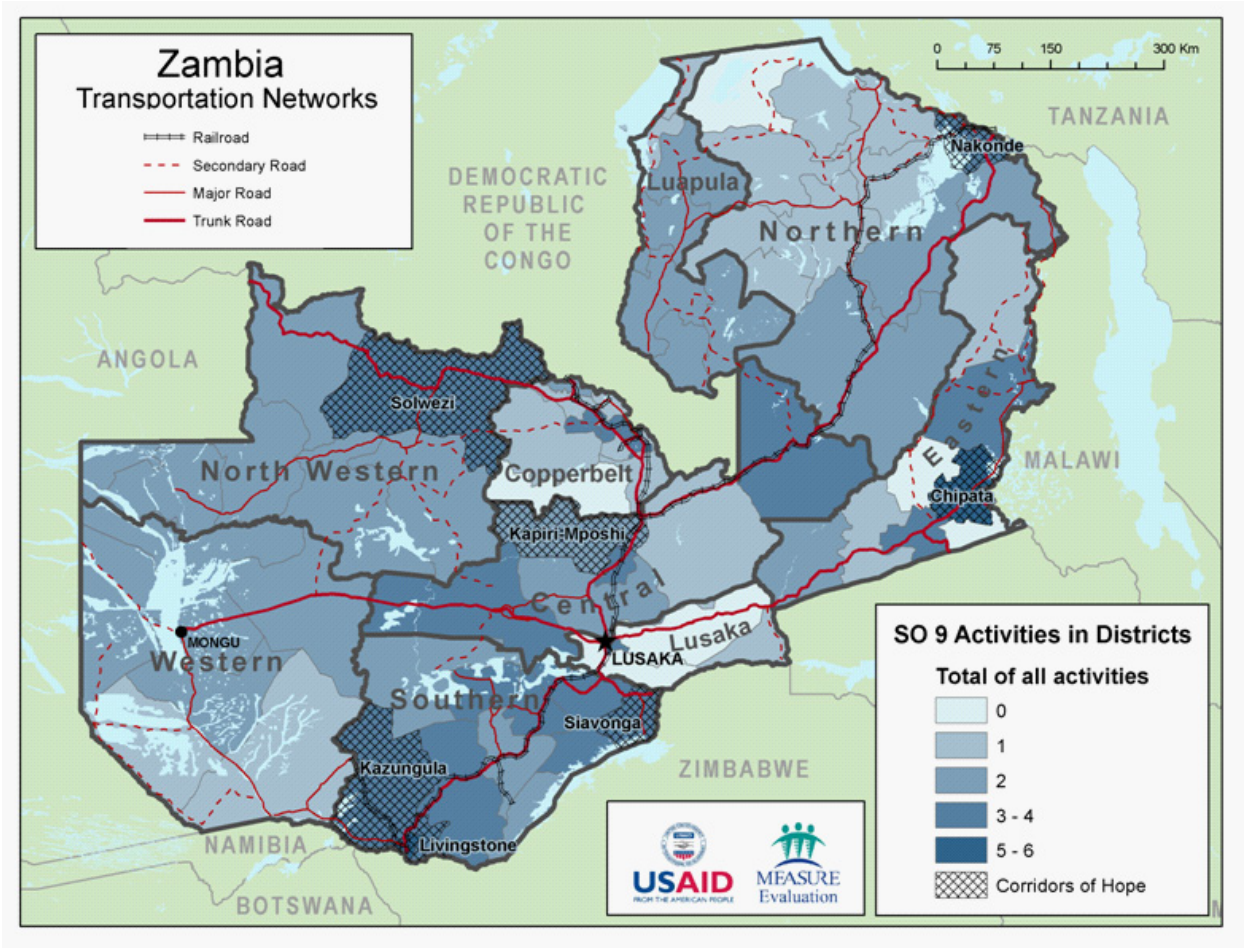


Figure 19. Map of Zambia's major transportation networks showing HIV prevention activities.*

* SO9 refers to USAID's Strategic Objective 9, "Reduced Impact of HIV/AIDS Through Multisectoral Response." These activities have recently focused on major transportation networks, as these networks are seen as transmission corridors for HIV/AIDS.

Conclusion

Geographic location is a convenient and effective common denominator that should be leveraged to the fullest extent when collecting, storing, and displaying data related to HIV treatment and prevention services. It is hoped that this guide will provide helpful guidelines for effective collection and organization of the geographic component of data, enhancing their potentially valuable role in evidenced-based decision-making. By including spatial data in data sets, key linkages can be made among a wide diversity of data such as demographic data, data from other public health programs, or even data from other sectors such as economic development, education, or agriculture. However, it is important to understand limitations of the data, data cycles, the existing existing data infrastructure, data schema, and data formats.

This document has also covered the importance of a well-defined spatial data hierarchy and how this relates to the data schema, and has described extensively special considerations for the collection and storage of GPS data in particular. The section Finding Spatial Data was included for the purpose of helping integrate third-party data into the data collection process. Spatial data carry a special set of considerations, including coordinate systems and datum, various file formats specific to spatial data, data accuracy and currency, and metadata availability.

Lastly, analysis and display of spatial data, particularly with regards to HIV-related health services, is discussed in this guide. Including the geographic component in data collection and storage makes unique methods of visual display possible. Maps can provide visualization of more than one variable at a time. Maps also allow for useful computations, such as the number of residents within a particular distance from a clinic, water source, or transportation route. The unique data visualization provided by maps can often lead to “aha” moments, revealing relationships between variables or anomalies in the data that were not previously apparent.

If we understand *where* things are happening, we are on our way to understanding *why* they are happening. Adding a geographic component to our data can not only help us visualize the data more easily, leading to effective decision-support, but also can serve as a common link across multiple data sets, making it easier to join data sets and synthesize information.

Glossary

Administrative unit — an area defined by a government to ease administration of services or to identify distinct geographic, ethnic, or cultural divisions within a country (useful in creating a choropleth map, administrative units often have a hierarchical relationship such as ward, district, province, country).

Buffer — an area created by specifying a particular distance from a point or line or polygon on a map.

Choropleth map — a visually informative thematic map that uses colors or shading to display hierarchical attribute data in geographic areas; to take into account visual confusion due to the size differences in the various areas, data should first be normalized (e.g., use population density rather than total population counts in calculating the display values).

Coordinate system — a particular reference system used to represent the locations of geographic features on the earth (i.e., latitude/longitude for points collected with a GPS unit).

Data dictionary — a list of variable definitions (often provides an explanation for the codes used for column headings in a table such as M=male, or N=total population, or A37=number of children in household ages 5-12).

Data schema — the arrangement of data in a database (involves choosing row and column headings, which will later determine the ways in which data can be displayed or joined with other data sets).

Data use cycle — a cycle of data collection and availability that leads to increased use of data, which in turn creates a demand for more data in the future.

Datum — a mathematical model of the earth used for mapping (must be specified as a setting on the GPS unit used to collect data, or in any geographic data file; multiple files used together must use [or be translated into] consistent datums).

Ecological fallacy — data error occurring from data misuse due to invalid assumptions about relationships at the individual level vs. the group level (i.e., extrapolating information on a single resident based on average values for a village).

Geocoding — the process of converting postal addresses into geographic coordinates.

Geodatabase — the method of geographic data storage and data management currently used by ESRI's ArcGIS program, which allows all pieces of a geographic data file to be stored in one folder and read by a variety of other relational database applications; geographic data in a geodatabase can be exported to and from the shapefile format.

Geo-display program — a mapping program that takes a table of data and displays it on a map (it is generally simpler to use than a full GIS, which can be useful in identifying relationships or problems in data without performing more complicated analysis).

Kernel density estimation — a geographic technique that provides a realistic representation of the spread of people and services across continuous space without the constraints of administrative boundaries.

KML file — a file format used to display geographic data in a Web-based application (i.e., Google Earth), it can be created using a simple text editor; a compressed version is known as a KMZ file.

Metadata — data about the data, telling when and how, and by whom, data were collected.

Network analysis — a method of geographic analysis that calculates measures along a network of linear entities such as roads, rivers, or railroads (can be useful in studying accessibility of health care services).

Random error — errors in a dataset introduced through user or equipment; involves problems in actual measurement often due to errors in coding or misunderstanding about questionnaires or equipment, and is an unpredictable data limitation.

Raster — spatial data stored in a grid pattern (cells) such as satellite imagery (generally requires more computer memory for storage, especially at higher resolution/smaller cell sizes); common formats include JPG and TIFF.

Relational database — a program that helps to keep data organized and manipulated by means of multiple tables and data types (very particular about data structures, which allows data to be easily imported into mapping and graphing programs).

Selection bias — bias introduced to a dataset either through the act of choosing which variables to initially include or by using a non-representative sample of a population.

Shapefile — a spatial data format developed originally by ESRI and in widespread use by other programs such as QGIS; it consists of several files including .shp and .shx (both containing geometric and positional information for vector data), .dbf (containing attribute data for features), and .prj (containing information on the coordinate system and datum).

Spatial data infrastructure — the aggregate collection of all potentially available spatially-based data within a given country (it is dynamic and ideally flexible, and includes data from diverse realms such as political or census boundaries, topography and environmental data, satellite imagery, and demographics).

Spreadsheet — an independent data table in which data can be grouped or sorted (can be simpler and easier to use than a relational database, but is not as particular about data structures and thus may not have as many flexible uses).

Systematic error — problems in data caused by mis-calibrated equipment or systematic collection problems; can often be corrected if discovered in time.

Unique identifier — a name or code that uniquely identifies a data entity (essential for distinguishing it from other entities in its own or other databases).

Vector — spatial data stored as points, lines, and polygons with various attributes (generally most effective method of spatial data storage for smooth display and analysis involving networking or buffering); common formats include KML and SHP.

Appendix 1: Mapping Software

The following provides the authors' opinions of benefits and drawbacks of commonly available mapping software. Authors are not endorsing any of these software products.

Name (Cost)	GIS Capabilities	Ease of Use	Data Input/ Output	Comments
ArcGIS[1] *(Fee-based license)*	Excellent; considered industry leader in GIS analysis	Quite complex, takes time to be proficient; very good online support and documentation	High quality; many formats available for input and output	More complexity than some organizations may need
Diva GIS *(Free)*	Several types of data analysis and display; limited data classification methods	Somewhat complex; but online documentation is available	Bitmapped output; limited layout and labeling; imports a wide variety of formats	Can be confusing; several versions available; limited output
Q-GIS[2] *(Free)*	Fairly comprehensive; good analysis available and customizable with Python language	Point-and-click interface will be familiar to users of other GIS software	Imports and exports in a wide variety of formats	Can be a good choice for experienced GIS users with limited resources
E2G Tool[3] *(Free)*	Not a true GIS; intended only for data classification and map display	Easy, good online documentation including video tutorials	Input process helps identify data problems; output is best for screen viewing	Limited number of countries available for mapping; easy and quick to learn and use
Dev-Info[4] *(Free)*	Limited; mainly a data display tool	Easy to use with available boundaries and data; free online support available	Easy to share maps and incorporate into publications; but few sub-country boundary files are available	Data formats not easily converted; limited advice on producing maps from own data
HealthMapper[5] *(Free)*	Limited; mainly a data display tool but can show buffers as well as class data	Moderately easy, with good online documentation and tutorials provided	Limited; difficult to update content, but maps are easy to read and share	No plans to update in future
EpiMap[6] *(Free)*	Limited, but with some shapefile editing capabilities and some data analysis	Moderately easy, may be a good choice for EpiInfo users	Inputs shapefiles and Access databases, outputs to bitmap	Limited output choices; boundary files provided have not been updated recently

Notes: 1. Available from Environmental Systems Research Institute. 2. Also known as Quantum GIS. 3. Excel to Google Earth (E2G) Thematic Mapping Tool, v. 2.0, is available from MEASURE Evaluation. 4. Available from the United Nations Children's Fund. 5. Available from WHO, World Health Organization. 6. Available from the U.S. Centers for Disease Control and Prevention.

Appendix 2: GIS and Mapping Resources

The following GIS and mapping resources are available from MEASURE Evaluation at the links described:

MEASURE GIS Working Group holds meetings twice a year. Past agendas have included the exploration of emerging free and low cost mapping options, using GIS to improve data collection and strengthen data infrastructure, and issues of confidentiality when using spatial data.

http://www.cpc.unc.edu/measure/approaches/gis/wg

The **HIV Spatial Data Repository** provides HIV data for over 50 countries aided by the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) and others. Data come from MEASURE DHS (Demographic and Health Surveys) and from the U.S. Census Bureau.

http://www.hivspatialdata.net/

The E2G Thematic Mapping Tool is a Microsoft Excel macro comes with extensive documentation, including tutorial videos on YouTube. The tool can be downloaded and used with very little training. Also available is a Global Positioning System Tool Kit, which provides practical advice on how to use a GPS. Both tools are available here.

http://www.cpc.unc.edu/measure/tools/monitoring-evaluation-systems/geographic-information-systems/geographic-information-systems

Summary paper reviewing a CODIST workshop: "Enlisting national mapping agencies in the fight against HIV/AIDS: building partnerships with ministries of health and social services, and national AIDS commissions."

http://www.cpc.unc.edu/measure/publications/ws-10-16

Paper from *Health Services and Outcomes Research Methodology* about spatial analysis technique, "Using kernel density estimation to assess the availability of health care services in Nicaragua."

http://www.cpc.unc.edu/measure/publications/ja-07-76

Working paper about how to identify uniquely (geographically) a health facility using a GPS: "The signature domain and geographic coordinates: a standardized approach for uniquely identifying a health facility."

http://www.cpc.unc.edu/measure/publications/wp-07-91

Appendix 3: Mapping Tools

The following free mapping tools are available at the links provided:

MEASURE Evaluation E2G Tool: The E2G (Excel to Google Earth) mapping tool is available at: www.cpc.unc.edu/measure/e2g. This tool, which runs in Microsoft Excel as a macro, was developed for data mapping at a sub-national level in 40 different countries and then displaying the maps in Google Earth. There is also a tutorial that introduces a step-by-step use of the tool, and a series of shorter videos to help with frequently-asked questions.

ESRI ArcGIS Explorer Desktop: ESRI offers a free GIS reader to help with visualizing and exploring GIS information: www.esri.com/software/arcgis/explorer/index.html.

WHO's HealthMapper: The HealthMapper tool, a public health surveillance and mapping application developed by WHO, can be downloaded here: www.who.int/health_mapping/tools/healthmapper/en/index.html.

CDC's EpiMap: EpiMap is the mapping part of Epi Info, CDC's communicative disease analysis tool, which can be downloaded here: www.cdc.gov/EpiInfo/.

DevInfo: DevInfo is a database reader and administration tool which is distributed by the United Nations for use with their Millenium Development Goals. It has the ability to create tables, graphs, and maps and export them in common formats. The main data reader and also the data administrator tool can be downloaded here: www.devinfo.org.

DIVA-GIS: DIVA-GIS is an open-source, free GIS program whose input choices include .shx, .dbf, and .txt files, and output choices include .bmp and .tif images. Label and legend placement are somewhat limited, but the program allows color selection and classes data according to quartiles. It is available for download here: www.diva-gis.org.

Geocommons: Geocommons is a simple, step-by-step, Web-based tool for thematic mapping and spatial analysis. Thematic mapping includes the ability to create choropleth (colored, shaded) or proportional symbol maps. Analysis includes the ability to map the difference or correlation between two variables. Data can be imported from CSV files, shapefiles, KML, GeoRSS, and plain text. Maps and data can be viewed within Google Earth or exported as CSV files, shapefiles, or KML. The global base map layers include imagery, roads, and hybrid views of the two from such sources as Google, Microsoft, Yahoo, NASA, and OpenStreetMap: www.geocommons.com. (Note: Data and maps must be shared with the entire Geocommons community in order for them to be available to others, which would preclude the use of Geocommons for mapping private or confidential data.)

QGIS: QuantumGIS has had multiple updates in recent years and also provides a fair amount of documentation. It supports vector, raster and other spatially enabled tabular data formats. It also has a number of plug-in modules to further expand its use, mainly with raster data, and is customizable, with the ability for programmers to create their own modules for performing other tasks, using an extensible plug-in architecture. It is available for download here: www.qgis.org.

Appendix 4: GIS Data Portals

Administrative Unit Boundaries

United Nations Second Administrative Level Boundaries Project: Good metadata and thorough tracking of administrative unit boundary changes down to the second level (province and district) by year, many quite current. Does not have currently updated data for all countries, however. Available at: http://www.unsalb.org/.

Global Administrative Areas: Created for the BioGeomancer Project (biodiversity catalog at the global level), coordinated out of the University of California at Berkeley. Thorough listing of most countries in the world, with previews of admin boundaries for each country; however, documentation and metadata is not reliable. Available at: http://www.gadm.org/.

Subject-Specific Data and Maps

HIV Spatial Data Repository: Sponsored by PEPFAR, this includes shapefiles and linked tables for mapping of DHS data for over 40 countries, which can be viewed online and downloaded. Good metadata, fairly current data (1999 to 2007) on population, administrative boundaries, and HIV statistics. Census projections are available to the year 2010. Available at: http://www.hivspatialdata.net/.

Center for International Earth Science Information Network: Located at Columbia University, resources include Gridded Population of the World (2000 data as of 2011) and the Global Rural-Urban Mapping Project (2005 data). This includes searchable information from many countries in the world, focusing on land use, environmental indicators, and population. Available at: http://www.ciesin.columbia.edu/.

World Health Organization (United Nations) Communicable Disease Global Atlas: This can perform data queries and generate reports, charts, and maps. Interactive mapping section allows creation of maps of diseases, location of health facilities, schools, roads, and physical features. Available at: http://apps.who.int/globalatlas/.

Geonetwork: The following three organizations all use the same search engine and interface to access their GIS data: Food and Agriculture Organization (FAO), WHO, and World Food Programme (WFP). Not all files are available for download in shapefile format — some are available only via interactive map or Adobe Portable Document Format (PDF), but there is good metadata and good search functionality, and a wide variety of maps available. Available at: http://apps.who.int/geonetwork/, http://www.fao.org/geonetwork/, and http://vam.wfp.org/geonetwork/.

Earth Science Data Directory: From the National Aeronautics and Space Administration (NASA), many datasets are available on a global scale. This large portal is searchable by geographical area or subject. Many other sites link to this one, some through defined portals. Considered a master directory for climate change information, it is available at: http://gcmd.nasa.gov/.

Famine Early Warning System (FEWS): This is sponsored by the U.S. Geological Survey (USGS) and USAID. Africa Data Dissemination Service. Environmental monitoring program serving Africa, Southern and South Central Asia, Central America, and more. Searchable by region and data type, downloadable shapefiles, includes metadata. Available at: http://earlywarning.usgs.gov/.

International Center for Tropical Agriculture. Poverty mapping case studies with detailed shapefiles and good metadata for a limited number of Central American and African countries are provided. Available at: http://gisweb.ciat.cgiar.org/povertymapping/.

DevInfo: This is a database reader and administration tool distributed by the United Nations for use with their Millenium Development Goals. It has the ability to create tables, graphs, and maps. Sub-country level maps can be downloaded from this link after registering as a DevInfo user. Available at: http://www.devinfo.org/di_digital_map_library.html.

CDC: Free shapefiles are available for download to use with EpiInfo, a mapping tool also available from CDC. Available at: http://www.cdc.gov/epiinfo/maps.htm.

HealthMapper: Information about WHO's HealthMapper tool is provided, as well as sample data sets available for download. HealthMapper comes with a core geographic database that covers areas in Africa and SouthEast Asia. Available at: http://www.who.int/health_mapping/tools/healthmapper/en/.

ESRI: ESRI has an extensive array of shapefiles and imagery from around the world, available as free downloads to ArcGIS users. Available at: http://www.esri.com/products/#data_panel.

Gazetteer: Standard spellings of foreign geographic names are available at the National Geospatial Intelligence Agency. Can be searched graphically or by text based query. Available at: http://earth-info.nga.mil/gns/html/index.html.

Geographic Information Support Team: The data repository of the Geographic Information Support Team at USAID was created to provide GIS support for humanitarian relief and emergency response. Upon registering at this site, the user can search an extensive database of shapefiles with good supporting metadata from around the world. Available at: https://gist.itos.uga.edu/.

Appendix 5: Monitoring and Evaluation Tools

MEASURE Evaluation: MEASURE Evaluation monitoring and evaluation tools are available at www.cpc.unc.edu/measure/tools. Among them are the following:

- Finn T. A Guide for Monitoring and Evaluating Population-Health-Environment Programs. Chapel Hill, NC: MEASURE Evaluation; 2007. Available at:

  http://www.cpc.unc.edu/measure/publications/ms-07-25/at_download/document

- M&E Fundamentals is an online mini-course that covers the basics of program monitoring and evaluation in the context of population, health and nutrition programs. It also defines common terms and discusses why M&E is essential for program management. The course is available at:

  http://www.cpc.unc.edu/measure/training/online-courses/certificate-courses

- Geographic Approaches to Global Health is an online course on how to use spatial data to enhance the decision-making process for health program implementation in limited resource settings. This course is available at:

  http://www.cpc.unc.edu/measure/training/online-courses/certificate-courses

UNAIDS: Several M&E publications, including a series of monitoring and evaluation fundamentals guides, are available at:

www.unaids.org/en/dataanalysis/tools/monitoringandevaluationguidanceandtools/

USAID | DELIVER PROJECT: This project, funded by the U.S. Agency for International Development, has produced the following M&E tools:

- USAID | DELIVER PROJECT. Logistics Indicators Assessment Tool (LIAT) [Task Order 1]. Arlington, VA: USAID | DELIVER PROJECT; 2008. Available at:

  http://deliver.jsi.com/portal/page/portal/44F65C7D4C9B01A3E040007F01001808

- USAID | DELIVER PROJECT. Logistics System Assessment Tool (LSAT) [Task Order 1]. Arlington, Va.: USAID | DELIVER PROJECT; 2009. Available at:

  http://deliver.jsi.com/portal/page/portal/44F65C7D4D0A01A3E040007F01001808

AIDSinfo: The AIDSinfo Web site allows interactive analysis of United Nations Global Assembly data. This includes a clinical guidelines portal, a drug database, search capabilities for vaccines and clinical trials, and also a glossary and searchable list of topics about HIV/AIDS. Available at: www.aidsinfo.nih.gov/.

Indicator Registry: The Indicator Registry is a central repository of HIV information on indicators used to track the AIDS epidemic and the national, regional, and global response. Available at: www.indicatorregistry.org.

Global HIV M&E Information: This Web site, with links to numerous resources for monitoring, evaluation, surveillance, health information systems, systems management, and capacity building, is available at: www.GlobalHivMEinfo.org.

Appendix 6: GIS Training

Introductory Information

How GIS works: A good general description of the way GIS works, written and updated over a 10-year period by geography professors at University of Texas and University of Colorado is available at: http://www.colorado.edu/geography/gcraft/notes/datacon/datacon_f.html.

GIS video: This two-minute video provides a brief but comprehensive introduction to GIS from the Earth Science Research Institute, makers of the most comprehensive and popular GIS software, ArcGIS. Links to other resources are also provide, at: http://www.esri.com/what-is-gis/index.html.

Introduction to GIS: Offered by the Chief Directorate, Spatial Planning & Information, Department of Land Affairs, Eastern Cape, South Africa, this introduction to GIS provides a user-friendly site with 11 modules as well as a number of video tutorials. Examples use an open-source (free) GIS software package called Quantum GIS. Videos are also provided, as well as links to the software download site, at: http://linfiniti.com/dla/index.html.

MapAction Field Guid to Humanitarian Mapping: This is an excellent field guide to mapping using Google Earth and an open source GIS called MapWindow, as well as collecting data using a GPS. This comprehensive site contains a number of useful links, in addition to basic instructions, available at: http://www.mapaction.org/resources.html.

Software-Specific Tutorials

ESRI training and education (ArcGIS): As the industry leader in GIS software, ESRI has an extensive selection of online courses and tutorials, some of which are free (usually the first module of a course) while others require payment. These are available at: http://training.esri.com/gateway/index.cfm. Some recommended courses at this site include:

- Getting Started with GIS (9 hours) at: http://training.esri.com/acb2000/showdetl.cfm?DID=6&Product_ID=915
- Learning ArcGIS Desktop (24 hours) at: http://training.esri.com/acb2000/showdetl.cfm?DID=6&Product_ID=870
- Understanding Geographic Data (18 hours) at: http://training.esri.com/acb2000/showdetl.cfm?DID=6&Product_ID=702

MEASURE Evaluation E2G Tool tutorial: The E2G (Microsoft Excel to Google Earth) mapping tool is available free and can be downloaded here: www.cpc.unc.edu/measure/e2g. The site includes a video tutorial with a step-by-step introduction to using the tool, and a series of shorter videos to help with frequently asked questions.

WHO's HealthMapper: This is a public health surveillance and mapping application developed by WHO, available at: http://www.who.int/health_mapping/resources/technical_documents/en/index.html.

References

1. Gould P. *The Slow Plague: A Geography of the AIDS Pandemic.* Oxford, United Kingdom and Cambridge, USA: Blackwell Publishers, 1993.

2. Shannon GW. *The Geography of AIDS: Origins and Course of an Epidemic*. New York: Guilford Press, 1991.

3. Weir SS, Morroni C, Coetzee N, Spencer J, Boerma JT. A pilot study of a rapid assessment method to identify places for AIDS prevention in Cape Town, South Africa. *Sex Transm Infect.* 2002;78:106-113.

4. Carrel M, Escamilla V, Messina J, Giebultowicz S, Winston J, Yunus M, Streatfield PK, Emch M. Diarrheal disease risk in rural Bangladesh decreases as tubewell density increases: a zero-inflated and geographically weighted analysis. *Int J Health Geo.* 2011;6:10-41.

5. Brooker, S. Spatial epidemiology of human schistosomiasis in Africa: risk models, transmission dynamics and control. *Trans Royal Society Trop Med Hygiene.* 2007; 101 (1):1-8.

6. Noor AM, Zurovac D, Hay SI, Ochola SA, Snow RW. Defining equity in physical access to clinical services using geographical information systems as part of malaria planning and monitoring in Kenya. *Trop Med Intern Health*. 2003;8(10):917-926.

7. Keinschmidt I, Pettifor A, Morris N, MacPhail C, Rees H. Geographic distribution of human immunodeficiency virus in South Africa. *Am J Trop Med Hyg*. 2007;77(6):1163-1169.

8. Kalipeni E, Zulu L. Using GIS to model and forecast HIV/AIDS rates in Africa, 1986-2010. *Prof Geog*. 2008;60 (1):33-53.

9. Ajjampur SSR, Gladstone BP, Selvapandian D, Muliyil JP, Ward H, Kang G. Molecular and spatial epidemiology of cryptosporidiosis in children in a semiurban community in south India. *J Clin Microb*. 2007;45 (3):915-920.

10. Cromely E, McLafferty S. *GIS and Public Health*. New York: Guilford Press, 2002.

11. Friedman DJ, Hunger EL, Parrish II RG (eds). *Health Statistics*. Oxford, United Kingdom: Oxford University Press, 2005.

12. Noor AM, Alegana VA, Gething PW, Snow RW. A spatial national health facility database for public health sector planning in Kenya in 2008. *Intern J Health Geograph*. 2009;8:13.

13. Spencer J, Angeles G. Kernel density estimation as a technique for assessing availability of health services in Nicaragua. *Health Serv Outcomes Res Methodology*. 2007;7:145-157.

MEASURE Evaluation
Carolina Population Center
University of North Carolina at Chapel Hill
206 W. Franklin Street
Chapel Hill, NC 27516 USA
919.966.7482 / measure@unc.edu
http://www.cpc.unc.edu/measure