Government of the People’s Republic of Bangladesh

Ministry of Water Resources

Guidelines on
NWRD Spatial Data Quality

July 2015

Water Resources Planning Organization

Acronyms and Abbreviations

ASCII	American Standard Code for Information Interchange
BBS	Bangladesh Bureau of Statistics
BTM	Bangladesh Transverse Mercator
BWDB	Bangladesh Water Development Board
CAD	Computer Aided Design
DEM	Digital Elevation Model
DGPS	Differential Global Positioning System
DLRS	Directorate of Land Records and Surveys
DTM	Digital Terrain Model
ESRI	Environmental Systems Research Institute
GCP	Ground Control Points
GIS	Geographical Information System
GPS	Global Positioning System
GW	Ground Water
HDOP	Horizontal Dilution of Precision
ICA	International Cartographic Association
ISO	International Standards Organisation
LCC	Lambert Conformal Conic
NASA	National Aeronautics and Space Administration
NWRD	National Water Resources Database
PC	Personal Computer
PDOP	Position Dilution of Precision
PWD	Public Works Department
RMS	Root Mean Square
SOB	Survey of Bangladesh
UTM	Universal Transverse Mercator
VDOP	Vertical Dilution of Precision
WARPO	Water Resources Planning Organization
WMIP	Water Management Improvement Project

Table of Contents

Chapter 1
Introduction

1.1 Background

The National Water Resources Database (NWRD) aimed to maintain the sustainable integrated GIS based database that would consolidate all previously collected data on environment and water resources. It is to be mentioned that First Phase of the NWRD was started 1998 and substantial amount of data were collected and processed during this phase of the project. Further, second phase of NWRD was also started during 2003 and substantial amount of data layers were also updated and created. Different quality procedures were reviewed and exercised during the second phase of the NWRD. Several discussion meetings were also organized for developing the standard spatial data quality guidelines. Different professionals from GIS data holding agencies have also given their consensus to develop a spatial data quality standard for using it at national level. A draft spatial data quality guideline was also prepared during the second phase of the NWRD.

Further, ongoing phase of the NWRD (3rd Phase) has initiated under WMIP project of BWDB/WARPO, where large amount data to be updated and created during the different time span of the project. Moreover, NWRD would collect additional data in the future and upgrade the database in a consistent manner. In the course of time, new sets of data will also be produced based on planning needs. In this context, the issue of data quality guideline has been prepared considering the following aspects:

- For better national water resource planning purposes it is essential to maintain and assure the data quality according to application needs in different fields;
- Assessing quality of all collected and produced data;
- Providing quality information on data so that users could decide whether or not to use the data for specific applications.

1.2 Objectives

The purpose of this document is mainly to construct a guideline on spatial data quality to adapt in the development of NWRD. However, the specific objectives of this document are:

- To develop the spatial data quality standards;
- To formulating data quality management guideline for sustainable NWRD;
- To develop of standard format of data quality reporting.
- To develop product specification for producing new and update data layers of NWRD in near future and as well as other organizations

This data quality report with guideline for NWRD would assure the expected data quality for all data integrated into the NWRD;

1.3 Output

The outputs of this initiative are as below:

- Data quality report with guideline for NWRD
- Product specification for developing any new and update spatial data layers

Chapter 2

Quality Management - Standards

2.1 Introduction

Data quality is the degree of data excellence that satisfies the given objective. In other words, completeness of attributes in order to achieve the given task can be termed as data quality. Data produced by the private sector as well as by various mapping agencies are assessed against data quality standards in order to produce better results. Data created from different channels with different techniques can have discrepancies in terms of resolution, orientation and displacements. Data quality is a pillar in any GIS implementation and application as reliable data are indispensable for obtaining meaningful results.

The International Cartographic Association (ICA) formed a Commission on spatial data quality during 1991 and published a book on 'Elements of Spatial Data Quality". According to the ICA there are seven major elements of spatial data quality, which are (1) Lineage (2) Positional Accuracy (3) Attribute Accuracy (4) Completeness (5) Logical Consistency (6) Semantic Accuracy and (7) Temporal Information. Later, the International Standards Organisation (ISO) under the technical committee ISO/TC 211 (Geographic Information Geomatics) took over the responsibility and finalized the report in 2001.

As per the ISO guidelines, this chapter of the report will focus on the following aspects of data quality:

- Elements of spatial data quality;
- Implementation strategy for spatial data quality;
- Data quality evaluation procedure;
- Data quality evaluation method;
- Aggregation of quality evaluation result;
- Documentation of data quality.

2.2 Elements of spatial data quality

It is very essential to categorize spatial quality elements. According to the ISO, the major category of spatial data quality can be (i) data completeness, (ii) data precision, (iii) data accuracy and (iv) data consistency. A very brief description of these categories is given below:

- Data completeness: It is basically the measure of the totality of features. A data set with minimal amount of missing features can be termed as Complete Data.

- Data precision: Precision can be termed as the degree of details that are displayed on a uniform space.

- Data accuracy: This can be termed as discrepancy between the actual attributes value and coded attribute value.

- Data consistency: This can be termed as the absence of conflicts in a particular database.

The ISO describes spatial data quality elements using two major components (i) Data quality overview elements, (ii) Data quality elements.

Data quality overview elements

Data quality overview elements provide general non-quantitative information while data quality elements, together with data quality sub-elements and their description, report how well a dataset meets the criteria set forth in its product specification. The ISO has identified the following data quality elements and their sub elements:

- ◊ Purpose
- ◊ Usage
- ◊ Lineage
- ◊ User defined
- ◊ **Completeness**
    - Commission
    - Omission
- ◊ **Logical consistency**
    - Domain consistency
    - Format consistency
    - Topological consistency
- ◊ **Positional accuracy**
    - Absolute or external accuracy
    - Relative or internal accuracy
    - Pixel position accuracy
- ◊ **Temporal accuracy**
    - Accuracy
    - Temporal consistency
    - Temporal validity
- ◊ **Thematic accuracy**
    - Accuracy of the value given to a quantitative attribute
    - Classification correctness

A brief description of these data quality elements is given below and the details of the ISO data quality elements and sub-elements are attached in **Annex-A.**

2.2.1 Purpose

Purpose describes the summary of the intentions with which the data set was developed. Purpose also indicates the main objective of the production specific data set. The purpose of the data set can be explained both the data producer and user perspective and quality of data set are also dependent on the purposes of the data production.

2.2.2 Usages

The usages of the data set describe the purposes and up to what limit e.g. scale (e.g. national, regional, local and up to 1: x scale etc.). It also indicates the constraints of data usages to assure the protection of privacy or intellectual property, and any special restrictions or limitations on using the dataset. It also describes the geographic domain and extent of the dataset to be used in for any planning and analysis purposes.

2.2.3 Lineage

Lineage means history of a dataset throughout its life span. This starts from collection to final compilation. Lineage information could be seen from two points of view, namely, that of the data producer and that of the data user.

The data producer requires documentation of each stage of data production to ensure that the organization’s standard has been maintained. This information is needed to estimate the production accuracy and to identify the largest source of error in the production process.

The data user must be able to assess a dataset from the point of his particular application to decide whether it is “fit for use”. To enable the user to make this decision, the lineage information must be detailed enough and understandable to the user.

Lineage information could be structured in the following way:

Source

Information on source data may consist of information on (i) Origin: individuals, institutions, funding organizations, collection dates, geographic coverage etc.; (ii) Reference field: any theoretical field for which the data was referenced e.g.: PWD datum, Everest 1830 datum; (iii) Spatial data characteristics: scale of source data; (iv) Coordinate system: used to reference location e.g. UTM, BTM, LCC; (iv) Map projection: in the case of map source, map projection information should be included in the lineage information. This should also include parameters of the projection; and (iv) Correction and calibration: description of any process applied to calibrate or correct the dataset, e.g. atmospheric or radiometric correction, digitizing correction, etc.

Acquisition, compilation and derivation

Acquisition and collection: information of initial acquisition. The acquisition phase of a dataset is similar to the Level-0 category, specified by NASA and defined as “Reconstructed unprocessed instrument data at full resolution, time referenced and annotated, with ancillary information including radiometric and geometric calibrations.”

Compilation: The compilation stage of a dataset’s life is the building of scientifically useful database. This would include geo-referencing, applying reference field, and integrating the dataset into a larger database. This is equivalent to levels 1b and 2 according to the NASA category. Level-1b describes a dataset as: “Data which has been processed into sensor units e.g. radar backscatter value, brightness, temperature etc.; the Level-2 definition is “Derived environmental variables (e.g. ocean wave height, soil moisture, etc.) at the same resolution and location of Level 1 source data.

Derivation: Derivation means generating output as a result of interpolation, averaging or other techniques that change the format or structure of the original data. This process may or may not be reversible. This is a key event in the dataset life cycle, and must be well documented. Because derivation of data in one method for a particular application, it may or may not be suitable for another. For example, A DEM generated for national level application using the 500X500 grid size may not be applicable for a local level application. This phase could be defined as NASA level-3 and Level-4. Level-3 is the “data mapped on uniform space time grid scales, usually with some completeness and consistency properties applied (e.g. missing points interpolated, complete region mosaic from orbital swaths etc.).

Conversion of data

The primary conversion process ranges from primary to secondary format. The primary source is usually a graphical form such as a printed map, while the secondary format is the initial digital form. Other conversion involves conversion from a secondary to tertiary format, and possibly a digital format as well. The conversion of primary to secondary format is usually either a process of raster scanning or manual digitizing. Conversion affects the positional accuracy that is affected by a number of factors:

Equipment used: resolution, repeatability, and accuracy across the full area of the source material, threshold, and radiometric resolution; (i) Operator policy: expertise of operator, and pointing error; (ii) Digitization policy: pixel size (for raster scanning), sampling interval between vector line points, and re-sampling method for raster scanning; (iii) Source material: stability of material, scale, and map projection.

Transformation and analysis of data

The typical transformations applicable to a dataset are: (i) Co-ordinate transformation: conformal, (ii) affine, rotation, polynomial, map projection, etc.; (iii) Interpolation: deterministic, statistical; (iv) Cartographic displacement: averaging, generalization, exaggeration, reclassification, merging, etc.; (v) Various algorithms could exist for the same types of transformation.

2.2.4 Data Completeness

Completeness describes whether the entity objects within a dataset represent all entity instances of the abstract universe. *The degree of completeness describes to what extent the entity objects within a dataset represent all entity instances of the abstract univer*se.

Here, the abstract universe or the universe of discourse is the subset of reality which is digitally represented by the dataset. The entity instance names a real world phenomenon and the entity object is its digital representation. Completeness could be understood in two ways: *data completeness* and *model completeness*. When the question is data quality, the abstract universe is specified through data capturing rules and usually describes within the meta data. This is called data completeness, the measure of which is application independent. When the fitness of use needs to be assessed, the abstract universe as specified by the dataset would be compared to the abstract universe defined by the requirements of the application. Completeness issues of that comparison are called model completeness. Data completeness could be *entity object completeness, attribute completeness*, or *formal completeness*. *Entity object completeness* specifies whether or to what degree all entity instances, implicitly defined by the data description, are really present in the dataset. Consider a dataset that was specified to represent all water bodies of a given characteristic in Bangladesh. The entity object completeness specifies to what degree the abstract universe (that all water bodies meeting the requirements in the specification) is present in the data.

Attribute completeness expresses partial omission of information about an entity object, i.e. some attribute values are missing for some entity objects.

Formal completeness specifies whether or to what degree the formal structure of dataset is complete. This includes information on whether all mandatory meta-information is available, whether the format corresponds to the standard or data format used, respectively, or whether the data is syntactically correct.

Completeness is considered in two ways: commission and omission. Commission means excess data present in a dataset while omission means data absent from a dataset. Both express the measure of difference between the number of items in the universe of discourse and the number of items present in the dataset.

2.2.5 Consistency

Consistency is the formal definition of comprehensive rules and their consistent application to ensure high quality data in the formal data resource. Inconsistency occurs when two facts (data, information or rule) do not agree. The following example shows inconsistency that could happen in a set of geographic facts: (i) Fact one: Terrain height at the known point P is 1520m (using DTM), (ii) Fact two: Road R1 passes over a bridge at point P, (iii) Fact three: Road height at point P is 1480m (using a 3D digital map), (iv) Fact four: All bridges are over the terrain surface.

At least one of these facts is not precise and causes inconsistency. If fact number four is considered as a consistency constraint then inconsistency exists between the other facts according to the consistency constraint (fact 4). Consistency in a dataset is usually examined by consistency constraints, and refers to an absence of conflict in the dataset, not the conflict between the dataset and the real world. Absence of conflict between the dataset and the real world is referred by correctness. There are four types of consistency: *domain consistency, format consistency, and topological consistency, data derivation, and data retention*.

Domain Consistency

Data value consistency specifies the allowable values for each data characteristic, and each relation between data characteristic in the formal data resource. Domain consistency has two aspects: *data consistency value and data consistency rule.* A data consistency value is an actual data value or data code value that is allowed whereas the data consistency rule is a statement that defines the actual data values or data code values that are allowed. For example, all thanas should have the correct geocode specified in the BBS code list whereas elevation cannot be negative. Data value consistency is defined in the data subject like any other data, and it becomes part of the formal data resource.

Format Consistency

Format consistency expresses the degree to which data is stored in accordance with the structure of a dataset. It specifies the data consistency for data relation. The existence of one data subject sometimes requires the existence of another while the existence of some data subject prevents the existence of another.

The other part of format consistency is the referential consistency that ensures a parent data occurrence that exists for each subordinate data occurrence. A subordinate data occurrence cannot be added if there is no parent data occurrence, and a parent data occurrence cannot be deleted if subordinate data occurrence still exists.

Topological consistency

Topological consistency expresses the degree of adherence to the characteristics of geometry remaining invariant under continuous mapping transformations so that a dataset's topological incidence relations remain unchanged.

2.2.6 Positional accuracy

Positional accuracy is how well data stored in the formal data resource represents the real world which includes current positional accuracy and the adjustment in accuracy to meet business needs. It has several aspects such as: *precision, scale, resolution granularity, significant digits, reliability, source, method of collection, type of equipment used, method of analysis, method of calculation or estimation, adjustments made, persons or organization involved, and confidence.*

Precision indicates how precisely the measurement is done and how many significant digits are in the measurement. For example, a distance could be set to the nearest hundred feet, the nearest one foot or the nearest .001 foot. Scale is the ratio of real world distance to map distance, such as 1:20,000 or 1:1,000,000. Resolution is the degree of granularity of the data, indicating how small an object could be represented with current scale and precision. Spatial datasets have three types of positional accuracy: absolute or external accuracy, relative or internal accuracy, and pixel position accuracy.

- Absolute accuracy expresses the closeness of reported coordinate values to values accepted as or being true.

- Relative accuracy expresses the closeness of the positional relationships for features in a dataset to the relationships accepted as or being true.

- Pixel position accuracy specifies closeness of position of the center of a pixel relative to ground truth with respect to a specified reference frame. This is limited to raster data.

2.2.7 Temporal accuracy

Every phenomenon in this dynamic world changes in time. For this, any real world phenomenon must be described with the three criteria, "when", "where" and "what". These three concepts are essential to identify any spatial objects. Consideration of time as a component of geographical object increases the possibility of analyzing objects along a time line and keeping track of their history.

Events occur in space at a particular time. Time is dynamic, passing continuously with events happening along this passage. In space, one could easily point where a certain object is located or measure its location either relative to another object or absolutely with respect to a predefined coordinate system. This measurement could be repeated again and again. The positional errors obey a well-defined error propagation law and, therefore, could be easily trapped. In the temporal domain, if something happens at a particular time it is understood only by evidence. Moreover, there is no way of repeating it.

For a geographic phenomenon, three types of time are important to note: *Event time* when the change actually happens; *observation time* when the phenomenon is observed; and *transaction time* when the information is recorded in the database.

A major issue concerning data quality is whether the temporal information adequately describes the geographic phenomena. Because each feature should have some temporal data, decisions must be made over the type of time (event, observation, and transaction) that should be recorded. Some units of temporal measurement must be chosen (years, days, etc.) Decision must be made as to whether it is the absolute measure of time that is required or merely an ordinal indication of a sequence of events. Similarly, it must be decided whether the description should account for the fact that no event is instantaneous and that every event has duration; or if it is the starting or the ending time or both that are required. Some features have inherently different time inertia and thus require different time intervals of inspection or revalidation of a given state. For example, the elevation of a geodetic control station is almost time invariant while the water level at a gauge station could vary rapidly.

Granularity is another issue that should be considered to determine consistency in the temporal aspect. Granularity of the temporal axis lets the events be distributed along the axis. This is very important to determine when the event occurred in reality. The required precision depends upon the application that is also dictated by the frequency of occurrences.

In temporal relations, “being at the same time” means occurrences of two events at the same division of granularity. If the granularity for example is a week based, the two events could be called simultaneous only if they happen in the same week.

A consequence of granularity is “indeterminacy” which could be characterized as “do not know exactly when”. In order to handle temporal information with a certain indeterminacy, one has to define a threshold. Historical indeterminacy occurs in valid time. Transaction times are always determinate since the time during which a transaction takes place is always known.

The quality aspect of temporal accuracy deals with three issues: accuracy, consistency and validity.

- Accuracy is the correctness of the temporal references of an item (reporting of error in time measurement);
- Consistency is the correctness of ordered events or sequences if reported;
- Validity indicates whether the dataset is valid with respect to time.

2.2.8 Thematic Accuracy

Thematic accuracy describes the accuracy of attributes and correctness of the classifications of features and their relationships.

An attribute in a spatial database could be defined as a fact about some locations. The fact could be an outcome of historical consensus like the name of an area, or it could be the result of some measurements like temperature, or the result of interpretation by a trained observer. Attributes serve to distinguish one location or set of locations from another. Attributes of features are represented in tables where each row corresponds to one feature and each column represents one attribute. Attribute accuracy has the following aspects:

Spatial representation: Attributes represent facts about space that is represented by a point, line or area. Spatial resolution is implicit in the case of point attribute. Attribute of a point may represent the exact location of that point, like elevation or the aggregated value of the area around the point, e.g., population of a city. Sometimes attributes may represent the value for a set of locations in the form of average or sums over the set. For example, the population density of a country is the average of the population over the total area.

Scale of measurement: Attributes could be divided as being qualitative (e.g. names or classes) and quantitative (e.g. measurement of GW level). Measurement could be nominal (identification number, soil type), ordinal (suitability class 1, 2,), interval (temperature in Celsius), and ratio (Literacy rate).

Sources of uncertainty: Accuracy could be defined as the closeness of the results of observations, computations, or estimates to the true values or the values accepted as being true. Attributes are extracted by observations or measurements using some instruments. Uncertainties are involved in these observations or measurements due to the accuracy and precision of the instrument, and professional qualification of the observer.

Thematic accuracy deals with two aspects: accuracy and classification correctness:

- Accuracy specifies the correctness of a value given to a quantitative attribute.
- Classification correctness is the comparison of the classes assigned to features or their attributes to a universe of discourse (e.g. ground truth or reference datasets)

2.3 Implementation of spatial data quality

Implementation of data quality has two perspectives: the producer's perspective and the user's perspective. The two perspectives may or may not be matched. To implement a data quality framework, the producer needs to define a conceptual model of the data quality.

2.3.1 Data quality: producer's perspective

From the producer's point of view, data quality starts from defining a product specification. A dataset is defined as an identifiable collection of related data. These data represent real or hypothetical entities of the real-world which are characterized by having spatial, thematic and temporal aspects. The universe of discourse is described by a product specification against which the quality content of a dataset is tested.

The quality of a dataset depends on the intended use of the geographic dataset and its actual use. Since a dataset is not generally produced for a specific application but rather for a set of supposed applications, the quality of the dataset could only be assessed by knowing the data quality elements and the data quality overview elements. Data quality elements evaluate the differences between the dataset produced and the universe of discourse (that is the perfect dataset which corresponds to the product specification). Data quality overview elements provide general, non-quantitative information.

The data quality conceptual model provides an important framework for data producers and users as well as the means to specify how well the mapping used to create a dataset reflects its universe of discourse. Data producers could validate how well a dataset meets the criteria set forth in its product specification.

2.3.2 Data quality user's perspective

The data quality conceptual model provides data users with the means to assess a dataset derived from a universe of discourse identified as being coincident with requirements of a data user's application. Data users could assess quality to ascertain if a dataset could satisfy the requirements of an application. The data quality conceptual model recognizes that not all data quality elements nor all data quality sub-elements and their subsequent means of measurement and testing are applicable to a particular type of dataset. Additionally, some data quality sub-elements are applicable to and measured or tested for a dataset while others are applicable to and measured or tested for various reporting groups in a dataset.

For a data producer, a product specification describes the universe of discourse and contains the parameters for constructing a dataset. For a data user, user requirements describe a universe of discourse which may or may not match the dataset's universe of discourse. The true quality of a dataset is how well it represents a universe of discourse.

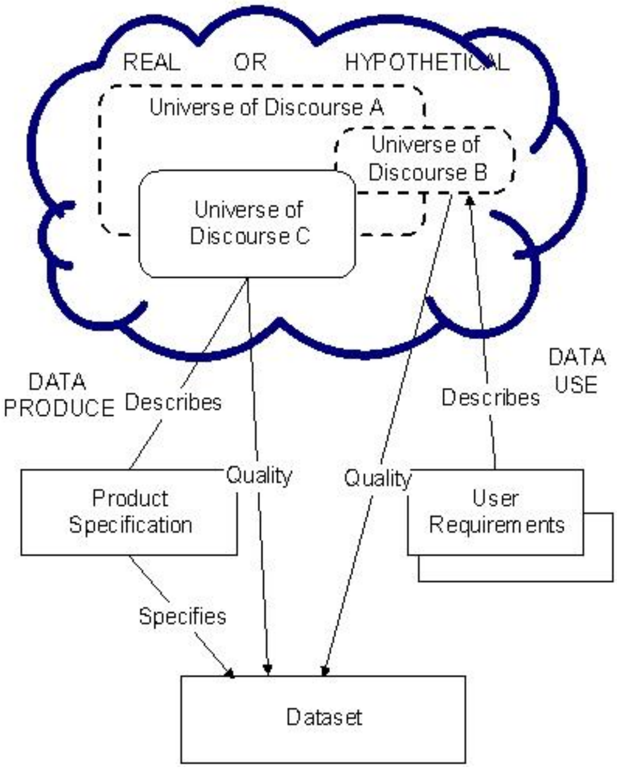


Figure 1.3.2: Schematic diagram of data quality conceptual model using producer and user perspectives

2.3.3 Assurance and assessment

Assurance is a proactive process of ensuring that the formal data resource contains appropriate data where assessment is a reactive process of determining the existing level of data quality. Assurance is the responsibility of data producers where the users need to assess data quality in the context of their application. To assure data quality, the producers implement a production strategy compliant to the product specification. A report is then prepared to describe the quality of the datasets.

Both the producers and users are required to follow an evaluation procedure to determine the quality status of specific datasets.

2.4 Quality evaluation procedure

The description of a dataset quality permits data producers to define how well their products meet their specification and to assist data users to assess how well a dataset meets their requirements. For this purpose, the quality of geographic data should be evaluated using consistent and standard methods. Levels of acceptable quality of a dataset are defined by the data producers' product specification and the data users' data quality requirements. The data producers are responsible for documenting the quality of a dataset and consequently, for selecting the specific evaluation procedure to determine data quality explicitly or implicitly given in the product specification. ISO TC 211 is preparing a standard for quality evaluation procedure to provide a standard for evaluating spatial data quality. Data quality is being assessed using different evaluation techniques by different users. Following are two major types of evaluation technique:

- The data producer performs the first level of assessment. This level of assessment is based on data quality check based on given data specifications.

- Second level of data quality assessment is performed at consumer side where feedback is taken from the consumer and processed. Then the data is analyzed / rectified because of processed feedback.

For NWRD data a user feedback form will be prepared to get feedback from the data users/consumers. Furthermore, an online feedback system will be developed under the third phase of the NWRD project.

2.4.1 Data producer and user perspectives

The ISO Standard for quality evaluation procedure recognizes that data producers and users may view data quality from different perspectives. Therefore, quality evaluation procedures for geographic data are expressed in two models, one for data producers and one data users. These two quality evaluation models are defined to specify quality evaluation methods and to report the results. The data producer model is for testing and reporting conformance to specification while the data user model is for evaluating how well the quality of the dataset meets user requirements. The data user model is intended for users who want to base a determination of fitness for use on an interpretation of quality information provided by data producers. However, when data users require more quality data information than is provided, they may follow the data producer model to obtain the additional information. In this case, data user requirements are treated as a product specification for purposes of using the data producer model.

2.4.2 Data producer quality evaluation model

The data producer quality evaluation model expresses a consistent and standard manner to determine and report the quality information of a dataset. The model consists of the following operations:

Select data quality element

Using the product specification, data producers shall identify the data quality element/s which need to be evaluated to determine conformance to the specification.

Select data quality sub-element: Using the product specification, data producers shall identify the data quality sub-element/s which need to be evaluated to test for conformance to the specification.

Define data quality scope: Data producers shall determine the scope of the quality evaluation on each sub-element that is necessary to test for conformance to the specification and intended product purpose. Each data quality sub-element may have a different scope or multiple scopes depending upon the product specification.

Select data quality measure: Using the product specification, data producers shall select data quality measures that would be calculated to test for conformance to the specification.

Choose data quality evaluation method

For each data quality measure, producers shall choose the quality evaluation method to be used.

Specify conformance quality level

If a conformance quality level is required, data producers, using the product specification, shall specify the conformance quality level for each data quality measure in order to establish conformance or nonconformance to the product specification. When the percentage or count of non-conforming items is specified as the conformance quality level, data producers shall specify a limit for determining conformance of each item.

Determine data quality values

Data producers shall compute data quality values by applying the chosen method to the related sub-element scope.

Assess conformance to product specification

If a conformance quality level is specified, data producers shall compare data quality values to the specified conformance quality levels for each selected data quality sub-element.

Report quality evaluation results

Data producers shall report the data quality results.

2.4.3 Data user quality evaluation model

The data user quality evaluation model expresses a consistent and standard manner by which to evaluate the quality information of a dataset against data user requirements. The model consists of the following operations:

Select data quality element

Using the user requirements, data users shall select the data quality element/s that need to be examined in order to determine how well the dataset meets user requirements.

Select data quality sub-element

Using the user requirements, data users shall select the data quality element or sub-element to be evaluated in the test on how well the dataset's quality meets the user requirements.

Define data quality scope

Using the user requirements, data users shall define the scope of the quality evaluation necessary to test how well the dataset's quality meets the user requirements.

Select data quality measure

Using the user requirements, data users shall choose the data quality measures to be evaluated.

Specify conformance quality level

Using the user requirements, data users shall specify the conformance quality level for the value reported in the dataset metadata or in the additional quality report, comprised of the selected data quality element, sub-element, data quality scope, and data quality measure.

Assess how well data quality meets data user requirements: Assessments based upon the user requirements shall be made for one or more sub-elements of the selected element. This is accomplished by comparing the data producer's reported data quality result value for the sub-elements to the data user's specified conformance quality level for the same sub-element. The data producer may not have reported a result value for the sub-element's data quality measure selected by the user, or may not have reported a result based on a measurement selected by the user or a method acceptable to the user. In either case, the user may choose to eliminate that element or sub-element from consideration or compute a quality result from the dataset itself.

2.5 Data quality evaluation method

There are two quality evaluation methods, direct and indirect. The direct method is accomplished by sampling or a full inspection of the data in the dataset. The indirect method is accomplished by evaluating data quality information from sources other than the data in the dataset. Choice of the quality evaluation method is determined by the data producer and user.

2.5.1 Direct evaluation methods

There are two types of direct evaluation methods, the full inspection evaluation method and the sampling evaluation method. The full inspection evaluation method involves the testing of 100 percent of the items in a population to determine quality result. The sampling evaluation method involves the testing of only a sample of all the items in a population to determine quality result. Inspection by sampling and full inspection evaluation methods may be accomplished by either automated or non-automated means. It is to be noted that some types of errors could be more easily detected automatically by a computer program than by non-automated means. Data quality elements and sub-elements, which are easily checked by automated means include, but are not limited to, logical consistency and format consistency. Some example could be as follows:

- Logical consistency: topological consistency, [example: polygon closure]
- Logical consistency: domain value consistency, [example: bounds violations, specified domain value violations]
- Completeness: omission, [example: comparison check of street names from another file]
- Completeness: commission [example: comparison check of street names from another file], and
- Temporal accuracy: temporal consistency.

Full inspection evaluation method

To apply the full inspection evaluation method, the population to be tested for quality shall be defined by the data quality scope. Every feature in the dataset that occurs within the data quality scope shall be tested according to the test defined by the data quality element, data quality sub-element, and data quality measure. It can be mentioned that this method of testing is more appropriate for small populations or for tests that could be accomplished by automated means. Because this is a test of 100 percent of the population within the scope, it gives a more reliable result since it lacks the ambiguity caused by sample variability in a result based on a sample.

Sampling evaluation method

Data producers shall devise sampling schemes and related values of the measure according to the conformance level being determined. In each instance, when utilizing a sample, the producer shall report the sampling evaluation method.

2.5.2 Indirect evaluation methods

Indirect evaluation methods are based on estimates of data quality measure values from sources other than the data items of the dataset. The variety of sources includes, but is not limited to, metadata, knowledge of the dataset's purpose, dataset lineage documentation, history of uses made of the dataset, and quality reports on the data used to produce the dataset. Knowledge of the production process and errors that may have been introduced or detected during production is useful.

EXAMPLE 1

Indirect evaluation methods are used when a digital terrain model file is obtained by means of stereo-correlation from images, and when experience tells the evaluator that the Root Mean Square Error is consequently estimated to be about 10 meters as long as data capture conditions were good.

EXAMPLE 2

Indirect evaluation methods are used when a dataset derived from digitizing a 1:25,000 map by conventional means is evaluated to conform to a user's requirements as background for an overview for town planning.

2.6 Aggregation of quality evaluation results

The quality of a dataset may be expressed by an aggregated quality result. This may require combining quality results from data quality evaluations based on differing data quality elements or sub-elements, each resulting in perhaps different meanings than others.

EXAMPLE

To determine the aggregate quality of a dataset containing road features, one may combine quality measures for the road centerline coordinate positional accuracy, the percentage of missing road features, the percentage of misnamed roads and the percentage of incorrectly spelled road names. Each data quality result has a different meaning.

The purpose of such an aggregation, even of dissimilar data quality results, is to provide a single measure of dataset quality. Although the aggregate methods described have been written in terms of aggregating quality evaluation results within a dataset, the same methods may be applied to provide an aggregate quality evaluation result for a dataset series. In the aggregation process, the significance of non-conformance of the data quality elements and sub-elements may be taken into account. A dataset may be deemed to be of an acceptable aggregate quality even though one or more data quality elements or sub-elements fail acceptance.

2.6.1 Interpreting an aggregate quality result

Because of the differences in meaning of the data quality results being aggregated, data producers and users should be certain that they do understand the meaning before they draw any conclusions about the quality of a dataset.

2.7 Documentation of data quality

According to ISO standards, it is mandatory for the data producer to report data quality results in the dataset's metadata. The data producer may create an additional data quality report containing more details than reported in the dataset's metadata. According to ISO standards, the data producer should cover full data quality information in both the metadata and the additional quality report. The report would have sufficient documentation and references on each quality evaluation procedure used, methods applied, values computed, and comparisons and interpretations made so that the evaluation might be repeated to arrive at the same results.

2.7.1 Maintaining up-to-date status of the quality report

Since datasets are continually created, updated, and merged, the quality or a component of the quality of a dataset might change. The quality information of a dataset could be affected:

- when any quantity of data is either deleted from, modified, or added to a dataset,
- when a dataset's product specifications are modified.

Modifications to a dataset might occur quite frequently, as many datasets are not static. If the reported quality of a dataset is likely to change with modifications, its quality should be reassessed and updated as required when the changes occur. Complete knowledge of all applicable data quality and overview elements, with the exception of the data quality overview element "usage", should be available when a dataset is created. Only the data producer's usage of a dataset could be initially reported. There is a reliance on data users to report unidentified uses of a dataset so that continual updates to this particular data quality overview element could be made to reflect the occurrence of unforeseen uses.

It is conceivable, however, that as a dataset is used its product specifications are updated for future modifications to better meet the actual need. As the product specifications change, the quality of the current dataset also changes. Quality information for a dataset should always reflect the current dataset given its current product specifications.

Chapter 3

Quality Management for NWRD Spatial Data

3.1 Introduction

In order to support the strategic planning capabilities of the Water Resources Planning Organisation (WARPO), the NWRD needs to ensure a consistent quality management system. The quality management for spatial data should be compliant with the currently developing international standard. The datasets in the NWRD are acquired either by own initiative or collected from secondary sources. From the data quality point of view, the NWRD is a producer as well as a user. The perspectives of both the producer and the user should be taken into consideration in the quality management of the NWRD. In this document the quality management of the NWRD is viewed from the producer's perspective. The NWRD should manage the quality of datasets collected from secondary sources as a representative of the primary data producer.

The quality management would include:

- Identification of sources of spatial data discrepancy
- Preparation of product specification
- Quality management of spatial data production
- Quality assessment and validation
- Quality control of data collected and incorporated from secondary sources
- Quality documentation
- Database management and data quality

3.2 Identification of Sources of Spatial Data Discrepancy

Sources of Spatial Data Discrepancy:

1. **Data Information Exchange:**

   Data information exchange is basically information about the data provided by the client to the organization. The degree of information provided by the client defines the accuracy and completeness of data.

2. **Type and Source:**

   Data type and source must be evaluated in order to get appropriate data values. There are many spatial data formats and each one of them has some beneficiary elements as well as some drawbacks. For example, in order to use CAD data on GIS platform, data must be evaluated and problems must be rectified otherwise resultant values will show high levels of discrepancies. Conventional data formats are quite specific to data storage technique and functional compatibilities. For example, topology cannot be created on shapefiles. It can be created only on the latest geospatial storage format i.e. geodatabase. Therefore, data type and source must be identified and evaluated before proceeding towards any analysis.

3. **Data Capture:**

There are many tools that incorporate manual skills to capture data using various software such as ArcGIS. These software allow the user to capture information from base data. During data capture, the user might misinterpret features from base data and captures the features with errors. For example: A user misinterprets two buildings as a single building and captures as a single feature. However in the real world, there are two features. So, the correct interpretation of features in base data must be performed. There are many tools that enable the user to find and fix those errors, but these tools are not used frequently due to lack of awareness. Data capture must be performed on a perfect scale where one must be able to view the features distinctly.

4. **Cartographic Effects:**

After capturing data, some cartographic effects like symbology, pattern, colors, orientation and size are assigned to the features. This is required for a better representation of reality. These effects must be assigned according to the domain of the features. For example, forestry domain specific cartographic elements must be used for forestry applications. Elements of any other domain used for a particular domain degrades the output of results.

5. **Data Transfer:**

Some discrepancies may occur while transferring data from one place to another, for example, while transferring data from a web source to a standalone, web disconnected machine. Sometimes, in order to make data more accurate than ever, the user tries to apply different advanced rectification techniques but it results in less accurate data changing into highly degraded data. There is no bad or good data. There are only data which are suitable for a specific purpose. So, data must be evaluated according to the domain for which it is supposed to be used.

6. **Metadata:**

Sometimes metadata is not updated according to original features. For example, a few features are edited on some software platform but the edited information is not updated such as name of the editor, reason for editing and some more relevant information. Therefore, metadata must be updated with original data.

3.3 Preparation of product specification

3.3.1 General principle

The first step of any project on spatial data acquisition is defining the product specification. The quality of the final product depends on intended use. However from the producer’s perspective, dataset is commonly produced for a set of applications and generally, applications for the production of any dataset in the NWRD are those related to WARPO’s planning needs. In addition to WARPO’s needs, similar applications in other planning organizations might also be considered.

The quality requirement specified in the specification of a product should be the minimum quality required for any of the target applications. Minimum quality needed for those applications does not vary too much. Target applications are those having similar characteristics regarding spatial, attribute and temporal aspects.

3.3.2 Guideline

The product specification should include the following items:

Item	Description
Product name	Name of the output dataset
Product date	Date of data production or modification or update
Product description	Purpose and use of the dataset
Spatial	
Scale	Which scale the application would work Local, sub-regional, regional, national, sub-continental, continental, global Example: coastal, national, regional etc.
Maximum output map scale	What would be the largest permissible map scale for any output from the dataset Example: 1: 250000
Extent	What is the largest area covered by any of the applications Example: South-west and south central part of Bangladesh
Resolution	Highest resolution needed by any of the target applications Horizontal and vertical Example: Mouza, 30 cm
Accuracy	What should be the minimum positional accuracy required at 95% confidence level Example: within 20 meter, within 0.3mm of true position at map scale
Temporal	
Extent	Cover total time period concerned for all target applications Example: Year 1960 to 2000
Resolution	Highest resolution needed for any of the target applications Example: weekly, monthly, yearly
Accuracy	Minimum temporal accuracy. Should not be less than the resolution Example: For a dataset with required resolution of 1 month should not contain data with more than 1 month of temporal error.
Feature	What features are to be included in the dataset. Feature specification should include feature name, definition, attributes
Attribute accuracy	Accuracy of attributes needs to be stated for 95% confidence level. Example: 95% of attribute value should match the source
Source	What are the data sources for the dataset
Quality elements	Data quality elements and sub-elements that need to be stated in the data quality report. These should be a subset of the international standard given in Annex-A

Item	Description
	Example: Elements: completeness, logical consistency, positional accuracy, thematic accuracy Sub-elements: Commission, omission, domain consistency, format consistency, topological consistency, absolute accuracy, classification accuracy
Evaluation procedure	What would be the evaluation procedure for each of the data quality elements and sub-elements For example: Positional accuracy would be assessed by direct methods comparing with DGPS measurement at common check points. The minimum number of check points should be 40, 10 at each quadrant of the dataset extent, the maximum distance between check points could be 5 km

3.4 Quality management for spatial data production

Production of any spatial dataset goes through different stages. Special measures need to be implemented at each stage to meet the product specification. The final product would have the accumulated error occurring at each stage of the dataset production.

Stages of spatial data production:

- Acquisition planning
- Acquisition of graphic data
- Geo-referencing
- Merging
- Aggregation and generalization
- Acquisition of attribute data

3.4.1 Acquisition planning

Planning is the most important part of any project, especially any data input project such as digitizing. Proper planning would save hours of editing. Planning includes the following tasks:

- Review product specification
- Assess available resources (data source, equipment and personnel)
- Decide on acquisition method for graphic data
- Quality monitoring

To comply with the product specification, the two tasks that have to be accomplished are (i) review of product specification and (ii) assessment resources. A brief description of these tasks is given below:

Review of product specification

Product specification must be reviewed thoroughly to have a proper understanding of the final output and quality requirement. The project member, coordinators and representative users should interact to clarify the exact data and quality needs. This would help the project members to get a feel of the task that they would perform.

Assessment of resources

A dataset cannot achieve better accuracy than its source. Moreover, the number of steps needed to compile the final product has a great impact on the accuracy of the dataset. Accuracy degrades every time a transformation takes place. For example, a topographic map with 10 meter contour intervals and a possible vertical error of 5 meters would not be used to delineate wetlands. A change of 5 meters in water level would result in major flooding. Often, spatial data from a digitizing project is merged with external data. The data could be a database such as Microsoft Access or some other set of spatial data. The first situation would require the existence of a common field to link the spatial information digitized to the database. The second would require the two sets of data to be of the same scale and in the same coordinate system with several matching points for geo-referencing.

The skill and precision with which the operator digitizes would be a factor in the accuracy of the final product. If possible, digitizing should be performed using the same GIS software that would be used later to display and analyze data. File formats created by digitizing with different GIS programs may not be compatible. If the GIS software does not have digitizing capabilities, digitizing should be performed with a compatible program or in the ASCII format, which is a generally readable format. Failure to digitize in a compatible format would make the work useless. At least one format would need to be translated to another, however, the translation process may introduce its own errors. The data sources need to be assessed in the context of accuracy requirement specified in the product specification to determine whether it is achievable or not. The following elements should be considered**:**

- Authenticity of data (map) source
- Scale of the map
- Geo-reference
- Physical condition of map (hard copy)
- Documentation of source data

Authenticity of data (map) source is an important factor for obtaining qualitative datasets. There may be many sources of database. However, not all sources are genuine in terms of quality, year of creation, data base preservation and acceptance of the source (organizations) by different levels of users. The source data organization either has the method to verify or check data quality control, or they use any standard procedure to maintain data quality.

Scale of Source Map is important in relation to data resolution, and consequently impact upon positional accuracy. It may be mentioned here that accuracy and precision are the functions of the scale at which a map (paper or digital) is created.

Quality of the Source Map includes age of the map, quality of the material used, distortion or fold in the map, and geo reference (latitude, longitude). In the case of point or elevation data, the map label should be clearly readable, and in the case of line or polygon, the category or classification of data should be clearly identifiable.

Paper maps should always be preserved so that they do not become crooked or distorted. Distorted or bad quality maps contain high RMS error and provide inaccurate data. In conclusion, if it seems that specified accuracy is unachievable with available resources, the following revisions could be adapted in the specification:

- estimate the maximum achievable quality with available source and resources, and compromise
- recommend special measures needed to comply with the specification and negotiate

Guideline for assessment of data sources

The following issues have to be considered in the assessment of data sources and resources:

Source description	**Describe the data sources in brief (title, organization, coverage)**
Features to be extracted	What features are planned to be extracted from this source according to product specification
Source map scale	Map scale of the source
Material used	Paper or Mylar (stability applicable for paper maps)
Distortion, folding	Existence of folding and distortion
Geo-referencing detail	projection, existence of graticules, etc.
Readability of features	Would it be possible to extract all target features from the sources
Readability of texts	Would it be possible to extract all text labels
Source positional accuracy	What positional accuracy is stated for the source? If it is not available, assume
Source data format	In the case of digital source, is the format compatible with the software to be used
Estimated achievable accuracy	What would be the final achievable accuracy within the constraints at 95% confidence level $E_{1.96} = \mathrm{SQRT}(E_s^2 + E_r^2 + E_d^2 + u^2) * 1.96$ *Where,* *$E_{1.96}$ = Total error at 95% confidence level* *E_s = Average error stated for the source* *E_r = mean Registration error* *E_d = Mean digitization error* *u = additional unexplained sources error* Example: Data source DLRS thana map E_s=20 meter

Source description	**Describe the data sources in brief (title, organization, coverage)**
	E_r=0.01 in X 63360 (map scale) =16 meter E_d=.1mm X 63360 =6.3 meter U = 5 meter (assume) $E_{1.96}$ = 52.7 meter
Recommendation if no match with the product specification	Check whether alternative feasible data sources are available or not. If not, then whether the achievable accuracy is workable. Is there any special measure for improvement

Choosing the acquisition method

Graphic data

The decision of data acquisition method primarily depends on available resources (data source, equipment and software). Accuracy requirement in the product specification also needs to be considered while choosing the methodology. Each methodology has its own advantages and disadvantages over the other.

The following are the probable sources of graphic GIS data:

Source	**Method**
Paper map	Direct digitization
	Scanning and automatic conversion
	Scanning and onscreen digitization
Aerial photo/Hardcopy satellite images	Scanning and onscreen digitization
Digital satellite images	Classification
	On screen digitization
GPS survey	Field survey

The number of processing steps required to reach the final output varies for different data sources. It should be mentioned that each of these steps would introduce some errors in the process, which would accumulate in the final error.

Guideline: The following guideline may be considered in choosing the right method based on the available sources. The matrix should be considered as a guideline. In some cases, a combination of methods should be applied. The GPS survey is also a growing method of data acquisition. The process of quality monitoring is shown in Table below:

Quality monitoring

		Available resource			Preferred method
		Scanner	Digitizer	Automatic data conversion software	
Available source	High quality single layer line work maps	**Y**	**Y**	**Y**	Scanning and automatic/Semi automatic conversion.
		Y	**Y**	**N**	Board digitisation, scanning and on screen digitisation
		Y	**N**		scanning and onscreen digitisation
		N	**Y**		Board digitisation
	High quality multi layer line maps with very low noise	**Y**	**Y**	**Y**	Scanning and semi automatic conversion
		Y	**Y**	**N**	scanning and onscreen digitisation,Board digitisation
		Y	**N**		scanning and onscreen digitisation,
		N	**Y**		Board digitisation
	Low quality line maps with lot of noises	**Y**	**Y**	**Y**	scanning and onscreen digitisation, tracing on mylar and board digitisation
		Y	**Y**	**N**	scanning and onscreen digitisation, tracing on mylar and board digitisation
		Y	**N**		scanning and onscreen digitisation
		N	**Y**		Tracing on mylar and board digitisation
	Low quality Isoline maps (contours)	**Y**	**Y**	**Y**	Scanning - image cleaning - semi automatic (interactive) digitisation. For portion with very bad quality, on screen digitisation needed
		Y	**Y**	**N**	Tracing on mylar and board digitisation (stream mode)
		Y	**N**		scanning and on screen digitisation
		N	**Y**		Tracing on mylar and board digitisation (stream mode)
	Aerial photo	**Y**	**Y**	**Y**	scanning and on screen digitisation
		Y	**Y**	**N**	scanning and on screen digitisation
		Y	**N**		scanning and on screen digitisation
		N	**Y**		Tracing of features and board digitisation
	Hardcopy Satelite images	**Y**	**Y**	**Y**	scanning and on screen digitisation
		Y	**Y**	**N**	scanning and on screen digitisation
		Y	**N**		scanning and on screen digitisation
		N	**Y**		Tracing of features and board digitisation
	Digital satellite image				On screen digitisation

Monitoring of data quality is essential to ensure expected quality, and to achieve this, the data quality supervisor needs to apply monitoring tools throughout the process. Forms should be supplied to digitizing operators to document the values of each parameter they might accept during digitization. The supervisor should check the accuracy by taking the check plot and matching it with the source. Software may be written to automate the checking of label errors, topological errors, etc.

The supervisor may keep the following questions in mind regarding monitoring:

Items of check plot to compare with the source data	Status
Does any of the digitizing parameters exceed the accepted parameter? What is the reason, and is it well documented? Example: Maximum acceptable RMS is 0.03. But it was not achievable for one of the sheets. It was accepted at 0.1	
All polygons were closed	
All features have labels	
Contours do not cross	
Contours has elevation labels	

3.5 Acquisition of graphic data

Digitizing is one of the key ways in which data could be acquired and stored in GIS. Apart from the manual technique of digitizing, geocoding could also be performed by scanning the source document.

3.5.1 Board digitization

Features have real world geographic coordinates associated with them as well as topological data that are fed as part of the digitizing process. There are two modes or methods by which features are captured – point mode and stream mode. Both have advantages and disadvantages and are better suited to capturing certain features than others. In a digitization job caution is needed in three phases: during pre-digitization, digitization and post digitization. It is to be mentioned that scanning with large size scanner are preferable than board digitization for efficient and quick data capturing. However, several important matter have to be followed before and during digitization tasks. These are described below:

Pre-digitizing tasks

a. Map preparation, including:

- Determining the number of map sheets needed to provide coverage for the region of interest.
- Working on one map sheet at a time while performing the following tasks:
  - ◊ smoothing out the map to remove folds or wrinkles;
  - ◊ placing a stable medium (e.g. Mylar) over the map (keeping it in place with paper clips around the edges);
- carefully tracing out on the Mylar all the significant features to be digitized with a sharpened pencil;
- marking out control points on the Mylar; and
- putting a starting tick mark on the redrafted trace for closed features like hill tops or lakes (this would also act as the finishing point during digitization).

b. Mounting the map on the digitizer, including:

- placing the map on the digitizing table and smoothing out any folds, bubbles or creases before taping it to the table (using tape that would neither tear the map nor leave a sticky residue on the table surface);
- always ensuring that the section of the map that would be digitized falls entirely within the active area; and
- aligning the Mylar sheet on which the redrafted features have been drawn on top of the map and taping the edges to the table.

c. Map registration

- This needs to be carried out for each new digitizing session, as well as each time a new section of the map is placed on the digitizer. This task would include:
    - ◊ entering at least four control points to register the map;
    - ◊ always using the same control points for each session of a particular map sheet;
    - ◊ clicking the lower and upper left corners of the region of interest to establish the size of the digitizing window; and
    - ◊ checking for scaling or inversion errors (in which case the map is either stretched or flipped) that may indicate an incorrect setup of the control points.

Digitizing tasks

During the actual digitizing process, the operator traces over the feature (e.g. river, soil polygon, wells) with a pointing device or cursor. The cursor has a number of buttons (4, 12, 16 or more) that could send a signal via the digitizing table to the computer to indicate that a point has been selected. The digitization can be carried out in two modes are:

Point Mode: This mode is useful for individual locations (e.g. elevation benchmarks) as well as for straight lines that only require a few points to be digitized.

Stream Mode: After pressing a button to begin the data collection, the digitizer continually collects points as the cursor is moved along a linear feature, until the operator presses another button to end the digitization.

Digitizing environment: Setting the tolerances (weed, node snap, edits) that give the standard values. The standard value of weed tolerance is the default of 1/1000th of the width or height of the current map extent. The standard value of the Node snap is the default of 1/1000th of the width or height of the edit coverage's BND, whichever is greater. The default edits the distance to 1/100th the width or height of the edit coverage's BND, whichever is greater. In the case of point data there is no need to set the weed, edit tolerances and node snap.

Digitization order and rules: Systematic digitization should be followed for the better capturing of data. Individual identification numbers should be introduced for the line or polygon during the digitization. In the case of dense features and large size maps, it would be better to divide the map into several parts and start digitizing it part by part. In that case, there would be less probability of missing any feature during the digitization.

Very long or complicated lines should be digitized in portions. Part of the lines should be traced with a subsequent halt in the data collection. Next, without moving the cursor, the start button should be pressed following which digitization of the subsequent portion of the lines should be recommenced. This procedure is to be repeated until the lines are complete. In order to be topologically correct while digitizing a network, lines should never cross each other. A node should represent the intersection of two or more lines.

The start and end nodes should coincide in the case of polygon features. If the start and end nodes do not coincide, the snapping or automatic closure should be used. Each area edge is to be digitized only once, even if it is a boundary between two polygons. The topology about the nature of the adjacent polygons could be added in a later editing exercise. While digitizing contour lines the following rules should be adhered to:

- determining in advance whether all of the contour intervals are to be digitized;
- in mountainous regions where the contours may be closely spaced, tracing out the selected intervals on the Mylar sheet with a sharpened pencil;
- starting at the innermost contour and working outwards;
- following the procedure for digitizing long or complicated lines;
- for enclosed contours, following the procedure for digitizing areas; and
- for lines that contain elevation labels, tracing through the middle of the label as digitization is carried out. Labels could be added later.

3.5.2 Scanning and on screen digitization

Scanning is increasingly becoming an effective means of automating spatial data. Traditionally, all spatial data was digitized manually by using a digitizing tablet or in some cases, keyed in by hand. However, in recent years, developments in hardware and software have made map scanning a viable tool for data automation. There are several different techniques that utilize map scanning as a means of recording spatial data. One method is to scan a map, and then use it as a background image for on-screen digitizing. Another method is to use sophisticated software to convert a scanned raster image into useable GIS data. Information regarding the processes and considerations involved with scanning:

1. Two decisions that must proceed a map scan:
   - # The image type needed as the result (i.e. grayscale or color).
   - # The desired output resolution measured in dots per inch (dpi).
2. The implications of dpi selection, as it relates to the output file size, and the desired spatial resolution requirements of the project:
   - # In any scanning software, selection of an appropriate dpi for the scan is in essence the determining factor of how many dots per inch the scanner would record.
   - # The more dots per inch, the more bits (binary digits) needed, and the larger the resulting image file.
   - # Generally, the limitation would be the possible scanning resolution of the scanner itself.
   - Sometimes the limit could be dependent on hardware issues and image file sizes.
   - Color scans (false or true) require a greater number of bits and, therefore, are greater in file size; the greater the dpi selection, the greater the spatial resolution and file size of the image.

3. Image enhancement:

Contrast and brightness may be adjusted to preserve the highest degree of precision

Gamma correction:

▪ The scanner analyzes a histogram based on the map's distribution along the gray or color scale.

▪ Points are strategically placed along the histogram to isolate data types.

▪ The scanner may then select the best areas to brighten or darken

Resolution considerations:

▪ Resolution is the density of the raster image measuring the number of pixels within a given distance.

▪ Since inches are a common unit of measurement, dots per inch (DPI) are commonly used.

▪ Since a scan line is one pixel in length, feed scanners are most efficient in handling large map surfaces.

▪ The user must balance a reasonable resolution to match expected resolution and storage capacities.

◊ A 100% (2X) increase in resolution yields a map image four times greater in physical storage size.

◊ It is recommended to use the following rule of thumb

- 200 dpi -- to view text and line art
- 300 dpi to view most photographs
- 400 dpi to view high quality orhophotos
- With the above considerations, a 600 dpi scanner should more than suit most needs.

4. Scale and Accuracy:

It is essential to remember that a scanned map could be no more accurate than the original map.

It is also essential to keep in mind that paper stretches as well as shrinks with changing humidity. A properly controlled environment might not be as important as previously thought. The best rule of thumb is to scan the maps in roughly the same temperature and humidity that they are stored and used in.

5. Map preparation:

Most people involved with scanning maps are interested in creating GIS data in geo-referenced coordinates. Control points should be transferred onto the map if they are not available on the printed media. It is easier to do this before the scan than afterwards.

Even with high resolution scanning, some features may not "show up" after scanning due to complex symbols, shading, or patterns. The source map may be enhanced by drafting in difficult-to-see areas and lines.

▪ The following rules may be followed:

◊ Never scan a map without control points

◊ Never scan a dirty map

◊ Never use petroleum based permanent markers. (These may destroy your scanner!)

◊ Never draft with graphite

◊ Avoid glossy finishes.

3.5.3 Digitizing from digital satellite images

Satellite images are one of the major sources of natural resources and environmental information. Information is sometimes automatically generated from satellite images by classification. Sometimes, features cannot be very well classified automatically. Visual interpretation and on screen digitization may be the best way of extracting some of the features. A lot of information in the NWRD (rivers, waterbody, settlement, bankline, etc.) is extracted by onscreen digitization from satellite images. Higher resolution image should yield more accurate and detailed information.

Information regarding the processes and considerations involve:

1. Accuracy of the collected data cannot be better than the accuracy of the image. So the geo-referencing should take into consideration that:

    # Images may be warped unevenly across the image extent.

    # Ground Control Points (GCP) must be established on the image to establish a link to real-world positions.

    # GCP information may be based on coverage or on map values.

2. Image correction:

    # Radiometric correction

    ◊ De-striping

    ◊ Speckle removal from RADAR images

3. Image enhancement:

Images could be enhanced for better interpretation using

Spectral enhancement

Spatial enhancement

Edge enhancement

Histogram stressing

Band composition

Multi sensor and multi resolution image fusion

3.5.4 Data collection by GPS

A Global Positioning System (GPS) is a set of hardware and software designed to determine accurate locations on the earth using signals received from selected satellites. Location data and associated attribute data could be transferred to mapping and Geographical Information Systems (GIS). GPS would collect individual points, lines and areas in any combination necessary for a mapping or GIS project. More importantly, with GPS complex data dictionaries could be created to accurately and efficiently collect attribute data. This makes GPS a very effective tool for simultaneously collecting spatial and attribute data for use with GIS. GPS is also an effective tool for collecting control points for use in registering base maps when known points are not available. The following issues need to be considered while collecting data using GPS:

- Equipment
- Accuracy
- Systematic Data collection

Equipment used

The systems hardware contains three parts; *antenna*, *receiver*, and *data-logger* -- sometimes called *data-collector*. The data-logger is a hand-held computer that contains software to coordinate signal collection and storage, file manipulation, and file transfer to and from a Personal Computer (PC). Some systems might combine some of these elements into a single piece of hardware. The systems software, which normally resides in a PC, has four primary functions: preplanning, post-processing correction of the raw satellite data, display/editing of the data, and converting/exporting of the data. A brief description of the different GPS components which should be taken care of for data quality aspects is given below:

GPS receivers: GPS receivers span a wide range of price and capability. The basic split of receivers is based on the number of satellites the receiver could track at one time. Each tracked satellite requires a channel. A minimum of four satellites need to be visible for position calculation. It is recommended that the receiver could track six or more satellites.

GPS Data-loggers: The major differences between data-loggers are the amount of memory and the type of keypad. GPS files are surprisingly small, but at least 1 Mb of memory is needed in the data-logger to do a full day of collecting. Some data-loggers do not have a full alphanumeric keypad. Because attribute data is key to GIS, the type of keypad is a consideration.

GPS Software: Many of the less expensive GPS units, especially those meant just for navigation, do not come with software capable of exporting files or using a data dictionary to define attributes. Without the ability to create data dictionaries and export files, these units are not functional for GIS.

Accuracy consideration

Two levels of accuracy should be maintained in using GPS. One is "Mapping Accuracy" which usually implies "corrected" accuracy within a few feet or less. This is usually sufficient for most GIS work. The other one is "Surveying Accuracy" which is usually advertised as having "sub-centimeter" accuracy. These units are significantly more expensive. The following issues have an impact on GPS accuracy:

Satellite Geometry: Satellite geometry affects the accuracy of reported locations. An unfavorable orientation could throw off accuracy even when satellite distance and location are excellent. Errors introduced by satellite orientation are referred to as Position Dilution of Precision, or PDOP. PDOP is composed of two pieces: HDOP or Horizontal Dilution of Precision and VDOP or Vertical Dilution of Precision. One may be small and the other large. The final effect of HDOP and VDOP depends on whether good horizontal accuracy, vertical accuracy, or both are needed**.** The recommended maximum PDOP is 6

Horizontal Error: HDOP is minimized when the position being located is in the middle of the satellites being tracked. With three satellites, the best case is each satellite equidistant from the position and at 120 degrees from each other. Furthermore, the satellites should be as low on the horizon as possible so that the elevation of the satellites above the position is zero. With four satellites, the best case is in addition to the above three, a fourth directly overhead. The worst HDOP occurs when all the visible satellites are on one side of the position and high above the horizon.

Vertical Error: VDOP is minimized when one satellite is above the position and another is below the position, which is impossible. The position could never be "between" the satellites, so elevation is more difficult to establish than position on the plane. The satellites would never be in a best orientation, just better or worse.

However, VDOP is not a severe problem since elevation is not usually as important as position on the plane. The accuracy is more than sufficient for most purposes, and if adequate time is given for the position to be fixed, it is possible to obtain the desired accuracy in elevation. Most receivers allow a maximum of HDOP and VDOP to be selected. A rule of thumb is that vertical values are about one-third as accurate as horizontal values.

Differential Correction: Differential correction is the cure for most GPS errors except for local errors such as multi-pathing. This correction should be applied to all reported locations when accuracy within 5 meter or less is required.

Datum Plane: The height above a surface must be measured relative to that surface. For example, most elevations are reported as being some number of feet or meters above the mean sea level. Unfortunately, the mean sea level is an average that changes locally because of gravity, spinning force, sun-moon-planet orientation, and a dozen other things. Most GPS receivers allow specifying the ellipsoid or Geoid to calculate elevation. Which one is chosen depends on what would be done with the information. If relatively correct elevations are required that would be consistent over an area, the ellipsoid must be used. GPS information should match the GIS datum, or at least a transformation to the needed datum must be available.

Coordinate Systems: The horizontal location, or the X,Y location on the earth's surface changes with the way the earth's surface is represented. The common coordinate systems for GPS work are Latitude and Longitude, Universal Transverse Mercator, and the State Plane Coordinate System. Most GPS software convert GPS data into the appropriate coordinate system.

Systematic Data collection

In a GPS data collection project, systematic data collection that involves preplanning would reduce a lot of errors and confusion in the collected data. Preplanning involves determining what data to collect, creating a data dictionary to organize that data, and then determining the best time to collect the data.

What to Collect: Two major types of data could be collected by a GPS. Feature Data and Attribute Data. Feature Data is location data-points, lines and areas; and Attribute Data is what is in those features.

1. Feature Data Types:

    Points: The location of one spot, a tube well, GCP point etc. A single collected point would have the most inaccuracy, but the accuracy could be improved by having the GPS make many individual readings and by having it average this data.

    Lines: Lines are a series of connected points used for features such as streams, roads, boundary lines, etc. Since many points are collected at a set interval by the GPS to form the line, the accuracy is better than that with individual points.

    Area: Areas are closed lines, and used for measuring property parcels, vegetation areas, etc. Most GPS software would automatically close an area so that there is no need to return to the starting point.

2. Attribute Data:

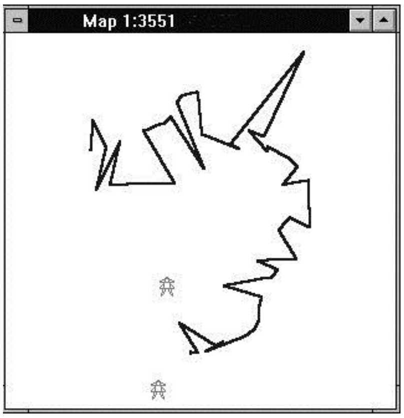
    Attributes to be collected in the fields should be well coded and prepared in the data dictionary. These would be determined in the planning process, while specific values would be entered in the field.

Data Dictionary: data dictionary is precisely that, a list of the types of data (both features and attributes) to be collected and the format in which they would be collected. This is determined before the collection begins and is usually created in the planning software and downloaded to the data logger. Properly designing a data dictionary is one of the most important steps in using GPS for GIS data collection.

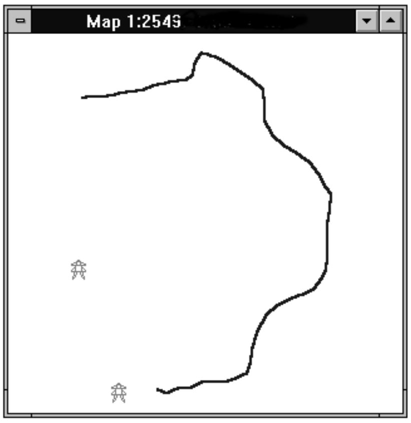
When to collect data--Determining Satellite Availability and PDOP: This step involves using the GPS software's preplanning module, which would bring up charts showing satellite availability, PDOP levels, and other information for a selected location and time. This is necessary to ensure that enough satellites are available in a suitable orientation to collect the data.

Collecting Data: While collecting data in the field, the following things should be checked: logging interval, PDOP, elevation mask, time and the coordinate system used for display.

Differential Correction: Because of the inaccuracies described above and the intentional fuzziness introduced by the US government, the data collected is probably only accurate to within 50 ft. To achieve better accuracy the file collected needs to be compared to the data collected at a known point, called a Base Station. See examples below of uncorrected (on the left) and corrected data (on the right).



Before differential correction



After differential correction

The Omnistar satellite differential services with the Trimble ProXRS receiver are of high precision and accurate for DGPS. The stated average horizontal accuracy of the system is 1 meter.

Exporting Data

Most GPS software programs allow the export of corrected data in several formats and coordinate systems. It is necessary to ensure that the GPS could export the data in a compatible format. There are many variables to choose to match output requirements. The keys to exporting are to choose the appropriate format and coordinate system to match the GIS or CAD program.

3.5.5 Data editing tasks

The following tasks are to be performed after a digitization work has been done:

- ◊ Error identification and elimination check the digitized map for the following errors
- ◊ Sliver polygons.
- ◊ Line segment errors or dangling chains, where lines in a network do not meet at the appropriate nodes. This is corrected by snapping.
- ◊ Polygonal errors which may include duplicate line segments, missing line segments, mislabeled line segments, unlabeled polygons and conflicting names.
- ◊ Strange polygons which do not appear on the original map, but which have occurred due to the operator's negligence.
- ◊ Check for positional accuracy by comparing the digitized map against an independent map of higher accuracy.

- ▪ Editing
    - ◊ review the digitized maps, checking for errors;
    - ◊ remove duplicate lines where they occur;
    - ◊ snap nodes for polygons that should be closed; and
    - ◊ snap lines to the relevant nodes where under- or overshooting occurs.
- ▪ Labeling of features

    Each digitized point, line segment or area edge must have a label or identifier associated with it. These feature identifiers represent a category of a feature (e.g. "1" for highways, "4" for dirt roads) and could either be added using an EDIT or LABEL menu in the post-digitizing phase, or during digitizing.

3.5.6 Geo-referencing

The first step of geo-referencing is to identify the projection of the source and decision of the projection to be used in the product dataset. The information would be used to geo-reference data into a real-world coordinate system. Registering all layers to a common coordinate system ensures data integrity during spatial joins and analyses.

Datums

A datum defines the reference spheroid (or ellipsoid) used to describe a portion of the earth. In Bangladesh, the Everest spheroid has historically been used in all projections. However, there are various Everest datums and there is a great deal of confusion about which to use. To date it has not been possible to completely clarify the confusion, but an attempt is made here in this regard.

The most commonly use spheroid and the one given in all manuals and textbooks for Bangladesh is the Everest (1830). The British Ministry of Defense, Directorate of Military Survey Plans lists five different Everest spheroids. However, in its table of grids they only refer to the reference spheroid as E", or Everest, not specifying anyone of the five. Most people have assumed that the Indian IIB grid (used on most of the Bangladesh SOB maps) datum is Everest (1830). Some reliable sources have indicated that the actual datum is Everest (Malaya RKT). The Everest (Malaya RKT) is about 5 m longer on both semi-major and semi-minor axis than the Everest (1830).

In addition there is the confusion about the Indian yard. The Indian yard is defined as 0.914395233 meters, slightly shorter than the international yard (0.9144 exactly). This small change makes a difference when considering the radius of the earth, which is over six million meters.

An Everest spheroid which takes some of these discrepancies into consideration is now being used by Finn map, F AP24 and SOB (but not on all projects, apparently). F AP 24 has called this the Modified Everest Modified. For simplicity this report will refer to it as Everest (Bangladesh).

The difference in semi-major axis lengths between the various Everest spheroids is less the 30 meters. Although there is no linear conversion some tests have indicated that the offset in coordinates is about half the difference of major-axis length in one direction and negligible in the other direction. (example: the difference in co-ordinates between using a spheroid with an axis length of 6,377,307 and one with a length of 6,377,297 is about 4.5 m in one direction and 0.15 m in the other.)

The WGS84 datum has become important in recent years because it is the basic datum for the GP WGS84 provides a best-fit spheroid for the entire earth. It is important to remember that the geographic co-ordinates (latitude and longitude) of any given point are different for each datum. The parameters for the different datums used in Bangladesh are as follow:

Everest 1830

Semi-major axi:	6,377,276.345 m
Semi-minor axi:	6,356,075.413 m
I/f (inverse flattening):	300.801700000
e2 (eccentricity squared):	0.006637847
Source:	Most textbooks and manuals

Everest (Bangladesh, or Modified Everest Modified)

Semi-major axi:	6,377,298.524 m
Semi-minor axi:	6,356,097.518 m
I/f (inverse flattening):	300.8017
e2 (eccentricity squared):	0.006637847
Source:	SOB, F AP24, Finn map

Everest (Malaya RKT, or 1948, or Modified)

Semi-major axi:	6,377,304.063 m
Semi-minor axi:	6,356,103.039 m

I/f (inverse flattening):	300.8017
e2 (eccentricity squared):	0.00663784663
Source:	British Military survey (may be datum for Indian IIB)

WGS84

Semi-major axi:	6,378,137.000 m
Semi-minor axi:	6,356,752.314 m
I/f (inverse flattening):	298.257223563
e2 (eccentricity squared):	0.006694380
Source:	Most textbooks and manuals

Practically the Everest (1830) is still the most commonly used datum in Bangladesh. However, a number of organizations which have recently investigated the datum and projection problem in detail are using the Everest (Bangladesh). The Everest (Malaya RKT) is not used today, but it appears to be the datum used in preparation of the Indian Zone IIB and IIIB maps.

Datum Shifts

Most GPS data is output in the WGS 84 datum, and most of the final output in Bangladesh is in one of the Everest datums. A number of programs require datum shift constants to make the conversion from WGS 84 to the datum being used. However, here too there is some confusion as there are a number of different constants being recommended. The table below gives four sets of constants that are recommended for Bangladesh.

Source	Everest	ΔX(m)	ΔY(m)	ΔZ(m)
Magellan User Guide (GPS units)	1830	289	734	257
JICA Study	1830	283.729	735.942	261.143
FAP 24	Bangladesh	288	735	255
Molondensky Constants (IDRISI & Geo. Calc.)	1830	282	726	254

Projection

Transforming geographic co-ordinates (latitude and longitude) to plane co-ordinates is known as projection. There are numerous map projections being used in Bangladesh. The following is a list of known map projections their parameters, and their known users, or uses.

Name:	**TM (Transverse Mercator)**
User/ers:	SPOT XS Satellite images, FAP6
Spheroid:	Everest (probably 1830)
Central Meridian:	90 ° E of Greenwich
Latitude of Origin:	0° (the equator)
False Northing:	0 meters
False Easting:	500,000 meters
Scale Factor:	0.9998

Name:	**UTM (Universal Transverse Mercator)**
User/ers:	GPS
Spheroid:	depends on set-up; may be WGS84
Central Meridian:	87d E (for zone 45, between 84E-90E)

	93d E (for zone 46 between 90E-96E)
Latitude of Origin:	0° (the equator)
False Northing:	0 meters
False Easting:	500,000 meters
Scale Factor:	0.9996

Name:	**BUTM (Bangladesh Universal Transverse Mercator)**
User/ers:	Finnmap, JICA, FAP3.1
Spheroid:	Everest (Bangladesh or 1830)
Central Meridian:	90°E of Greenwich
Latitude of Origin:	0° (the equator)
False Northing:	0.0 meters
False Easting:	500,000 meters
Scale Factor:	0.9996
Notes:	Regular UTM splits Bangladesh in two at 90° E; a modified system is used to cover all of the country - BUTM.

Name:	**BTM (Bangladesh Transverse Mercator)**
User/ers:	EGIS, FAP24, SWMC, FAP3.1
Spheroid:	Everest (1830) [FAP24 uses Everest (Bangladesh)
Central Meridian:	90^0 E of Greenwich
Latitude of Origin:	0^0 (the equator)
False Northing:	-2000,000 meters
False Easting:	500,000 meters
Scale Factor:	0.9996
Notes:	Same as BUTM except False northing is -2,000,000.

Name:	**LCC1 (Lambert Conformal Conic - One Standard Parallel)**
User/ers:	Panchromatic Sat-photos SSC, LGED maps
Spheroid:	Everest
Central Meridian:	90^0 E of Greenwich
Latitude of Origin:	26^0 (Standard Parallel)
False Northing:	914,395 meters
False Easting:	2,743,186
Scale Factor:	0.9988
Notes:	This is the same as the Indian Zone IIB Grid. According to Northwest Hydraulic Consultants & Finn map this is a LCC of the one standard parallel type.

Name:	**LCC2 (Lambert Conformal Conic - Two standard Parallels)**
User/ers:	aviation charts
Spheroid:	(Everest 1830)
Standard Parallels:	17^0 20’ N 22^0 40’ N
Convergence Factor:	0.34215
Notes:	Information on aviation charts from Northwe t Hydraulic Consultants

Name:	**India Zone IIB Grid (is a LCC projection)**
Uses/ers:	SOB Maps
Spheroid:	Everest 1830 or Malaya RKT
Len. semi-maj axis:	6974310.6 yards
Sq. of eccentricity:	0.00663784663
Central Meridian:	90^0 E of Greenwich
Latitude of Origin:	26^0 N (Standard Parallel)
False Northing:	1,000,000 Indian yards
False Easting:	3,000,000 Indian yards
Scale Factor:	0.998786408
Notes:	This is the same as LCC1 (with a single standard parallel) Information provided by British Directorate of Military Survey Plans (as recorded in Northwest Hydraulic Consultants documents). It is not clear which Everest spheroid is used many assume it is the 1830 but some think it is Malaya RKT. The Indian yard is equal to 0.91439523 meters. The southern tip of Bangladesh (Cox's Bazar south) maybe in the IIIB Grid (according or aviation charts but not SOB maps).

Name:	India Zone IIIB Grid (is a LCC projection)
Uses/ers:	SOB maps
Spheroid:	Everest 1830 or Malaya RKT
Len. semi-maj axis:	6974310.6 yards
Sq. of eccentricity:	0.00663784663
Central Meridian:	100^0 E of Greenwich
Latitude of Origin:	19^0 N (Standard Parallel)
False Northing:	1,000,000 Indian yards
False Easting:	3,000,000 Indian yards
Scale Factor:	0.998786408
Notes:	This grid may only apply to the southern tip of Bang1adesh (Cox's Bazar south), according or aviation charts, but not SOB maps. This is also a Lambert Conformal Conic with one standard parallel Information provided by British Directorate of Military survey Plans (as recorded in Northwest Hydraulic Consultants documents) It is not clear which Everest spheroid is used many assume it is the 1830, but some think it is Malaya RKT. The Indian yard is equal to 0.91439523 meters.

Digitizer units vs. real world coordinates

Coverage may be digitized in digitizer units or real-world units. The digitizer is based on a type of rectangular coordinate system with its origin in the lower-left corner. On the digitizer surface, moving one inch up or down covers the same distance as moving one inch left or right. Anywhere on the table’s surface, an inch is an inch. Whether the unit of measure is inches or centimeters, when associated with digitizing it is called a digitizer unit.

Location and distance are the key to mapping geographic features. The real world has a curved surface and is often measured in feet or meters. Both feet and meters are standard units of measure, but they do not have an obvious global origin. Therefore, a system of spherical coordinates (referred to Measurements on the Globe) locates specific locations on the surface of the globe. The units of this reference system are degrees of latitude and longitude. However, the distance represented by a degree depends upon its location on the globe; that is, the Geographic Coordinate System is not a rectangular coordinate system. Thus, a coverage could be digitized in meters, but not in degrees. The advantages and disadvantages of digitizing a coverage in either digitizer units or in real-world units are outlined in the following table and explained in the paragraphs that follow.

Digitizer Unit	Real world units
Easy to create edit plots at scale of source map	Maps need to be plotted at a precise scale to overlay edit plots
Digitizing staff has less to learn and understand	Digitizing staff should understand transformation and projection concepts
Coverages are not spatially referenced and cannot be displayed simultaneously	Allows multiple coverages to be shown, such as background or adjacent coverages
Inconvenient for update	Usually used for update
Less concern over bad projection information	Must have correct projection parameters
Do not know whether initial digitizing of tics was accurate	RMS error indicates actual tic accuracy in real-world units

A common step in quality assurance when initially developing a database is to compare a digitized file to its source map. This is most commonly done by creating an edit plot and overlaying it on a light table. If a map is digitized in real-world units, it may have been stretched and scaled so that it will no longer register accurately with the source map, even if the file was digitized accurately. If coverages are to be digitized in real-world units, the digitizing staff should understand how to project and transform coverages. This requires some knowledge of projection concepts, as discussed in Map projection concepts. The ability to display adjacent and background coverages can be quite helpful, particularly when updating maps. This is not possible when maps are digitized in digitizer units.

Coverages should not be edited, cleaned, built, or buffered, nor should any spatial analysis be performed when they are stored in reference units (latitude-longitude). The algorithms that perform snapping functions or maintain topology use a measurement of length or area that is based upon Cartesian coordinates. The length of a line of latitude between two meridians varies with latitude, and area is confusing when measured in square degrees. Considering the above, the guideline below could be followed:

- For new database creation, digitize feature in digitizer unit. However, before digitizing all the features make a primary transformation into real world coordinates to see the RMS error in real world unit. This will enable to have the original for future reference.

- For database update, digitize feature in projected coordinates.

Establishment of tics

A tic is a registration or control point for a coverage. Tics register coverage coordinates to a common coordinate system (e.g., BTM meters, LCC feet etc.) and, therefore, relate locations of features in a coverage to locations on the earth’s surface. Points representing tics could be located on maps by identifying the x- and y-coordinates from a known real-world coordinate system. These are usually stated in degrees of latitude and longitude. At least four points common to each map sheet should be found whose location and coordinate values on the map could be easily determined for each layer of the database. These would be used as control points when digitizing and transforming the map. It is recommended to establish a master tic file - a file to which all layers are referenced.

A consistent procedure should be maintained to number tics. One common guideline is to use a regular numbering system, always increasing in the same direction, and not back and forth across the maps. If map sheet corners are used as tic locations, the coincident locations on adjacent maps are commonly given the same Tic-ID. The Tic table for the source should be stored for future reference.

Transformation

Transformation converts data from one coordinate system to another. It is often used to convert coverage created in digitizer units (most often inches) into the real-world units represented on the map manuscript. These real-world locations are represented by a particular projection and some units. Normally, the real locations are marked on the map sheet described in meters or feet or at the latitude-longitude intersection.

The transformation function is based on comparing the coordinates of the input coverage tics with the coordinates for corresponding output coverage tics. The Tic-IDs are used to identify the tics to be compared. The minimum number of tics needed depends upon the transformation option used.

Transformation uses one of the three types of equation: Affine, Projective, Similarity. Affine is the default method that uses a set of two linear equations having six unknown parameters. At least three Tics are needed to perform an Affine transformation. The Projective method uses a projective equation that needs eight parameters to resolve and at least four tics. The Similarity method is similar to the Affine method, but needs four unknown parameters to resolve and a minimum of two tics.

A Root Mean Square (RMS) error is calculated for each transformation performed and indicates how good is the derived transformation. The transformation is derived using the least squares, so more tics than necessary could be given. The RMS error for each transformation is reported in both input coverage units (e.g., 0.031 inches) and output coverage units (e.g., 37.465 feet). RMS Error (input, output) = (0.031,37.465)

A perfect transformation produces an RMS error of 0. Although a value of 0.000 would not be obtained, the RMS error should be kept as low as possible. It is best to establish and maintain a maximum acceptable error. Acceptable values would vary depending on the accuracy of the original data and the scale of the source map. High RMS errors indicate that the old and new tics do not correspond to the same relative locations. If the RMS error is higher than the established value, it could indicate conversion problems which might incorrectly stretch output coverage. If a minimum number of tics is provided, the transformation would produce 0 RMS error.

3.5.7 Merging

Usually map data is collected as a series of adjacent map sheets - each sheet is mounted on the digitizer and digitized one sheet at a time. In order to be able to combine these smaller sheets into larger units or study areas, the coordinates of each coverage must be transformed into a single common coordinate system. Once in a common coordinate system, attributes are associated with features. Then, as needed, map sheets for a layer are edge-matched and joined into a single coverage for the study area. Since the accuracy varies for different sheets, the methodology adopted during merging of these sheets influence the ultimate accuracy of the combined dataset.

3.6 Attribute data acquisition

When geographic information is stored in a GIS, each geographic feature has two types of data stored:

- Location information in the form of explicit geographic coordinates
- Descriptive information in the form of standardized descriptive attribute values

For geographic features to be properly described, the descriptive attribute information must be explicitly associated with the location information. This is accomplished by storing location information in one file and descriptive attribute data in another. In real-time, a relational link is used to associate the attribute data with the location data.

3.6.1 Data conversion

Using attribute data assigned by some other reputable organization saves man-hours, therefore, whenever possible attribute information that comes with existing data should be used. If reliable attributed data is available, it should be used. Each and every feature and attribute should not be visited unless the validity of the new data is in question. A software conversion of the attributes should be performed to reformat them to the current project specification. Then only a quality control check should be performed rather than full labeling and quality control sessions.

One way to preserve pre-existing attribute information is by importing existing data through software filters, which reformat or convert existing attribute data to the finished format of the project. Usually, these filters would have to be custom designed for each job. They would be developed based on a set of attribute conversion rules defined in the project data dictionary.

3.6.2 Automatic Label Point Generation

Location information on area features is usually composed of a set of coordinates defining the area feature boundary and the coordinates of a label point, as well as the entity that stores the attribute data. Prior to attributing area features a check should be conducted to ensure that all area features contain a corresponding label point. If the label point is not present, it is impossible to assign attributes to that area feature. To ensure that attributes could be assigned to all area features, label points should be added automatically to all area features that do not contain them. After new label points are installed in the database, a check should be conducted to ensure that a valid polygon topology exists.

3.6.3 Default Attributes Assignment

There are times when a discrete attribute value cannot be determined for a feature or attribute to be calculated or derived based on other attribute values. These derived, computed, or assumed attribute values are called default attribute values. During the planning phase, a data dictionary should be developed. This document defines those attributes that would be used in the database. It also identifies the range of valid values that could be stored for each field. For each feature classification the following information is provided in the data dictionary:

- Allowable feature types
- Classification code
- Default attributes
- Optional attributes

Using the data dictionary as a guide, software should be developed that would automatically generate acceptable attribute values based on assumptions from other attributes or information. These default values would not necessarily be the final value to be stored, but could later be changed based on the operator's interactive attribute assignment.

3.6.4 Interactive Attribute Assignment

Generally, interactive attributing is used most often in database generation. In this process, the analyst must select each and every feature and manually change the attributes. Although this process is time consuming, there are situations in the database generation process where there is no other way to add attribute information.

3.7 Quality assessment and database validation

Database validation is the process of determining if database values are accurate, complete, and logically consistent as specified in the product specification. Validation would often consist of several steps, including logical checks, accuracy assessments, and error analysis. In this context, accuracy assessment is the determination of spatial and thematic accuracy relative to a known standard, and error analysis involves the evaluation of data with regard to measurement uncertainty and includes source errors, use errors, and process errors.

Validation could be considered as a two level process, with primary validation looking at the directly verifiable properties of the data, and secondary validation involving various levels of user judgment during the assessment. Different types of validation are described in the following table:

Validation type	Validation process	Example
Primary Validation	Logical cartographic consistency	Closed polygons.
		One label for each polygon.
		No duplicate arcs.
		No overshoot arcs (dangles).
		Similar features use similar symbols.
	Logical attribute consistency	
	Values within logical range (look for illegal values).	Dates (e.g. month less than or equal to 12).
		Time of day less than 24:00 hours.
		Nominal data illegally re-sampled into ratio data.
		Rainfall values equal to or greater than zero.
	Linkage of features with attribute fields.	Is a polygon feature properly linked with it?

Validation type	Validation process	Example
Secondary Validation	Logical query and statistical tests of the spatial and attribute data (look for unlikely values).	Points placed in distant locations on the map
		Elevations with reasonable values.
	Ground truth or comparison to known standards specified in the product specification	
	Sample ground areas and compare to database. Completeness of data (relative to product specification).	Evaluate positional accuracy (detail is given in the following section).
		Evaluate attribute accuracy.
		Are all the unions included in a union map?
	Sensitivity analysis	
	Change the data, and see if those changes affect the results of your application.	

3.7.1 Positional Accuracy assessment

The positional accuracy of a dataset should be tested by comparing it with an independent source of higher accuracy. This source should be of the highest accuracy possible and practicable for evaluating the accuracy of the dataset. The steps for accuracy testing are as follows:

Determine the extent of testing: This should be the extent of the dataset. If the dataset has varying accuracy, the whole extent should be divided into zones of uniform accuracy. Accuracy assessment should be performed for each zone, and the overall accuracy should be later derived from the zonal accuracy.

Decide independent source: The independent source should be one of the following: Geodetic Network, Global Positioning System (GPS) survey, Dataset of higher accuracy (the accuracy of which is well known and meet the expected quality of the product specification)

Identify check points: These are points visible on the ground, on the independent source of dataset and on the product itself.

Check point distribution: The maximum distance between two checkpoints should be less than 10% of the diagonal extent of the dataset. At least 20% of a checkpoint should be located in each quadrant of the dataset.

Compare position: The position for the checkpoints derived from the dataset and independent source would be compared and the accuracy estimated.

3.8 Quality control for data from secondary sources

A major part of the NWRD data is produced by different agencies. These databases are of varying quality and need to be incorporated in the NWRD. The database producers produce data according to their own needs and specification. The NWRD may use the same data for some different needs. Data quality is a big issue in this context where the quality specification of the producer may vary with that of the NWRD.

- Product specification in the NWRD:

  Like its own production, the NWRD should prepare product specification before collecting data from other organizations. Specification is needed to assess the collected data in the context of planning needs in WARPO.

- Collect product specification in the producing agency:

  The dataset product specification (if any) should be collected from the producing agencies to cross match with NWRD specification. If their respective specifications do not match, the effort to collect the datasets would be a fruitless exercise.

- Collect data quality documents from the producing agencies

  The data quality document (if any) should be collected from the producing agency for primary assessment of dataset quality. In the absence of such documents, discussions should be held with relevant people to collect as much information as possible.

- Validate data:

  Validate the collected dataset according to Section 2.6. During validation, identify any need for refinement. The quality aspect of data refinement is discussed in Section 2.7.1.

- Incorporate in the NWRD:
  If the quality meets the product specification incorporate the dataset into the NWRD. If needed, apply the special process for quality improvement of the dataset before adding it to the NWRD.

3.8.1 Data quality in refining data from other sources

It is necessary to identify and define the quality of the collected datasets. The desired quality of the data is defined in the product specification. The best way to manage data quality is to identify it and implement a plan to improve it. This section investigates how data quality is considered in refining secondary data.

Refining Consistency

Collected data frequently have incomplete and conflicting data domains for the same data characteristic, since data products are developed by different people for different uses. Conversion programs may be developed to convert by aggregation from the existing domain to the NWRD specified domain. Data may have different structure consistency that must be aggregated to obtain the NWRD data structure consistency. The NWRD data structure consistency is developed to meet the planning needs of the dataset. Data often have different data derivation rules. Each data product characteristic needs to be analyzed to determine whether it is derived. If data product characteristic is derived, the data derivation algorithm must be obtained and included in the data characteristic which contributes to the derivation. The different data derivation algorithms are then reviewed and discrepancies are resolved to meet the product specification.

Refining Accuracy

Data have different degrees of accuracy, although the accuracy is frequently unknown and not readily apparent. The accuracy cannot be changed or improved during refining. It could only be identified and documented to increase understanding of the data. The identification of accuracy could be done by people’s knowledge about the data. Data accuracy should be identified and documented at the most detailed level. Clients could determine for themselves the levels of accuracy to be combined to meet their needs. The most detailed levels of accuracy should be documented during data refinement.

Achieving Completeness

Generally, data product characteristics have many different data uses. Data completeness could be ensured by identifying the data needed to comply with the product specification. If parts of the data are missing, it is necessary to determine whether the data exist in some organization that could be acquired through data sharing arrangements. Data that are needed to support the product specification but do not exist anywhere or cannot be acquired through data sharing must be collected by own initiative.

3.9 Data quality documentation

The product specification lists data quality elements and sub-elements that need to be described in the data quality documentation. The evaluation procedures for each element are also specified in the product specification. Documentation would be prepared following the format given in Annex A.

3.10 Database management and data quality

The NWRD would be continuously upgraded since it aims to meet all data requirements for the planning needs of WARPO. The NWRD would continue to provide data according to research needs, and in this context, database management and quality assurance is of utmost importance. The following issues need to be considered:

Maintaining quality of the dataset during updates: The original product specification needs to be followed while updating any dataset. If there should be any need to modify the product specification, the operation should follow the latest specification.

Maintaining up-to-date status of the data quality document: Any update in the dataset should be followed by updating the quality document. This issue is described in Section 1.7.1 in detail.

References

Stephen C. G., Morrison J. L (Editors) 1995: "Element of Spatial Data Quality", International Cartographic Association.

Goodchild M., Gopal S. (Editors) 1989 : "Accuracy of Spatial Database", Taylor and Francis

FGDC, 1994 : "Content Standards for Digital Geospatial Metadata

ISO 1998, Geographic information Geomatics, Part 13: Quality principles

ISO 1999, Geographic information Geomatics, Part 14: Quality evaluation procedures

ISO 1998, Geographic information Geomatics, Part 15: Metadata

Brackett M. H., 1994 : "Data sharing using common Data Architecture", Wiley Professional Computing.

NCGIA Core Curriculum in Geographic Information Science, URL: www.ncgia.ucsb.edu/giscc

NCGIA, GIS Core Curriculum for Technical Programs, URL: www.ncgia.ucsb.edu/cctp

Tony O'Dempsy, 2000, Managing Accuracy of Spatial Data, White Paper, ESRI South Asia

FGDC, 1998, Geospatial Positioning Accuracy Standards, Part 3, National Standards for Spatial Data Accuracy

Annex A: Data quality elements and sub elements as per ISO standard

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
COMPLETENESS					
Commission	Pass/fail	Pass/fail	Boolean variable	100 items in a data quality scope, 10 excess items, 5 missing items, FAIL (assuming conformance quality level is 5 in terms of excess items)	Dataset series
Excess data present in a dataset	Measure of difference between the number of items in a universe of discourse (e.g. ground truth or reference dataset) and the number of items.	> Existence of excess items	Boolean variable	1: signifying the existence	Dataset
		> Number of excess items	Number	10 (number of excess items)	Reporting group - feature types, feature attributes, or feature relationships: degree to which the types defined in the product specification are present as types in a dataset
		> Percentage of excess items	Ratio	10 percent of excess items = 10/(100)	Reporting group – feature instances, attribute values, or single occurrences of feature relationships: degree to which single occurrences are present as occurrences in a dataset
Omission	Pass/fail	Pass/fail	Boolean variable	100 items in a data quality scope, 10 excess items, 5 missing items, FAIL (assuming conformance quality level is 5 in terms of excess items)	Dataset series

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
Data absent from a dataset	Measure of difference between the number of items in a universe of discourse (e.g. ground truth or reference dataset) and the number of items.	> Existence of omission	Boolean variable	1: signifying the existence	Dataset
		> Number of missing items	Number	10 (number of excess items)	Reporting group - feature types, feature attributes, or feature relationships: degree to which the types defined in the product specification are present as types in a dataset
		> Percentage of missing items	Ratio	5 percent of excess items = 5/(100)	Reporting group – feature instances, attribute values, or single occurrences of feature relationships: degree to which single occurrences are present as occurrences in a dataset
LOGICAL CONSISTENCY					
Domain consistency	Pass/fail	Boolean variable	Boolean variable (Boolean variable indicates for a given number of items whether there are any failure)	100 attribute value in a data quality scope, 3 attribute values violating value domain Pass (assuming 5 in terms of violation)	Dataset series

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
Adherence of values to the value domains	Degree of inconsistency	> Existence of non conforming items	Boolean variable	1 : signifying the existence	Dataset
		> Number of items which do not adhere to value domain	Number	3 (number of attribute values violating value domain)	Reporting group – features or feature attributes: degree to which the values fall within their value domains.
		>Percentage of items do not adhere to value domain	Ratio	3 (percentage of attribute values violating value domain 3/100)	Reporting group – features or feature attributes having derived value domains; that is, a value derived from encoded information of other items in a dataset consistent with the bounds of the other items.
Format consistency	Pass/fail	Pass/fail	Boolean variable (Boolean variable indicates if the format of the product specification was followed)	100 records in the data quality scope, 2 records not adhering to the prescribed structure Pass (assuming conforming threshold = 5)	Dataset series

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
Degree to which data is stored in accordance with the structure of a dataset	Degree of inconsistency	> Existence of non conforming items > Number of items which are not in accordance with the structure of the dataset > Percentage of items do not adhere to value domain	Boolean variable Number Ratio	1 signifying the existence 2 (number of records not in accordance) 2 (percentage of records not in accordance)	Dataset
Topological consistency	Pass/fail	Pass/fail	Boolean variable	1000 connections between road segments in a data quality scope - 10 connections are found incorrect FAIL assuming conformance threshold is 5 in terms of number of violations	Dataset series
Degree of adherence to characteristics of geometry remaining invariant under continuous mapping transformations so that a dataset's topological incidence relations remain unchanged	Degree of inconsistency	> Existence of violations of prescribed topologic relations > Number of violations of prescribed topologic rule Percentage of violations of	Boolean variable Number Ratio	1 : signifying the existence 10 (number of violations) 1 (percentage of violations)	Dataset The following characteristics of a data quality scope are measurable: a. Closure of geographic boundaries

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
		prescribed topologic rules			b. Consistency of Arial geocoding c. Planar topology by detection of intersections d. Validation of nesting/hierarchical relations
POSITIONAL ACCURACY					
Absolute or external accuracy	Pass/fail	Pass/fail	Boolean variable	10 point features (items) in the data quality scope Elevation error Number of features -3 m 1 -2 m 2 -1 m 0 0 m 3 1 m 3 2 m 1 3 m 1 PASS in terms of conformance RMSE = 2m	Dataset series
closeness of reported coordinate values to values accepted as or being true	error statistics - standard deviation, RMSE - % confidence level error statistic dimension - one dimensional - two dimensional - three dimensional	> existence of items with larger error than prescribed limit > number of items with larger error than the prescribed limit	Boolean variable number	1: signifying existence 5 with larger errors than the prescribed limit of 1 m	dataset reporting group - feature

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
	measure components - base error - proportional error	>Percentage of items with error larger than the prescribed limit >mean error >standard deviation, variance >variance covariance matrix error ellipsoid root mean square error >percent confidence level+C23	ratio angle change in position over time distance distance ratio	50 Mean error = 0.4m Variance = 3.04m2 standard deviation = 1,74m RMSE=1.79m	reporting group: feature attributes: applicable only to feature attribute that supply coordinate values
Relative or internal accuracy	Pass/fail	Pass/fail	Boolean variable	10 distance relationships (items) between 5 point features in a data quality scope Distance error Num. of items -3 m 1 -2 m 1 -1 m 0 0 m 3 1 m 2 2 m 2 3 m 1 FAIL (assuming conformance RMSE = 1)	dataset series

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
– closeness of the positional relationships of features in a dataset to the relationships accepted as or being true	error statistics - standard deviation, RMSE - % confidence level	> existence of items with larger error than prescribed limit	Boolean variable	1: signifying existence	dataset
	error statistic dimension - one dimensional - two dimensional - three dimensional	> number of items with larger error than the prescribed limit	number	+E18	reporting group - feature
	measure components - base error - proportional error	>Percentage of items with error larger than the prescribed limit	ratio	5 with larger errors than the prescribed limit of 1 m	reporting group: feature attributes: applicable only to feature attribute that supply coordinate values
		>mean error >standard deviation, variance >variance covariance matrix error ellipsoid root mean square error >percent confidence level	angle change in position over time distance distance ratio	50 Mean error = 0.1m Variance = 3.29m2 standard deviation = 1,81m RMSE=1.82m	
Pixel position accuracy	Pass/fail	Pass/fail	Boolean variable	10 pixel items with relative errors to ground truth in a data quality scope Distance error Num. of features 0 m 5 1 m 2 2 m 2	

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
				3 m 1 PASS (assuming conformance threshold = 2.0m standard deviation)	
	error statistics: - standard deviation, RMSE, error ellipse or error ellipsoid - % confidence level	> existence of pixels with larger error than prescribed limit	Boolean variable	1: signifying existence	Limited to raster data
		> number of pixels with larger error than the prescribed limit	number	3 with larger errors than the prescribed limit of 1 m	
		>Percentage of pixels with error larger than the prescribed limit	ratio	30 percent	
		>mean error >standard deviation, variance >variance covariance matrix error ellipsoid root mean square error >percent confidence level	distance angle ratio of distance	Mean error = 0.4m Variance = 3.04m2 standard deviation = 1,74m RMSE=1.79m	

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
TEMPORAL ACCURACY					
Accuracy	Pass/fail	Pass/fail	Boolean variable	10 feature (items) with temporal references in the data quality scope Error Number of features - 2 days 1 - 1 day 1 0 day 5 1 day 2 2 days 1 conformance level is 3 in terms of absolute error more than 1 day Two items has error more than 1 day PASS	Dataset series
correctness of the temporal references of an item (reporting of error in time measurement)	error statistics: - standard deviation - % confidence level	> existence of items with larger error than prescribed limit	Boolean variable	1: signifying existence	dataset reporting group – features, feature attributes, or feature relationships
		> number of items with larger error than the prescribed limit	number	3 with larger errors than the prescribed limit of 1 m	
		>Percentage of items with error larger than the prescribed limit	ratio	30 percent	
		>mean error	Time unit	Mean error = 0.9m	

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
		>standard deviation >variance >percent confidence level		Variance = 1.09m2 standard deviation = 1,04m RMSE=1.38m	
Temporal consistency	Pass/fail	Pass/fail	Boolean variable (Boolean variable indicates if ordered events or sequences are correct)	5 events in a data quality scope; A,B,C,D,E is the correct sequence. -In the dataset, the five events are recorded; A,B,D,E,C -the individual event is defined to be an item. Fail (assuming conformance quality level is 50% in terms of percentage of items prescribed event sequences or order)	Dataset series Dataset
correctness of ordered events or sequences, if reported	degree of inconsistency	> existence of items with larger error than prescribed limit	Boolean variable	1: signifying existence	reporting group – features, feature attributes, or feature relationships
	error statistic: standard deviation % confidence level	> number of items with larger error than the prescribed limit	number	3	
		>Percentage of items with error larger than the prescribed limit	ratio	60 percent	

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
Temporal validity	Pass/fail	Pass/fail	Boolean variable (Boolean variable indicates if ordered events or sequences are correct)	100 features (items) with the collection date of 1995 in a data quality scope - 95 features actually collected in 1995 - 5 features actually collected in 1990 Pass (assuming conformance quality level 3 in terms of invalid items)	Dataset series Dataset
validity of data with respect of time	degree of inconsistency	> existence of items with larger error than prescribed limit	Boolean variable	1: signifying existence	reporting group – features, feature attributes, or feature relationships
	error statistic: standard deviation % confidence level	> number of items with larger error than the prescribed limit	number	5	
		>Percentage of items with error larger than the prescribed limit	ratio	5 percent	
THEMATIC ACCURACY					
Accuracy of value	Pass/fail	Pass/fail	Boolean variable	10 point-features (items) with temperature in a data quality scope Temp. error Number of items -3 C 1	Limited to feature attributes defined by a value

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
				-2 C 0 -1 C 2 0 C 5 1 C 1 2 C 1 3 C 0 Pass (assuming conformance quality level is 3 in terms of items with larger error than the prescribed specification limit)	
Accuracy of a value given to a quantitative attribute	error statistic: standard deviation, RMSE % confidence level	> existence of pixels with larger error than prescribed limit	Boolean variable	1: signifying existence	
		> number of pixels with larger error than the prescribed limit	number	2 items larger errors than the prescribed limit of 1^{0}C	
		>Percentage of pixels with error larger than the prescribed limit	ratio	30 percent	
		>mean error >standard deviation, variance >variance >percent confidence level	number type	Mean error = 5 ^{0}C Variance = 1.35 ^{0} C standard deviation = 1.16 ^{0} C 80 percent confidence: from -1 to +1 ^{0}C	

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
Classification correctness	Pass/fail	Pass/fail	Boolean variable (Boolean variable indicates for a given number of items whether there are any misclassifications)	In the universe of discourse; 10 features of class A, 5 features of class B 5 features of class C. In a data quality scope; 7 class A features are classified as A, 2 as B and 1 as C 1 class B feature is classified as A, 2 as B and 2 as C. 1 class C feature is classified as A, 1 as B and 3 as C. Fail (assuming conformance quality level is 80% in terms of percentage correctly classified (PCC))	dataset series
Comparison of the classes assigned to features or their attributes to a universe of discourse (e.g. ground truth or reference dataset)	Percent correctly classified (PCC)	Percentage of misclassification (percent correctly classified)	Percentage	12 features are correctly classified out of 20 features. PCC = 60%	dataset
		Producers accuracy	Percentage	Class A features: 7/10 = 70% Class B features: 2/5 = 40% Class C features: 3/5 = 60%	reporting group: feature or feature attribute
		Consumer's accuracy	Percentage	Class A features: 7/9 = 78% Class B features: 2/5 = 40% Class C features: 3/6 = 50%	
		Misclassification matrix	Matrix of percentage	Matrix of misclassified features A B C	

Data quality sub-element	Types of data quality measures	Example	Data quality value types	Computational example	Types of data quality scopes to which a type of data quality measure can be applied
				A 7 2 1 10 B 1 2 2 5 C 1 1 3 5 9 5 6 Matrix of misclassification (%) A B C A 70 20 10 100 B 20 40 40 100 C 20 20 60 100	

Annex B: Data quality implementation for Union Database

B.1 Product Specification

Dataset name	Administrative boundary up to union level
Product description	The administrative boundary (union) is a general purpose digital database to support GIS applications. Any other data that can be aggregated at union level or higher administrative unit can utilize this database for spatial analysis. The database will contain items (national geocode) that can be used to link any other data to the dataset. The database will use ESRI ArcInfo geo relational data model to support a vector GIS. The database will have polygon topology that will contain union polygon and their attributes. The database will also contain arc topology where arcs will be coded based on the highest administrative boundary they represent. (International, Coast line, District, Thana, union)
Types of features (name, definition)	Administrative boundary, Administrative area (union)
International boundary	Definition: International boundary of Bangladesh Source: SOB topographic map
Coastline	Definition: Coast line of Bangladesh Source: LANDSAT image
Other administrative boundary	Definition: Administrative boundary of districts, thana, union, pourshava Source: DLRS thana map
Universe of discourse	All union of Bangladesh listed in the 1998 BBS geomaster file and identified on the DLRS upazila map should be included.

Data dictionary for features			
Features Description			
Name	Feature type	Feature code	Default attribute
Administrative Boundary	Coastline	9	
	International Boundary	10	
	District Boundary	30	
	Thana Boundary	40	
	Union Boundary	50	
	Pourshava Boundary	70	

Attributes for Features			
Attribute Name	**Descriptions**	**Data Type**	**Width**
Geocode	Geocode according to BBS (Bangladesh Bureau of Statistics) Geomaster database	Numeric	10
District Code	District Code according to BBS Geomaster database	Numeric	2
Thana Code	Thana Code according to BBS Geomaster database	Numeric	2
Union Code	Union Code according to BBS Geomaster database	Numeric	3
District Name	District Name according to BBS Geomaster database	Character	25
Thana Name	Thana Name according to BBS Geomaster database	Character	25
Union Name	District Code according to BBS Geomaster database	Character	25

B.2 Data Quality Specification

Selected data quality elements	Completeness, Logical consistency, positional accuracy, thematic accuracy
Selected data quality sub elements	Commission, omission, domain consistency, format consistency, topological consistency, absolute accuracy, classification accuracy

Detail specification of data quality sub elements

Data quality sub elements	Product specification	Scope
Completeness (Commission)	Requirement: All union of Bangladesh listed in the 1998 BBS geomaster file and identified on the DLRS upazila map should be included. Assurance: Exhaustive Comparison with the geomaster file and DLRS maps will be conducted. A cross check by DLRS will be arranged. BBS will delineate additional boundaries.	The dataset
Completeness (Omission)	Not included	
Logical consistency (Domain consistency)	Requirement Unit: Units of measurement should be English measurement system Assurance: All attribute names and definition are verified	The dataset
Logical consistency (Format	Datum	The dataset

Data quality sub elements	Product specification	Scope
consistency)	Horizontal: Everest Projection: BTM Database structure: ArcInfo NT and ArcView Shape file format Assurance: Verification and review	
Logical consistency (Topological)	Requirement: Lines: All lines should have an ID value Polygon: Polygon should have only one level point. Assurance: Verification	
Positional accuracy (Absolute)	Requirement: The positional accuracy should be same of that source map (SOB topo map, LNADSAT image, DLRS thana map). In the digitizer unit lines should not vary more than 0.25 mm. Assurance: Making check plot and verifying. If needed check plot should be taken more than once. A cross check will be done by the DLRS for authorization	
Positional accuracy (Relative)	Not included	
Temporal accuracy (Accuracy)		
Temporal accuracy (Consistency)		
Temporal accuracy (Validity)		
Thematic accuracy (Classification accuracy)	Line: Arcs that will be given code for the highest administrative boundary it encloses. For example: If a line represents international boundary, district boundary and union boundary, it will be assigned code for international boundary Assurance: Verification	The dataset

B.3 Quality management in production

Assessment of source

Documentation

Map title	Union Database of Bangladesh
Index of map sheet if part of a map series	N/A

Organization	DLRS (Department of Land Records and Survey)
Project under which the map was prepared	Revenue surveyed by DLRS under the authority of Government in the season of 1922 - 1964.
Name of individual who authorizes the map (id any)	Settlement Officer under the authority of Government of Bangladesh
Scale of the map	1 inch = 1 mile (1 : 63,360)
Year of publication	1927 (updated in 1981)
Source of the content (in detail) as described in the map sheet	Revenue surveyed by the government in the season of 1922-23. The content of the map District, thana, mauza/village and municipal boundary, metalled/unmetalled road, railway line, thana, district headquarter, post & telephone office, factory & coal pit, market place etc.
Material used (paper, Mylar)	Paper
Distortion, folding…	Little distortion due to folding of the map
Geo-referencing available	Yes
Color or black and white	Black and white
Readability of features	Good
Readability of texts	Good
Final grading according to table 1	II

Grading

Grade	**Descriptions**	**Grade Name**
Grade I	Map features, data information everything is quite good.	Very Good
Grade II	Everything is ok, but need some check.	Good
Grade III	Many errors but can be used giving little effort.	Fair
Grade IV	Quality is very bad which can’t be usable as source data.	Poor

Data Capture		
Map digitisation		
Map preparation	Folds or wrinkles removed	Removed by putting paper or any suitable weight
	Feature transferred to stable medium or digitized from source directly	Source directly

Registration	Establishment of tic points	Graticules specified in the source map
	Number of tic points used	13
	Spread of tic points over the map sheet	Evenly spread
	RMS error recorded during registration	0.002 inch (input), 2.88 meter (out put)
Digitization environment		
	Digitization mode	Stream Mode
	Snap distance	0.005 inch
	Weed tolerance	0.002 inch
Data editing environment	Fuzzy tolerance	0.002 inch (without projection)
	Dangle length	0.005 inch (without projection)
	Number of missing features	No
	Number of excess features	No

Checking	
Check plot made and compared with the original	Yes
Checked for dangle node errors and corrected	Yes
Checked for duplicate line segments, missing line segments mislabelled line segments, unlabelled polygons and conflicting names.	Yes/No
Compared with check plot for strange polygons which do not appear on the original map, but which have occurred due to operator negligence	Yes/No

Editing	
duplicate lines removed	Yes
polygons are closed	Yes
lines snapped to the relevant nodes	Yes
Labelling	
All lines and polygons have labels and corresponding ID	Yes

Geo-referencing	
Tics	
Minimum number of tics per sheet	
Minimum number of common tic in adjacent sheets	
Maximum RMS in transformation	
Input	0.015
Output	25
Minimum RMS in transformation	
Input	0.001
Output	2.2

Edge matching	Procedure of edge matching	First of all the RMS errors for all the thanas of the district were listed to get a quick picture of that district. The boundary of less RMS errors value containing thanas considered as the common boundary of adjacent thanas was kept and in this way the edge match of all thanas was done. Shape and aerial variation were also considered from the source map during edge matching.