UNCLASSIFIED

AD-A264 125

AR-006-971

DEPARTMENT OF DEFENCE

DSTO

Information Technology Division

RESEARCH NOTE
ERL-0632-RN

DATA QUALITY STATEMENTS FOR SPATIAL DATABASES

by

R.J. Williams

ELECTRONICS RESEARCH LABORATORY

APPROVED FOR PUBLIC RELEASE

93-10331

UNCLASSIFIED

93 5 11 097

UNCLASSIFIED

AR-006-971

ELECTRONICS RESEARCH LABORATORY

Information Technology Division

RESEARCH NOTE
ERL-0632-RN

DATA QUALITY STATEMENTS FOR SPATIAL DATABASES

by

R.J. Williams

SUMMARY

This paper overviews contemporary issues in incorporating data quality statements into spatial databases. The paper includes discussion of two approaches; one emanating from the Digital Chart of the World Project and one through a working party within the International Cartographic Association.

JUL 92

APPROVED FOR PUBLIC RELEASE

POSTAL ADDRESS: Director, Electronics Research Laboratory, PO Box 1500, Salisbury, South Australia, 5108. ERL-0632-RN

UNCLASSIFIED

CONTENTS

FIGURES

ABBREVIATIONS

ADF	Australian Defence Force
AGPS	Australian Government Publishing Service
AR	Australian Report
ASDTS	Australian Spatial Data Transfer Standard
CD-ROM	Compact Disk - Read-Only Memory
DCW	Digital Chart of the World
DMA	Defense Mapping Agency (USA)
DSTO	Defence Science and Technology Organisation
DSTOS	Defence Science and Technology Organisation Salisbury
ERL	Electronics Research Laboratory
GIS	Geographic Information System
GPS	Global Positioning System
ICA	International Cartographic Association
NCDCDS	National Committee on Digital Cartographic Data Standards (USA)
ONC	Operational Navigation Chart
SDTS	Spatial Data Transfer Standard (USA)
VPF	Vector Product Format

1 INTRODUCTION

Digital processing of spatial data brings immense benefits in the form of rapid, precise and sophisticated analysis, but reveals weaknesses which may not otherwise be apparent. Computers are very precise machines, and errors and uncertainties in data can lead to serious problems, not only in the form of inaccurate results but in the consequences of decisions made on the basis of poor data. Capabilities that excite enthusiasm among potential users are the ability to change scale and the ability to overlay different themes of information at random. These capabilities are indeed exceedingly useful; they constitute much of the comparative advantage geographic information system technology (commonly referred to as GIS) holds over spatial analysis based on analog maps (Goodchild, 1991; Abler, 1987).

These capabilities, however, can also mislead decision makers who are unaware of the imprecision inherent in all cartography and who are untutored in the ways errors compound when map scales are changed or when maps are merged. Burrough (1986) observes "a false lure in the attractive, high quality cartographic products that cartographers, and now computer graphics specialists, provide for their colleagues in environmental survey and resource analysis. ... Many scientists and geographers know from field experience that carefully drawn boundaries and contour lines on maps are elegant misrepresentations of changes that are often gradual, vague or fuzzy".

Goodchild (1991) warns that "if the burgeoning GIS industry is indeed driven by false perceptions of data accuracy, then the truth will be devastating: even the simplest products will be suspect. The best insurance at this point is surely to sensitise the GIS user to the accuracy issue, and to develop tools which allow spatial data handling systems to be used in ways which are sensitive to error". That is, systems that use digital geographic information require a method to maintain and manage their contents and processes over the long term.

Up until just a few years ago, the description of data quality and associated issues have been neglected topics. Fortunately, however, the topic is now being recognised as one of importance and the issue of the description of data quality is being addressed by a number of research organisations and professional bodies throughout the world. The catalyst for this work is because of incomplete coverage, variable accuracy, inconsistencies in standards and inadequate sources.

Two approaches are worthy of assessment. One approach emanates from the Digital Chart of the World (DCW) project while the other emanates from the Scientific Advisory Board of the International Cartographic Association.

1.1 Digital Chart of the World (DCW)

The DCW Project is a United States Defense Mapping Agency research and development effort (to which Australia, via the Royal Australian Survey Corps, is a cooperative partner), whose ultimate objective is the promulgation of standards for the exchange of digital spatial information and the development and distribution of a global topographic database on compact disk (CD-ROM) (DMA, 1991).

DCW will be a new product of the Defense Mapping Agency (DMA). It will provide worldwide coverage using a topologically based vector data structure to digitally represent the earth's land surface information on a microcomputer accessible storage media. The 1:1000000 scale Operational Navigation Chart (ONC) series will provide the majority of the information to produce the DCW. The Jet Navigation Chart (JNC) series will provide the information over the Antarctica. Features will be collected and stored along with their attributes at the level of detail provided on the ONCs.

The purpose of the project is twofold:

- To develop, refine, and establish a suite of standards that enable the exchange and utility of spatial information; and
- To perform the necessary research and development steps to produce the DCW in compliance with these standards.

In order to insure the suite of standards will be compatible with the international community, as well as the US Department of Defense; allied partners, namely Canada, United Kingdom and Australia, are participants in the overall research and development.

Standards to be developed for the DCW include format standards, media standards, a DCW product specification, and a data directory standard to include tiling, coverage index, thematic index, gazetteer index, and spatial query index.

The data structure used for DCW is a topologically structured vector structure in a relational model and is known as Vector Product Format (VPF). VPF also contains data quality information so that users may evaluate the utility of the data for a particular application.

1.2 International Cartographic Association initiative

Technological issues such as those concerned with digital data quality are also receiving attention from working groups within professional organisations. Perhaps the lead professional body in the disciplinary area concerned with spatial data is the International Cartographic Association (ICA). The ICA has within its organisational structure a number of commissions and working groups whose terms of reference, amongst other things, includes "undertaking efforts on critical topics of research". The Scientific Advisory Board of the International Cartographic Association has produced a set of guidelines as its contribution to a clear and consistent approach to the assessment of data quality. These are presented in Section 3 ICA Data Quality Proposal.

2 VPF DATA QUALITY STATEMENT

The data structure used for DCW is a topologically structured vector structure in a relational model and is known as Vector Product Format (VPF). VPF is a generic geographic data model designed to be used with any digital geographical data in vector format that can be represented using nodes, edges, and faces. VPF is based upon the georelational model, combinatorial topology and set theory. VPF also contains data quality information so that users may evaluate the utility of the data for a particular application.

VPF contains data quality information at a number of different levels within the database with the detailed description being modified from the Spatial Data Quality section (Section 4) of NCDCDS Report #7 (Moellering, 1986).

2.1 Data quality hierarchy

The VPF model is a hierarchical one with information held at database, library, coverage, feature and primitive levels. Data quality information at the database level applies to all libraries of the database, except where those libraries contain their own data quality information of the same kind. Similarly, data quality information at the library level (which may have been inherited from the database) applies to all coverages within the library, except those that contain their own data quality information of the same sort. Coverage level data quality information applies in the same manner to features. Feature level data quality information in turn likewise applies to both spatial primitives and attributes that compose them.

2.2 Data quality encoding

Data quality information is represented as attributes or as a coverage. If as attributes, it may be either added to an existing VPF table, or as an independent table residing at the appropriate level. If a coverage, it shall be a coverage whose area or complex features designate areas with uniform data quality information of specified types. Figure 1 depicts the attribute and coverage locations of data quality information through the database.

2.3 Types of data quality information

There are seven types of data quality information:

- Source. Source describes the origin or derivation of a single feature, primitive or attribute. This includes any processing techniques applied to the data, as well as the data source.
- Positional accuracy. Positional accuracy provides an upper bound on the deviation of coordinates in VPF from the position of the real world entity being modelled. Positional accuracy must be specified without relation to scale and shall contain all errors introduced by source documents, data capture, and processing.
- Attribute accuracy. Attribute accuracy describes the accuracy or reliability of attribute data.
- Currency. Currency represents the date at which the data was introduced or modified in the database. This date of entry is used as a proof of modification for a single data element, permitting statistical interpretation of groups of data elements.
- Logical consistency. Logical consistency describes the fidelity of relationships encoded in a VPF data set. Logical consistency requires that all topological foreign keys match the appropriate primitive, that all attribute foreign keys match the appropriate primitives or features, and that all tables described in feature class scheme tables do indeed have the relationships described.

VPF Data Quality Information

Level	Quality Attributes	Quality Coverages
Database	In a table within the database directory	Within the database directory
Library	In a table within the library directory, or within the library attribute table or a table related to it	Within the library directory
Coverage	In a table within the library directory, or within the library attribute table or a table related to it	Within the same library directory as the coverage to which the quality information applies
Feature	In a table within the coverage directory, or within the feature's feature table	Not applicable
Primitive	In a table within the coverage directory, or within the feature's feature table	Not applicable

Figure 1 VPF Data Quality Information

- Feature completeness. Feature completeness indicates the degree to which all features of a type for the area of the data set have been included.
- Attribute completeness. Attribute completeness indicates the degree to which all attributes of a feature have been included for that feature. Actually, since this information can be derived from the feature itself, simply by counting null values, this particular form of data quality information should not need to be explicitly included.

These types of information above are VPF's standard types of quality data. Product specifications, such as the Digital Chart of the World, call for additional types of data quality information as well.

2.4 DCW metadata

The DCW is one database with two libraries. The database level includes three tables: a database header table, a database description table and a library description table. The database header table contains metadata pertaining to the DCW data and includes information on security and release information.

The DCW library is a directory containing VPF tables, coverages and index tables. One table, known as the library header table, identifies the data set, sources, extent, projection, security, and data quality information in the library (Figure 2).

As the Digital Chart of the World is available for public release from February 1992, the schema will be the first containing a 'data quality statement' that will be supported as a 'standard'. Therefore, future defence data (in vector format) should include, as a minimum, that information as shown in Figure 2[1].

It seems unfortunate that, although 'data quality statements' have been identified as being important, the implementation in VPF (and therefore in DCW) is somewhat simplistic and poorly described in accompanying documentation. This component of VPF (and DCW) is clearly one needing further development and enhancement.

1 It is apparent that the developers of the 'data quality' module of DCW lacked experience and/or knowledge in cartography and surveying. In the draft documentation there are errors and uncertainties. Firstly, the projection is noted to be 'Unprojected' with decimal degrees but horizontal unit of measure is given as 'Meters' (possibly should be expressed in arc units). Secondly, the vertical unit is expressed as 'Meters' but the source material was an aeronautical chart with elevation in 'Feet'. The absolute horizontal accuracy was given as +-2040 meters (perhaps +-2 KM might have been more commensurate with the source scale).

VPF column name	DCW column name	Record entry
Undefined	ID	1
Product type	PRODUCT_TYPE	DCW
Name	LIBRARY_NAME	DCW
Data Structure Code	DATA_STRUCT_CODE	1, 2 and 6
Series	SOURCE_SERIES	ONC
Source Identification	SOURCE_ID	Complete ONC series
Edition	SOURCE_EDITION	Varies with source map sheet
Source Name	SOURCE_NAME	Operational Nav Charts, Jet Nav Charts
Source Date	SOURCE_DATE	1989
Ellipsoid Name	ELLIPSOID_NAME	WGS
Ellipsoid Code	ELLIPSOID_CODE	None
Vertical Reference Name	VERT_REF_NAME	Mean Sea Level
Vertical Reference Code	VERT_REF_CODE	MSL
Vertical Datum Code	VERT_DATUM_CODE	Unknown
Geodetic Datum Name	GEOD_DATUM_NAME	Unknown
Geodetic Datum Code	GEOD_DATUM_CODE	Unknown
Longitude of SW Corner	LON_SW_MBR	0 Longitude
Latitude of SW Corner	LAT_SW_MBR	90 South Latitude
Longitude of NE Corner	LON_NE_MBR	0 Longitude
Latitude of NE Corner	LAT_NE_MBR	90 North Latitude
Longitude	LON_BOUND_FACE	+- 180 degrees
Latitude	LAT_BOUND_FACE	+- 90 degrees
Projection Name	PROJECTION_NAME	Decimal degrees (Unprojected)
Projection Code	PROJECTION_CODE	Unknown
Security Classification	SECURITY_CLASS	U
Downgrading	DOWNGRADING	No
Date	DOWNGRADING_DAT	N/A
Releasability	RELEASABILITY	Unrestricted
Feature Completeness	FEATURE_COMPLETE	100% of ONC
Attribute Completeness	ATTRIBUTE_COMPL	100% of ONC
Consistency	LOGICAL_CONSIST	TBD
Edition Number	DATASET_ED_NO	1
Creation Date	CREATION_DATE	TBD
Revision Date	DATASET_REV_DATE	TBD
Recompilation Date	RECOMP_COUNT	0
Revision Count	REVISION_COUNT	0
Specification ID	PRODUCT_SPEC_ID	MIL-D-89009
Date	SPEC_DATE	April 29, 1991
Amendment	SPEC_AMENDMENT	N/A
Earliest Source	EARLIEST_SOURCE	1971
Latest Source	LATEST_SOURCE	1989
Quantitative Attribute	QUANT_ATTRIBUTE	Unknown
Qualitative Attribute	QUAL_ATTRIBUTE	TBD
Collection Criteria	COLLECTION_SPEC	ONC Spec and DCW Design Criteria
Absolute Horizonatal Accuracy	ABS_HORIZ_ACC	+- 2040 meters
Unit of Measure	HORIZ_UNITS	Meters
Absolute Vertical Accuracy	ABS_VERT_ACC	+- 610 meters
Unit of Measure	VERTICAL_UNITS	Unknown
Relative Horizonatl Accuracy	PT_PT_HORIZ_ACC	N/A
Relative Vertical Accuracy	PT_PT_VERT_ACC	N/A
Comments	COMMENTS	Source map editions from 1971 to 1989

Figure 2: Schema for DCW Library Header Table

The second initiative referred to earlier, that by the International Cartographic Association, offers an approach to improve on the weakness in the VPF 'data quality statement'.

3 ICA DATA QUALITY PROPOSAL

The International Cartographic Association (ICA) through its Scientific Advisory Board, has developed a set of guidelines as its contribution to a clear and consistent approach to the assessment of data quality.

The guidelines are intended to satisfy certain basic requirements:

- Defensible. Qualitative rating schemes like 'high', 'medium' and 'low' would be difficult to defend because of subjectivity, in the form of inconsistency between assessors, and confusion over what the terms mean. The guidelines emphasise objective measurement, with summaries as simple, unambiguous choices.
- Informative. The purpose of a rating should be to give the user the greatest possible amount of useful information. If ratings are to be designed by a testing scheme, they should be designed to pass as many detailed results of testing as possible on to the user. They should reflect likely uses by anticipating what the user will be doing with the data.
- Definitive. It is important that the differences between ratings be as definitive as possible, and not based on subjective scales of assessment.

Rather than attempt to assess quality in an absolute sense, the quidelines emphasise the quality of data relative to user needs and anticipated uses, by comparing reality to likely expectations. In many cases spatial databases are assembled from well known and widely distributed sources, so an important measure of quality is the degree to which the information content of the source has been captured accurately in the database: this relative measure may be more useful to the potential user than an absolute measure of quality.

The guidelines use certain terms which require definition:

- Reality: independently verifiable ground truth; an item of information that can be verified by visiting the appropriate place on the earth's surface and making a measurement or observation;
- Source: the documents (often maps) from which the database was built. The source is assumed to be available for assessment of the quality of the database;

- Database: the product being tested; a set of digital records organised in some appropriate structure. The assessment of quality extends not only to the records themselves, but also to information that can be deduced from the records by simple processes. For example, a user may wish to know the accuracy of the length of a digital line, whether length is stored explicitly in the database or computed from the line's coordinates;
- Source errors: inaccuracies apparent in the source when its contents are compared to reality. These may include the uncertainties due to different interpretations of ground truth;
- Processing errors: inaccuracies introduced by digital processing (including digitising) and thus apparent in the database when its contents are compared to the source.

The guidelines describe two distinct approaches, and each has two levels: overall summary rating, and detailed assessment. In the latter area sections of the guidelines have been adapted and modified from the Spatial Data Quality section (Section 3) of the proposed US National Spatial Data Transfer Standard (SDTS). This standard will be the basis of the proposed Australian Spatial Data Transfer Standard (ASDTS)(Moellering, 1986).

The intent of the guidelines is they be used to assemble an informative Data Quality Statement to accompany the database.

3.1 Overall Summary Rating

A summary rating is assessed using one of two methods, depending on whether accuracy is determined with respect to source document or ground truth.

3.1.1 Method 1

Method 1 is used to assess databases with respect to source documents, but also must address the quality of the source document itself, usually by reference to independent reports. A Method 1 rating has two parts, e.g. A1 C, denoting a database that captures accurately the entire contents of a source document of unknown quality. These parts are:

- A measure of the relationship of the digital database to its source;
- A statement of the quality of the source;

Refer to Figures 3 and 4 for measure of rating for Method 1.

SUMMARY RATING - METHOD 1 - Relationship between database and source

A		The database is not significantly less accurate than the document from which it was derived. Source map features are within 0.5mm of their source map positions at the scale of the source map (e.g. the location of a feature in the database is within 50m of its marked position on a 1:100 000 scale source map). Features attributes are coded correctly in all cases tested. Details of the testing procedure must be supplied (i.e. sampling design, numbers of features tested) as all data cannot be guaranteed error-free.
	A1	The database is a complete representation of the information on the source map: all facts that could reasonably be deduced from the source map can also be obtained from the database.
	A2	The database is an incomplete representation of the information on the source map. Certain classes of features are missing, or certain relationships between features cannot be obtained from the database. The data quality statement should specify the missing classes and relationships as precisely as possible, since users may otherwise assume the database to be a complete representation of the source.
B		There is a substantial loss of accuracy with respect to the source document: some parts of features are more than 0.5mm from their correct positions, and feature attribute coding errors are present. The data quality statement should include a detailed assessment of the severity of errors. In the case of certain types of errors (e.g. unsnapped junctions) it is important to know whether these errors could be removed by appropriate processing by the user. If so, the user should be supplied with appropriate parameters (e.g. the size of the required snap tolerance).

Figure 3 Summary Rating - Method 1 - Relationship between database and source

SUMMARY RATING - METHOD 1 - Quality of the source

A	Known and documented. The source document has been tested in a rigorous process of quality assessment, or has been certified by an identified group, and a data quality statement is either present on the source document or available. Note that the quality of the source document will have been assessed against the uses for which it was originally intended, e.g. visual communication of information about geographic variation. Unfortunately this will not be sufficient to ensure quality for many applications of the digital database. For example, information about the quality of contour positions is not useful in determining uncertainty in estimates of slope obtained from a contour database.
B	Implied, or documented, but not met in fact. Testing has shown that the actual quality of the source is not equal to reasonable expectations, or to the information in the supplied data quality statement. Details of testing should be provided if available.
C	Unknown. No information is available about the quality of the source.

Figure 4 Summary Rating - Method 1 - Quality of the source

3.1.2 Method 2

Method 2 assesses the quality of the database by direct reference to ground truth, and has only one part: that being a statement of the relationship of the database to ground truth. Refer to Figure 5.

3.2 Detailed Assessment

Spatial databases frequently contain multiple themes, often from different sources. A detailed assessment of data quality must address each theme individually, particularly in comparisons with ground truth. Detailed assessment is relevant in two cases:

- In determining the accuracy of the database in relation to its source (Method 1 above); and
- In determining the accuracy of the database in relation to ground truth (Method 2).

There are significant differences in the approaches in the two cases.

3.2.1 Method 1

Each assessment consists of five sections:

- Lineage;
- Positional accuracy;
- Attribute accuracy;
- Logical consistency; and
- Completeness.

3.2.1.1 Lineage

The lineage portion of a quality report includes a description of the source material from which the data were derived and the methods of derivation, including all transformations involved in producing the final digital files. The description should include the dates of the source material and the dates of ancillary information used for update. The date assigned to a source should reflect the date that the information corresponds to the ground; however, if this date is not known, then a date of publication may be used, if declared as such.

Any database created by merging information obtained from distinct sources should be described in sufficient detail to identify the actual

SUMMARY RATING - METHOD 2 - Quality assessment through ground-truthing

A	The relationship to ground truth has been tested by a scientific body appropriate to the data's theme A formal program of testing was carried out using either: • A source of higher accuracy; or • Ground checks. Accuracy is described in accordance with the procedures for detailed assessment.
B	The relationship to ground truth has not been tested directly. Information on data quality must be inferred as for Method 1, i.e. the accuracy of the database with respect to its source, and the quality of the source with respect to reality.

Figure 5 Summary Rating - Method 2 - Quality assessment through ground truth

source for each element. In these cases, either a lineage code on each element or a quality overlay (source data index, etc) should be provided.

The lineage report should include information on all coordinate transformations applied to the data, including changes of projections, and the parameters used in each transformation (e.g. figures of the earth).

3.2.1.2 Positional accuracy

Descriptions of positional accuracy should consider the quality of the final product after all transformations. The information on transformations forms a part of the lineage portion of the quality report.

Measures of positional accuracy may be obtained by one of the following optional methods:

- Deductive estimate: an estimate of positional accuracy based on knowledge of the errors introduced in each production step. Any deductive statement should describe the assumptions made concerning error propagation (e.g. independence);
- Internal evidence: an estimate based on repeated measurements, e.g. by having several operators digitise the same source material;
- Comparison to source: an estimate based on graphic inspection of results and comparison with source ("check plots"); and
- Independent source of higher accuracy: the preferred test for positional accuracy is a comparison to an independent source of higher accuracy. The number of test points and sampling design should be reported.

3.2.1.3 Attribute accuracy

Accuracy assessment for measures on a continuous scale (interval/ratio) should be expressed in terms of a numerical estimate of expected discrepancies (standard or RMS error). Accuracy for measures on a discrete scale (nominal) should be given as percent correct, which could be expressed in the form of a misclassification matrix with summary statistic for the classified attributes. Sampling design and sample size should be reported.

3.2.1.4 Logical consistency

A report on logical consistency should describe the fidelity of relationships encoded in the data structure of the database. Tests for permissible values may be applied to any data structure. Such a test can detect gross blunders, but does not ensure all aspects of logical consistency. A data base containing lines may be subjected to general questions such as 'Do lines intersect only where intended? Are any lines entered twice? Are all areas completely described? Are there any overshoots or undershoots? Are any polygons too small, or any lines too close?

For exhaustive areal coverage data transmitted as chains or derived from chains (see the layer model discussion below), it is permissible to report logical consistency as 'topologically clean' under the condition that an automated procedure has verified the following conditions:

- All chains intersect at nodes;
- Cycles of chains and nodes are consistent around polygons. Or, alternatively, cycles of chains and polygons are consistent around nodes; and
- Inner rings embed consistently in enclosing polygons.

3.2.1.5 Completeness

The quality report should include information about selection criteria, definitions used and other relevant rules used to capture features from the source. For example, geometric thresholds such as a minimum area or minimum width should be reported.

The report on completeness should describe the relationship between the objects represented and the abstract universe of all such objects present in the source. In particular, the report should describe the exhaustiveness of a set of features.

3.2.2 Method 2

Two different strategies are acceptable, depending on the nature of the theme:

- Layers; and
- Objects.

Each assessment consists of five sections:

- Lineage;
- Positional accuracy;
- Attribute accuracy;
- Logical consistency; and
- Completeness.

3.2.2.1 Layers

The theme represents a single variable with a value everywhere, e.g. a map of soil class, land use, or elevation. The database will likely be expected to provide estimates of the value of the variable at specific points, and the measure of accuracy should inform the user of the uncertainty involved in determining such values.

3.2.2.2 Objects

The theme consists of a set of well-defined geographic features with associated attributes. Features should be sufficiently well-defined to be identifiable on the ground, allowing a test of positional accuracy to be made with respect to ground truth. Building footprints, shorelines, rivers, mountain peaks, bridges and roads are examples of well-defined geographic features. In cases where the object is highly interpreted and thus not suitable for ground truth (an independent observer could not reasonably be expected to identify correctly whether an arbitrarily chosen point was located inside the object or not), accuracy cannot be evaluated (e.g. location of object 'The Top End' of the Northern Territory).

Accuracy should be assessed using the same five categories identified above (lineage, positional and attribute accuracy, logical consistency and completeness). For the layer model, positional accuracy should be omitted as it is not relevant, but attribute accuracy is particularly important, and attention should also be paid to the data structure aspects of logical consistency. For the object model positional accuracy is particularly important, but the data structure will likely impose few logical consistency conditions.

3.2.2.3 Lineage

The lineage portion includes a description of the entire process of data handling from raw ground observations through to the digital database, including all transformations involved in producing the final digital files. The description should include the dates of raw observations, and the dates of ancillary information used for interpretation or update.

Any database created by merging information obtained from distinct sources should be described in sufficient detail to identify the actual source for each element. In these cases, either a lineage code on each element or a quality overlay (source data index, etc.) should be provided.

The lineage report should include information on all coordinate transformations applied to the data, including changes of projections, and the parameters used in each transformation (e.g. figures of the earth).

3.2.2.4 Positional accuracy (object model only)

Descriptions of positional accuracy should consider the quality of the final product after all transformations. The information on transformations forms a part of the lineage portion of the quality report.

Measures of positional accuracy may be obtained by one of the following optional methods:

- Deductive estimate: an estimate of positional accuracy based on knowledge of the errors introduced in each production step from raw observations to digital database. Any deductive statement should describe the assumptions made concerning error propagation (e.g. independence);

- Internal evidence: an estimate based on repeated measurements, e.g. by having several operators collect and process the same data; and
- Comparison to ground truth: an estimate based on actual ground check of the positions of objects, e.g. using GPS.

3.2.2.5 Attribute accuracy

Accuracy assessment for measures on a continuous scale (interval/ratio) should be expressed in terms of a numerical estimate of expected discrepancies (standard or RMS error). Accuracy for measures on a discrete scale (nominal) should be given as percent correct, which should be expressed in the form of a misclassification matrix with summary statistic for classified attributes. Sampling design and sample size should be reported. Attribute accuracy may be assessed by comparison to ground truth, internal evidence or deductive estimates.

3.2.2.6 Logical consistency

A report on logical consistency should describe the fidelity of relationships encoded in the data structure of the database. Tests for permissible values may be applied to any data structure. Such a test can detect gross blunders, but does not ensure all aspects of logical consistency. A data base containing lines may be subjected to general questions such as 'Do lines intersect only where intended? Are any lines entered twice? Are all areas completely described? Are there any overshoots or undershoots? Are any polygons too small, or any lines too close?

For exhaustive areal coverage data transmitted as chains or derived from chains (see the layer model discussion below), it is permissible to report logical consistency as 'topologically clean' *under the condition* that an automated procedure has verified the following conditions:

- All chains intersect at nodes;
- Cycles of chains and nodes are consistent around polygons. Or, alternatively, cycles of chains and polygons are consistent around nodes; and

- Inner rings embed consistently in enclosing polygons.

3.2.2.7 Completeness

The quality report should include information about selection criteria, definitions used and other relevant rules used to capture features from the source. For example, geometric thresholds such as a minimum area or minimum width should be reported.

The report on completeness should describe the relationship between the objects represented and the abstract universe of all such objects in reality. In particular, the report should describe the exhaustiveness of a set of features.

4 IMPLEMENTATION STRATEGY

As digital geographic data has not usually contained details of data quality explicitly within its structure or in associated documentation, there is a requirement to formulate an implementation and management strategy to incorporate this form of information. Such a strategy needs to take into consideration the diversity of forms and formats currently in existence as well as the sheer magnitude of the task if fine detail is required immediately for all data assets (not only from within Defence but also the wider community).

An implementation strategy is complex and involves knowledge of digital data requirements, production and acquisition priorities, and coordination through a number of ADF organisations. It is therefore the subject of another study. A strategy would include, however, a number of steps:

- Compilation of a register of digital data assets of defence and civilian agencies;
- Assembling an overall summary rating of the data sets; and
- Producing detailed descriptions for the data sets.

Any implementation plan, however, involves a 'cost'. But such a 'cost' should not only be considered in terms of dollars and manhours, it should also be evaluated against benefits to Defence systems. As technology evolves, future weapons systems, navigation systems, command and control, targeting, and intelligence systems will become 'smarter'; and the 'smarter' the systems become the more reliance there will be on the data on which they base their 'decisions'. This means that the systems will require detailed knowledge of the 'quality' and reliability of the data (similar to those discussed in the ICA Data Quality Proposal).

In the meantime, there are in excess of thirty separate projects (that need to access digital geographic data in one form or another) being staffed in the Forces Executive, Navy, Army and Air Force acquisition programs. It, therefore, seems appropriate to commence the implementation process of applying 'data quality labels' to existing data sets and those in current production and to guidelines compatible with our Defence partners. For example, a number of systems (such as the F/A-18 Mission Data Planning Facility, Electronic Chart Display and Information System, Mine Warfare Systems Centre Information System, Australian Army Tactical Command Support System, and Operational Movements Planning System) require digital feature data for a range of analyses, and it seems appropriate to format these data and include 'quality statements' that are being introduced as MILITARY STANDARDS by other ABCA organisations. As such, the VPF Data Quality Statement should be used as Stage One of an implementation strategy.

REFERENCES

Abler, R.F. — "The National Science Foundation National Center for Geographic Information and Analysis", **International Journal of Geographical Information Systems**, 1,1987.

DMA — **Vector Product Format - Draft**, Prepared for Defense Mapping Agency Systems Center by Environmental Systems Research, Institute, Inc., dated November 1990, [US] MIL-STD-600006

DMA — **DCW Product Specification - Interim**, Prepared for Defense Mapping Agency Systems Center by Environmental Systems Research, Institute, Inc., dated April 1991, [US] MIL-D-89009

Goodchild, Michael F. — "Keynote Address", **Symposium on Spatial Database Accuracy**, Department of Surveying and Land Information, The University of Melbourne, June 19-20, 1991.

ICA — "Guidelines for Quality Evaluation of Spatial Databases", Draft report by Scientific Advisory Board, International Cartographic Association dated December 19,1990.

Moellering, H. (ed) — **Issues in Digital Cartographic Standards**, Report #7, National Committee on Digital Cartographic Data Standards [US], The Ohio State University, OH., 1986.

DISTRIBUTION

	Copies
Department of Defence	
Director General, Communications and Information Systems, HQADF	1
Hydrographer, Royal Australian Navy	1
Director of Survey-Army	1
Directorate of Operational Support Services, Air Force	1
Director Geographic Information, Defence Intelligence Organisation	1
Director Intelligence Development, Defence Intelligence Organisation	1
Scientific Adviser, Defence Intelligence Organisation	1
Defence Science and Technology Organisation	
Chief Defence Scientist)	
Central Office Executive)	1
Counsellor, Defence Science, London	Cnt Sht
Counsellor, Defence Science, Washington	Cnt Sht
Scientific Adviser, Defence Central	1
Naval Scientific Adviser	1
Air Force Scientific Adviser	1
Scientific Adviser, Army	1
Electronics Research Laboratory	
Director	1
Chief, Information Technology Division	1
Chief, Communications Division	1
Chief, Electronic Warfare Division	1
Chief, Guided Weapons Division	1
Research Leader, Command and Control and Intelligence Systems	1
Research Leader, Military Computing Systems	1
Research Leader, Human Computer Interaction	1
Head, Command Support Systems Group	1
Head, Image Information Group	1
Head, Information Acquistion and Processing Group	1
Head, Information Management Group	1
Head, Systems Simulation and Assessment Group	1
Head, Exercise Analysis Group	1
PRS Intelligence Development	1
Head, C3I Systems Engineering Group	1
Head, Software Engineering Group	1
Head, Trusted Computer Systems Group	1
Head, Computer Systems Archirecture Group	1
R. Williams	9
Publications and Component Support Officer	1
Media Services	1

Libraries and Information Services	
Australian Government Publishing Service	1
Defence Central Library, Technical Reports Centre	1
Manager, Documentation Exchange Centre (for retention)	1
National Technical Information Service, United States	2
Defence Research Information Centre, United Kingdom	2
Director Scientific Information Services, Canada	1
Ministry of Defence, New Zealand	1
National Library of Australia	1
Defence Science and Technology Organisation Salisbury, Research Library	2
Librarian Defence Signals Directorate, Melbourne	1
British Library Document Supply Centre	1
Spares	
Defence Science and Technology Organisation Salisbury, Research Library	6

Department of Defence

DOCUMENT CONTROL DATA SHEET

Page Classification: UNCLASSIFIED

Privacy Marking/Caveat (of Document): N/A

1a. AR Number	1b. Establishment Number	2. Document Date	3. Task Number
AR-006-971	ERL-0632-RN	JULY 1992	

4. Title

DATA QUALITY STATEMENTS FOR SPATIAL DATABASES

5. Security Classification

Document	Title	Abstract
U	U	U

S (Secret) C (Confi) R (Rest) U (Unclass)

* For UNCLASSIFIED docs with a secondary distribution LIMITATION, use (L) in document box.

6. No. of Pages 24

7. No. of Refs. 6

8. Author(s)

R.J. Williams

9. Downgrading/Delimiting Instructions

N/A

10a. Corporate Author and Address

Electronics Research Laboratory
PO Box 1500
SALISBURY SA 5108

10b. Task Sponsor

11. Officer/Position responsible for

Security.......N/A.......

Downgrading.......N/A.......

Approval for Release.......N/A.......

12. Secondary Distribution of this Document

APPROVED FOR PUBLIC RELEASE

Any enquiries outside stated limitations should be referred through DSTIC, Defence Information Services, Department of Defence, Anzac Park West, Canberra, ACT 2600.

13a. Deliberate Announcement

No limitation

13b. Casual Announcement (for citation in other documents)

- [√] No Limitation
- [ ] Ref. by Author , Doc No. and date only.

14. DEFTEST Descriptors

Spatial data, Quality, Geographic information systems, Digital image processing

15. DISCAT Subject Codes

0802, 120902

16. Abstract

This paper overviews contemporary issues in incorporating data quality statements into spatial databases. The paper includes discussion of two approaches; one emanating from the Digital Chart of the World Project and one through a working party within the International Cartographic Association.

Page Classification

UNCLASSIFIED

16. Abstract (CONT.)		
17. Imprint Electronics Research Laboratory PO Box 1500 SALISBURY SA 5108		
18. Document Series and Number ERL-0632-RN	19. Cost Code	20. Type of Report and Period Covered ERL RESEARCH NOTE
21. Computer Programs Used N/A		
22. Establishment File Reference(s) N/A		
23. Additional information (if required)		