

## **DATA QUALITY - A MANAGEMENT PHILOSOPHY**

GM Kennedy-Smith  
Military Survey

Systems & Techniques Unit RE  
Elmwood Avenue, Feltham  
Middlesex, TW13 7AE, UK

### **ABSTRACT**

Military Survey is developing a 2nd generation digital production system known as the Military Survey Geographic Database. The system will use multi-product data sets. In such an environment it is essential that data integrity is maintained to the highest possible standards. To meet this requirement, Military Survey is developing a concept known as the Source ID database. Although the origins of the concept are rooted in data integrity, it is proposed that the Source ID database is developed as a tool for the management of both data integrity and quality. The Moellering Committee recommendations for data quality have been adopted as the design criteria for the Geographic Database. This paper outlines the status of development and describes proposals for an implementation of the Moellering recommendations.

### **INTRODUCTION**

Military Survey is a production organisation. It is responsible for the provision of geographical materials to the Services. These materials include conventional printed maps and charts, digital products and digital data and are produced using either in house production resources, external commercial contracts, or are acquired through exchange agreements with other NATO nations. Since the early 1970s, Military Survey has been involved in the production and supply of digital geographical materials. The military requirement for digital geographical materials is expanding and Military Survey is developing a 2nd generation digital production system to meet the demand. Data quality is a key aspect of the development. This paper describes how the adoption of a prototype development strategy has affected development, and reviews current plans for the management of data quality.

### **DATA STRUCTURES STUDY**

#### **Background.**

In 1984, the Director of Military Survey established a data structures study. The decision to form this study was taken amid moves within NATO to develop a more structured format for the exchange of digital geographic data. The aim of the study was to identify a structure to support digital geographic data production in Military Survey in the 1990s.

Economy of production was a key requirement. The first phase of the study included a review of requirements, an analysis of the lessons learnt in production over the previous decade and a review of data structure developments. Two factors were clear in these early stages; the first was the need to develop a system based on an edge node structure - for exchange compatibility with other NATO map production agencies; the second was the adoption of a prototype development strategy to test and evaluate system requirements. It was recognised that the requirement for data quality was a key factor in the overall system design. At that stage of development neither the requirement for data quality nor the mechanisms to support it were fully understood.

To improve economy of production it was decided to test and evaluate the concept of multi-product operations (MPO). In its simplest form this requires the design and development of a data set capable of supporting production of two or more products. The production and maintenance costs for a multi-product data set should be less than the sum of the costs for all the component products. The concept of MPO is simple; but it needs to be treated with caution if the maintenance overheads are not to outweigh the advantages. The trials of the MPO concept were to illustrate the need to maintain data integrity.

From experience of the Military Survey 1st generation systems (Howman 1983), it was known that a multi-product data set should be maintained by "continuous" revision. If Military Survey was to embark on an expansion of the MPO philosophy, then it would be necessary that data integrity was maintained. For example, a data set that had taken many man years of effort to produce would be a valued asset. If something were to go wrong that might cast into doubt the integrity of that data, and the producer were without a system of recovery, then the loss would be catastrophic. In such circumstances the availability of an audit trail to trace the origin of all erroneous data would be essential. Many things could cause such problems; software errors, procedural errors, equipment faults, or operator errors. Maintenance of data integrity was identified as an essential design requirement (to protect the investment in data capture). To support this requirement, the concept of a Source Identification (ID) database was established (see below).

#### Moellering Committee Recommendations.

The final phase of the data structures study, the preparation of the final report, began in Apr 85. About this time, a copy of the 6th Moellering Committee report was received (Moellering, 1985). The data quality section of that report was of particular interest. There were two reasons for this; first, the report was considered an effective and complete statement of the requirement for data quality; and second, and probably the catalyst that awakened

our interest, was the recognition that the Source ID concept had many similarities with the Moellering concept for a "lineage code". Although the origins of Source ID were rooted in data integrity the approach was similar and it was decided that the Moellering Committee recommendations for data quality be incorporated in the data structures study final report.

#### Data Structures Study Recommendations.

The data structures study final report recommended that:

1. A 2nd generation digital geographic data production system be developed. The system was to be based on an edge node data structure and was to be known as the Military Survey Geographic Database (Geo DB).
2. The Moellering Committee recommendations on data quality be adopted as the design criteria for the Military Survey Geo DB system.
3. The Source ID concept be developed as a tool for the management of both data integrity and quality.

#### **MILITARY SURVEY GEOGRAPHIC DATABASE DEVELOPMENT**

The Military Survey Geo DB development was started in Dec 85 and will run until Oct 87. By Oct 87 it is planned that the full technical specification will be completed. The development will include a detailed assessment of the requirement for, and operation of, data quality within the production environment. The prototype development philosophy adopted during the data structures study will be maintained. The remainder of this paper describes the proposed Source ID concept and illustrates how it will be used within the production environment.

#### **PRODUCTION SYSTEM**

To understand the basics of the proposed system, it is necessary to outline the 5 stages of production recommended for the Military Survey Geo DB. These are:

1. Requirements Assessment. Military user requirements (national and inter-national) are assessed and grouped into families of requirements to form the most cost effective data sets.
2. Source Materials Acquisition. The most effective source materials to support the specified data set are identified and acquired. The available source materials are collated into a Source Material(s) File (SMF) and delivered to the production group with recommendations for use of the material.
3. Data Capture. The data capture process includes the analysis, capture and edit of the specified data (set).

4. Storage and Maintenance. This includes the management and maintenance of the specified data set.

5. Product Generation. Product generation includes the process by which data is extracted from the (multi-product) data set and transformed to meet the specification of the required product.

It should be noted that many data sets may be maintained within the Military Survey Geo DB. As stated above it is considered essential that the quality and integrity of all data is maintained. The basic tool for the management of this requirement is the Source ID database.

### **SOURCE IDENTIFICATION DATABASE**

The Source Identification (ID) concept was developed to support the data integrity problem. As such it is necessary that an audit trail be provided to trace all data within the database to the level of individual attribute values or in the case of spatial data to the most fundamental data element, the edge or node. With attribute data, it is proposed that all attribute values are described as follows:

**Attribute Value** - the means to interpret the attribute value will be defined.

**Confidence Factor** - a coded statement of the analyst's confidence in the interpretation or accuracy of the attribute value.

**Security Tag** - a coded tag denoting the (military) security level of the attribute value.

**Source ID Code** - a unique key to a record within the Source ID database.

The Source ID code is recorded at data capture and will only be altered during subsequent revision.

An example of information that might be held within a Source ID record is shown at Figure 1. Prior to data capture, details of the Source Materials File would be entered in the Source ID database. The record is allocated a unique Source ID number when it is first opened. At data capture, when the operator logs on, the system will "know" the date, equipment and operator. In addition, the system will require input of the task and Source Materials File numbers. These will be validated. Thereafter all data recorded will be tagged automatically with the appropriate Source ID Code (Source ID number + Suffix) - see figure 1. The Source ID code may be considered synonymous with the Moellering concept of a "lineage code".

In practice it is proposed to extend the Source ID concept further, to record for example the data set specification,

## SOURCE ID RECORD

<b>&lt;Source ID Number&gt;</b>	
<b>Source Material File (SMF) number and date</b>	
<b>Description of source material (from SMF)</b> - names - scale - classification - dates - accuracy assessments <b>and/or cross references to other files/databases</b>	
<b>Ancillary information used</b>	
<b>Control information used</b>	
<b>Transformations applied to source data etc.</b>	
01	<b>Captured by:</b> <b>On date:</b> <b>Equipment used:</b> <b>Task Number:</b> <b>Checked by:</b> <b>Entered date:</b>
02	
03	
04	
05	

↑  
 MAIN RECORD  
 ↓

↑  
 DATA CAPTURE RECORDS  
 ↓

"Suffixes"  
 ↙  
 ↘

**Note:**

The SID code used to access this record comprises:

<Source ID Number> <Suffix>

Figure 1

product specification, product history and transformation software used (including the spheroid and datum values). Version data will be recorded where appropriate.

It would be difficult and limiting to restrict the Source ID database to a concept of fixed format records. The Source ID record may be considered to contain free format text. Current proposals intend the use of hardware search devices to support a powerful and flexible search capability. Real time responses to on line queries are not envisaged. The system is designed to provide an extended audit trail capability to support a wide range of queries. In this role it is only expected to be used at specific stages of the production process. Examples of queries might include:

Plot all changes since <date>.

List all Source ID codes generated on equipment <number> between <date> and <date>.

List all Source ID codes for Source ID records containing <source material>.

Downgrade all attribute security classifications with <Source ID code> from <classification> to <classification>.

## **DATA QUALITY**

### Introduction.

It is surprising just how contentious data quality is and how difficult it is to reach agreement on its definition. Much time has been spent in discussion of this most elusive of subjects; discussion of terminology; of such terms as completeness, currency, horizontal and vertical accuracy (relative and absolute), of precision, and standard error. Perhaps this is the appeal of Moellering; the concept that "truth in labelling", properly applied, can provide the recipient with the information needed to determine "fitness for use". It permits the pragmatic to precede the numeric. It provides flexibility and permits the user an achievable approach to data quality that may, in the fulness of time, be supplemented by a more rigorous approach using "fixed sets of numerical thresholds".

The 6th Moellering Committee report proposes standards for data quality. No examples of an implementation of these proposals are known to the author. The producer must therefore derive and agree a philosophy; then apply that philosophy to production. In 1983, the major military map production agencies established the Digital Geographic Information Working Group (DGIWG) to develop military standards for the exchange of digital geographic data. Among its many tasks, this group is assessing the Moellering proposals. Some items referred to in this paper are under discussion by the DGIWG. In the remainder of this section,

some ideas for implementation are discussed. The 5 stages of production are used as the basis to describe the proposed implementation.

### Proposed Implementation.

1. Requirements Assessment. Development should be requirements driven. In general this simple statement is true; most production agencies have limited resources and production work is restricted to those requirements for which there is most demand. As soon as the producer has identified a requirement he will assess the best means to meet it. This may involve the extension of an existing data set, or the development of a new set to meet the required demand. This process can entail considerable work. For example, the requirement must be mapped to the feature coding system, validation rules must be established, and for all but the simplest tasks, a sample area must be tested to prove the proposed solution. It is proposed that the Source ID database will be used to record the data set specification, data capture specification and validation rules.

2. Source Materials Acquisition. Following agreement of the data set specification, a work plan and schedule will be required. Suitable source materials will be identified which are capable of supporting the data set specification and accuracy requirements. Details of the source materials will be recorded on the Source ID database and used during data capture to provide the "lineage" of the data.

3. Data Capture. Data capture is a key process. In the proposed system it is planned that only "clean" validated data will be stored. To attain this, a judical mix of procedural and automated methods must be developed. Some of the issues under discussion are listed below:

Spatial Data. The Military Survey Geo DB will handle vector data in an edge node structure. Topological relationships will be used to test for connectivity, adjacency and containment. Logical "errors" in the data will be flagged. It is intended that the topology will be used to support checks for logical consistency.

Attribute Data. During data capture, the analyst will be supported by a system controlled input - to ensure valid and complete entry of data. The DGIWG International Feature Attribute Coding Catalogue (IFACC) will be used to ensure the fidelity of encoded data. This is essential if coding ambiguities are to be avoided (particularly important for international military users). The IFACC system is based on a catalogue of international military feature and attribute codes. The catalogue will contain:

A unique feature/attribute code.

A definition including - the meaning of the code.

- the application of the code.
- the unit(s) of measure.

Each national (military) map production organisation will allocate a national descriptor for each code. The accuracy with which the feature and attribute codes are applied will be defined by the data set or product specification.

Data Organisation. A data set may contain several layers of data. To maintain the logical consistency of the data set it is necessary to maintain inter-layer consistency. Where the data set is split into "tiles" it will also be necessary to edge check all data. Logical errors will again be flagged.

Quality Control. The Moellering recommendations for testing positional and attribute accuracy are being assessed. Agreed standards will need to be implemented. The recommended methods are:

For positional accuracy and continuously variable attributes:

- Deductive estimates
- Internal evidence
- Comparison to source
- Independent source of higher accuracy

For categorical attribute values:

- Deductive estimates
- Tests based on independent point samples \*
- Tests based on polygon overlays \*

\* These tests include the use of a mis-classification matrix.

Production Standards. All data will be held in geographical coordinates to a specified precision, datum and spheroid. Production standards, for example, for feature codes, software and query languages will aid operator familiarity and minimise operator error.

4. Storage and Maintenance. All the above data capture criteria apply equally to maintenance. In addition, software may be required to maintain the integrity of unique feature numbers, seed coordinates or feature IDs. The precise requirements are still to be determined (see "delta" file). Features deleted during maintenance will be retained and flagged as deleted records.

5. Product Generation. During product generation the data will be extracted and transformed to meet product specification requirements. Many changes may be made to the data; the history of these changes comprise the lineage of the product. This history must be recorded and may be

required should the "legality" of the data be called into doubt. Product data must be tested for completeness against the product specification. Each product should be supported by validation software, and prior to release must meet the specified criteria.

#### Other Issues.

"Delta File." Digital geographic data is, on its own, of limited value. It is the addition and integration of user data, referenced to the geographic data, that gives the system its purpose (and value). The methods used to "hook" such user data to the geographic data are not standard. Some users may use feature numbers and others, seed coordinates. It is important that the integrity of these "hooks" is considered throughout the production process. The provision of a "delta file", listing all differences between one edition and the next, is still under discussion. The "delta file" would help the user to maintain compatibility between editions of geographic data.

System Access. The Military Survey Geo DB will operate within a secured environment. Access to the data will be controlled.

Level of Recording. The Source ID database will be used to record a wide variety of information. It is proposed that the lineage, positional and attribute accuracy records are related to individual items of data. Logical consistency and completeness records are more likely to be used globally to describe sets of data (eg by "tile" or "window").

#### Moellering Recommendations.

If the Moellering Committee recommendations for lineage, positional accuracy, attribute accuracy, logical consistency and completeness are to be managed successfully, then it is necessary that the requirements be integrated within the overall production system design. The Military Survey Geo DB uses the Source ID database as a tool for management of data quality. As noted in the preceding paragraphs, a full implementation will require careful integration of both automated and procedural methods.

### **GEO DB DEVELOPMENT PLAN**

The Military Survey Geo DB has been developed using past experience and the results of the data structures study. The Geo DB development began in December 1985. During the development period, it is planned to run a pilot study to test and evaluate the full production system. The Pilot Study will include an evaluation of the requirements for data quality. The cost and overheads of a full implementation will be assessed during this trial. The Geo DB development is planned to be completed by October 1987.

## CONCLUSIONS

Data quality is a key requirement that must be considered at the earliest stage of system design. This is particularly true for all digital geographic production systems.

Military Survey has adopted the Moellering Committee recommendations for data quality as the design criteria for the Military Survey Geographic Database. There were two reasons for this; first, the Moellering recommendations were considered a sound and complete statement of the requirement for data quality; and second, there were many similarities between the Military Survey Source ID concept and the Moellering concept of a "lineage code".

It is proposed that the Military Survey Geographic Database, a 2nd generation digital production system, be implemented by 1988/89. Current developments are seeking to establish a practical implementation of the Moellering recommendations. In the proposed system it is considered that both automated and procedural methods will be used. One of the "tools" under development to support this requirement is the Source ID database. Much work is still required to finalise the design, however plans are in hand to test and evaluate a pilot implementation during 1986/87. The results of this work will be used to define the specification for data quality management within the proposed Military Survey Geographic Database.

## ACKNOWLEDGEMENTS

This work was undertaken when in the employment of Military Survey (UK MOD). The views expressed are however those of the author and do not necessarily represent those of Military Survey.

## REFERENCES

- Moellering, H. (1985) Digital Cartographic Exchange Standards - An Interim Proposed Standard for Digital Cartographic Data Quality. Report No 6. Ohio State University, USA.
- Howman, C. (1983) The Production of Automated Charts of Europe (PACE) - Conference of Commonwealth Surveyors.