4 DATA QUALITY-ASSURANCE

4.1 Overview

Data quality is a relative term, for which there are no absolute measures. In practice, data quality is a measure of the **fitness for use** of a dataset for a specific purpose, and cannot be determined before that purpose is known. For example, a topographic map at a scale of 1:500,000 might be considered 'high quality' for national-level planning purposes, but 'poor quality' for local planning. Thus, the quality of a dataset is clearly affected by its accuracy and validity, but is not necessarily defined by it.

The complexity of natural phenomena means that many environmental measurements are uncertain or subject to error. For example, it is inevitable that some species will be mis-identified in a large-scale biological inventory, even if the highest professional standards are employed. Similarly, the inference of vegetation categories from remotely-sensed satellite imagery will never be 100 percent accurate. Such uncertainties may or may not be a cause for concern, depending on the intended use of the data. Box 2 distinguishes three common forms of deficiency in environmental datasets which may affect data quality.

Recognising that most environmental datasets contain deficiencies, it is vital for custodians to pass on an understanding of these when a dataset is distributed for external use — otherwise users may not be able to derive the maximum benefit from it. Clearly, a description of known deficiencies is only one item of information required by users to employ the dataset fully and safely. Others issues to document relate to the accuracy of the data, the standards which have been followed, and the processing techniques which have been applied (see Section 4.5).

Procedures aiming to improve the quality of a dataset can be applied from the moment it is collected through to the time that it is distributed for use. These procedures, which are collectively known as **quality-assurance procedures**, are designed to satisfy the needs and expectations of users.

4.2 Quality-assurance Procedures

Quality-assurance refers to the overall process governing the quality of a product, from the time that it is originated to the time that it is used. In the present context, the

Box 2 Forms of deficiency in environmental datasets

- **Limitations**

  Limitations are structural deficiencies in a dataset which become clear when it is used for purposes other than originally intended. A good example is the use of a map with an inappropriate scale.

- **Uncertainties**

  Uncertainties are introduced when variables are measured against a non-objective standard, for instance when an area is classified as belonging to a particular habitat type which, itself, may be poorly defined.

- **Errors**

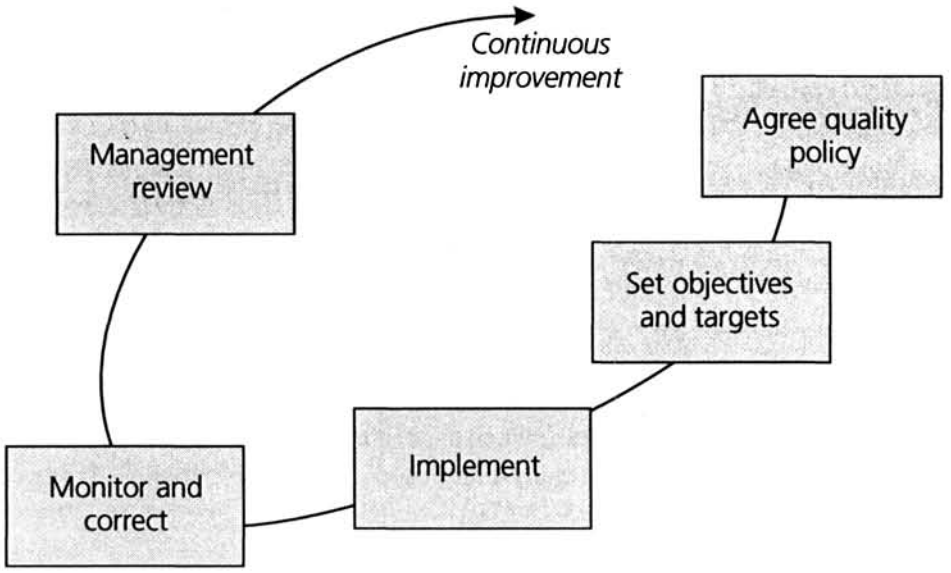
  Errors are introduced when variables are measured incorrectly against an objective standard, for instance when the depth of a lake is recorded with the digits in the number accidentally transposed, or with the wrong units (e.g. feet instead of metres).

process begins with data collection and ends with distribution of information to users. Quality-assurance procedures can be applied during all stages of this cycle. These include procedures to validate, maintain, document and secure data. It is the responsibility of custodians to ensure that these procedures are implemented in line with accepted standards and user demands (see Volume 5). Policies, judgements and decisions all depend on them doing so.

Within an organisation, quality-assurance procedures should be defined within a **quality policy** that is well understood by appropriate staff. The policy should set challenging objectives and targets for staff to achieve, such as specific levels of numerical or spatial accuracy in data collection, allowable error rates during validation, or consistent standards of documentation. The targets need to be consistently applied across the organisation and be measurable for monitoring and review. As well as internal review, organisations should also seek feedback from users of its products and services. The combination of internal and external reviews

allows the organisation to correct deficiencies in data quality and continuously improve its quality-assurance procedures. Figure 1 illustrates the essential steps of the quality-improvement process (adapted from BSI 1994).

Figure 1 Quality-improvement loop



4.3 Validation

Uncertainties and errors are introduced into a dataset in the natural course of data collection. The aim of validation is to eliminate these completely or reduce them to a background level where they do not interfere with the use of the data. Validation can be a labour-intensive and tedious task, but it is nevertheless a critical quality-assurance procedure. Key activities include:

- testing the accuracy and reliability of data prior to storage; and
- introduction of tools and methods to regulate data entry.

Basic tests should be run on data items before they are permanently stored (e.g. before new data items are added to existing datasets). These enable suspect or unusual data items to be identified and brought to the attention of experts for independent assessment. Box 3 describes some basic tests applied to species distribution records prior to inclusion in a large national dataset in Australia (Chapman and Busby 1995). Another good example of the expert assessment process is the validation of bird distribution records in East Africa. Here, national experts validate the vast majority of bird distribution records generated by field survey activities, but very unusual records are processed at the regional level by the Ornithological Sub-committee of the East Africa Natural History Society (Reynolds *et al.* In press).

Box 3 Example validation procedures for species dataset

- Records checked to see that all required data fields are present.
- Scientific names checked for validity.
- Grid references of terrestrial species checked for being over land, not water.
- Presence of a species in a certain location tested against a prediction based on bioclimatic factors, and outliers selected for further investigation.

Errors can be introduced into a dataset when it is stored, for instance in a computer. Common errors include the entry of incorrect numbers into a spreadsheet or incorrect boundaries into a map. As an illustration, take the entry of species data into a computer database. Suppose that a particular data entry screen has 10 fields (e.g. family, genus, species, common name, threat category, etc.), each taking, on average, 8 characters to fill. If the success rate of the typist is 99 percent, then the probability of the whole screen being completed correctly is, surprisingly, only 45 percent.[3]

3 If the probability of a single character being typed correctly is 99 percent (0.99), then the probability of 10 fields, each with 8 characters, being typed correctly is $0.99^{(10x8)} = 0.45$, which is 45 percent.

Such errors result largely from lack of care and attention by human operators, and training will help to reduce these. However, they can be reduced even more effectively through the introduction of tools and methods to regulate data entry. These promote consistency and enable operators to identify errors at the earliest detectable moment, so that they do not propagate or become buried in large volumes of other data.

A key feature is automatic validation, which involves performing 'reasonableness' checks on data items as they are entered, such as the geographic feasibility of a grid reference or the physical possibility of a particular measurement. Unreasonable values (e.g. a land-based animal observed at sea) can then be reported to the data entry operator, who can correct simple mistakes or seek expert advice as required. Even more effective at reducing errors are tools which allow the operator to select values from a set of pre-defined choices, eliminating the possibility of typographic errors completely. Automatic validation is especially useful in situations where consistency of data entry cannot be guaranteed, for instance when data are entered into large datasets by many different staff.

4.4 Maintenance

Most datasets become obsolete if they are left unmanaged for long periods of time. Measuring techniques may be improved, leading to more accurate and reliable data collection; new standards may be agreed, meaning that old structures and assumptions are no longer acceptable; and new formats, media and technologies may be evolved to manage data more efficiently. Unless a dataset is actively maintained, it may simply be overtaken by events leading to a gradual reduction in its usefulness. Key activities include:

- keeping it up to date;
- making sure it is kept abreast of significant standards; and
- adapting its structure, format and storage medium in line with user's needs.

Keeping data up to date involves establishing a routine for continuous, or at least regular, **enrichment of a dataset with new data**. Many projects fail to take account of this, with datasets being created to serve only immediate project objectives, rather than long-term capacity needs. This is inefficient, since new projects may have to build similar datasets from scratch. One of the distinguishing characteristics of a

professionally-managed dataset is that it is maintained not only for immediate uses, but also for other applications — now or in the future — which could potentially benefit. As with other strategic approaches, this can create a funding challenge in the short term.

Earlier sections revealed the importance of data standards. These also evolve over time as new opportunities for standardisation are created through information networks and individual partnerships between organisations. Where relevant standards exist, they may be applied to datasets in order to ensure consistency and reduce transaction costs; where they evolve, datasets should evolve with them to maintain these advantages.

Over time, increasing numbers of users may apply a dataset to their tasks. Feedback from users, for instance their impressions of the strengths, weaknesses and overall usefulness of the dataset, can be used to adapt the structure, format and medium in which it is made accessible. Note that the dataset itself can be **managed** in whatever form is discovered to be most efficient by the custodian, but it should be made **accessible** in the form which is most acceptable to users (see Section 3.2).

The opportunities created by rapidly-changing information technologies, storage media and low-cost communications, impose a continuous challenge on those attempting to maintain datasets. However, it is far more important for data managers to maintain the content of their data than worry about keeping up with the latest technology; from a user's perspective, all that is required is a simple and cheap source of quality-assured data.

4.5 Documentation

When a dataset is released to an external user, knowledge of its limitations, uncertainties and errors is lost unless this understanding is passed on in the form of documentation. As well as knowledge of its deficiencies, users may require a host of other items of information in order to employ the dataset fully and safely.

In the past, custodians rarely devoted much attention to documenting their datasets. This was because the latter were usually built for one specific project by people who well understood the nature of the data, including its deficiencies. At the end of the project the data were archived, filed or neglected. Today, however, datasets may be used many times for many purposes, and documentation is regarded as a

strategic asset enabling custodians to maximise the value they derive from a specific data source. One of the driving forces of this change is the growth of information networks which depend on organisations being granted simple and cost-effective access to data.

In summary, custodians document their datasets for two important reasons:

- to increase internal effectiveness by clarifying the function and quality of their datasets; and

- to facilitate use of their data by others.

Box 4 lists some potential aspects of a dataset to document. The fundamental principle to follow is **truth in labelling**. This means that the dataset should be exactly as described and of a quality which is suitable for its stated and implied uses. Assessments of the completeness and accuracy of documentation should be undertaken periodically, especially in the case of essential datasets (see Volume 3), preferably by an independent auditing team.

4.6 Data Security

A range of operational procedures are necessary to guarantee the security of a dataset. This applies whether or not data have been computerised. Indeed, if they are not in electronic form, then it may be considerably more difficult to manage them securely.

In general, threats to electronic data security tend to be greatest where the physical environment is hostile to computing equipment (e.g. extremes of temperature, high humidity or dust), where electronic interference is strong (e.g. in hospitals, industrial plants, locations near transmitters), where power supplies are uneven or unpredictable, and where informal and therefore virus-prone computer networks are the primary means of data transfer.

The most important requirement is to protect data from accidental erasure, which may occur due to human error in copying and reorganising files, updating records or other 'maintenance' procedures. Erasure may also occur due to mechanical failure of disk drives, or logical faults caused by power failures or fluctuations. Computer viruses also pose a threat to data security, although this is often greatly over-estimated (they certainly are a nuisance however).

Box 4 Aspects of a dataset to document

- Title/theme.
- Contact details of custodian.
- Intended/unwise/improper uses.
- Accuracy/resolution/scale.
- Data collection methodology (or original sources of data).
- Data structure/model.
- Data management standards followed.
- Processing and interpretation techniques applied.
- Known limitations, uncertainties and errors.
- Currency of data.
- Life expectancy (e.g. date of next update).
- Quality-assurance procedures applied.
- Quality targets.
- Access conditions/procedures/costs.
- Available formats and media.

Box 5 describes a number of protective measures which help to combat threats to data security. Such procedures can be elaborated within the overall quality policy of the organisation, or be prepared separately in the form of an operating manual. Specific plans to cope with emergencies should also be considered, for instance hardware malfunction, fire or theft. Organisations should accord a high profile to data security. On occasion, an entire project or programme has been forced to close due to loss of essential data. This occurred once in the South Pacific when a freak wave

struck the office of a custodian, eliminating its data. No copy of the data was maintained off-site.

Box 5 Procedures for protecting data

- Regular (daily, weekly and monthly) backup of all critical data on removable electronic media (magnetic tape or optical disk).
- Storage of backup media off-site (away from the workplace) in order to restore data after damage or theft of key equipment.
- Periodic test restoration of backed-up data to ensure that the procedure is effective.
- Periodic test recovery from simulated virus attack, hardware malfunction or other disaster.
- Regular virus-checking with up to date software.
- Avoidance of unlicensed or borrowed software, computer games or other personal software.
- Power regulation via the use of uninterruptable power supplies, surge protectors and radio interference filters.