3 DATA STANDARDS

3.1 Overview

Standards enable people to communicate with each other in recognisable ways: languages are a good example. In the present context, data standards refer to agreed methods of collecting, managing and accessing data amongst a group of organisations. In the same way that language standards enable more efficient (and cheaper) communication, data standards enable more efficient use of data. The chief advantages of data standards are as follows:

- **Lower transaction costs**

  If data are available in standard formats, based on standard collection methodologies, users can absorb them more easily into their work. However, if standards are not applied then data may be perceived as incompatible, inappropriately focused or otherwise unusable. In summary, **lower transaction costs** are associated with accessing and using data when they are managed according to recognised standards.

- **Comparison of results**

  Without agreement on data standards, organisations tend to employ their own methods of collecting and managing data which, due to differences, complicate integration of the data with other sources at a later stage. Even within an organisation, methods may be applied inconsistently by different groups, or at different points in time. Data standards overcome this problem by enabling **comparison of results** in space and time, and between different sources.[1]

Admittedly, reaching agreement on data standards is a time-consuming, largely intellectual, activity requiring concrete and determined action to succeed. However, there is no other realistic way of reducing transaction costs or maximising the value

1 This is particularly relevant to the study of natural phenomena which, due to their incremental nature, tend to reveal themselves over long periods of time.

of expensively produced data. Data management is a resource-intensive activity, and it can be disappointing to discover that otherwise well-managed data are unusable due to lack of standardisation. This can be avoided by building in conformity from the early stages of a data management project, with the intention of widening the potential range of uses of data which are developed.

3.2 Types of Standard

Standards may be applied to all aspects of data management, from data collection and storage, to quality-assurance and distribution. They define accepted formats, structures, systems and procedures for managing data. Mostly, they define only minimum requirements, allowing those following the standards to exceed the requirements as appropriate. For example, a standard method of recording species' distributions might require observers to provide a location, date and species name for each observation. Observers would be welcome, however, to record any number of other factors in addition to this minimum set. Some of the potential range of data standards is described below.

- **Collection**

  Recording/measuring techniques for specific themes (e.g. biological records, human impacts, policy performance, sustainability); classification systems (e.g. soils, vegetation, climate, species names); criteria for assessing threats to biological resources.

- **Storage**

  Core data models/database structures; storage formats and media; methods of data retrieval; use of information technology; maintenance procedures.

- **Quality-assurance**

  Validation, maintenance and security procedures; documentation formats.

- **Distribution**

  Product definitions (e.g. map keys, acknowledgements, symbols); reporting formats; data transfer formats (interchange standards); protocols for electronic communication of files.

It is not always feasible or even desirable for organisations to adopt every type of standard. They may have their own, highly effective ways of managing data which could become compromised, possibly disrupted, by the blanket introduction of new and unfamiliar standards. Where increased efficiency is unlikely to follow the introduction of standards, they should not necessarily be pursued.

There is one group of standards which will almost always bring efficiencies, at very little cost. These are **interchange standards**, whose purpose is to streamline the transfer of data. The introduction of interchange standards has very little impact on the way organisations collect and manage their data internally, but has a strong impact on data mobility. Interchange standards focus mainly on the formats and media in which data are transferred. Because they apply solely to the distribution of data, interchange standards are simpler and cheaper to implement than wider-ranging standards.

A number of interchange standards already exist for the transfer of biological and geographic data. For example, the International Working Group on Taxonomic Databases for Plant Sciences has developed an International Transfer Format for Botanic Garden Records (Hollis and Brummit 1992). Interchange standards also exist to regulate the transfer of spatial datasets, which are often highly complex due to the varied nature of the data.[2] Most of these are based on the formats developed by the manufacturers of geographic information systems (GIS), for example the export formats of ARC/INFO and AutoCAD software. Such standards are privately controlled (i.e. they are **proprietary**) and may not necessarily reflect the needs of data users.

Non-proprietary standards have been developed to address this concern, although they are not yet in widespread use. For example, the Spatial Data Transfer Standard

---

2 For example, raster data, vector data, three-dimensional data and attribute data.

(SDTS), which is coordinated and promoted by the United States Geological Survey (USGS), consists of specifications for the organisation and structure of digital data transfer, definitions of spatial features and attributes, and encoding instructions for data transfer (Wortman 1992). The SDTS was approved by the United States Department of Commerce in 1992 (NIST 1992), and has been adopted by several other countries.

3.3 Development of Standards

Information networks provide an opportunity to reconcile existing standards — and agree new ones — in the interests of mobilising data for collective goals. Their development can be facilitated by one or more technical teams, arranged by the network's hub, who are tasked with reviewing and agreeing standards covering essential themes, and for publishing them for use by the network's partners (see Volume 4). Data standards are so important to a network that they cannot be overlooked, taken for granted, nor left to specialists who do not fully represent the network's interests.

Recognising that progress towards formally accepted data standards can be slow, organisations often develop their own, interim standards. The latter, sometimes referred to as *de facto* standards, are commonplace across many scientific themes, often having arisen to suit particular data collection and management priorities. Wherever possible, interim standards should build on existing standards within their theme, rather than risk duplication. For example, international initiatives have so far proposed at least seventeen definitions of sustainable forest management, many of which could be translated into national standards for forest monitoring (WBCSD 1996).

As the profile of interim standards is raised, and increasing numbers of organisations begin to adopt them, they may be vetted by the organisations concerned and formalised through publication. A good example of this process is the East African Biodiversity Network, now in its seventh year, which has successfully developed biological recording standards at a regional level. The network, which brings together biodiversity professionals from Kenya, Tanzania and Uganda, was originally established "in response to the need for biologists and conservationists to develop compatible working systems, from the database formats themselves to having common lists of scientific names" (NMK 1995).

A number of working groups were set up to develop data standards appropriate to the region. Taxonomic working groups are developing checklists and other standards for birds, mammals, reptiles, amphibians, fish, aquatic invertebrates and plants. A working group on important bird areas is responsible for listing, prioritising and surveying key sites; other groups are developing a regional gazetteer, habitat classifications, database structures, and policies on data exchange and training.

At the international level, the International Working Group on Taxonomic Databases for Plant Sciences (TDWG) was established by the International Union of Biological Sciences in 1985 to explore standardisation and collaboration between major taxonomic databases (Hollis and Brummit 1992). The group brings together all the major working taxonomic databases into a loose confederation. Through a series of international workshops, TDWG has developed a number of standards including the International Transfer Format for Botanic Garden Records (ITF) and a World Geographical Scheme for Recording Plant Distributions. The latter provides four nested levels comprising continents, regions, countries and botanical recording units.