Defra/Environment Agency
Flood and Coastal Defence R&D Programme

Scientific Data Management by Project Consortia:
Best Practice Guidelines

R&D Technical Report FD2110/TR2

Defra / Environment Agency
Flood and Coastal Defence R&D Programme

Scientific Data Management by Project Consortia:
Best Practice Guidelines

R&D Technical Report FD2110/TR2

Authors:
Keiran Millard and Claire Brown

Dissemination Status
Internal: Released Internally
External: Released into Public Domain

Statement of use
This document is intended to provide guidance upon best practice for scientific data
management during the lifecycle of a project. It focuses on key data management issues
to consider during all phases of the project. These are illustrated through a series of
examples taken from the EMPHASYS and ERP Phase 1 Uptake projects, during which
a GIS database of environmental data for UK estuaries was developed and subsequently
updated.

Keywords – GIS, Data Management, Best Practice, Metadata, Estuaries, Estuaries
Database

- Name, address and contact details of the research contractor –

EXECUTIVE SUMMARY

Many estuary management projects require the collation of scientific data and this guide has been produced to assist organisations working on such projects. The need for such guidance was highlighted during the Estuary Research Programme Phase 1 (ERP1) EMPHASYS project completed in 2000 (Pye, 2000). Whilst collating data on the physical processes of British Estuaries, the EMPHASYS project reported the need to develop a standard framework for future projects. The aim of this guide is therefore to assist organisations both *commissioning* projects as well as those actually *undertaking the work* to have a reference describing what is involved in work of this kind.

Creating a dataset is more than an integration of data, it is also an integration of the policies and procedures of the organisations that own and manage the data. Accordingly the successful creation of a dataset demands much wider thought than what data should go into the dataset and what technology should be used for the database. Addressing legal and financial issues can be fundamental to a successful outcome. Furthermore, many such issues exhibit a dynamic nature which must also be acknowledged and appropriately considered. For example will the technology be valid in five years?

To enable project consortia to effectively understand this complex picture, this guide presents principles for data management for each stage in the lifecycle of a dataset. Good data management is implicit in knowing the lifecycle state of your data at any time, and these principles give pointers for action (a framework) as to what more detailed issues should taken into account. Examples are given from the experience of compiling the database as part of the EMPHASYS project and its subsequent update, together with 'key tips' for each data lifecycle stage. The content of the guide is kept to a generic level to prevent specific recommendations becoming obsolete.

CONTENTS

1. INTRODUCTION

1.1 Context For These Guidelines

This guide has been produced to assist organisations that are required to collate scientific data as part of an estuary management project. The aim is to provide both those commissioning the project and those actually undertaking the work with a reference manual to assist them through the process, outlining issues for consideration and providing examples and tips. This should help to ensure that experience previously gained 'the hard way' can be captured and utilised to guide others undertaking similar work. The guide should help other project teams to rapidly gain an appreciation of the task in hand, maximising the quality and value of the resultant dataset. The need for such guidance was highlighted during the ERP1, EMPHASYS project completed in 2001. Whilst collating data on the physical processes of British Estuaries, the EMPHASYS project reported the need to develop a standard framework for future projects. In particular, this was considered most relevant amidst concerns about the restrictions placed upon data exchange and the lack of general information on data availability and acquisition.

Within the UK, there are over 250 government agencies and local authorities with responsibilities related to estuarine and coastal management, all requiring data in one form or another. Private institutions such as dredging companies, port and harbour authorities; universities, research institutions and consulting engineers also undertake coastal data collection. However, it remains a difficult task to identify and acquire all of the potential sources of data for a given location. Many organisations with an interest in the estuarine management would accept that there are insufficient data readily available upon which to base management decisions. This is because existing data may be either confidential, of insufficient quality or stored in a format which is not easily accessible. There are currently no recognised standard procedures for the collection, processing, storage and distribution of data. Consequently most organisations have developed different strategies.

The restrictions that current data management practices place on data sharing are illustrated not only within projects such as EMPHASYS, but also by the implementation of Catchment, Shoreline, Estuary and Coastal Habitat Management Plans. The development of these integrated management initiatives could be greatly improved by scientific data exchange between responsible public and private authorities. In practice, however, this is not often achieved. Until recently, much of this scientific data was collected, used as required and then 'lost' either through deletion or ineffective archiving. There is now an increasing demand for effective data management that will not only improve the overall quality of data but also ensure that data remains available for future use.

Progress is being made to improve some of these data exchange issues. However, for the foreseeable future, managing data collection, processing, storage and distribution will not often be straightforward, particularly when involving numerous organisations. Hence this guidance document should prove to be of direct benefit to anyone involved in the collation and application of scientific data.

1.2   Data Use And Exchange

It is recognised that data sharing leads to data integration and hence more meaningful analysis and use. It also reduces the duplication of effort and is more cost effective. This is applicable to both individual organisations and consortia based projects. However, given the complex and varied range of uses of a particular data set, it is not always possible to foresee it's full potential value. All organisations anticipate a degree of value from the creation of a database[1]. However, this is not simple to achieve and requires a transparent and open approach.

Such an open and transparent approach was recently published by CIRIA in a document entitled "*Maximising the Use and Exchange of Coastal Data: A guide to best practice*" (CIRIA, 2000). This guide illustrated that data exchange and reuse is bounded by a combination of policy and technological barriers. Some of these barriers are within the control of organisations and some are not. Accordingly the guide illustrated how to manage these barriers from the perspective of maximising data value. The CIRIA guide introduced three main concepts to assist data management. These are the data processing chain (DPC), data lifecycle (DLC) and five management principles to govern them. These form the basis of the guidance in this report.

The data processing chain (DPC) is the most familiar to an organisation as this is the processing underpinning data (or database) creation. Within the data processing chain of a project, data are collected or sourced and integrated to produce information as required under the terms of the project contract. What is less familiar is how to manage the DPC in the context of the data lifecycle and this is fundamental in good data management practice.

1.3   Data Management And The Data Lifecycle

1.3.1   The Data Lifecycle

Consortia based projects operate within the frame of a project contract that has a defined lifecycle beginning with a signature on the project contract and ending with the client's sign-off. As such it is useful to focus this guidance on the lifecycle of the data that develops within the project lifecycle. The data lifecycle considers what happens to an individual data item, such as a measurement, from its creation. A particular dataset, at a particular time, will be in one of six states that comprise the data lifecycle. These states are:

- ❑   Creation
- ❑   Storage
- ❑   Access
- ❑   Update
- ❑   Retention
- ❑   Deletion

1.1.1   ───────────────

[1] This includes not only the organisations creating the dataset, but also the organisations that supply the data.

The order, duration and repetition of the states in which data exist at any one time will vary. However, the data life cycle always starts with 'creation' and always ends with 'deletion' and these two phases can only occur once[2]. It is important to note that whilst all stages of the lifecycle should be treated individually, consideration should be given to subsequent stages, particularly at the creation stage. Good data management is implicit in knowing the lifecycle state of your data at any time, and ensuring that deletion is only carried out as specified by the contract.
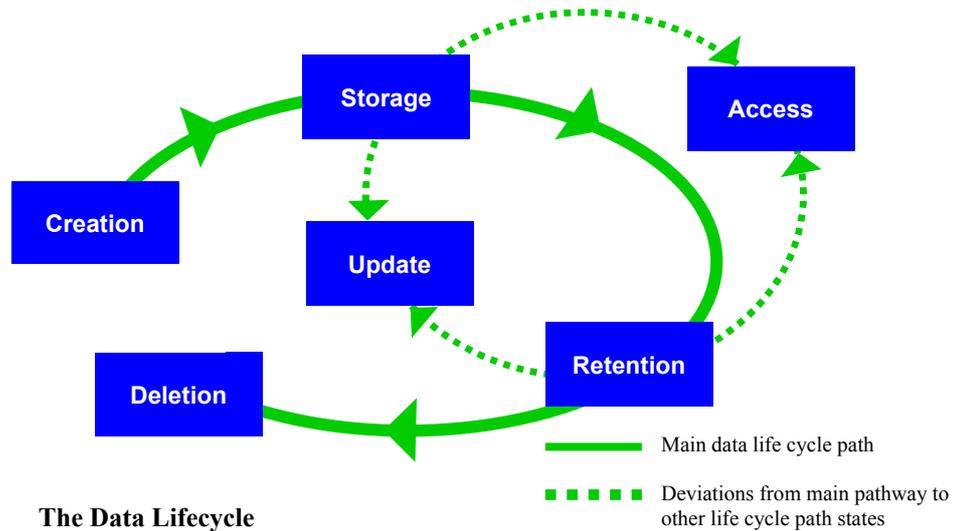
Figure 3.1   The Data Lifecycle

The data lifecycle may exist wholly within a single organisation, or distributed amongst many institutes. For example, a survey company may 'create' and 'store' the data, which are then passed on to a client for 'access' into their Data Processing Chain (DPC). The DPC is by definition very closely aligned with the project lifecycle and this usually clearly defines the first four elements of the data lifecycle. However, the 'retention' of the data sometimes falls outside a DPC and this can result in the 'deletion' of the data. This is illustrated in Figure 3.2. Indeed, one of the major issues raised during the EMPHASYS project is how to manage the data lifecycle beyond the project lifecycle.

1.1.1 _____

[2] Data can be duplicated, but then this is the creation of a new branch to the data lifecycle. The more identical data that is in existence the less chance that the data will be lost; although quality control becomes a greater problem.
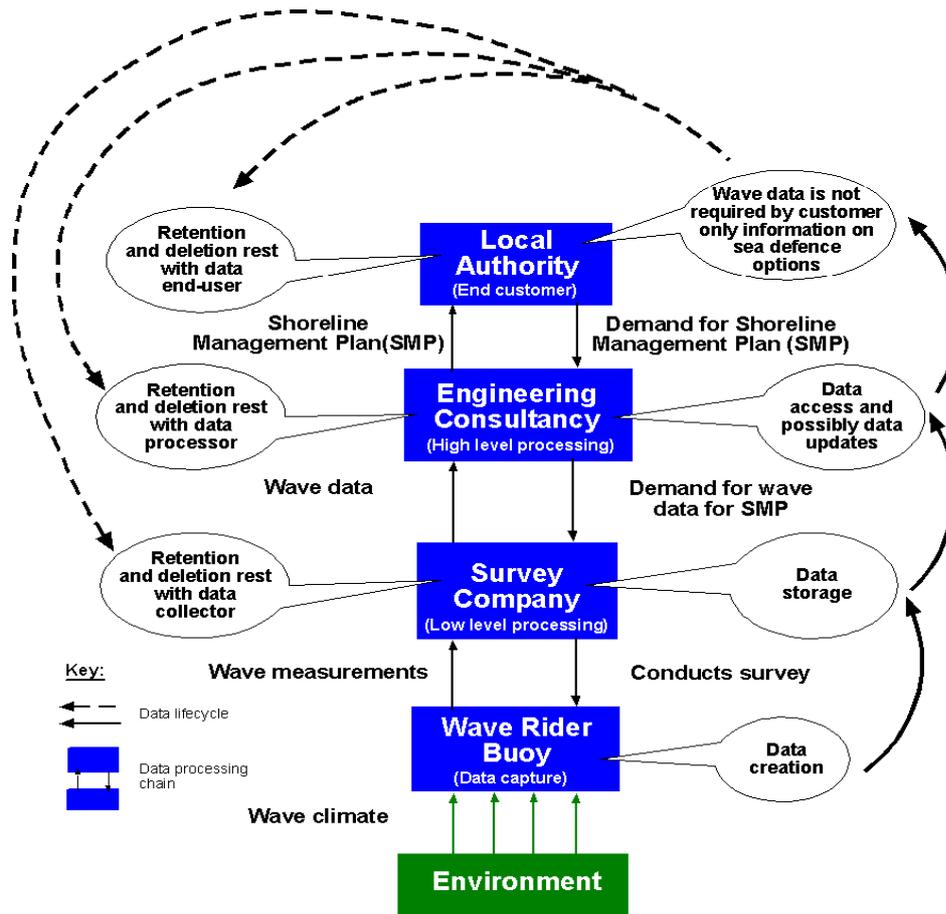
Figure 3.2 Data Lifecycle and Processing Chain (© CIRIA 2000)

1.3.2   Data Management Principles

Organisations operate different procedures and rules regarding data management. These often allocate roles and responsibilities to individuals in those organisations and define the DPC. However, when dealing across and between organisations at a strategic level it can be more effective to deal in terms of principles. Principles provide the flexibility to enable organisations to use their own procedures in order to meet best practice standards. The CIRIA guide recommended the adoption of five principles of good data management. These are based on principles adopted by the British Standards Institute for the management of electronic documents (Mayon-White and Dyer, 1997) and are essentially pointers as to how to manage the data lifecycle. These principles are technologically and politically independent, ensuring they will remain valid in the future. The five principles of good data management are shown below.

1.4   How To Use This Document

This document considers the five principles of good data management in the context of each stage of the data lifecycle. This forms a powerful and easily applicable approach to support data management decision-making (Dyer & Millard 2002). Each of the six main chapters of this report covers a stage of the data lifecycle and outlines the main issues that need to be considered, drawing on the aims and decisions taken within the

scope of the EMPHASYS project and the subsequent update of the database in ERP1 UPTAKE. Best practice data management for these issues is presented for each lifecycle stage in the form of the data management principles, i.e. "what should be done at each stage of the data lifecycle".

The Five Principles

Data Understanding	Roles and Responsibilities	Processes and Procedures	Enabling Technologies	Audit
Recognise, understand and describe all data used, needed and available	Understand legal and contractual issues and execute responsibilities.	Identify and specify organisational processes and procedures.	Identify & implement appropriate technologies for data management & processing	Audit and monitor processes for data use and exchange
Document and Describe	Understand and Execute	Identify and Specify	Identify and Implement	Audit and monitor

Figure 3.3   **Five Principles of Data Management (© Mayon-White & Dyer 1995 and 1997)**

2. CREATION

This is the first stage in the lifecycle of all data and is the point at which data management should begin if the full value of the information is to be realised. Data creation is governed primarily by the requirements of the client but is restricted by the technological capabilities of those undertaking the work. Figure 4.1 below identifies the issues that need to be considered at the creation stage of the data lifecycle
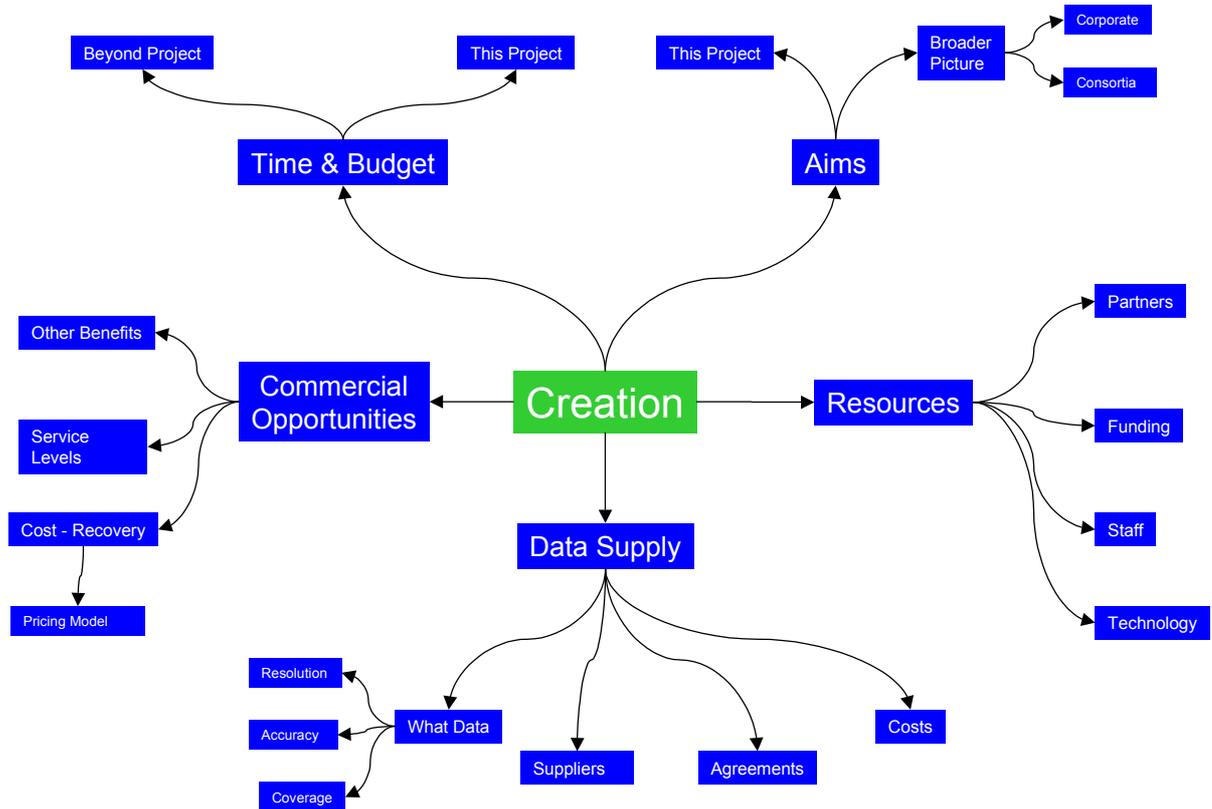
Figure 4.1 **'Mind Map' of Issues to Consider at the Creation Stage of the Data Lifecycle**

2.1 Data Supply

The most fundamental issue at this stage is knowing exactly what data is required to fulfil the aims of the project. This should be predetermined by the client but factors such as spatial and temporal resolution and accuracy must be established in order to apply good management practice. For example The EMPHASYS project aimed to create a database of environmental datasets that could be used to model morphological changes in estuaries. An exercise was carried out in order to examine the type and extent of data available for each major UK estuary, which led to 6 estuaries being chosen for inclusion in the database. An assessment of the resources required to collate the data and how the operation is to be funded is also necessary at this stage. Arrangements may be made to sell the data or as was the case in both EMPHASYS and ERP1 UPTAKE, the data suppliers received copies of the outputs in return for the use of the data. Whilst project budgets and timescales are generally fixed, data collection can be co-ordinated across

several different projects or data management issues extend beyond the life a specific project. Hence, the extent of the data lifecycle must also be considered at the creation stage.

2.2 Legal And Commercial Issues

The legal issues associated with data management must be considered from the outset i.e. who will own the copyright and who will have ultimate responsibility for each stage of the data lifecycle. The duration of responsibility should also be predetermined with the client to allow for adequate support and maintenance of the database during its expected lifecycle. It is essential at this stage to decide who will ultimately have access to the data. For example, are the data intended solely for use during a particular project or will it be available to external organisations at a later stage. If this is to be the case, what steps should be taken to allow for this at the creation stage? This issue is particularly pertinent if the dataset contains safeguards for sensitive information, to prevent unauthorised access, which may affect the distribution process later on. During the EMPHASYS project, the legal requirements were fairly straightforward as the data were only being used by members of the project team for their own research. As such, this simply involved the setting up of user licenses in order to govern use of the database by the consortium members. However, the placing of the data into the public domain when the database was updated for ERP1 UPTAKE meant the establishment of a legal framework became more complex. Given that different data suppliers often stipulate different agreements, it is important to establish with the client standard conditions of distribution and access at the creation phase.

2.3 Data Processing And Analysis

Having decided what data needs to be collected, the following issues must be considered. Firstly, what techniques have been used to collate the data? Suppliers should provide details of fieldwork programmes, literature studies etc but as occurred frequently during the EMPHASYS project, no such information was supplied. Secondly, the suppliers should also provide information about any processing or correctional procedures that have been applied to the raw data. In terms of the data lifecycle, it is essential that all analytical procedures are described and supplied with the data. It is also important to know whether any analysis was carried out by the data provider prior to treatment by the project team. Thirdly, it is advisable to standardise the analytical procedures prior to data collection, particularly if data from different sources is to be merged as different analysis techniques may result in data incompatibility. Recently, national (e.g. NGDF) and international (e.g. ISO19115) standards have been developed for the documentation of data. Although these were not available during the EMPHASYS project, such a standard developed for the Geographic Information Community was adopted for ERP1 UPTAKE. Finally, agreement on metadata standards is also crucial at this time in order to ensure that the data lifecycle is fully transparent.

2.4 Economical And Technological Issues

Consideration must be given to the cost of producing the data as a deliverable, particularly if the data is to be widely distributed. Given that technology is continually advancing, the data should be supplied in such a way that it does not become obsolete.

Additionally, there may be scope for updates to the data itself and some provision for this should be made at the creation stage. Having gathered a set of data, it should be processed to the format and delivered on the media agreed by the client. This effectively reduces the risk of technological obsolescence and minimises conversion time. The software/hardware used to display the data must also be easily transferable in order to alleviate compatibility problems. Where possible, the use of generic industry standard formats / languages such as those being developed by the Open GIS Consortium (e.g. Geography Markup Language, GML) will make data exchange simpler and less vulnerable to out-dated formats.

Creation

Data Understanding

- Clearly describe what information the database is going to convey and therefore what data needs to be included to provide this information.

- Document the attributes of the required parameters including required accuracy, resolution, spatial coverage and temporal coverage. This includes ensuring consensus of parameter definitions and units of measurement.

- Describe where the data is to come from, whether from a dedicated measurement campaign, a computer model or a third party supplier etc.

- Determine the spatial coordinate system to be used, including vertical datum. This should be specified by the organisation commissioning the work and ideally consistent with available base-mapping data

- Determine the temporal referencing system to be used, for example GMT or UTC.

Roles and Responsibilities

- Check all organisations and staff involved in the project are made aware of their contractual obligations. Ensure that funding covers the activities to be undertaken.

- Agree which organisations (both inside and outside the project) are responsible for data supply and identify the individual staff undertaki9ng the task. Establish data supply agreements where necessary.

- Consider future stages of the project lifecycle taking account of not only the project contract but also industry practice and 'the bigger picture' e.g. developments e-government.

Processes and Procedures

- Document the procedures to be used for populating the database (including data pre-processing) based on Data Understanding and Roles and Responsibilities. Ensure these are communicated to all relevant organisations to ensure data harmonisation.

- Agree a metadata standard and accompanying glossary for use in the project as a formal method for documenting the database, constituent data sets and any associated data processing. This will save time and effort at subsequent stages in the data lifecycle.

- Specify acceptance criteria for data to be included in the datasets. This may be based on accuracy, size, precision, format etc.

Enabling Technologies

- Select a suite of technologies that satisfies this stage, and the next three stages (storage, access and update), of the lifecycle. For example, if the data has a strong geo-spatial component then GIS is likely to be appropriate, or if distributed access is key then a web-based solution may be considered. However, take full account of industry practice to ensure technology does not limit unnecessarily database use, i.e. do potential users have web-access?

- Consider the validity of these technologies for future obligations during the retention phase of the lifecycle and re-appraise if necessary. Look to make the dataset as 'future proof' as possible.

Audit

- Determine how progress of the database is to be monitored to ensure assumptions made at the creation stage are correct. This may be based on 'data sets received by a given date' (particularly if data is reliant on a measurement campaign or third-party supplier) or 'data sets pre-processed and meeting acceptance criteria' by a given date.

3.   STORAGE

Having created a data set, the information must be stored ready for access. How the data is stored is again governed primarily by the requirements of the client and by the storage capabilities of the data processors in the project team. If this stage is not managed effectively, access problems or even data loss are liable to occur. The major issues to be considered during this stage of the data lifecycle are shown in Figure 5.1.
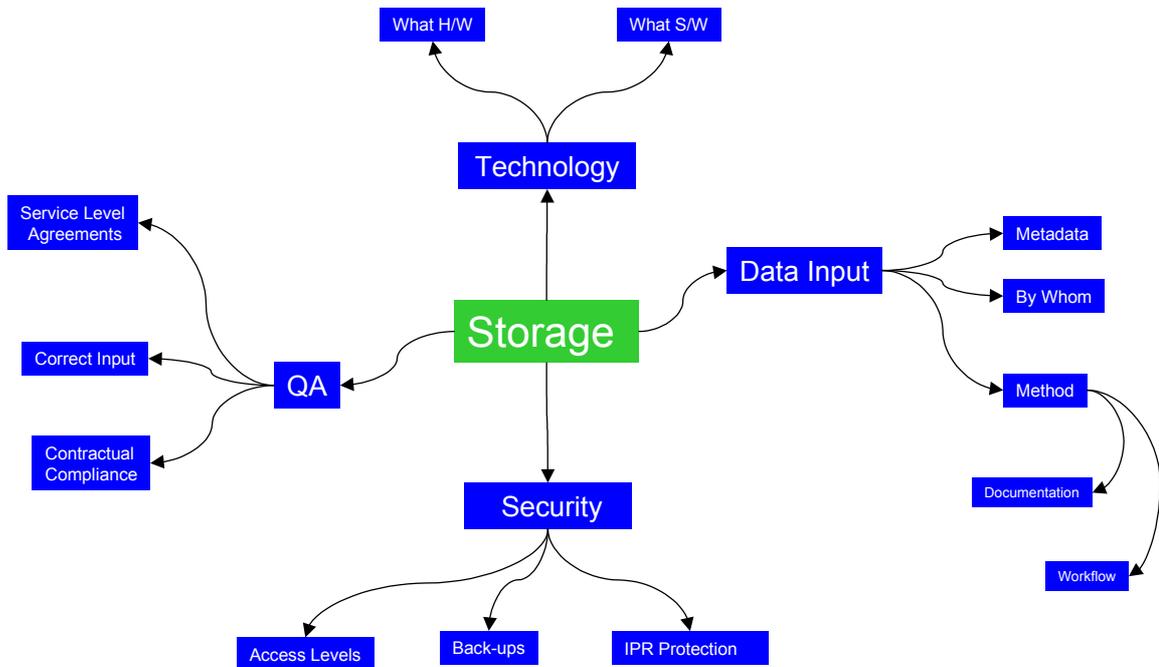
Figure 5.1   'Mind Map' of Issues to Consider at the Storage Stage of the Data Lifecycle

3.1   Cataloguing Data

The primary requirements when storing data are security and ease of access. The content of the database should be predetermined and the compiler should ensure that only relevant information forms the main part of the database but that there are adequate links to the original data sets or references. Of utmost importance to improved data management and exchange is the inclusion of a metadata description of the data content. When created, metadata should be stored with the data to facilitate future access. The EMPHASYS database was delivered using a GIS package called STEM, which had facilities for uploading, storing and retrieving datasets from the database using a map viewer. During EMPHASYS, STEM had very limited metadata capability. This software did not have the functionality to display this type of information other than in a rudimentary way. Therefore, it was difficult for users to access any information about the data directly. The ERP1 UPTAKE project used an updated version of STEM which contained new metadata tools, with links to organisations and their websites as well as detailed descriptions of the content which is available at the time of use.

3.2   Security

When storing a compiled dataset prior to use, it is important to consider the liabilities associated with storing the data and the implications of how the data is stored. Both IPR and Data Protection[3] issues should also be addressed so that the data can be adequately protected. For example who would be held responsible if the data became corrupted during storage? Storage of the EMPHASYS data was considered at the outset and it was agreed that the data would be stored on CD and only accessible to the Consortium members and members of project teams arising out of the Estuaries Research Programme. If the data is likely to become public domain, in the future, it may be necessary to obtain licences or pay Royalties for the inclusion of unpublishable data such as Admiralty Chart information in the case of ERP1 UPTAKE.

3.3   Data Entry

An estimate of the time, cost and manpower required to input the data is necessary at this stage. This should include a risk assessment of likely reformatting and compatibility problems and time delays incurred whilst waiting to receive data from the supplier. When entering data, all storage procedures must be clearly documented in order to simplify access and future data use. The compiler should establish with the client, the ultimate fate of all the data, not just that specified as a deliverable as this may affect the method of storage. Where several versions of the same dataset exist, each should be clearly labelled with a creation date, version number and contents list. Storage of the EMPHASYS data was largely dictated by the STEM software and the input methodology was held in reports and journal notes. Limited metadata was also made available. By keeping careful records of the data input and storage process, it is easy to verify that the data is being stored correctly and to identify which project tasks still remain to be completed.

3.4   Technology

The issue of data lifecycle vs. project lifecycle is extremely important when considering how best to store data. The accessibility of the data and the needs of the client should undoubtedly be of primary concern but it is also important to consider future distribution requirements. Hence, the data should be stored in a format that is easily accessible and is suited to the aims of the study although in reality, the storage mechanism is likely to be governed by the size of the user group. Attention should be given to the technology used to collate and store the data. For example, is the storage medium likely to become obsolete during the lifespan of the data? If so, can the software or hardware be upgraded to reflect customer requirements? In EMPHASYS, the STEM architecture (or Water Information System (WIS) as it was then known) was chosen as the preferred means of storage because, the 4D functionality allowed both spatial and temporal data to be stored and queried within a GIS. The software suite also contained a Publisher, enabling the database to be recorded onto CD and distributed with a free viewer. This allowed users to manipulate their own version of the database. STEM also underwent further development specifically for the EMPHASYS project and was updated again for ERP1 UPTAKE.

1.1.1   _____

[3] In particular where scientific data gives reference to identified individuals, particularly contact details in metadata etc.

Storage

Data Understanding

- Describe what data is to be stored and in what form. For example as time series, summary statistics or an image. Crucially, consider how metadata is to be stored and associated with its respective dataset.

- Anticipate the size of the database in terms of number of datasets, number of data base records and storage (Mb) together with the potential for future increases.

- Determine what information about the database is to be documented in metadata and what is to be included in supporting reports. Ideally most should be included in the metadata

Roles and Responsibilities

- Specify who is responsible for storage and checking of the data within the project team. Scientists and engineers should be responsible for creating metadata records for data they originate, but collation of the database should be performed by a central team

- Ensure staff are fully aware of responsibilities towards the legal aspects of database collation including protection of intellectual property, requirements under the Data Protection Act (where data or metadata identifies individuals), and liabilities accepted for the quality of the data.

Processes and Procedures

- Document the tasks required for data storage such that they can form the basis of guidelines for data updates later on.

- Ensure there is a procedure for version control following data-input and for disaster recovery should corruption of the database occur.

Enabling Technologies

- Determine what storage technology is to be used (e.g. database), including data format of items in the database.

- Determine technology requirements for both data visualisation and maintenance.

- Specify back-up technology to be used during storage

- Consider if one technology is to be used for compiling the database within the project and another for distribution of the database outside of the project.

Audit

- Determine procedures to check the integrity of the database. This should be done at identifiable steps in the database development, e.g. after the addition of a new dataset.

- Monitor the size of database to ensure it is consistent with initial expectations. Significant changes in dataset size may be due to data errors or accidental deletions.

4. ACCESS

Effective management of data access is fundamental to deriving the maximum value from a dataset. This sometimes requires a fine balance between promoting data exchange and ensuring that copyright/ licensing agreements are not compromised. This is essentially the public face of the data lifecycle and by ensuring that data are easily and reliably accessible and that the information content is relevant, data products will be well received. The prominent issues associated with data access are shown below in Figure 6.1.
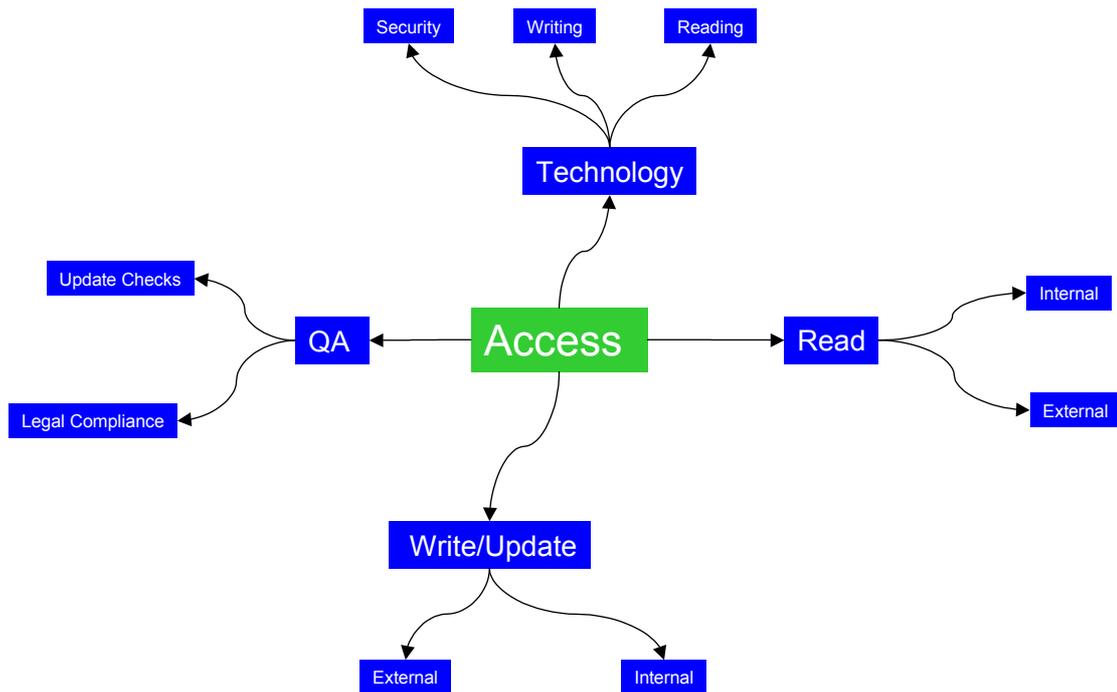
Figure 6.1   'Mind Map' of Issues to Consider at the Access Stage of the Data Lifecycle

4.1   Access Requirements

The access limitations for a given dataset should be predetermined at the creation stage and agreed by all parties. For example, during EMPHASYS, it was agreed that only members of the project team or future teams working on EMPHASYS related projects would have access to the data. However, as the database was delivered upon a CD with the viewing software, it could potentially have been accessed by anyone within the recipient organisation. Details of restrictions upon use were not readily transparent to someone loading the database for the first time, which could have lead to misuse of the data. The implementation of the metadata describing the data and how to use it along with any restrictions are therefore essential at this stage. There are also different levels of access to consider such as who can read the data and who can write or edit it. Hence, understanding the exact requirements of the data user is fundamental to obtaining the maximum value from the data. It is also important to consider how the data might need to be modified if the lifecycle extends beyond that of the project.

4.2   Legal Issues

The legal conditions of data access including onward transmission and sharing should be predefined in order to clarify the issue at the outset of the project. Provision should also be made for the wider distribution and usage of the database at the end of the project lifecycle. This may include lifting access restrictions or the payment of royalties. Within the EMPHASYS project, the legal implications of accessing the data were not considered a major concern because this was a research project with a fairly limited number of users. Therefore, individual license agreements were signed by the partners as part of the project. However, for the ERP1 UPTAKE version, this issue had to be addressed in much more detail as a wider audience was intended. Issues relating to access varied between different scales of data and therefore, more stringent conditions of user documents were required for some datasets. Permission and terms and conditions letters were incorporated into a single document by solicitors specialising in copyright and IPR issues.

4.3   Levels Of Access

All access procedures should be carefully documented along with information on what levels of support will be provided with the data. It may be necessary to offer different levels of access to data and perhaps different degrees of service to different types of user? The fate of all the data collected should have been predetermined at the creation stage including the level of access to any data not specified as a deliverable. However, the main dataset should contain links to any unpresented data and any metadata should contain information about all the data collected for the project. With regard to the EMPHASYS project, procedures for accessing the data were again largely dictated by the STEM software and were covered in the project literature. There is only a limited number of ways of using the software to access the data. However, the information returned from the database can vary considerably depending upon the user's query. The help files supplied with the free viewing software guided users through the query process using a series of tutorials. Having understood how the system works, users could then develop their own queries.

4.4   Access Technology

It is essential to determine early on what technology will have the capability to display data repeatedly and reliably and attention must be given to technologies suitable for protecting the data from illegal access or duplication. When the EMPHASYS database was created, the software did not include facilities for data protection, allowing users unrestricted access to the data. However, this functionality was available for ERP1 UPTAKE, which enabled tight controls to be placed upon database access. Different levels of security were applied to different datasets, which helped in copyright negotiations as data suppliers, were happier to include their data if they thought that it was secure. However, access restrictions may need to be lifted upon completion of the project lifecycle and steps should therefore be taken to determine how best to release the data into the public domain, if the need arises as was the case for ERP1 UPTAKE. The creators of the end-user products should ensure that all associated software and hardware are compatible with the data customers. For this reason, Spatio-Temporal

Environment Mapper (STEM) was chosen as the technology with which users could access the EMPHASYS database because it avoided the need for them to have to convert the database to their own native formats. If the data is stored in electronic format, any related software should be easily accessible and, if possible upgradeable to reflect advances in technology. The format of the data should be agreed prior to creation in order to ensure that compatible technology is used and that conversion time is minimised.

Access

Data Understanding

- Document the access requirements for the database (i.e. what content is to be accessed) both for within the scope of the project team and also possible wider dissemination in the future.

- Describe conditions of access, e.g. free to all, pay per use, subscription etc.

Roles and Responsibilities

- Determine who can access the data and whether any differentiation needs to be made, e.g. by business type (academic users only).

- Determine who is responsible for providing support for access, e.g. which organisation should be contacted if there is a problem or query. Such support does not need to be provided and the data can be supplied 'as is'.

Processes and Procedures

- Produce instructions for data access, specifying terms and conditions for use, including any disclaimers that may apply.

- If metadata only is to be made available, ensure the metadata describes how an application can be made to obtain the actual data set,

Enabling Technologies

- Specify the technologies to be used to access the data and ensure they are available to users of the database. If necessary supply the tools for access along with the database.

- Select appropriate technologies for security and access control. Password protection and software keys can be used to restrict access to only bona fide users. However, ensure the costs of access control are not disproportionate to the value of the data set.

Audit

- If access support is provided, then support logs can provide a useful method to indicate performance of the database and areas where improvements are required.

- The demand for access can be measured in terms of 'request for CD-ROM' (if database is distributed off-line) or 'request for password' (if database is distributed on-line)

- Identify any reasonable steps that can be taken to ensure conditions of access are being adhered to.

5.   UPDATE

It is common for a database to be updated to reflect new research or technology. Good management practice should determine when and how such updates are to be applied and to ensure that data users are kept abreast of changes to the data sets and that the lifecycle continues as intended. The major issues for consideration during the update phase of the lifecycle are shown below in Figure 7.1.
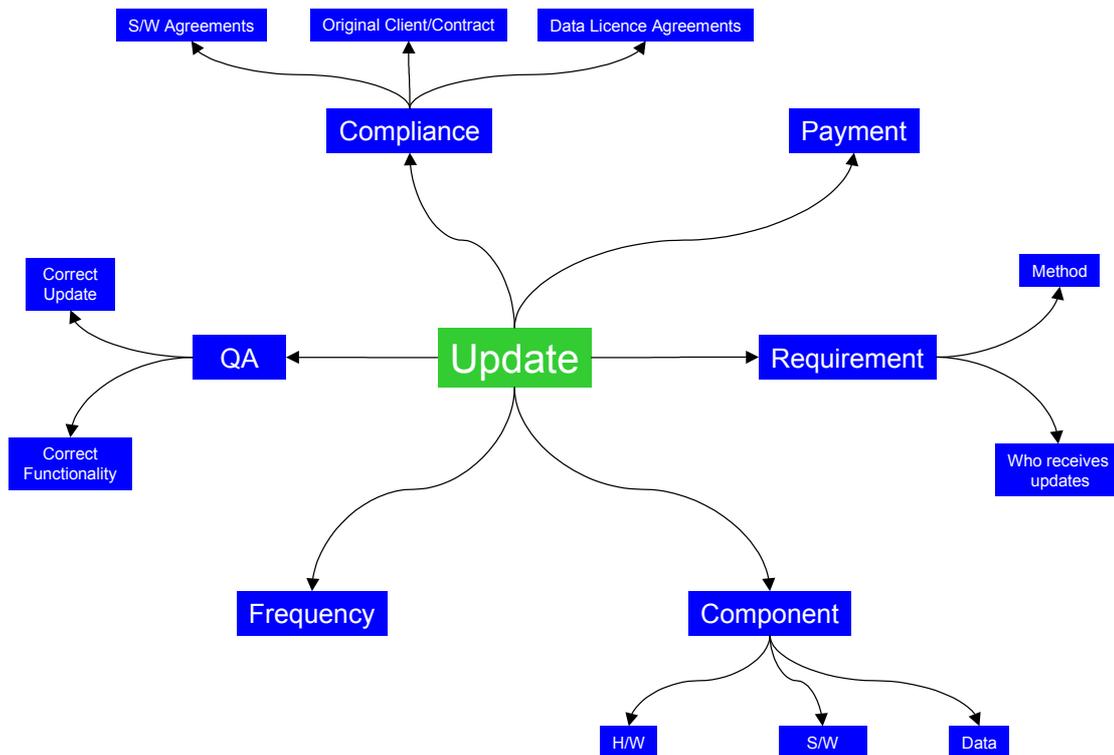
Figure 7.1   'Mind Map' of Issues to Consider at the Update Stage of the Data Lifecycle

5.1   Updates and Upgrades

Before attempting to update an existing dataset, it is necessary to ensure that the product is upgradeable and, how much effort is necessary to carry out the work. For example, The ERP1 UPTAKE Database is an   of the EMPHASYS database. Although, a key aim of the new version was to make a public domain version of the database, the content was largely determined before the project started, in that several additional datasets were identified during the EMPHASYS project which were not included on the database. When upgrading a dataset to include new information, care should be taken to ensure that the existing data is neither lost nor becomes incompatible, which is especially important since the data may be archived, (see next section on data Retention).

5.2   Legal And Commercial Issues

All copyright and licensing issues, including provision for upgrades and distribution, should be predetermined at the creation stage. The fact that the ERP1 UPTAKE database was to be placed into the public domain, led to much greater emphasis upon copyright issues. This is because the general public are normally charged to receive some of the data within the system such as Ordnance Survey and Admiralty data. Fortunately, there has been a culture change in the past few years regarding the use of data and data sharing. As a result, organisations previously reluctant to any such use of their data, are becoming more amenable to data sharing. It is however, important to consider the implication of updating the database with information from different suppliers as this may affect the distribution policy. Also, if the database is to be made available to outside users, should they have the option of purchasing upgrade licences? If so, who would fund and carry out the continued distribution once the project lifecycle is complete and how would access restrictions be controlled The improved metadata support supplied with the STEM software allows very detailed copyright information to be included. Furthermore, this is accessible at the time of usage, rather than being held separately offline. In addition, the database CD is supplied to the user with a detailed 'Terms and Conditions' license, in order to ensure that the user is in no doubt what they may and may not do with the data.

5.3   Update Management

All modifications and upgrades to the data should be clearly documented and supplied as metadata. Care should be taken either that any new data is presented in the same format as the original or the original format is migrated to some new format for continuity purposes. Version management of datasets is a key aspect of the update cycle due to the fact that it may be necessary to keep copies of superceded datasets, either because they have attained some value as an historic dataset or because they provide an easy method for data recovery. Updates to the EMPHASYS database were carried out as follows: firstly a scoping exercise was undertaken for ERP1 UPTAKE in order to find out from the partners, which datasets from the original database could be included in the public version. Secondly, the partners were also asked to supply additional data which they held and which could be included. Datasets with more involved copyright issues were tagged. A list of the required metadata was supplied in advance to the data providers, in order to ensure that we received key information about each dataset. The requested metadata was formatted based upon the NGDF national metadata standard. Documentation was kept of which datasets were input into the database and any modifications which may have been made. This was recorded in an electronic log by the database administrator, in order to provide a permanent record of changes and to display version management histories.

5.4   Technology Requirements

Those responsible for upgrading a dataset should aim to incorporate new information or technology without restricting access to the data. It is inevitable that storage media will ultimately become obsolete, but steps can be taken to minimise this by using software or hardware that is easy to upgrade. If the data is stored in electronic format, the compiler should be make every attempt to ensure that it can be modified when new software becomes widely available if the format is not upwardly compatible and clients should

look to provide funds for this purpose. STEM was chosen as the format for the updated ERP1 UPTAKE database, largely for the same reasons as it was used for EMPHASYS. However, STEM had evolved considerably, making it even more attractive for the delivery of a database project. Firstly, enhanced metadata capabilities were made available to the user based upon the NGDF standard and contains information about the data, such as scale, date of capture etc and the full contact details of the organisation that captured the data, including website links and email addresses. Access to the data could also be restricted, and different levels of access set for different users. Finally, new generic export routines made it much easier for users to export data from the database for use in their own modelling software.

Updates

Data Understanding

- Fundamentally, determine whether updates to the database are to be provided, both in terms of content and the database technology itself.

- Ensure updates can be performed without loss or corruption of data in original dataset

Roles and Responsibilities

- Determine whether updates to the database require changes to existing contracts and licencing agreements.

- Ensure that adequate funding is in place to cover any upgrades to the dataset

- Specify whom is responsible for updates to the dataset content, database software and any viewing tools that may be required.

- Specify responsibilities for distributing updated datasets

Processes and Procedures

- Provide documented procedures for update of database content, database software and any viewing tools that may be required. This includes distribution procedures for the updates

- Ensure there is adequate version control in place to ensure traceability of updates. Use metadata to document where and when changes have been made.

Enabling Technologies

- Ensure technologies used are updateable (and upwardly compatible) and that updates can be performed at minimum cost both in terms of software and staff time.

- Select technologies to support update procedures e.g. remote access, permit updates only by certain users, distribute updated version of the dataset

Audit

- Determine conditions for when an update is required, e.g. significant change in composite dataset, after a given time period, following depletion of existing stocks, obsolence of software

- Check that update procedures are in accordance with project contract

6.   RETENTION

Data that is not in active use may be archived prior to some envisaged future use or deletion. Once archived, the data may become 'lost' and hence, steps should be taken to ensure that this stage of the lifecycle is managed correctly. This is especially true given that many projects will finish and a decision has to be made about what happens to the data then outside the current project. Again provision for this should be made at the outset of the project. See Figure 8.1 below.
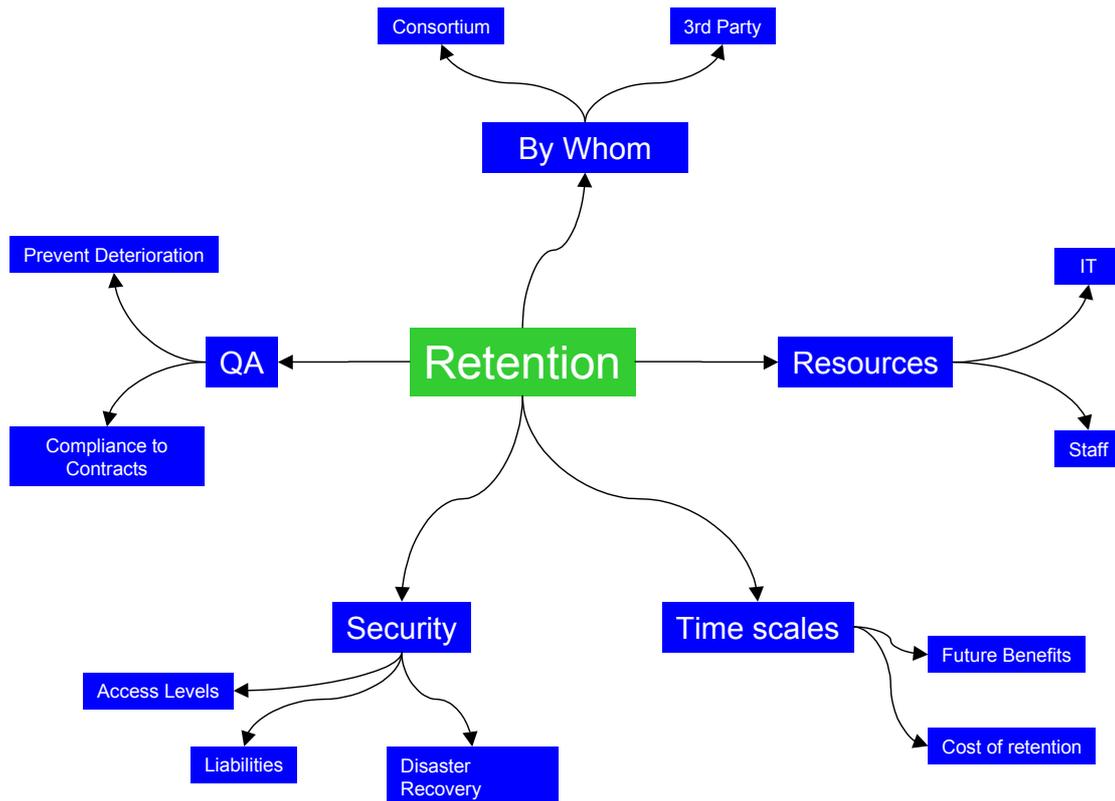

Figure 8.1   'Mind Map' of Issues to Consider at the Retention Stage of the Data Lifecycle

6.1   Defining Retention

At the outset of this phase of the data lifecycle, it must be determined what data is to be retained and by whom. Data must always be accompanied by a detailed description of its content that, in turn, must be identifiable to the data. A copy of this metadata containing details of where the data is stored should be independently held so that the data can be easily located and accessed. Data retained by the ERP1 UPTAKE project but not actually used, including source data from which final datasets were derived. Future databases may want to include such source data or the derived outputs may need to be recreated should a problem occur with the database. As data administrators, it is sensible to keep a copy of the source CDs in order that they can be rebuilt should the database need to be further developed or rebuilt. However, this could only be done if the suppliers had given their permission for a copy of the data to be retained. The extent of

the data lifecycle in relation to the project lifecycle should also be considered at this stage, as it may be that data have been updated and old versions of data need to be stored. If so, it is crucial to know for how long and who has responsibility for this, as this is likely to have associated costs.

6.2   Legal And Security Issues

Anyone holding data should take reasonable action to ensure that it remains easily accessible for future use. If the data is being held on behalf of a client then the duration of this retention period and associated obligations should be determined Details of copyright agreements and distribution restrictions should be clearly defined and kept with the data in order to prevent unauthorised use. Moreover, liabilities should be identified, should the data become lost or corrupted while it is being archived.

6.3   Retention Procedures

When archiving data, it is essential to ensure that all procedures are clearly documented and stored along with the data. This will ensure that the data is not 'lost' or accidentally deleted whilst in storage. Procedures (and accompanying legal framework) also need to be set in place for recovery and re-use of this data, should access be required to the archive. This may include a 'disaster recovery' procedure should data be destroyed, the way in which data were retained for the ERP1 UPTAKE project, reflected 'in-house' archiving procedures. A record is kept of each project on a catalogue. Therefore, it is possible to recover work by browsing information about the data and then locating these CDs and the information contained upon them. This means that should the database become corrupted or it is required for a further update, the appropriate staff involved can easily access the archived data, even if they were not part of the original project team. Alternatively, a third party may be charged with archiving the data or taking responsibility for the database at some future date. For example, a government organisation that funded the database's creation may seek to extend its life beyond the project and may retain copies for future use.

6.4   Archive Technology

Attention should be given to the proposed lifespan of the data and consideration given to the fact that the storage medium may deteriorate over time. Suitable back up procedures for data recovery should be instigated at this stage. Provision for upgrading the data is also necessary to reflect technological advances should the dataset be required for future use. The in-house archiving system used for ERP1 UPTAKE was Reference Manager, which provides links to archived information. The archive itself has 2 copies of all data relating to a project upon CD, unless this was expressly forbidden by the data providers. One of these copies is kept off-site, as part of the 'disaster recovery' plan. CDs / DVDs were considered to be the current standard for archiving data offline. Both can be read by DVD players and as a format this should be around for several years to come. In addition, the delimited text format required by STEM is sufficiently generic to allow easy transfer into other systems. Finally, MS Windows is likely to be the standard platform, for the foreseeable future and was the only one in wide use. Therefore, it represented the obvious choice of operating system.

Retention

Data Understanding

- Describe what the retention policies for the produced dataset are, particularly as this may extend beyond any original contract.

- Describe retention policies for source data and processing algorithms used for the production of the dataset.

- Determine the duration of the retention period and associated obligations during this period, including what happens at the end of the retention period

Roles and Responsibilities

- Determine if retention is to be the responsibility of an organisation in the project team, or a third party.

- Clarify obligations for the dataset during retention including preservation of the integrity of the dataset, access procedures and retention procedures. Produce a new agreement to outline these responsibilities if this is not already documented.

- Clarify responsibilities for the retention of source data used for compiling the dataset. Licence agreement may indicate that such data cannot be retained and is to be returned to the supplier

- Ensure there is sufficient funding to cover all costs during the retention period.

Processes and Procedures

- Ensure there is a detailed metadata explaining exactly what data is retained, who is responsible for the retention and access and update instruction during this retention period. Communicate this to all interested parties

- Ensure there are procedures to cover both back-up and disaster recovery during the retention period.

Enabling Technologies

- Select technologies for retention appropriate not only to the length of the retention period, but also for any access or updates during the retention period

- Ensure technologies do not become obsolete during the retention period through a regular technology audit

Audit

- Check that data is stored in a suitable environment and periodically check the media to ensure that there has been no deterioration.

- Demonstrate the length of the time the data has been retained and indeed when the retention period has expired.

7.   DELETION

Deletion is the final stage in the data lifecycle (see Figure 9.1). Data may be physically deleted but it may also become 'lost' due to ineffective archiving. Data should only be deleted as specified in the contract, if it becomes technologically obsolete, or if it is replaced by newer information. Even then it may be that the data retains some value as an historical dataset and as such archiving may be more appropriate. This stage of the lifecycle requires careful management in order to ensure that valuable data are not destroyed, particularly where one holds the *only* copy of the data.
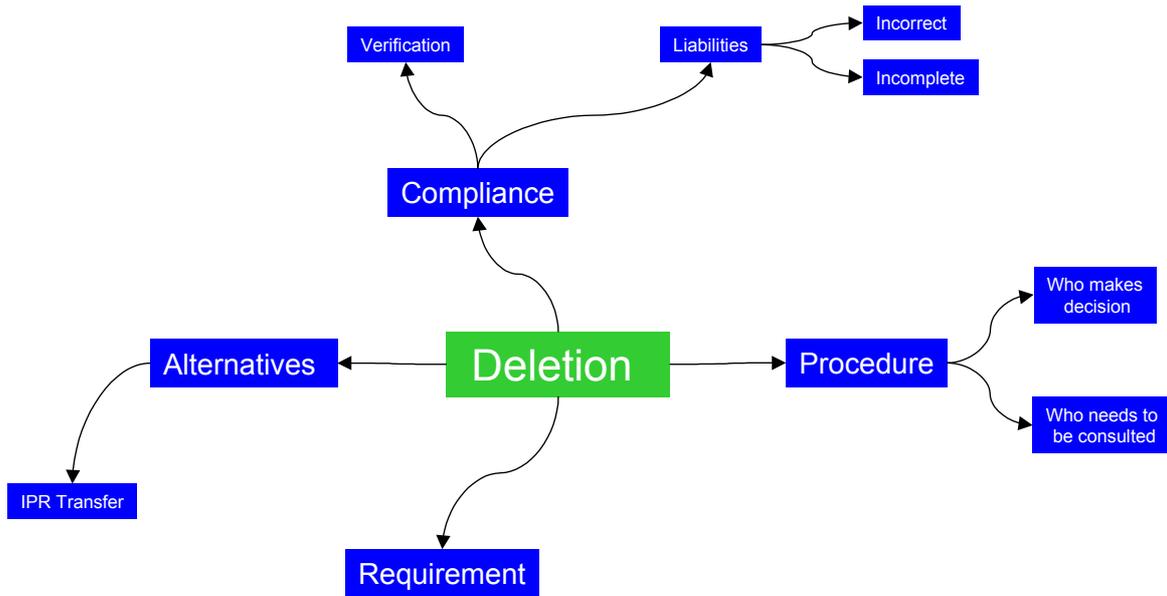

Figure 9.1   'Mind Map' of Issues to Consider at the Deletion Stage of the Data Lifecycle

7.1   To Delete Or Not

When considering deleting a dataset, careful consideration should be given to the contractual requirements to ensure that the data is of no further use, or that it cannot be accessed using current technology. The content of the data must be studied before a final decision is made and options such as archiving and storage medium transfer considered. It is also important to know which data can be deleted and who has the ultimate decision about whether a dataset is deleted? In addition, how does one determine how this happens and ensure compliance? Does the data need to be deleted or will a third party be willing to take the dataset, in exchange for say, the IPR of the data? For ERP1 UPTAKE some data had to be deleted due to the fact that data suppliers although happy for the data to be available to the project team in EMPHASYS, did not want the data to be released in the public domain as the data were either sensitive or usage undermined suppliers' own sales of data. Where datasets had to be deleted a record was kept in the project data catalogue of its previous existence. Moreover, permission was sought to include either samples of the data or metadata indicating its existence. Therefore, even though the data could not be included, users were made aware of its existence and they had the contact details of the relevant people.

7.2   Compliance Issues

The ultimate fate of a dataset should have been predetermined at the creation phase although the owner retains overall responsibility. Provision should be made during the project for either the release or deletion of any data sets after a given time. One should also be aware of one's liabilities for the wrongful deletion of a dataset? As data administrators for the ERP1 UPTAKE database, it makes sense to keep a copy of the source data CDs in case the database or one of its datasets becomes corrupt. This saves a great deal of time, re-requesting the input data. Additional copies of the ERP1 UPTAKE database may need to be created after the formal project is over. Therefore, a secure version of the database needs to be available, in order to facilitate this. Officially, the data are still being used for the project but the project is officially over. Hence it is important to consider the data lifecycle in terms of the project lifecycle and beyond.

7.3   Deletion Procedures

If a data set is to be deleted for whatever reason, care should be taken to ensure that its contents are at least documented. Any metadata pertaining to the data should ideally be stored in order to provide a permanent record. This will prevent people searching for data that no longer exists and also be of use if similar data is to be collected in the future. Furthermore, any organisation should advertise their intention to delete data to determine if there is another organisation willing to act as a custodian for the data. Copies of the data CDs should also be returned to the supplier on completion of the project if this is explicitly stated in the terms of supply. With regard to the EMPHASYS project, there was no formal in-house procedure for the deletion of the data, the CDs were destroyed or rendered unusable and recycled. However for the ERP1 UPTAKE project, documentation of deleted data was held within the project diary. The aim was to track any changes to the original database and to prevent users searching fruitlessly for data that no longer exists. Within the documentation, the metadata and contact details of the suppliers of the deleted data were retained, should anyone wish to use the data again.

7.4   Disaster Recovery

It is important to ensure that data is fully deleted and cannot be 'accidentally' accessed. If the client/supplier agreement stipulates that the data should be destroyed after a certain period of time, it must be properly disposed of rather than simply stored and forgotten about. This prevents anyone not involved with the original project accessing the data at some future time and using it outside the project. The data held for the ERP project were stored on a server during the construction of the database. At the end of the project any data not retained was deleted and un-recoverable other than from backup tapes. Backups to the server system are made on a daily basis and structured such that they can be recovered at intervals up to 3 months ago. However, this is a rolling programme and the last of these backups are overwritten after 3 months. It is not possible to access this data unless explicitly requested for use on the original project. It would be very time consuming to recover and delete individual datasets from all the backups held, which is why it was left to be deleted automatically. Source data CDs

were destroyed where it was expressly forbidden for copies to be retained under the supplier's 'terms and conditions'.

Deletion

Data Understanding

- Recognise that deletion of data is a valid stage in the data lifecycle, provide it is executed as a conscious and considered task.
- Determine the value of any dataset and appraise if deletion is necessary as costs of retention exceed any future benefit

Roles and Responsibilities

- Clarify if there is a responsibility to delete data, in particular source data supplied for dataset creation
- Specify who has responsibility for deleting data within an organisation, i.e. whom can authorise a 'request to delete'.

Processes and Procedures

- Provide a data deletion procedure that must be followed once the retention period is over.
- Ensure there are procedures to demonstrate that data deletion has taken place. Ensure that is reflected in the metadata belonging to the data. Metadata should not be deleted as it contains a reflection on the value derived from the data, e.g. to determine if the data may be used on a future exercise.
- Provide options for transfer of database rights to third parties as an alternative to deletion

Enabling Technologies

- Ensure that any deletion is permanent, i.e. the data itself is deleted and not simply an index pointing to the data.
- Ensure the technology is 'fail safe' in that it helps prevent accidental deletion and ensures confirmation from the user that deletion should occur

Audit

- Provide verification that data has either been deleted or returned to the supplier
- Be able to demonstrate that a deletion procedure has been followed for any deleted data set

8. CONCLUSIONS AND RECOMMENDATIONS

This guide presents the five principles for data management at each stage in the lifecycle of a dataset and provides examples from the EMPHASYS and ERP1 UPTAKE projects as way of illustration. The guide also highlights the need to know the lifecycle state of data at any one time since this is implicit in good data management. The content has been deliberately kept to a generic level to prevent specific recommendations becoming obsolete, and these are presented as 'key tips' at each stage of the lifecycle.

9. REFERENCES

CIRIA (Construction Industry Research and Information Association) 2000, *Maximising the Use and Exchange of Coastal Data: A guide to best practice*, CIRIA (Construction Industry Research and Information Association) Publication C541

Dyer B & Millard K, 2002, A Generic Framework for Value Management of Environmental Data in the context of ICZM, *Journal of Ocean and Coastal Management*, 45, pp59-75, Elsevier

Pye K, 2000, Recommendations for Phase 2 of the Estuaries Research Programme – Final Report, Report TR 113. Estuaries Research Programme, Phase 1. pp 19, Department of Geology, Royal Holloway, University of London.

Mayon-White W & Dyer B, 1997 *Principles of Good Practice for Information Management*, Version 2.0 IMDA, London School of Economics, Published by British Standards Institution, ISBN 0 580 26855 1

Fax:	023-8033 8040
To:	Claire Brown ABPmer
Subject:	Scientific Data Management Guidelines Draft 2.1
From:	
Organisation:	
Phone:	
Email:	

Please remove this page to let us know your comments on this document. Please reply by 01/02/03.

Overall I found the guide
Not useful
Useful as a 'memory jogger'
Useful as a 'guide to follow'

Primarily I am involved with
Commissioning work involving data management
Undertaking work involving data management

The following changes to the guide would help me:

The opportunity to comment on this document was given at the Estuary Data Management workshop held by CIRIA on 14 January 2003. Participants were provided with this form on which they could fax back comments. No comments were received.