The Royal Statistical Society
& the UK Data Archive

Preserving & Sharing Statistical Material

Working Group on the Preservation and Sharing of Statistical Material: Information for Data Producers

First Published 2002

Published by:
UK Data Archive
University of Essex
Wivenhoe Park
Colchester
Essex
CO4 3SQ

ISBN: 0-906805-02-3

Designed and typeset by Nimbus Design, Farnham, Surrey
Printed and bound by Blackmore Limited, Longmead, Dorset

Welcome letter to the reader

The Royal Statistical Society and the UK Data Archive are pleased to have worked together on this document to promote the need for the preservation and dissemination of statistical material.

The last two or three decades have witnessed an unprecedented increase in the collection of data as a means of commenting on and interpreting social, economic and political phenomena. This increase is true also in areas of research and investigation in the humanities, sciences, medicine and commerce. Equally, there have also been advances in computing, and developments in statistical techniques, that allow information to be collected, managed and interpreted from a far greater range of sources than was possible previously.

However, despite recent technical advances, indeed in part as a result of them, computerised data are potentially at greater risk of loss and redundancy than ever before, if only because of the sheer pace of technical change. Equally, although data storage costs have fallen dramatically, data collection remains a costly undertaking. Thus, it could be argued that the need to be aware of the benefits of sharing and preserving data are greater now than ever. The concept of re-using data is particularly well developed in the social sciences where data archiving for secondary use has been commonplace for over 30 years. Sharing and preserving should not, however, be limited to the social sciences. The sharing of data within this area has proven to be of demonstrable advantage to the research community, to policy makers, to special interest groups and to the public. There are equivalent, potential advantages in the re-use of other types of data. Consequently, the Royal Statistical Society and the UK Data Archive have jointly produced this document to encourage a stronger ethos of preservation and sharing amongst data producers.

Our aim is not only to encourage this ethos in the many organisations in which data are collected, but also to take a pragmatic view in recognising the constraints on sharing and preserving and, by addressing these, offer strategies for overcoming them. We also recognise that the requirements for preserving electronic material may, on first consideration, appear daunting. In response to this we offer a code of good practice which we believe is a realistic first step for data preservation and sharing. We exhort readers to implement this and, at the same time, remind readers of the many expert organisations that specialise in data preservation and sharing and whose staff are willing both to discuss the issues and problems, and to offer advice and support.

We hope you will find this document useful.

President, Royal Statistical Society

Director, UK Data Archive

Acknowledgements

I would like to thank successive Presidents of the Royal Statistical Society for their support of this work - Robert Curnow, Denise Lievesley and Peter Green. I would also like to thank Kevin Schürer, Director of the UK Data Archive, for his support of this project and his input into the finalisation of the text. The work would not have been achieved but for the contribution of members of the Working Group on the Sharing and Preservation of Statistical Material. In particular, I would like to thank James Denman, Alison Macfarlane, Andrew Westlake and Neil Walker who all made significant contributions to the final document. Thanks also to Bev Botting and Raivo Ruusalepp who were involved during the latter stages of production.

Membership of the working group altered over time. I would like to thank all of the following for their contributions:

Sheila Anderson
Kevin Ashley
Richard Blake
Louise Corti
James Denman
Susan Healey
Paul Hunter
Ho Law
Alison Macfarlane
Neil Walker
Andrew Westlake

Hilary Beedham
Chair, Royal Statistical Society Working Group on the
Sharing and Preservation of Statistical Material.

Contents

Executive Summary

This document has been written on behalf of the Royal Statistical Society (RSS) and the UK Data Archive (UKDA) to raise organisational awareness of the value and benefits of preserving and sharing statistical material.

There is a growing recognition that valuable statistical material is being lost in preventable circumstances. The increasing rate of technological development, organisational changes, the disintegration of project groups and a failure to implement simple procedures all contribute to the loss of valuable material.

This document seeks to move beyond the short to medium term protection of material, characterised by ‘backing up’, that ensures that nothing is lost whilst it is still needed for a primary purpose, towards a more strategic approach to long-term data preservation. Its main concern is with the preservation and sharing of material which has often served its immediate purpose but which is sufficiently valuable for society to gain, possibly in ways that are not currently apparent, by having the opportunity to re-use it. The document does not assume that all data should be freely-available, but does argue for more open access to data.

Both the RSS and UKDA recognise that there are many reasons that seem to prevent organisations from ensuring that their statistical material is preserved in the longer term. These include the perceived costs to the organisation, technological constraints and concerns about data confidentiality. We believe, however, that many of these difficulties could be reduced or eliminated if organisations were to implement a system of best practice which will ensure, from the outset of the data collection process, that the data gathered, along with accompanying source and contextual information, will be preserved and shared.

This document provides advice and information for those wishing to take this route. It includes the RSS Code of Best Practice, which is intended for those who are responsible for data collection, but who may lack experience of organising data preservation. This document takes the view that preservation of digital material is a vital issue, not only from the point of view of sharing it, and should be considered by everyone who creates data and needs to maintain them over medium to long periods of time. It also incorporates a list of a number of specialist organisations that are willing to share their expertise in the preservation of statistical material along with their contact details. Finally, it provides a bibliography of key documents that address the most common issues raised when long-term preservation and sharing are being considered.

Introduction

- Do you think you have some statistical data that may be of interest to others?
- Are you collecting, creating, processing, or planning to collect or collate statistical material?

If your answer to either of these questions is **YES** then:

- Have you thought about sharing your data?
- Are you doing something to preserve the usability of your datasets?
- Are you aware of any risks, problems and issues related to digital preservation?

If you answered **NO** to any or all of these questions then this booklet is for you and we invite you to read further!

The Royal Statistical Society (RSS) and the UK Data Archive (UKDA) have produced this document in order to demonstrate the many benefits of keeping and sharing statistical data. We would like to encourage you to share your data with the statistical and other scholarly communities. At the same time, we would like to draw your attention to the issues related to long-term digital preservation. It will not be possible to share or even use the digital resources in the medium to long-term future unless they are properly maintained and preserved since the original methods of accessing data files may become obsolete through technology development. It is, therefore, important to promote best practice in data preservation and establish policy guidelines that specify responsibility for preservation procedures. An example of a guideline that incorporates most of these aspects is the RSS Code of Best Practice on the Preservation and Sharing of Statistical Material which is included in this booklet as Appendix 1.

The rationale behind this document has been to produce a basic set of guidelines that is both informative and persuasive. The target audience will be those who are responsible for data collection but may have little or no experience of data preservation. Key emphases are the need to prevent the loss of valuable statistical material and the wisdom of consultation between data producers and experts who can provide specific, up to date information. Institutions exist throughout the UK and the world that specialise in both preserving and disseminating digital research material to the commercial and academic communities and to the general public. These agencies can be contacted to learn about their expertise in handling the practical and organisational issues of preserving digital datasets and providing access to them. This document includes appendices that give information about, and contact details for, a number of such organisations in the UK. Many of these organisations have web sites which provide links to equivalent or similar organisations outside the UK.

This document focuses on four aspects of sharing and preserving digital statistical resources. First, an argument is made for sharing digital statistical material with a number of significant reasons why re-using adds value to the data. In this age of over-abundant information it would be impractical to preserve all digital data and principles are required for deciding what kind of statistical material should be preserved. These are discussed in the second chapter of this document. Once the decision to keep and share data has been made, procedures are needed to preserve the data and all the contextual information that go with them. Strategies for implementing such procedures are offered in the third chapter, while the fourth chapter offers suggestions for overcoming barriers to digital preservation and sources of professional help. The RSS has produced its own code of best practice for sharing and preserving statistical material and is promoting it amongst its members. The RSS Code of Best Practice on the Preservation and Sharing of Statistical Material is included as Appendix 1 of this document. A further two appendices include a list of institutions that have considerable expertise with digital preservation and the dissemination of electronic research data, plus a carefully selected bibliography of useful publications on issues related to digital preservation.

1. Why preserve and share statistical material?

The reasons for preserving statistical material are diverse. In some data collection settings, such as clinical trials, data preservation is a legal requirement. In others, such as areas of social science, the requirement to preserve for re-use is often a condition of funding. For many others, however, a case needs to be made as to the value and purpose of keeping and sharing their data.

1.1. Custodial responsibility

Individual data collections are unique entities. They take place within, and capture circumstances that are socially or scientifically fixed in time and they can rarely be reproduced in exactly the way they were first undertaken. At best, the instruments of data capture can be reproduced but the specifics of the cases for which the data are collected a second time almost invariably change. We live in a world in which increasing emphasis is placed on statistical measurement as a means both of measuring and testing the effectiveness of scientific and medical techniques and of measuring the effectiveness of political and business decision making. Thus data collectors have a responsibility to ensure that the materials which contribute so emphatically to the circumstances of our lives are preserved for further use and historical record.

1.2. Data collection costs

Data are expensive to collect, and become increasingly so if the collection process is to provide high quality, validated data. Computer-based technology has had a dual effect on the data collection exercise. It has made it easier, quicker and in some ways cheaper to collect and analyse large amounts of data, but by increasing the amounts of data that we can manage, it has also raised the expectations of users. Users not only demand more data but also have considerably higher expectations when it comes to quality, speed of delivery and processing of data. This has led to a need for more thorough descriptions of data resources to facilitate and speed access to them.

Users do not always need the most recent material. When data have been collected, at great expense to supplying companies or the public, they ought to be preserved for use on those occasions where they can contribute either to the solution of new problems or to the needs of those involved in teaching, learning and training. In this way, re-use of data gives added value and reduces the need to duplicate the original collection of data.

1.3. Legal requirements

For some data collectors, there are legal requirements to preserve the statistical material that they have collected. In these instances, there are usually well-defined procedures and clearly-identified organisational responsibilities for the preservation of the relevant material.

The revisions to the Data Protection Act may also have implications for data collectors in that they have to be in a position to supply recorded information on an individual if requested to do so by that individual. This may be possible while databases are actively being used but some consideration ought to be given to how such requests might be fulfilled in future if existing hardware and software systems become outdated.

The 1998 Data Protection Act[1] stipulates that anyone processing personal data must comply with the eight enforceable principles of good practice. These principles of data protection state that data must be:

- fairly and lawfully processed;
- processed for limited purposes;
- adequate, relevant and not excessive;
- accurate;
- kept for no longer than necessary;
- processed in accordance with the data subject's rights;
- secure;
- transferred only to countries with adequate protection (i.e. not to countries outside the European Economic Area).

Personal data cover both facts and opinions about the individual. They also include information regarding the intentions of the data controller towards the individual, although in some limited circumstances exemptions will apply. With processing, the definition is far wider than was previously the case. For example, it incorporates the concepts of 'obtaining', 'holding' and 'disclosing' (see also section 4.3 below).

The Data Protection Act encourages the timely destruction of individually identifiable data but this should not prevent the preservation of anonymised statistical data which fall outside the provision of the Act.

This document is suggesting that organisations develop a long-term preservation policy and incorporate data preservation needs into their codes of practice in a way that directly allocates responsibility for preventing data from becoming unusable.

1.4. Secondary use and analysis

Many datasets are collected with a specific goal in mind. Once that goal has been met, there is a tendency amongst some collectors and sponsors to consider the life of the study to be complete. In practice, there are numerous examples where statistical material has been subject to secondary analysis and has proven valuable (such as national census material). Time constraints also tend to limit the amount of work that can be undertaken on a study by the principal investigator. By preserving material and making it available for further analysis, the full potential of a study can be achieved.

The increasing emphasis that is laid on statistics as measures of success for phenomena as diverse as social policy and drug effectiveness means that those who collect and report on the data have a significant responsibility to make their work transparent to those whose lives they are affecting. By making data available for re-use, the methods, techniques and results can be discussed in an atmosphere that is open and constructive.

1.4. Secondary use and analysis

1.5. Business imperatives

However, material preserved must include information about their context. These metadata make it possible to understand and re-use the data correctly. Suggested procedures for successful preservation of data and providing datasets with necessary metadata are given in Chapter 3 and in the appendices of this document.

1.5. Business imperatives

Often, data that have been collected for a business imperative have great potential for other purposes. These other uses are frequently unfulfilled because the data have a commercial value in the business world. This may restrict their use to the principal investigator. We would like to encourage organisations with such material to reconsider their policies on sharing such data. Researchers can often make excellent use of data which may be out of date for business purposes. It may also be possible to create subsets or aggregates of the material which would permit further use without jeopardising the commercial benefits to the collector. The costs associated with preparing a dataset for deposit with a data archive are usually minimal and should not discourage anyone from offering their data for re-use.

It is clear that even if there is no direct legal requirement or funding condition binding the data creator or principal investigator to keeping and sharing the data, there is demonstrable value in making the data available for re-use and further analysis by the statistical community. Justifications for keeping and sharing statistical data can be found among the five reasons offered above or, in a combination of them, as well as other factors that data creators and funders may find applicable. The decision to preserve a statistical dataset involves several choices that need to be made. Two of the most important are: what to preserve and how.

2. What should be preserved?

It would be impractical to preserve all material of a statistical nature but it is important that the potential value of all such material is considered as part of the selection process. The selection process, therefore, requires qualitative assessment of the value of information contained in the datasets. The creator of the information, and often sponsors of research, are best placed to assess its true value, but any decision is inevitably subjective and it can be difficult to predict the future usefulness of a dataset. In order to avoid serious risks of censorship or misjudgement, organisations may wish to establish their own selection policy for the material they would like to keep or they may wish to consult expert organisations. Examples of such collection policies are referred to in Appendix 3.

The accurate later re-use of data assumes that an adequate description of each dataset and its context is preserved alongside the data. This is particularly vital where research data are preserved and used for secondary analysis or to replicate experiments. Not all such contextual information is necessarily obvious and easy to select and preserve. Research projects regularly generate considerable amounts of ‘electronic paperwork’, such as administrative files, research diaries, automatic output, descriptions of research instruments and electronic mail. This would necessarily all be part of contextualisation of the data selected for long-term preservation, but may well include documentation that is vital for interpreting the dataset and its contents. They should, therefore, be appraised for relevance and, if deemed valuable, preserved together with the dataset.

If no clearly defined common selection procedures or agreed criteria for selection exist, it should be the responsibility of the data creators to establish such guidelines or to arrange for qualified appraisal of their datasets. Data producers are encouraged to discuss which data should be preserved both within their own organisations and with external specialists. Most dedicated archives will have a clearly defined collection policy, which will aid in the decision-making process (see Appendices 2 and 3).

Once a decision has been made on what statistics will be preserved, it is easier to specify the individual components that need to be archived to form a sufficiently clear and re-useable dataset.

2.1. Data

It is self-evident that when preserving statistical material, digitised data should be preserved. There are a number of issues associated with this that should be considered by those who wish to preserve their material for the long-term themselves. As Chapter 4 explains, so far, there are no 100 per cent guaranteed solutions to the problems of digital preservation but there are several approaches that solve most of the technical issues related to technology obsolescence and media deterioration. There is a rapidly growing body of publication on these topics and some are listed in Appendix 3.

This document outlines some of the issues but the RSS and UKDA recommend that those wishing to preserve their statistical material should consider discussing the process and procedures with a relevant organisation offering specific advice. Specialist data archives that have been dealing with digital preservation for decades, and more recently have engaged in digital archaeology and emulation of computer environments, are available for consultation.

The ideal situation would be for a data-producing organisation to have in place a preservation policy that sets out procedures and responsibilities for preserving the material that might be of future use. Such policies have to be updated regularly to accommodate all data formats that are generated and collected as well as to reflect any organisational changes that may leave data unsupervised.

2.2. Explanatory documentation

Data alone are of little use. They need to be explained by accompanying documentation, which has to be of sufficient quality and detail to permit the accurate use of the material in the future. Various internationally agreed systems and standards for describing archival material exist. Describing digital collections is usually regarded as necessarily more complex and detailed than constructing the metadata required for paper documents. References to detailed resources of metadata and description of digital records can be found in Appendix 3 of this document, but in outline the vital elements of explanatory documentation that must accompany datasets selected for preservation are as follows.

- *provenance of the dataset* – from where does the dataset originate? what was the purpose of its creation? who are the authors and/or contributors? how were the data collected? what, if any, are the legal conditions for using the data?

- *technical description of the dataset* – what information is contained in its variables? what is the data type of each variable? what are the relationships between tables and whole datasets? what coding schemes have been used?

- *preservation metadata* – what file format is the dataset stored in? what hardware and software are required for using it? has the application software been upgraded or the dataset converted from one format to another and when? is the dataset still complete after conversion? what storage medium is used?

2.2. Explanatory documentation

2.3. Identifiable data

Some of this contextual information may be in paper documents, which the research project or organisation has produced, most often for its own internal use. These should be collected and preserved either as part of the archive of the research project or directly together with the dataset. One possibility is to digitise the relevant paper documents and store these digital metadata documents alongside the statistical data. Before deciding to do so, however, data creators must carefully consider the legal implications of preserving digital images rather than paper originals. In particular, they should consider whether they could guarantee the authenticity of digital copies. Long-term preservation of digital images of documents, as opposed to digital datasets, adds to the complexity of the preservation process and is very likely to increase the costs involved. Organisations with no prior expertise with digitisation of paper documents should also consider the resource implications that a digitisation programme will incur.

Data collection exercises and research projects that are only at the early stages should consider creating and maintaining such descriptive documentation digitally from the outset. It will then be easier to distribute, update and preserve this metadata information alongside the datasets created. However, it may require a conscious decision or a clause in the preservation policy to arrange for collecting and storing this information at regular intervals.

As with the data, those who are considering preservation of digitised material are encouraged to consult with specialist organisations for advice about what documentation should accompany a study and what formats are appropriate. Precise requirements are likely to differ and depend on the rationale for preservation. Additionally, specialists are likely to have experience of the use to which data are put and can help in deciding, for example, what the frequency of preservation should be for studies that are constantly being updated.

There is a trend towards increased standardisation and international agreement about the minimum metadata elements necessary for successful long-term preservation of digital material. UK archival, library and academic institutions are heavily involved in the international metadata standardisation process and are themselves implementing these description systems. Examples of metadata required for depositing a study with a data archive, description elements of research datasets and guidelines for digitisation of paper documents are noted in Appendix 3.

2.3. Identifiable data

As statistical methods for data linkage become more commonly available, there will be an increasing need for data to exploit the benefits offered by these techniques. A number of organisations are already applying these methods and are creating single datasets of linked data in environments that ensure that individual confidentiality is not compromised. It is worth noting that in some cases, anonymised versions of these datasets are available under special conditions.

There is a recognised need to preserve both existing linked data files and those data which may usefully be linked with other material, even if access to these materials is indefinitely embargoed or subject to restrictions. Technical restrictions such as the availability of computer disc space and the expense of preserving large volumes of data that hampered the sharing of linked data banks in the past are no longer serious obstacles. Equally, efficient methods are now available for accessing large datasets. If data volume is considered to be a problem, readers are encouraged to discuss this with a specialist organisation.

2.3. Identifiable data

2.4. Data sampling

2.5. Data at risk of loss

2.5.1. Paper records

2.4. Data sampling

It is clear that not everything could and should be preserved because the volume of digital material is growing rapidly and universal preservation would be impossible. This means that some selection for preservation is inevitable. If datasets are very large and there are no precise legal requirements or guidelines for preservation, preserving a sample from the dataset can be considered. Random sampling has been suggested as a useful approach for preserving certain types of digital information. Snapshots of dynamic datasets that are being constantly updated can be taken and stored at regular intervals. For non-cumulative datasets where individual records are changed or deleted, an audit trail can be kept of the changes made over a period of time. This will help to preserve a comprehensive view of the data held. Organisations should be aware of the management costs associated with a selective preservation policy.

2.5. Data at risk of loss

While the data component of a complete dataset that is being preserved is usually well maintained, other types of information receive less attention and are at risk of being lost either because of inadequate practices, incomplete collection and preservation policies, or simple human error. A few examples of endangered statistical material are listed below.

2.5.1. Paper records

There is much concern that as we shift from a world of paper and pencil records to one of keyboard and computer disc, material collected and stored in paper form is being lost.

Organisations in both the public and private sectors that have a responsibility for such material are urged to ensure that it is safely stored and catalogued for future use or, where possible, is converted and preserved on an electronic medium. Where legal requirement exists for keeping the paper originals of documents, digital dissemination copies can be made and preserved alongside the authentic paper originals.

Where organisations merge or are superseded by newly-composed equivalents, there is a danger that material can be lost as a result of clearance. This applies equally to electronic material as to paper. The RSS and UKDA recommend that all data producing organisations introduce systems of best practice in data collection and preservation on the premise that loss is much less likely to occur where formal systems are in place.

The danger of lost records is of particular concern in the public sector where responsibility for data provision may be passed to individual organisations with devolved budgets for which spending on data preservation is not viewed as a high priority. The RSS Code of Best Practice in Appendix 1 suggests the appointment of a ‘Champion’ at senior management level. This person will have responsibility for ensuring that the organisation provides policy guidelines that specifically address the issue of data retention and preservation and should be allocated sufficient resources to implement this policy.

2.5.2. Hastily-migrated electronic records

The speed of technological change is accompanied by risks to statistical material. The key to long-term preservation is the need to ensure that material is protected from hardware and software obsolescence. There is increasing evidence that as organisations update their software and hardware, important material that is no longer of immediate use is being set aside for future conversion. Unfortunately, these conversions are not always being scheduled. If they are attempted later, software may no longer run on new systems, or accompanying paper documents have been destroyed or lost and the work cannot be completed.

Data migration decisions always carry certain risks and should, therefore, be carefully timed, considered and documented. There are a number of publications setting out the best available practice in data migration (see Appendix 3). If organisations establish and follow an efficient preservation policy that specifies responsibilities for preservation actions, risk of irreversibly damaging datasets through conversion should be minimised.

2.5.3. Scientific and experimental data

Significant amounts of data are now being produced in support of scientific and experimental work in a number of fields. One example is the computer generation of facial images. This is common practice now for a number of purposes including aiding surgeons in the reconstruction of facial features following damage, constructing images that demonstrate the impact of ageing on the features of missing children, and reconstructing features onto skulls in forensic and archaeological work. Data from astronomical observations, geophysical surveys and other similar areas are traditionally not widely shared with other scholarly communities and little of this material is being systematically preserved.

2.5.4. Medical records

The loss of this type of material is of particular concern. Data collected as part of ‘research’ are of personal clinical relevance to the subject, and may be required for clinical audit. Secure preservation of data in the medium term is a regulatory issue in some cases. The ethical and legal constraints placed on data collectors often limit secondary use of the material, even when anonymised, if prior legal consent has not been obtained.

There is some evidence to suggest that the emphasis on medical confidentiality can result in an assumption that once clinical review is complete and statute is exhausted, confidentiality constraints prevent any further use of the material. Consequently, there is either no reason to preserve the material or even a positive reason to destroy it. Whilst acknowledging that historically, studies have not been designed with data re-use in mind, researchers should now be able to anticipate likely requests to share data. As part of the process of informed consent, study subjects could be asked for permission to make their data available to ethically-approved third parties.

It is worth noting that a draft regulation is expected, which will support Section 60 of the Health and Social Care Act. This section of the Act gives the Secretary of State powers to require identifiable information to be passed on without informed consent, in certain circumstances. For example, the passing on of information on communicable diseases may be in the public interest.

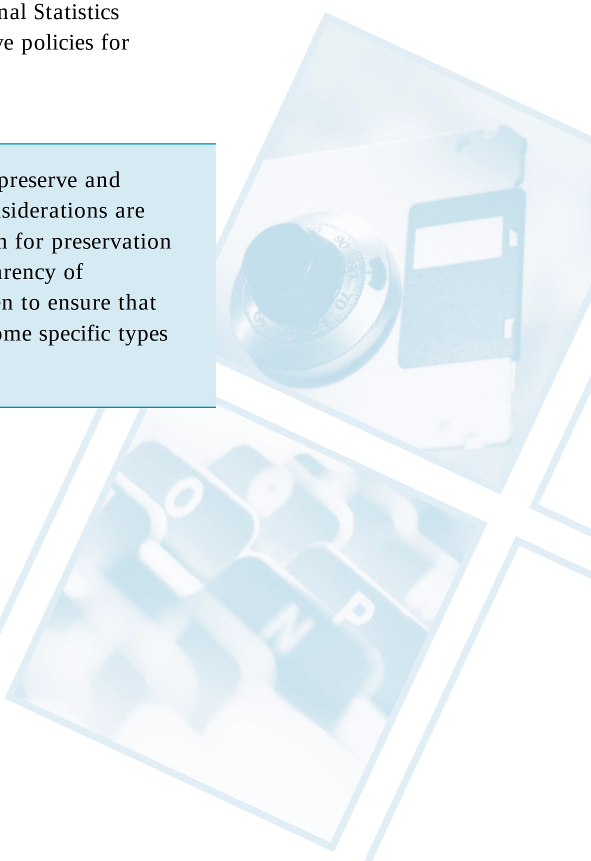
2.5.5. Data from clinical trials

Similarly, there is growing concern that data from clinical trials and pharmaceutical product development are not available for secondary analysis. The RSS and UKDA see no strong reasons for restricting access to such data for bona fide statistical purposes and recommends that the granting of licences for new products should be framed to promote more open access to associated statistical material.

2.5.6. Financial and economic data

The RSS Working Group was specifically asked to mention the need to encourage national financial organisations to consider the importance of preserving and sharing their data. The UK Office for National Statistics should be applauded for having had the foresight, many years ago, to ensure the preservation and sharing of data which contribute to the Retail Price Index (RPI) and their macro-economic time-series data, the National Statistics databank. Many regular producers of financial statistics still do not have policies for preserving and sharing such material.

There are a number of considerations when choosing what data to preserve and share. Legal, qualitative, organisational, technical, and resource considerations are all equally important. Setting out principles and criteria for selection for preservation in a collection policy document is a useful way of providing transparency of reasoning behind decisions that are made. Special care must be taken to ensure that sufficient documentation is preserved alongside the data and that some specific types of data that are particularly at risk of loss are being maintained.



3. A strategy: how to preserve material for sharing

Preserving data for later re-use and sharing is not only a matter of a decision to do so, nor is it a technical issue alone. It should also be understood as an organisational and strategic issue, which is best approached by establishing appropriate policies and adopting best practice in this area. Long-term digital preservation is a problem that every research project must address if it has data older than five years, or if the research project is expected to last for five years or longer. Even if sharing the collected data with other scholars and institutions is not possible or desirable, procedures need to be put in place to keep the statistical material accessible and useable throughout its life-cycle. A three-stage strategy for doing so is proposed below.

3.1. Best practice

One of the three most important elements of any strategy for the preservation and sharing of material is the adoption, within an organisation, of a code of best practice. The RSS Code of Best Practice is included as Appendix 1 of this document as a recommended model. It has been widely discussed amongst data producers, data archivists and data disseminators and presents a sound basic framework for those organisations wishing to share and preserve their data.

The key recommendation that the RSS Code of Best Practice makes is that responsibility for preservation and sharing should be specified at the managerial level. This ensures that policy guidelines are established and sufficient resources are guaranteed for implementing the policy within an organisation. The appointment of a Preservation Champion or Data Custodian, who assumes the overall data stewardship within the organisation, is the first step towards reliable preservation management.

3.2. Adoption of the preservation and re-use principle

The second most important element is for organisations to make a positive assumption that material will be preserved and made available for appropriate re-use. In this way, the processes required for the preservation and sharing of material will be incorporated routinely into data collection and management systems.

Decisions to archive a dataset should preferably be made at an earlier rather than later stage in the life-cycle of a data resource. This is because long-term preservation needs impinge upon methods of creating, describing, managing, storing and disseminating digital resources. If there is a danger that the short-term focus on cost efficiency is dominant during creation of a data resource, it should be the responsibility of the Preservation Champion to advocate the value of the long-term preservation and to demonstrate the benefits of additional investment that may be required.

The following three elements of the life-cycle of a digital resource are critically affected by long-term preservation requirements.

3.2.1. Design criteria

The design criteria for data collection exercises should incorporate elements that will facilitate the preservation and re-use of material. This could be as simple as the inclusion or extension of a consent form at the beginning of a questionnaire or interview or it could involve small alterations to the structure of the material within the collection or analysis software to facilitate the output of anonymised files.

3.2.2. Data management

Throughout all the various stages in the life-cycle of a digital resource, the quality and integrity of the dataset must be controlled so as to maintain its accessibility and to provide for exchange and transfer of datasets. There are commendable ways of ensuring that the data are not inadvertently put at risk or corrupted. These include the use of standardised formats for storing data and complying with national and international protocols, classifications and codes of practice relating to harmonisation, standardisation and integration.

The rapid rate of technological development necessitates regular reviews of data management procedures and formats in use. These may lead to a decision to migrate the data. Such decisions need to be fully documented as part of the regular metadata of the resource.

3.2.3. Metadata management

An assumption that material will be re-used, combined with an emphasis on best practice and quality control, will naturally result in the production of the descriptive documents and metadata that are essential to effective preservation. Throughout the collection, validation and analysis processes, organisations need to ensure that the various steps are fully documented and recorded, to industry standards where these exist. This is not only good practice, it will also enable subsequent users to gain a full understanding of the material, its evolution and the conclusions based upon it. All that is required is that these documents are kept up to date and are collated and that, where appropriate and possible, provision is made for digitising them.

3.3 Identification of sites for preservation and dissemination

Thirdly, organisations need to identify an appropriate site or sites for their data preservation and dissemination. The two may be different. If these functions are to be undertaken internally, then appropriate technical and procedural systems will need to be established and a cost analysis carried out. Where organisations prefer to use the facilities offered by outside agencies, then it is critical that effective liaison is established between organisation and agency to ensure that the requirements of both parties are met.

3.2.1. Design criteria

3.2.2. Data management

3.2.3. Metadata management

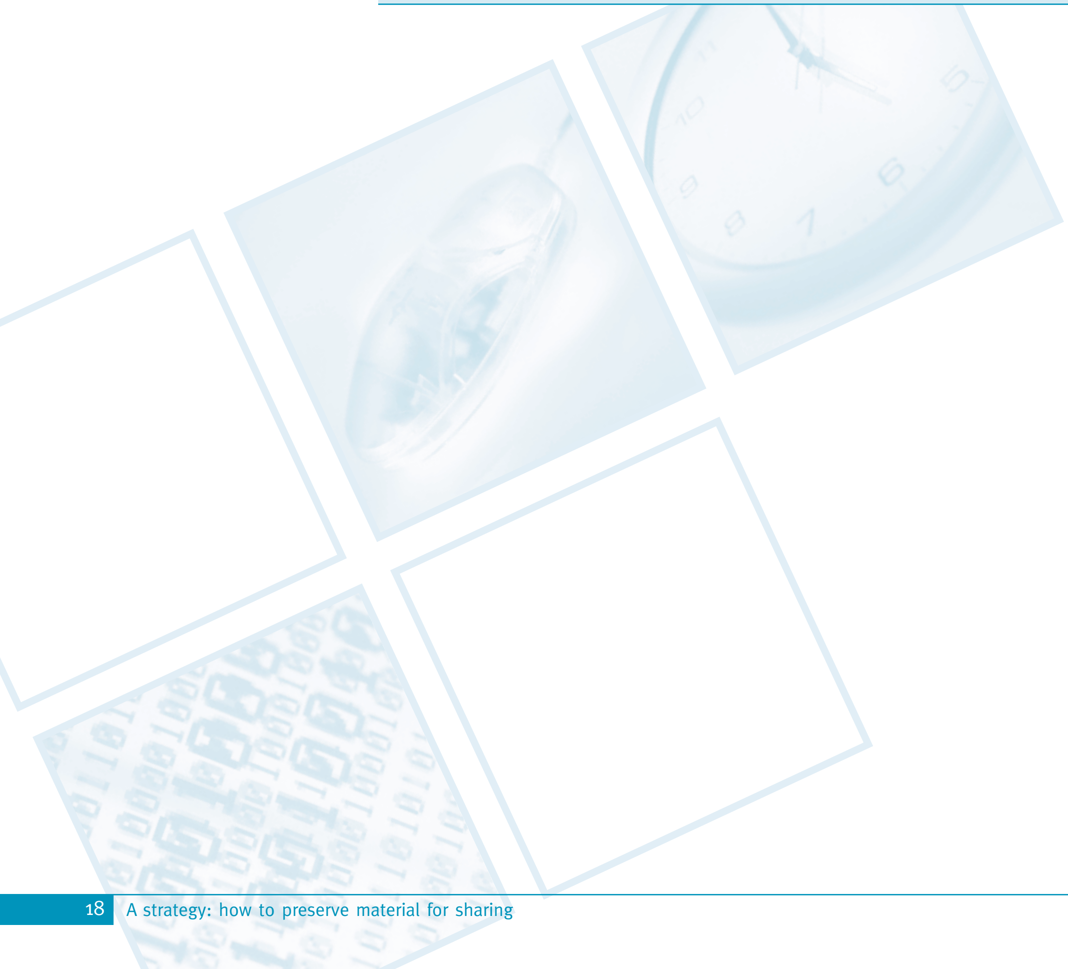
3.3 Identification of sites for preservation and dissemination

3.3 Identification of sites for preservation and dissemination

Any decision to disseminate a statistical dataset or to deposit it for re-use with a specialised agency must be based on a clear understanding of data ownership and the responsibilities that come with it. When contracts with outside agencies are drawn up for rendering preservation and/or dissemination services, it is crucially important to determine who should have the responsibility for preserving and disseminating a dataset. The responsibility may lie with the data creator alone. It may be shared with a funding agency, or it may fall into the policy framework of the peer institution of the research project (see also section 4.2 below).

Data archives exist in public and private, as well as academic, sectors. They have built up considerable expertise in technical aspects of long-term digital preservation and many of them have established secure systems for legitimate dissemination of research data. Please refer to Appendix 2 for further details.

Approaching the broader organisational issues of preserving and sharing statistical data and establishing policy guidelines to ensure awareness of preservation needs at all stages of the data management should lead to a solid and systematic foundation for mastering the technical issues linked with long-term digital preservation.



4. Overcoming the barriers to preservation

As currently used, the term digital preservation means ensuring the usability of a digital resource through changing technological regimes with a minimum loss of the resource's intellectual content.[2] Although the importance of technology cannot be underestimated in the preservation of digital datasets, previous sections of this document have already indicated that it is not solely a technical issue. It also has organisational and strategic aspects, which play equally crucial roles. This section outlines how to deal with these aspects.

4.1. Data currency

It is not unusual for data producers to be justifiably reticent about making their material more widely available because they wish to exploit the full potential of the data themselves. This can apply equally to an academic, who has had a research grant to collect the material or to a commercial company with a clear commercial interest in maintaining control of the contents of its files. In both cases, there may be possible solutions for sharing the data. These can include the creation of public use files or the application of legal restrictions whereby potential users agree to use or not to use the material for specific purposes. Similarly, conditions can be imposed requiring users not to publish results without prior agreement from the data creator. The UKDA and its major funder, the Economic and Social Research Council (ESRC), offer extensive information on the copyright protection of data and both use conditional agreements between depositors and users of data.[3]

Readers should note that neither the UKDA nor the RSS are arguing that all data should be freely-accessible in terms of cost. Instead, they advocate the establishment of systems that will permit more open access to data.

4.2. Ownership of data and copyright

On occasions when preservation is to be undertaken by an agent or a third party preservation service provider, there may be difficulties in deciding who has the right to deposit a dataset. This is most likely to occur where a study has been undertaken across a number of sites or organisations, where the collection has been undertaken on behalf of another organisation or when the origins of the material are not sufficiently well recorded. Destination organisations are often able to offer advice on this issue. It is fairly safe to assume that the legal copyright holder/s or his/her representative has the right to pass on data for further use. When an employee has produced the work as part of their employment, the employer is entitled to hold copyright. There are occasions, however, when this may not be the case and anyone in doubt is advised to seek legal advice. Where an organisation has been commissioned to produce material, the commissioning organisation will not automatically hold copyright unless an assignment of copyright has been made.

Moral rights of authors and owners of digital material remain a problematic area in the world of digital information: the right to be named as author, protection against misattribution and protection against corruption of the work are, as yet, not very well

4.2. Ownership of data and copyright

4.3. Confidentiality, disclosure, ethics and data protection

4.4. Technical barriers

covered in legislation. A comprehensive guide to legal and copyright issues for social science research projects has been prepared for the ESRC.[4] Agencies that disseminate research data usually require the data creator or data depositor to ensure that all necessary copyright permissions have been cleared prior to deposition and that ownership of the materials has been established. In case of uncertainty as to whom copyright should be assigned, we suggest contacting the Registrar or Copyright Officer in your institution or organisation for advice.

4.3. Confidentiality, disclosure, ethics and data protection

The RSS and the UKDA are both aware of the importance of this issue for those who may want to preserve and share their material. Detailed consideration of the issues lies outside the scope of this document and we recognise that legal and ethical responsibilities are paramount. Nevertheless, we are convinced that while challenging, these constraints are not insurmountable in the correct ethical framework and with appropriate informed consent. The UK Information Commissioner's Office[5] offers advice on principles of good information handling, data protection and freedom of information legislation. A special section in Appendix 3 has references to printed and on-line materials related to data protection and copyright issues.

Where consent has been gained for the sharing of anonymised data, every effort must be made to ensure that identifying information is not disclosed. A number of web sites already exist to demonstrate the effectiveness of the measures and methods available.

Three possible controls are:

- anonymising material prior to preservation and sharing using methods such as perturbation;
- the restriction of access by requiring users to sign up to legally binding conditions of use;
- technological controls which prevent unauthorised users from accessing sensitive material.

In some cases, only aggregated data should be made available and, in all cases, the anticipation of sharing and the request for informed consent is likely to ease the difficulties. A further possibility is the creation of public use datasets from which sensitive information is suppressed.

Part IV of Schedule 8 of the 1998 Data Protection Act lists conditions under which collected data are exempt from data protection clauses for the purposes of research. The data custodian must, nevertheless, ensure that personal data are not passed to a country without legal protection for personal data equivalent to that in the UK, unless the custodians first assure themselves that the data will be adequately protected in practice.

4.4. Technical barriers

The nature of digital technology dictates that it is not feasible to simply hand over stewardship of a digital resource at some point in the future, without having actively managed it sufficiently to maintain its usability. Because of the ever-changing technology needed for their storage and use, digital data are susceptible to two types of threats, physical and logical loss of accessibility.

4.4. Technical barriers

The physical medium on which digital data are stored can be damaged or degrade or become obsolete over time. Digital data and the storage media must be checked for readability and accessibility at regular intervals and, if necessary, copied to new storage media to prevent their loss.

More fundamentally, digital information is designed to be interpreted by computer programs that render it understandable to human users and are, by their nature, software-dependent. The logical format of the data is hence endangered by loss and obsolescence of the hardware and software environment on which access to data depends. It is also endangered through loss of the metadata that describes how to access it. Three main strategies are in wide use for preserving the usability of digital material over time.[6]

- Technology preservation takes the approach of preserving both the original data and the computer platform necessary to interpret them. This does not necessarily guarantee future usability, unless the particular technology on which they depend is also preserved and can be made to work over time. This strategy of creating what is essentially a computer museum can only be regarded as viable for the short to medium term, as a relatively desperate measure in cases where valuable or very complicated digital resources cannot be converted to hardware and/or software independent formats and migrated forward.
- Technology emulation involves developing emulator programs on current and future computer platforms and programming them to mimic the behaviour of old hardware platforms and to emulate specific operating system software. This is a specialist strategy that requires extremely detailed specifications for the outdated computer platforms and is used only where there is a specific need to maintain the look and feel of the original digital resource. Technology emulation as a preservation strategy is also very much dependent on the technical ability of the software engineers and does not entirely avert the risk of having to convert the resource at some time in the future.
- Data migration is the most widely used preservation strategy at present. It relies on conversion from one file format and/or computer platform to another, more accessible one. Despite the backward compatibility that many application software packages offer, and interoperability between rival popular software programs, a safer option is conversion to standard formats that most software is capable of interpreting and that are suitable for data interchange. The aim of the strategy is to migrate digital resources from the multiplicity of formats used to a smaller, more manageable number of standard formats that can still encode the complexity of structure and form of the original. The formats chosen for conversion will be determined by the structure of the digital resources themselves, the objectives set by the collection policy and users' requirements. For example, the decision will be influenced by whether priority is given to preserving the ability to process the content of the digital resource, or to preserving its format or visual presentation. Data migration always carries a potential danger of loss of information since conversion from one format to another will always be a translation during which some interpretation of data will take place. Data migration is also likely to be a continuous process, with material transferred to a series of new formats over time, and thus potentially an expensive option.

4.4. Technical barriers

4.5. Costs

4.6. Establishing exit strategies to deal with organisational change

In the past one problem has been file sizes and the amount of data that can be stored on, and handled by, particular devices. This is no longer a significant problem and the associated costs of storage space have reduced over recent years.

Perhaps the biggest problem lies in deciding which are the appropriate formats for the material being preserved and the timing of conversion decisions. This is an area of on-going research and development involving computer scientists, those responsible for preserving statistical material and data providers.

As with other complex issues, specialist organisations are always willing to share their expertise to find the most appropriate preservation strategy and format. Lists of specialist services and further reading are included in Appendices 2 and 3.

4.5. Costs

Digital preservation is essentially about preserving usability of and access to a digital resource over time. This makes it difficult to neatly segregate costs that relate solely to digital preservation from costs for other tasks and aspects of digital resource management. Integrated cost models have been developed to define costs at all of the various stages of the life-cycle of the digital resource, of which preservation is one. Clearly, the cost of preservation will vary, depending on the strategy that is adopted for digital preservation. In most cases it is likely to be higher at the initial stage of creating preservation copies of resources. After a period of low-cost maintenance, costs can rise again when format conversions or new emulators become necessary due to technology obsolescence.

Costs are a commonly stated reason for the failure to preserve and share statistical material. Whilst costs will be necessarily incurred when preservation is undertaken, we argue that these are likely to be significantly greater when the work is undertaken as a task completely separate from the collection process. The costs will be minimised if, from the outset, organisations plan to preserve and share their material. When this route is taken, procedures can be implemented as early as the initiation stage, which will ensure that necessary information is collected and collated.

Experts in preservation are available for advice on planning and preparation for archiving. Agencies exist that offer specialist digital preservation services, while the data creator maintains all ownership rights for the data. References to publications discussing costs involved in digital preservation can be found in Appendix 3.

4.6. Establishing exit strategies to deal with organisational change

A key factor in preventing inadvertent loss of material is the establishment of an organisational exit strategy to ensure that valuable material is not lost as a result of organisational change, mergers or the disintegration of project groups. In these circumstances, the preservation of important material may rely on nothing more than serendipity because the ownership of statistical material and the continuing responsibility for it, may be unclear. The adoption of a code of best practice and the establishment of an exit strategy allow for the systematic transfer of material to a pre-arranged place of preservation. If necessary, a data custodian should be designated. This custodian should be charged both with responsibility for safeguarding the data until their deposit with a specialising agency, and for ensuring that information is treated in confidence where relevant.

4.7. A role for funding bodies

Research funding bodies have a crucial role to play in the implementation of codes of best practice and the adoption of the re-use principle for statistical material. Large sums of money are spent on research, which requires the collection of data. Much of the material gathered could be made more widely and easily available for secondary analysis. Funding bodies are urged to make it a contractual requirement that material collected under their auspices is offered for secondary use and some already do this. For example, the Economic and Social Research Council (ESRC) and the Natural Environment Research Council (NERC) have established firm policies and guidelines on the preservation of electronic material. They require that grant recipients are willing and committed to either preserve data themselves or, following a discussion on the need for preservation, to make arrangements with an appropriate data centre to take responsibility for their data at a specified point in time.

There are a number of models under which data can be made available under given circumstances. The model chosen will reflect the type of data and their content as well as the interests of, and the resources available to, particular funding organisations. Nevertheless, we urge these organisations to implement a positive strategy for data preservation and sharing, as soon as is practicable.

Decisions about preservation of digital resources cannot wait until continued use of the materials has proved they are worth keeping. Postponing preservation decisions can, and most often will, result in preservation actions that are more complex, more labour intensive, and more costly. A resource can be held hostage by an obsolete piece of software if it is not converted to a new format at the right time. There is little likelihood of digital information surviving through benign neglect. Technical as well as legal obstacles to preservation can be overcome when considered strategic decisions are correctly timed, however. Preservation requires active management that begins at the time of creation of the material and depends on a pro-active approach and co-operation of all the stakeholders involved including funders, data creators, users and digital repositories.

5. Conclusion

The main aim of this document has been to raise awareness of the value of preserving and sharing statistical material. While arguing for benefits arising from re-use of statistical data, it identifies several difficulties that seemingly prevent organisations from ensuring that their statistical material is preserved in the long-term. These include:

- lack of well-defined selection criteria for collecting and preserving statistical resources;
- data ownership, copyright and confidentiality rules;
- inadequate data documentation and metadata;
- technological constraints;
- perceived costs.

This document has made some suggestions about how to overcome or reduce these obstacles and has offered a strategic approach for considering all of the issues involved in a systematic way. Organisations should implement a system of best practice that will ensure, from the outset of the data collection process, that the data gathered, along with contextual information, will be preserved and shared.

As a conclusion, we recommend a short checklist for establishing the organisational framework for the long-term preservation of your data:

- make a positive decision to keep and share your data;
- develop and implement a code of best practice that takes into account long-term data preservation needs;
- keep adequate metadata;
- decide on a comprehensive collection policy;
- choose an appropriate preservation strategy;
- clarify data ownership, copyright, access rights to data and confidentiality circumstances;
- seek professional help and consultation from specialist agencies;
- deposit your materials with institutions that specialise in dissemination of research data to the scholarly community.

References to such institutions and to publications on all issues discussed in this booklet can be found in Appendices 2 and 3.

Endnotes

1. See http://www.legislation.hmso.gov.uk/acts/acts1998/19980029.htm

2. For example, M. Feeney (ed.), Digital Culture: Maximising the Nation's Investment, (National Preservation Office, 1999), p. 41

3. See http://www.esrc.ac.uk/esrccontent/DownloadDocs/wwwcopyrightandconfidentiality.doc and http://www.data-archive.ac.uk/creatingData/legal.asp

4. See http://www.esrc.ac.uk/esrccontent/DownloadDocs/wwwcopyrightandconfidentiality.doc

5. See http://www.dataprotection.gov.uk

6. cf. M. Feeney (ed.), Digital Culture: Maximising the Nation's Investment, (National Preservation Office, 1999)

Appendix 1. The Royal Statistical Society Code of Best Practice on the Preservation and Sharing of Statistical Material

This Code sets out a number of data stewardship and data preservation principles which corporate and individual members of the Royal Statistical Society are encouraged to adopt and uphold in order to encourage the widest possible use, re-use and exploitation of their datasets by their own and future generations. It is compatible with, and complements, the RSS's main Code of Conduct.

Explanatory notes:

1. The following principles discuss data 'providers' and 'users'. In many cases, they will be the same teams, and data may be offered to external or secondary users on a collaborative rather than service provision basis.

2. The principles are based on the premise that effort expended on preservation is recouped through re-use. In some circumstances preservation might be a sufficient goal.

3. Some people may feel that many of these principles should carry an implicit caveat, 'providing it does not jeopardise the core goals of the data collection.' Our contention is that, within the correct framework, adoption of these principles will enhance and not detract from the ability to deliver on those goals.

The Code is based on a recognition that statistical data are a valuable resource, are often produced at considerable expense, can be regarded as non-renewable, may be irreplaceable if neglected, damaged or lost, and can only realise their maximum value if exposed to widespread and long-term use. It is also based on the fundamental belief that archiving should be viewed as an integral part of the whole data management process rather than an isolated activity. In other words, archiving decisions should be undertaken right at the beginning, and throughout every subsequent stage, of the information life-cycle rather than left until the end.

The RSS Code should be viewed, therefore, as a set of guiding principles covering data stewardship in general, as well as a code of practice, which relates specifically to data archiving per se.

Key principles

Framework

Organisations involved in the business of collecting, compiling, processing, analysing, interpreting or disseminating statistics should appoint a 'champion' at senior management or board level with responsibility for ensuring that the organisation:

- provides policy guidelines which specifically address the issue of data retention and preservation for their members or employees;
- sets aside sufficient resources to implement this policy;
- establishes links and maintains liaison with other data creators and custodians who have experience and expertise in data preservation.

Organisation and planning

Individuals involved in the business of collecting, compiling, processing, analysing, interpreting or disseminating statistics should ensure, at a very early stage, that any arrangements they make for planning and organising a data collection exercise should incorporate:

- the data preservation imperative;
- their parent organisation's policy on data preservation;
- the likely costs of archiving their data.

Thereafter, they should keep these requirements at the forefront of any subsequent data management decisions.

Data management

Organisations and individuals, both, should ensure that arrangements are made to cover each of the following aspects of information management. All of which will help to maximise the current and long-term value of any dataset they preserve for posterity:

- ***consultation*** – from the outset, establish the views and needs of peers, providers and potential users in order to ascertain whether there is a genuine business justification for any proposed data collection exercise and no possibility of duplication of effort. Thereafter, be open with peers, providers and users and be responsive to their needs at every subsequent stage of the data life-cycle;

- ***professionalism*** – manage the processes of collection, compilation, storage and dissemination in a way that best meets user requirements for coverage, relevance, accuracy, timeliness, comparability, consistency, coherence and disaggregation;

- ***cost effectiveness*** – reconcile the aims of minimising the costs to providers and maximising the benefits to users with a view to achieving the best possible value within the technology and resources available;

- ***documentation*** – fully document all the various steps in the data life-cycle, and retain and store any contextual and allied material, in order to preserve knowledge and expertise about the data system and its processes and outputs, maintain their functionality, and facilitate any subsequent data audits;

- ***metadata management*** – provide easily accessible and comprehensive guidance material and interpretative text in order to foster awareness and understanding of the dataset and limit the possibility of misinterpretation;

- ***statutory compliance*** – fulfil all relevant statutory or legislative obligations in order to ensure the security, confidentiality, legality, acceptability, and usability of their dataset, and safeguard any intellectual property rights attached thereto;

- ***quality assurance*** – ensure that all the various stages in the data life-cycle meet appropriate methodological criteria, quality assurance standards and change control requirements in order to guarantee the statistical integrity of the dataset;

- *harmonisation* – comply with national and international protocols, agreements, understandings, classifications and codes of practice relating to harmonisation, standardisation and integration in order to ensure full inter-operability of datasets;

- *accessibility* – disseminate the data and accompanying metadata/documentation using formats and access routes which will ensure that they reach the widest possible number of authorised users;

- *review* – subject any regular or continuous data collection exercise to systematic and periodic reviews in order to maintain their relevance, reliability and fitness for purpose;

- *risk and destruction control* – take steps to ensure that the data are not inadvertently put at risk or corrupted at any stage in the life-cycle and destroy data only where it can be shown to providers, users and archiving bodies that the intrinsic value of the data is insufficient to justify the costs of preservation and retention;

- *preservation* – where archiving is merited, choose whichever managerial, organisational and methodological arrangements which are most likely to guarantee the long-term preservation and re-usability of a dataset, taking into account the RSS's own recommended standards relating to the preparation of data in readiness for its preservation.

Appendix 2. List of expert organisations in the field of data preservation and sharing

This alphabetical list has been collated with the intention of providing readers with contact information for a number of organisations that offer information or advice on the archiving and long-term preservation of primarily electronic material. Several of these institutions also act as distributors of statistical and other categories of data to the research community and wider public.

The list does not include organisations simply because they contain references to archiving in their titles or accompanying literature. Many such sites exist but often function as a collection point for materials of a particular type or covering a particular disciplinary area. Sites that provide on-line access to their material via databases will have efficient systems of internal backup to prevent information loss for their own immediate purposes. This list includes specialist archival agencies which demonstrate systematic procedures for the long-term preservation of their material in files that are both hardware and software independent.

Where possible, regulations for depositing data with these institutions have been referenced here, but further information and guidance on preparing data for deposit is provided in the bibliography in Appendix 3.

Arts and Humanities Data Service (AHDS)

75-79 York Road, 8th Floor, King's College London, Waterloo, London SE1 7AW
w: http://ahds.ac.uk **t:** +44 (0)20 7928 7371

The mission of the AHDS is to work on behalf of the academic community to collect, catalogue, manage, preserve and promote the re-use of scholarly digital resources. It is established as a geographically-distributed service comprising a managing executive and five service providers devoted to collecting and preserving a range of digital material under the broad heading 'arts and humanities'. The service providers specialise in archaeology, history (literary, linguistic and other textual studies), the visual arts, and the performing arts. Links to these can be found on the central web site referred to above. This page also provides access to all AHDS data via a common catalogue gateway.

AHDS promotes awareness of the importance and value of digital information and provides guidance in its effective creation and use through a series of 'good practice' guides as well as advice (both free and at cost).

The five AHDS service providers all have clearly outlined collection policies and well-defined requirements for the deposit of data.

Edinburgh Data and Information Access (EDINA)

Edinburgh University Data Library, Main Library Building, George Square, Edinburgh EH8 9LJ
w: http://www.edina.ac.uk **t:** +44 (0)131 650 3302 **e:** edina@ed.ac.uk

EDINA provides national on-line services for the UK tertiary education and research community. All EDINA services are available free of charge to members of UK tertiary education institutions for academic use, although institutional subscription and end-user registrations are required for most services. EDINA carries out a number of development projects as part of the JISC's programme to build the DNER for UK tertiary education.

Manchester Information and Associated Services (MIMAS)

MIMAS, Manchester Computing, Kilburn Building, University of Manchester, Oxford Road,
Manchester M13 9PL **w:** http://www.mimas.ac.uk **t:** +44 (0)161 275 6109 **e:** info@mimas.ac.uk

MIMAS at the University of Manchester provides networked access to key data and information resources to support teaching, learning and research across a wide range of disciplines.

The British Library National Sound Archive (NSA)

The Recorded Sound Information Service, The British Library National Sound Archive
96 Euston Road, London NW1 2DB **w:** http://www.bl.uk/collections/sound-archive/nsa.html
t: +44 (0)20 7412 7440 **e:** mailto:nsa@bl.uk

The NSA's catalogue of sound recordings can be accessed on-line. This includes information on two-and-a-half million published and unpublished recordings. The NSA also conducts laboratory tests on optical discs, digital audio tape and other sound storing formats with a view to establishing the best archival medium for the future. The NSA is able to advise, free of charge, on the storage, cataloguing and restoration of most types of sound recordings. It played a key role in the development of CEDAR (computer enhanced digital audio restoration) technology now used by commercial record companies to retrieve sound from obsolete or damaged sound carriers. The web site provides links to related sites worldwide.

The organisation can provide general information as well as consultation for those who may wish to preserve their own material. It also accepts material into the Archive as long as it falls within the criteria of their acquisitions policy.

The Public Record Office (PRO)

Head of Records Management Department, Public Record Office, Ruskin Avenue
Kew, Surrey TW9 4DU **w:** http://www.pro.gov.uk **w:** http://ndad.ulcc.ac.uk (NDAD)
t: +44 (0)1684 585299 **e:** enquiry@pro.gov.uk (general)
e: records-management@pro.gov.uk (records selection)
e: archive-inspection@pro.gov.uk (policy)
e: support@ndad.ulcc.ac.uk (NDAD)

The PRO is the repository of the national archives. It was founded by act of Parliament in 1838 to preserve and make available the records of central government and the courts of law. Today, the Public Record Office advises government departments on best practice in records management as well as selecting those which will be kept in perpetuity.

The PRO Records Management Department (RMD), formerly Government Services Department (GSD), provides advice and guidance to government departments and other public record bodies on the management of records, and their selection and transfer to the PRO. Besides several policy documents and other initiatives, it has produced comprehensive Guidelines on the Management, Appraisal and Preservation of Electronic Records through its Electronic Records from Office Systems (EROS) programme.

The PRO also operates an archive service for public records that take the form of computer datasets which originated in central government departments and agencies. The University of London Computing Centre and the University of London Library host this service, the UK National Digital Archive of Datasets (NDAD), on behalf of the PRO. NDAD holdings comprise data scheduled for preservation in accordance with the PRO's collection policy. It provides open access to the catalogues of all its holdings, and free access to open datasets following a simple registration process. The data in the transfer procedures and requirements are listed in the NDAD Contributors section of their web pages.

UK Data Archive (UKDA)

UK Data Archive, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ
w: http://www.data-archive.ac.uk **t:** +44 (0)1206 872001
e: archive@essex.ac.uk **e:** quali@essex.ac.uk (Qualidata enquires)

The UKDA is a specialist national resource containing the largest collection of accessible computer-readable data in the social sciences and humanities in the United Kingdom. Founded in 1967 and funded by the ESRC, the University of Essex and the JISC, it has one of the longest records in preserving and providing access to digital material in the UK. Its holdings include data from many government departments as well as data resulting from academic research.

The UKDA is part of a worldwide network of data archives with which it has reciprocal data exchange agreements. It acquires, preserves and disseminates data in support of research, learning and teaching. Material held in the UK Data Archive is available in a variety of formats and on various media. For many studies there is also a free, direct download facility for registered users.

The web site provides detailed information, including file formats, for those depositing data with the Archive. There is a searchable catalogue for finding data, information on copyright and access conditions for users of the material in its holdings.

The UK Data Archive also incorporates Qualidata, which aims to facilitate and document the archiving of qualitative data arising from research, whilst also drawing the research communities' attention to its existence and potential. The Qualidata catalogue (Qualicat) is available on-line, from which researchers can search and obtain descriptions of qualitative research material, its location and accessibility.

The Qualidata web site also provides guidelines for archiving qualitative data and information on informed consent and copyright. It maintains links to an extensive list of sites at which qualitative material can be preserved.

Appendix 3. Annotated bibliography of selected documents relating to data preservation

The following, alphabetically listed, selection of documents is intended as an aid to organisations wishing to preserve and archive digital material. It is split into sections that reflect the topics covered in this booklet. References are given to legal acts and standards, guidelines and documents of broad interest and to some detailed discussions of specific issues. The last three sections offer references to web sites or information portals of a more technical nature, including, for example, information on standards initiatives for data interchange which are often a precursor to preservation activity.

Relevant general legislation

The following is a brief selection of legislative texts that may be useful for organisations that collect, process or store statistical material.

Copyright, Design and Patents Act 1988

http://www.legislation.hmso.gov.uk/acts/acts1988/Ukpga_19880048_en_1.htm

This Act governs the creation, ownership, transfer and exploitation of copyright, including database copyright. Broadly, an owner of copyright in a 'work' has the right to prevent others from copying it or otherwise exploiting it. Infringement of copyright can lead to legal action resulting in the infringer being ordered by the court to stop the infringement and/or being forced to pay damages. Infringement can also constitute a criminal offence. Use of a copyright work for certain purposes including research or private study may not amount to infringement where the use constitutes 'fair dealing' with the work; whether this will apply depends on the circumstances.

Data Protection Act 1998

http://www.dataprotection.gov.uk/

This Act implemented the European directive 46/96/EC which requires EU member states to protect the fundamental rights and freedoms of natural persons, in particular their right to privacy with respect to the processing of personal data. It strengthens the previous legislation of the 1984 Data Protection Act by extending the individual's say over what information is collected and how it is used. The Act includes eight data protection principles that define duties of people who collect and provide personal data. The Act came into force on 1 March 2000.

Draft Code of Practice on the Discharge of the Functions of Public Authorities, Under Part I of the Freedom of Information Act 2000

http://www.lcd.gov.uk/foi/dftcpOO.htm (The text of the Code of Practice)

This Code of Practice sets out good administrative practice for public authorities when handling requests for information. It protects the interests of applicants by setting out standards for the provision of advice.

Electronic Communications Act 2000

http://www.legislation.hmso.gov.uk/acts/acts2000/20000007.htm

This Act provides for legal recognition of electronic signatures and makes provisions to facilitate the use of electronic communications and electronic data storage.

Freedom of Information Act 2000

http://www.dataprotection.gov.uk/dpr/foi.nsf
http://www.legislation.hmso.gov.uk/acts/en/2000en36.htm (for the text of the Act)

The Act gives a general right of access to all types of 'recorded' information held by public authorities (and those providing services for them), sets out exemptions from that right and places a number of obligations on public authorities. Only public authorities are covered by the Act.

Public Records Acts 1958 and 1967

http://www.pro.gov.uk/about/act/default.htm

These Acts cover records of government departments and any other body established under Acts of Parliament to discharge public functions. The Public Record Office expects certain types of records produced by statutory agencies to be transferred to their custody once the agencies no longer require them for their ongoing business needs.

Standards

The following standards are relevant to managing and preserving digital data.

- *DISC PD 0008:1999* Code of Practice for Legal Admissibility of Information Stored on Electronic Document Management Systems, BSI, 1999
- *BS 5454:2000* Recommendations for the storage and exhibition of archival documents, BSI, 2000
- *ISO 15489-1* Information and documentation. Records management. General, ISO, 2001
- *ISO/TR 15489-2* Information and documentation. Records management. Guidelines, ISO, 2001
- *ISAD(G)* General International Standard on Archival Description, ICA, 2000
- *ISAAR(CPF)* International Standard Archival Authority Record for Corporate Bodies, Persons and Families, ICA, 1996
- *IS0 11179-3* Information Technology – Specification and standardisation of data elements, ISO, 1994

Policies and guidelines

This potentially very long list of references points to a few policy documents that organisations could refer to or use as examples when considering their own preservation and data management policies. A useful starting point for this is also the RSS Code of Best Practice (see Appendix 1).

English Heritage Centre for Archaeology, *Digital Archiving Strategy*

http://www.english-heritage.org.uk/knowledge/archaeology/digitalarchiving.asp

This document describes the strategy which the Centre for Archaeology has developed for managing its digital archives. The strategy is intended to establish best practice for the preservation of, and provision of access to, the full range of digital archaeological data of long-term value for which the Centre for Archaeology is responsible.

Mirjam M. Foot, *Building blocks for a preservation policy,* (National Preservation Office, 2001).

http://www.bl.uk/services/preservation/npo8.pdf

This publication of the National Preservation Office explains the need for a preservation policy for an organisation and gives guidelines for developing one.

Public Record Office, *Corporate policy on electronic records,* (Public Record Office, 2000).

http://www.pro.gov.uk/recordsmanagement/eros/RMCorpPol.pdf

A detailed guide to developing policy documents, including a preservation policy, and how to implement and audit policies.

Society of Archivists, *Best Practice Guideline 4: preservation and conservation,* (Society of Archivists, 1997).

http://www.archives.org.uk/publications/inprint.asp#bestpractice (for information)

A comprehensive guide to setting up and maintaining an archivally-sound preservation system. Includes references to relevant standards and examples of preservation and conservation policies.

UK Data Archive, *The UK Data Archive preservation policy,* (Colchester, 2002).

http://www.data-archive.ac.uk/depositingData/ukdaPreservationPolicy.doc

This document details the UKDA's preservation policy and describes the actions it takes to safeguard the continuing accessibility of the digital resources that the Archive has accepted for preservation.

Data management and evaluation

This section includes resources that point out vital issues of managing digital resources and the risks that are involved in this from the long-term preservation point of view.

Arts and Humanities Data Service, *Creating a viable data resource*

http://ahds.ac.uk/viable.htm

This text aims to inform researchers about steps they can take to ensure that data resources they create today remain accessible in the future. It includes a check-list for evaluating the importance of a resource's viability and assessing long-term value of a dataset.

Arts and Humanities Data Service, *Managing digital collections*

http://ahds.ac.uk/managing.htm

The web site above offers a collection of links to various AHDS documents discussing the issues of managing collections of scholarly data and metadata. The series is primarily aimed at curators and other information managers of digital collections. The AHDS is also publishing the Guides to Good Practice series (16 titles) and AHDS Information Leaflets to provide the humanities research and teaching communities with practical instruction in applying recognised standards and good practice to the creation and use of digital resources. Together these publications provide a complementary overview of standards and best practice for the creation, management and use of digital collections in the arts and humanities. Users of this bibliography are advised to visit this web site for the most up to date material.

DLM Forum, *Guidelines on best practices for using electronic information. How to deal with machine-readable data and electronic documents,* (Luxembourg, 1997)

http://europa.eu.int/ISPO/dlm/documents/guidelines.html
(English version, also available in French and German)

The guidelines are the result of an EU-funded forum to discuss the growing need to consider the impact of the increase of electronic documents and databases on current practice and rules for document use and archiving. The guidelines are multidisciplinary and whilst not claiming to answer all possible questions, the guidelines give examples of best practice and offer advice to help organisations define their own strategy for electronic information.

Inter-University Consortium for Political and Social Research, *Guide to social science data preparation and archiving*

http://www.icpsr.umich.edu/ACCESS/dpm.html

The Inter-University Consortium for Political and Social Research (ICPSR), University of Michigan, USA has produced this guide and aimed it at those engaged in the task of preparing data for deposit in a public archive, but most of the topics covered are of interest to anyone creating a dataset. Intended to be an advisory/informative document rather than a prescriptive/ regulatory document, the guidance offered is sound and universal, and the manner of its offering so persuasive that it could be accepted as a 'mandatory' standard. It provides advice and guidance on how to manage a data collection project from the outset and would be of use to data collectors/administrators, data managers, data depositors, archivists and researchers. The material focuses on survey datasets but the issues covered are equally applicable to other data types – qualitative, administrative etc. It also includes a useful glossary of relevant terms.

Public Record Office, *Guidelines on the management and appraisal of electronic records*

http://www.pro.gov.uk/recordsmanagement/eros/guidelines/default.htm

This document was written by the UK Public Record Office, in response to the public sector need for advice in managing electronic records in an office environment. It offers guidance on the core design requirements for IT systems to support electronic document management (EDM) and the requirements for the transfer of records in electronic form to the PRO. The principles and practice are applicable to wider management applications of EDM & ERM (Electronic Records Management) and wide use and dissemination of the material is encouraged. The document is designed for managers of government business, Departmental Records Officers, heads of IT, strategy and planning managers, project managers and PRO Inspection and Documentation Officers in public sector organisations. It offers an extremely useful, practical approach to records management, in particular it impresses in its recognition of the problems associated with resources and legacy material.

UK Data Archive, *Good practice in data documentation*

http://www.data-archive.ac.uk/creatingData/goodPractice.doc

In addition to enabling the long-term preservation process, good documentation also ensures that the research community can re-use collected data to the full. This good practice guide explains both the value of documenting the data and how best to create and keep them.

Depositing data

This section includes general texts about where data can be deposited and how they should be prepared prior to depositing. URLs of detailed data deposit requirements of several UK data collecting institutions are also given in this section.

Arts and Humanities Data Service, *How to deposit digital data with the AHDS*

http://ahds.ac.uk/depositors.htm

This web site describes the benefits of depositing data with the AHDS and how this can be done. Each of the five AHDS service providers has its own collection policy and data deposit rules (links to these are at http://ahds.ac.uk/dephow.htm). Contact details for help and advice as well as information on rights management can be found at http://ahds.ac.uk/depositors.htm

National Digital Archive of Datasets, *Transfer procedures*

http://ndad.ulcc.ac.uk/contributors/

This is a prescriptive document, as one would expect, given that NDAD preserves computer-based datasets from UK government departments and agencies on behalf of the Public Record Office. This series of documents has been written for archivists, government departments, records officers and IT staff concerned with the statutory preservation of the material described. Some of the information provided is in the form of instructions but other information is more about guidance and advice.

National Library of Australia, Digital Services Project. *Request for tender – digital collection management system*

http://www.nla.gov.au/dsp/rft/

The NLA does not propose this as a model but it is very instructive to any other organisation contemplating developing the infrastructure for managing and preserving digital collections. The draft contract at Attachment 1 clearly includes elements specific to Australian Government requirements but also includes many generic elements applicable to similar organisations in any country.

Susan Shaw, *Beyond the PRO: public records in places of deposit. Guidance about preserving and making available to users public records kept outside the Public Record Office,* (Public Record Office, 1994).

A comprehensive guide to legal aspects of depositing public records, as well as requirements that depositories have to fulfil.

UK Data Archive, *How to deposit?*

http://www.data-archive.ac.uk/depositingData/howtoDeposit.asp

The UKDA offers a set of guidelines for data depositors. Although the web page mentions only social science digital resources, the same deposit rules apply for every kind of scholarly data. The information in these pages explains why one should consider depositing a dataset, what one has to do to deposit a dataset, the legal aspects, and what happens to the dataset once it is lodged with the UKDA.

UK Data Archive/Qualidata, *Guidelines for depositing qualitative data*

http://www.qualidata.essex.ac.uk/depositingData/introduction.asp

This set of web pages is intended as a guide for those who are considering preparing qualitative data and any associated contextual documentation for deposit in a repository via Qualidata. The benefits of and processes involved in depositing material are also discussed.

Copyright and confidentiality

There follows a short list of guidelines that outline main issues of copyright and confidentiality when depositing datasets with archival agencies.

Arts and Humanities Data Service, *Copyright FAQ*

http://ahds.ac.uk/copyrightfaq.htm

A brief list of questions and answers that the AHDS has compiled for data collectors and depositors.

Economic and Social Research Council, *Guidelines on copyright and confidentiality: legal issues for social science researchers*

http://www.esrc.ac.uk/esrccontent/downloaddocs/wwwcopyrightandconfidentiality.doc

This guide introduces legal issues that are relevant to most social science researchers in some degree. It mainly concerns copyright and confidentiality and refers to defamation and data protection. The purpose of the guide is to help researchers deal with the often complex legal framework surrounding the collection, dissemination and use of data. Many researchers will recognise that the guidelines have much in common with statements of best research practice that can be obtained from learned societies and other bodies.

Medical Research Council, *Personal information in medical research,* (Medical Research Council, 2000)

http://www.mrc.ac.uk/pdf-pimr.pdf

This document provides guidance on the use of personal information, together with legal and ethical principles that govern the use of such data within medical research. The maintenance of safeguards over confidentiality throughout the collection, use and re-use of medical data are also discussed.

The Department of Health

http://www.doh.gov.uk/ipu/confiden/index.htm

The Department of Health web site includes a series of pages devoted to issues relating to patient confidentiality. Included on the site are the following documents:

HSG (96)18/LASSL(96)5 - Health Service Guideline - The Protection and Use of Patient Information;
The Protection and Use of Patient Information - Guidance from the Department of Health;
HSC 2000/009 – The Data Protection Act 1998: protection and use of patient information.

The Samples of Anonymised Records (SARs)

http://les.man.ac.uk/ccsr/cmu/index.htm

Provides information on a number of protective measures that were incorporated into the specification for the Sample of Anonymised Records from the recent UK censuses.

UK Data Archive, *The legal aspects of data creation and data deposit*

http://www.data-archive.ac.uk/depositingData/legalIssues.asp

An explanation of the legal issues connected with data depositing and how the data creator or data depositor can ensure that all necessary copyright permissions have been cleared prior to deposition.

UK Data Archive/Qualidata, *Confidentiality and informed consent*

http://www.qualidata.essex.ac.uk/creatingData/confidentiality.asp

This document provides links to helpful sets of guidelines on informed consent and respondent confidentiality for researchers and data providers involved in the collection of primarily qualitative data. It covers both ethical and legal aspects and specifically discusses the protection of individual confidentiality when material is being archived.

Data Protection

Some useful resources for compliance check with the regulations of the 1998 Data Protection Act.

BSI-DISC, *Data Protection Guides*

http://www.bsi-global.com/DISC/Working+Withyou/DataProtection+Guides.xalter

These guides give advice to organisations about how to ensure compliance with the Data Protection Act 1998 when managing their information processing operations.

Council of Europe, *Personal Data Protection*

http://www.legal.coe.int/dataprotection/Default.asp

This general web site contains European legal acts regarding data protection. It also includes studies into sensitive data and anonymised data.

Public Record Office, *Data Protection Act 1998 – A guide for records managers and archivists*

http://www.pro.gov.uk/recordsmanagement/dp/default.htm

Written in association with the Office of the Data Protection Commissioner, the guide offers a framework of issues to be addressed by records managers and archivists. It would also serve as a useful introduction to the Data Protection Act for individuals or organisations preparing data for despatch to a data repository.

Digitisation

This section includes discussions of benefits and pitfalls of digitising paper documents and how to manage document imaging projects.

Arts and Humanities Data Service (AHDS), *Digitisation: A project planning checklist*

http://ahds.ac.uk/checklist.htm

This short document offers practical guidance to those considering a digitisation project in the form of a check-list of strategic issues which need to be addressed in a project's design phase.

Paul Ayris, *Guidance for selecting materials for digitisation,* (RLG/NPO, 1999)

http://www.thames.rlg.org/preserv/joint/ayris.html

Discusses different approaches to starting a document digitisation project and offers a decision-making matrix to support selection activities in the digitisation process.

Higher Education Digitisation Service (HEDS)

http://heds.herts.ac.uk/

The HEDS not only advises its clients on how to digitise their material in the manner most fitting to their needs but also provides a full production service. This removes the need for organisations to procure their own equipment or to research every aspect of the digitisation process. The HEDS is a JISC service offering advice and provides complete digitisation services.

Steven Puglia, ‘The costs of digital imaging projects’, *RLG DigiNews,* vol. 3, no. 3, (1999)

http://www.rlg.ac.uk/preserv/diginews/diginews3-5.html#feature

Discusses the components that need to be considered when planning and budgeting for a digitisation project.

Maxine K. Sitts, (ed.), *Handbook for digital projects: a management tool for preservation and access,* (NDCC, Andover, Mass., 2000)

http://www.nedcc.org/digital/dighome.htm

The Northeast Document Conservation Centre (NDCC) is the largest regional conservation centre in the US that runs a successful series of scanning courses. The handbook provides a collection of tips, guidance and advice from institutions that have engaged in digital projects. It brings together best practices and summarises lessons learned from many experiences. The approach, while managerial, is also practical and based on actual projects.

Sean Townsend, Cressida Chappell, Oscar Struijvé, *Digitising history: a guide to creating digital resources from historical documents,* (Arts and Humanities Data Service, 1999)

http://hds.essex.ac.uk/g2gp/digitising_history/index.asp

This guide to creating, documenting and preserving digital resources is intended as a reference work for individuals and organisations involved with, or planning, the computerisation of historical source documents. It aims to recommend good practice and standards that are generic and relevant to a range of data creation situations. The guide focuses specifically on the creation of tabular data, which can be used in databases, spreadsheets or statistics packages.

Digital Preservation

This section includes references to introductory guides on digital preservation. Over the past decade, different scholarly communities have published many articles, studies and other texts on the subject of digital preservation. The concerns of the archival and library professions dominate this literature, and many recommendations given as little as five years ago are already out of date. However, the problem of guaranteeing the longevity of digital material is the same for everyone writing on the topic and an overview of the main problems and strategies for solving them, can be gleaned from most publications.

The UK Digital Preservation Coalition

http://www.jisc.ac.uk/dner/preservation/prescoalition.html

The UK Digital Preservation Coalition was established in 2001 to foster joint action to address the urgent challenges of securing the preservation of digital resources in the UK and to work with others internationally to secure our global digital memory and knowledge base.

Neil Beagrie, Maggie Jones, *Preservation management of digital materials: a handbook,* (British Library, 2001)

http://www.jisc.ac.uk/dner/preservation/workbook/

This handbook provides an internationally authoritative and practical guide to the subject of managing digital resources over time and the issues in sustaining access to them. It should be of interest to all those involved in the creation and management of digital materials.

Charles M. Dollar, *Authentic electronic records: strategies for long-term access,* (Cohasset Associates Inc., 1999)

http://www.rbarry.com/CSum1.html

Despite being primarily oriented towards the archives profession, this comprehensive text covers, in detail, strategies that are currently used for overcoming technological obsolescence – the main problem that hampers the long-term preservation of digital material: media renewal, data migration, emulation. The discussion of best practices of digital preservation is instructive and detailed, as are the more technical appendices.

Mary Feeney, *Digital culture: maximising the nation's investment,* (National Preservation Office/British Library, 1999)

http://www.bl.uk/services/preservation/report.html

A synthesis of NPO and JISC studies on the preservation of electronic materials, this document offers a good general introduction to the subject, discusses stakeholders and costs involved with the digital preservation and gives some recommendations for managing the digital preservation process.

Edward Higgs, (ed.), *History and electronic artefacts,* (Oxford University Press, 1998)

A collection of papers from a workshop held at the British Academy in 1993 where the main topic was the research community's concerns with the advancing information age. Articles in this book present views from creators and users of digital resources, and from technology, standards, as well as library and archives communities.

Museums and Galleries Commission, *The care of photographic materials and related media,* (MGC, 1998)

http://www.museums.gov.uk/books/pubs/intro.html

This publication describes, among other types of media, the preservation risks and hazards of magnetic and digital materials and offers advice on storage conditions and optimal care of these.

Seamus Ross, *Changing trains at Wigan: digital preservation and the future of scholarship,* (The British Library, 2000)

http://www.bl.uk/services/preservation/occpaper.pdf

Originally delivered as a paper at a digital preservation conference, this booklet examines the impact of the emerging digital landscape on long-term access to material created in digital form and its use for research. It identifies and examines challenges, risks and expectations.

Internet portals and information resources related to statistical data and metadata standards

The resources listed in this section offer information on standards and developments in metadata modelling as a means of describing data.

Digital Libraries: Metadata Resources

http://www.ifla.org/II/metadata.htm

An extensive list of on-line resources related to metadata from the International Federation of Library Associations and Institutions (IFLA).

W Grossman, Metadata archive and metadata tutorial

http://www.univie.ac.at/dac/projects/ismis.html

The compilation is part of an inventory about metadata-related material in statistics produced for EUROSTAT within the framework of a SUPCOM project. Publication is expected to be on CD-ROM and, at the time of writing, is imminent. For those interested in documenting electronic material using metadata, it is an invaluable resource, providing a look up list for a comprehensive collection of references to papers, publications and projects that focus on metadata issues. The initiative was financed by EUROSTAT.

Information Age Champions (IAG)

http://www.iagchampions.gov.uk/publications/frameworks_index.htm

IAG is an initiative to formulate e-government strategy which has as one of its key aims the provision of metadata as the main communication bridge between government and the citizen and of turning raw data into information. The IAG Working Group has, as one of its goals, the establishment of a government-wide metadata standard. This is being achieved through the establishment of a number of complementary working groups. One of these, on interoperability, will take forward an international drive to encourage greater integration of national information resources using communication technologies and transfer protocols such as eXtensible Mark-up Language (XML) and Z.39.50 and information structures such as the Resource Description Framework (RDF). As part of its remit, it has also produced a series of framework policies including one for electronic records management.

Metadata Watch

http://www.schemas-forum.org/metadata-watch/

The Metadata Watch is a broad and comprehensive list of projects, programmes, software tools and guidelines that use, or describe how to use, different metadata schemas.

MetaNet

http://www.epros.ed.ac.uk/metanet/resources/resources.htm

MetaNet is a European funded network of excellence for harmonising and synthesising the development of statistical metadata. It will develop proposals for standards in the methodology used for describing statistical metadata and statistical information systems and will make recommendations on the metadata objects in a common conceptual model of statistical metadata. It will then disseminate these proposed standards to the relevant user communities and standards bodies. The MetaNet web site includes a useful resource section that provides information on current standards.

StatBase

http://www.statistics.gov.uk/statbase/mainmenu.asp

The StatBase system is managed by National Statistics and sponsored by the thirty or more government departments that make up the Government Statistical Service (GSS). It is mainly concerned with the documentation of official statistical resources. It does, however, include a certain amount of complementary metadata material from non-governmental organisations. It incorporates four metadata templates. The first, with 47 metadata fields covers ‘Sources’, that is, any means of gathering data, for example, the decennial Population Census. The second has 32 metadata fields and covers ‘Analyses’ including the monthly Retail Price Index. The third covers ‘Products’ such as any Database, Publication, CD-ROM or Enquiry Service and covers 28 fields whilst the fourth, with 11 metadata fields, covers any associated ‘Contacts’. Of all these StatBase metadata fields, a total of 41 relate directly to the Dublin Core. The full StatBase metadata specification can be found in the ONS document, ‘StatBase – Metadata Assistant – A Guide to Populating Metadata Screens’. At the time of writing, it contains information relating to over 1,000 separate GSS ‘Sources’, over 100 ‘Analyses’ and nearly 2,000 ‘Products’ – more than half of which have been designated as ‘National Statistics’.

The Clinical Data Interchange Standards Consortium (CDISC)

http://www.cdisc.org

The CDISC is a multidisciplinary organisation committed to the development of industry standards to support the electronic acquisition, exchange, submission and archiving of clinical trial data and metadata for medical and biopharmaceutical product development. The aim of the group is to lead the development of global, vendor-neutral, platform independent standards to improve data quality and accelerate product development in the industry.

The National Geospatial Data Framework (NGDF)

http://www.ngdf.org.uk/Metadata/NGDFmeta.htm (Metadata Service Project)

The NGDF Gateway has been established to provide a central point of access over the internet to a wide variety of information (metadata) about spatially-related data owned by public and private sector organisations. NGDF recognises that the Gateway will not be the key impetus for agencies to collect metadata but instead will assist in educating the industry on the importance of, and issues related to, metadata. The NGDF Gateway is fundamentally linked to the availability of an organisation's metadata repositories. To facilitate linking, they are producing a set of guidelines for documenting data resources that are referenced in some way to the earth's surface.
The guidelines are based upon the draft ISO metadata standard 15046-15.

Internet portals and information resources related to digital preservation

The resources listed in this section offer information on digitisation for long-term preservation.

D-Lib Magazine

http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib.html

D-Lib Magazine is a journal of the Digital Libraries Initiative that regularly features discussions of current issues in digital preservation.

Preserving Access to Digital Information (PADI)

http://www.nla.gov.au/padi/

The National Library of Australia's PADI initiative aims to provide mechanisms that will help to ensure that information in digital form is managed with appropriate consideration for preservation and future access. The PADI web site is a subject gateway to digital preservation resources. It has an associated discussion list padiforum-l for the exchange of news and ideas about digital preservation issues.

The Research Libraries Group (RLG) DigiNews

http://www.rlg.org/preserv/diginews/

RLG DigiNews is a web-based newsletter with a focus on issues of particular interest and value to managers of digital initiatives with a preservation component or rationale. It provides filtered guidance and pointers to relevant projects to improve awareness of evolving practices in image conversion and digital archiving.

The Cedars project

http://www.leeds.ac.uk/cedars/
http://www.ukoln.ac.uk/metadata/cedars/ (For information on metadata development)

This project has as a major component of its work, the long-term development of a metadata framework, which will enable long-term preservation of digital materials. These metadata are required to support meaningful access to the archived digital content and includes descriptive, administrative, technical and legal information.

UK Office for Library Networking

http://www.ukoln.ac.uk

UKOLN is a national focus of expertise in digital information management. It provides policy, research and awareness services to the UK library, information and cultural heritage communities. UKOLN is based at the University of Bath.

Information on technical standards and standards initiatives

The following section offers links to sites providing information relating to standards in data interchange.

The Data Documentation Initiative (DDI)

http://www.icpsr.umich.edu/DDI/codebook/index.html

The DDI is an effort to establish an international criteria and methodology for the content, presentation, transport and preservation of metadata in the social and behavioural sciences. The DDI Document Type Definition (DTD) version 1 employs XML and was published in March 2000.

The Dublin Core

http://dublincore.org/

The Dublin Metadata Core Element Set is a core list of metadata elements, now applied extensively to metadata initiatives for the description of networked electronic information. It is a simple information resource description but also aims to provide a basis for semantic interoperability between other formats. It also seeks to provide the basis for resource embedded description, initially with HTML documents.

The Information Asset Register

http://www.inforoute.hmso.gov.uk

The IAR system, which is managed by HMSO and sponsored by the Cabinet Office, was originally designed to catalogue all unpublished Government material. However, its coverage has since been extended to embrace published material as well. The IAR Metadata template has a total of 17 metadata fields, 15 of which relate to the Dublin Core. The full IAR metadata specification can be found in the HMSO document entitled ‘Guidelines for the Preparation of IAR Records’.

The Organisation for Advancement of Structured Information Standards (OASIS)

http://www.oasis-open.org

This is an international consortium with an interest in promoting the adoption of product-independent formats based on public standards such as SGML, XML and HTML.

XML.ORG

http://www.xml.org

This site is hosted by OASIS, which collects, manages, and distributes information about XML applications, including vocabularies, schemas and DTDs. It allows groups to register their XML data exchange specifications and makes these available to others working in similar areas.

Our intention is to maintain an up-to-date version of these appendices on our web sites.
Please check the RSS and UKDA web sites for version changes.

http://www.rss.org.uk
http://www.data-archive.ac.uk

UK Data Archive

University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK
Tel: 01206 872001 Fax: 01206 872003 email: archive@essex.ac.uk web: www.data-archive.ac.uk

Supported by

Joint Information Systems Committee

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

University of Essex