MRC POPULATION DATA ARCHIVING AND ACCESS PROJECT

Consultants' Report

Louise Corti and Melanie Wright
UK Data Archive

September 2002

ACKNOWLEDGEMENTS

The authors would like to acknowledge Hamish James of the History Data Service for his sterling work on the case study visits and write-ups, and for his sole authorship of Appendix 6; Karen Dennison for her work on Appendix 2 and authorship of Appendix 5, Nadeem Ahmad for his work on Appendix 2 and assistance with the case study visits, and Jack Kneeshaw for his authorship of Appendix 4. We would like to acknowledge Pam Miller, Lorna Balkan, Sharon Jack and Anne Etheridge for proofreading, and Diane Geraci and Kevin Schürer for their valuable comments. We would also like to thank Peter Dukes and Yanine Jairazbhoy of MRC Head Office for organising, participating in and writing up the case studies and for valuable discussion and comment on the issues encountered. Any remaining errors are, of course, our own.

Executive Summary ***5***

Developing an MRC Policy for Population Data Archiving and Access ***9***

1. Introduction **9**
1.1 Rationale 9
1.2 Re-use of medical population-based data 10
1.3 The DAA Project 10

2. Phase I: Survey, Questionnaire and Working Group **12**

3. Phase II: Field Study, Further Research and Council Recommendations **13**
3.1. Site selection 13
3.2 Methods 14
3.3. Comparative models research 15
3.4 Council recommendations 15

4. Site Visit Results **16**
4.1. Varied study types 16
4.2. Varied data types 16
4.3. Current data management provision 17
4.4. Researcher Opinions 1: long term value of data 19
4.5. Researcher Opinions 2: attitudes towards sharing data and accessing others' data 23
4.6. Researcher Opinions 3: access control 28
4.7. Researcher Opinions 4: services desired 31

5. Informed Consent and Implications for Sharing Population Data **33**
5.1. Introduction 33
5.2. Guidelines for good ethical conduct 33
5.3. Information Sheets and Informed Consent Forms from Case Studies 34
5.4. Storage and retention of study documents: legal and consent considerations 36
5.5. Summary 39
5.6. References and organisations' URLs 40

6. Existing Metadata and interoperability standards **42**
6.1. Introduction 42
6.2. The Biomedical Sciences 43

7. Data Sharing Policies **45**
7.1. Introduction 45
7.2. The UK scene 46
7.3. The United States scene 50
7.4. The international picture 53
7.5. Conclusion 55
7.6. References and organisation URLs 55

8. Models of Service Provision **58**
8.1. Fully centralised: the national archive model 58
8.2. Centralised infrastructure/distributed expertise: the AHDS/ESRC model 58
8.3 Subject-focused distributed service: the NERC model 59
8.4. Virtual integration only: the GRID model 60
8.5. The portal approach: the RDN model 61

9. Recommendations **62**

9.1. Encouraging the culture of data sharing 62
9.2. Standards and guidelines 62
9.3. Rights management framework 62
9.4. Consent 62
9.5. Selection of data for DAA 63
9.6. Desired services 63
9.7. Service provision 63
9.8. Policy and resources 64

APPENDIX 1: Site Visit Summaries ***65***

APPENDIX 2: Dataset Descriptions ***66***

APPENDIX 3: Metadata Standards and Interoperability in the Biomedical Sciences ***79***

APPENDIX 4: Secondary Usage of Medical-related Data ***85***

APPENDIX 5: Data Archiving & Access: Case Study Tool ***93***

APPENDIX 6: Good Practice in Managing Data for Sharing and Preservation ***102***

Glossary ***114***

EXECUTIVE SUMMARY

DEVELOPING AN MRC POLICY FOR POPULATION DATA ARCHIVING AND ACCESS

Data archiving is not simply about best practice in storing primary records, although that is important. It is also about making the necessary long-term strategic investments to ensure that datasets can be assessed and are interpretable by current and future generations of researchers.

Rationale

There are several important motivations for preserving research data: scientific, historical, economic, and legal. There is also significant demand for the re-use of population-based data in the medical research area, as usage of medical/health studies at the UK Data Archive (UKDA) show.

The MRC has recognised the importance of developing a policy towards data archiving and access, but also understands that it should not do so in a vacuum. What is required is a better understanding of the range and variation of current MRC-funded data creation activities, the existing data management infrastructure and practice in MRC-funded contexts, and the views and opinions of those likely to be most affected by the establishment of such a policy.

Data Archiving and Access Project

To this end, the MRC Data Archiving and Access Project (DAA) was established in 2001 to gather information, consult widely, and at the end of the Project, to make recommendations to Council concerning data archiving and access policy. Phase I of the Project involved conducting a broad but general survey and convening a working group of interested experts. Phase II of the Project incorporated both wider and more narrowly focused work. On the one hand, a series of site visits were undertaken in order to collect in-depth case study information on the conduct of population-based data creation and management; and on the other hand, the "Horizons" workshop was convened to locate the current inquiry in the broader context of developments in e-science more generally.

DAA Phase II

This paper is a consultants' report growing out of Phase II. It presents the results of the case studies and lessons drawn from them, and analyzes current DAA provision within MRC-funded units and projects. It further identifies cultural issues involved, looks to other research data organisations, services, and funding bodies internationally for existing policies and practices, and presents a number of different possible models for DAA provision.

Case studies

A small team visited seven study sites and examined eighteen population-based datasets. Semi-structured interviews were conducted with principal investigators, project managers, laboratory managers, statisticians, geneticists and data managers, attempting to uncover the current situation "on the ground" with regard to data management, preservation, and access. Researchers views were sought on the philosophical and practical issues involved in developing a data archiving and access policy, and what services they themselves might like to see supported.

One of the findings of the case studies was that even under the rubric "population data" a very broad church was represented. The eighteen studies investigated revealed a lot of variation in study types, methods, data types, and formats. There was also great variation in current data management and data sharing provision, ranging from excellent to less than ideal. The attitudes of researchers towards data sharing were also variable, and although there was widespread support for data sharing on a conceptual level, in many cases researchers were concerned about secondary usage being carried out without appropriate control over the legal, intellectual and analytical frameworks.

Additional investigations
To augment the case studies, additional investigations were undertaken. An analysis was undertaken of current consent arrangements and their impact on data sharing. Research on existing standards for metadata and interoperability in the medical and health-related data areas was undertaken. Data sharing policies from organisations in the UK, the USA, and internationally were also examined.

Models for service provision
Building on the outcome of these investigations, a series of five models for possible architectures of service provision have been drawn, ranging on a continuum from fully centralised to fully distributed. ***Model one*** is fully centralised, with dataset preservation, provision for secondary use, and support of secondary users all located in a single centre. ***Model two*** has a centralised infrastructure, centralised preservation and dissemination, but user support is disseminated among specialist centres. ***Model three*** further distributes the service, with data preservation and dissemination as well as user support being devolved to distributed subject centres. ***Model four*** is a completely distributed one, with integration happening only at a virtual level on the end-user's desktop, and each individual data creator responsible for making their data available via common interoperability standards. ***Model five*** suggests a portal service, offering standardised information about, and pointers to, data resources, but leaving each researcher to negotiate access individually. The advantages and disadvantages of each model were assessed.

RECOMMENDATIONS

It is beyond the scope of this report to present specific recommendations to the MRC concerning the particular form and content of its data archiving and access policy and service provision. However, this report offers some general observations and lessons learned from the course of the site visits and accompanying research.

1. Encouraging the Culture of Data Sharing

It is clear that most researchers support the idea of data sharing. However, in discussion of the mechanisms of how access would operate, there is a reluctance to relinquish control. There are significant cultural barriers that need to be overcome, and researchers need to be encouraged to think in terms of sharing data as a normative activity, as it is in other areas of MRC science. Formalising access control procedures and protocols might serve to make researchers more comfortable with data sharing. It is important that a policy which mandates data sharing not be confused with a requirement to open all data to free and uncontrolled access. Researchers will understandably feel far more comfortable if there is a sense that secondary users have been in some measure "vetted" before access is allowed to potentially sensitive or particularly complex data. Depending upon the access control procedures implemented, such protocols, however, may come to seem unnecessarily onerous to the secondary analyst, who may quite reasonably feel that the peer review process already in place for research proposals and resulting publications would obviate such a requirement. Enforcing the sharing of data may require both "carrots" and "sticks" in the form of additional funds to researchers to prepare their data for secondary analyses, and perhaps making the incorporation of data sharing into the research plan a condition of the grant award.

2. Standards and Guidelines

It is clear that there is a demand for the re-use of population data in the medical research field, as well as a good scientific and economic rationale for its preservation and dissemination. It is also equally clear that in order for data to be reasonably re-usable and preservable, certain standards, particularly of dataset documentation, must be met, and that researchers would welcome assistance in meeting them, particularly in the form of published guidelines and advice.

3. Rights Management Framework

A good rights management framework is essential for the establishment of a data archiving and access service. Rights management in its broadest sense covers not only the relationship between the data creators and secondary data users, but also protects the rights of institutions, funders of research and even research subjects. A unified rights management framework for MRC data would greatly improve a data access service, no matter what the particular model of service provision.

4. Consent

Consent arrangements need not preclude the use of data for secondary analysis, and it is recommended that the MRC draft appropriate guidelines for wordings for consent agreements which would allow secondary analysis. It is important when issuing guidelines on consent and data sharing that the MRC educate both researchers and other bodies concerned with the ethics of research with human subjects.

5. Selection of Data for DAA

All data are not created equal, and in a world of limited resources, priorities for DAA must be drawn. Investigators' views should be of great value in helping MRC to consider a strategy for both assessing and prioritising datasets for archiving. In summary, the following criteria may be most appropriate for making these decisions, some of which were previously suggested from the MRC DAA Phase I survey:

- the degree of uniqueness or the size/longevity of the study/dataset;
- the reputability and quality of the study methods and dataset;
- the potential to answer new important research questions cost-effectively;
- the likelihood that a dataset can be pooled or combined with other data to provide explanatory power that individual datasets cannot provide;
- the degree to which the study fulfils ethical and legal requirements to enable the re-use of personal information;
- the degree of MRC sponsorship;
- the anticipated cost of preparing a dataset to professional archival standards;
- the recency of the study;
- building in prospective data preparation and documentation plans that enable secondary access for all new studies .

A set of criteria such as these would provide the framework in which to apply independent evaluation of MRC datasets for archiving.

6. Desired Services

There was nearly unanimous support among researchers for published guidelines for good practice in the management of data and their preservation. Many also supported the idea of a data archiving and access advisory service. A central registry of research instruments also received strong support, particularly among younger researchers and project managers. Support for a freestanding data archiving and dissemination service was mixed, as many researchers feared losing control altogether; support was stronger for a preservation service.

7. Service Provision

There are a number of different possible models for the architecture of service provision for data archiving and access, ranging along a continuum from completely centralised to completely virtual and distributed. Each model has advantages and disadvantages which must be weighed, including cost effectiveness, quality and focus of service and support, and ease of use and navigation for both data creators and data users, and support among the research community. Centralised services are often most cost effective because of lack of replication of infrastructure and expertise, and may be an easy focal point or "one stop shop" for secondary users and data creators, but may also offer the least focused user support, and may face the greatest resistance from researchers concerned about access control. Different models also require, and enable, different degrees of data and metadata standardisation. The MRC must determine the relative priority of these competing factors in choosing a model for service provision.

8. Policy and Resources

Regardless of the model of service provision, funding is required for long-term preservation and to facilitate data access and user support. Any data archiving and access policy needs to be adequately resourced, and may also require some contractual "teeth" to overcome cultural barriers and ensure compliance.

DEVELOPING AN MRC POLICY FOR POPULATION DATA ARCHIVING AND ACCESS

1. *INTRODUCTION*

This paper is a consultants' report growing out of the Medical Research Council's Project on Population Data Archiving and Access (DAA). It presents the results of a series of case studies and lessons drawn from them, analyses current DAA provision within MRC-funded units and projects, identifies cost drivers and cultural issues involved, looks to other research data organisations, services, and funding bodies internationally for existing policies and practices, and presents a number of different possible models for DAA provision.

1.1 *Rationale*

To quote from the Project's briefing papers:

> *"Over the decades the Medical Research Council has funded the construction of a large number of population-based studies. This includes several longitudinal studies and a significant number of cross sectional studies and clinical trials. Many have a unique value for research that runs well beyond their use by the Principal Investigators who created the studies.*
>
> *This Project aims to develop a policy for the Council that will identify its role in archiving population datasets, as well as that of investigators and the institutions which the MRC funds.*
>
> *Data archiving is not simply about best practice in storing primary records, although that is important. It is also about making the necessary long-term strategic investments to ensure that datasets can be accessed and are interpretable by current and future generations of researchers.... A policy for archiving will allow Council to manage its strategic investment in these resources more effectively for the benefit of research and for the health of the public."*

There are several important motivations for preserving scientific data: scientific, historical economic, and legal.

1.1.1. Scientific

Scientific knowledge creation is a cumulative process, and better access to well documented previously collected data can only facilitate this process. New data which are collected to be comparable with existing data increase the explanatory power of both.

Data also provide important baselines to track rates of change and capture the frequency of rare events. Data collected for one purpose may actually prove relevant to other scientific investigations, and the re-analysis of existing data with better tools and techniques may lead to new and different conclusions. Thus, access to archived data enables the formulation of new hypotheses and may unexpectedly change the relative importance of data previously collected.

1.1.2. Historical

Individual data collections are unique entities. They both take place within, and capture circumstances that are socially or scientifically fixed in, time. They can rarely be reproduced in exactly the way in which they were first undertaken.

Increased emphasis is being placed on empirical data 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 to these processes are preserved for further use and historical record.

1.1.3. Economic

Data are expensive to collect, and become increasingly so if the collection process is to provide high quality, validated data. The costs of preserving and archiving data are relatively small in comparison with the costs of acquiring scientific records through observation or experimentation. There is an economic imperative on the funder of data creation to maximise their investment, and derive maximal scientific (and at times commercial) value from the data sources they help generate.

1.1.4. Legal

For some data collections, there is a legal imperative to hold and preserve data, and in some cases provide an adequate "audit trail" in case of later legal challenges. This certainly applies in the area of clinical trials, and is increasingly the case with Data Protection and Freedom of Information legislation.

1.2 *Re-use of medical population-based data*

There is clearly scope for the re-use of medical population-based data. The UK Data Archive (UKDA) currently holds a number of such studies and provides them to the academic and research community nationally and internationally for purposes of secondary research.

Over 300 studies in the UKDA's collection (about 7%) contain medical or health-related data, largely health and lifestyle data (as one might expect from a social science data archive). These studies jointly accounted for an average of 400 orders for data each year for the past 10 years – a small but significant proportion of total data archive usage (see Appendix 4).

Different data types and methodologies engender different kinds of re-use. Some data may be most useful when first released and their value may decline fairly steeply thereafter, though their preservation may be necessary for legal reasons (e.g. clinical trials data). Other data, for example longitudinal data, increase in richness and value through time.

Clearly any data archiving and access policy must take into considerations the differing uses and requirements of different kinds of medical research data.

1.3 *The DAA Project*

The MRC has recognised the importance of developing a policy towards data archiving and access, but also understands that it cannot do so in a vacuum. What is required is a better understanding of the range and variation of current MRC-funded data creation activities, the existing data management infrastructure and practice in MRC-funded contexts, and the views and opinions of those likely to be most affected by the establishment of such a policy.

To that end, the MRC Data Archiving and Access Project was established in 2001 under the leadership of Peter Dukes to gather information, consult widely, and at the end of the Project, to make recommendations to Council concerning data archiving and access policy.

The DAA Project has focused on population-based datasets created as a result of MRC resources, but realises that any policy recommendations in this area must join up with other areas of MRC science, and should also fit into the broader scientific information landscape and be mindful of developments in other disciplines. To that end, information has been gathered not just from population-based research and researchers, but equally the attempt has been made to cast a moderately wide net for our investigations.

The project has been a staged one, with Phase I conducting a broad but general survey and convening a working group of interested experts. Phase II on the one hand tightened the focus by commissioning a series of site visits to collect in-depth information on the conduct of population-based data creation and management; and on the other broadened the focus by convening a "Horizons" workshop which attempts to locate the current inquiry in the broader context of developments in e-science generally.

2. *PHASE I: SURVEY, QUESTIONNAIRE AND WORKING GROUP*

Last year, Phase I of the Population-Data Archiving and Access project set out to identify the broad principles of a data archiving policy for MRC. It comprised: (1) a survey of the key characteristics of 95 datasets (40 survey respondents); (2) a questionnaire about archiving policy (40 respondents); and (3) the first meeting of the Data Archiving Working Group, chaired by Professor Sally Macintyre (June 2001). The conclusions at the end of this Phase can be summarised very briefly as follows:

1. MRC supported datasets vary significantly in their size, methodology, the variables, and the format and media used to collect and store data.

2. There are considerable challenges in putting order into, and making sense of, "old" datasets. Ensuring continuing participation of the original investigators is one solution.

3. Guidelines to help data creators prepare their datasets to appropriate archiving standards as part of routine data management would be helpful. However, the necessary standards may not currently exist.

4. Archiving will require specific expertise, infrastructure and financial resources.

5. Principal investigators creating datasets have important concerns about the quality of secondary research uses, loss of control over the data and interpretations of them. Some are uncertain about the consent constraints that might apply to new uses.

3. *PHASE II: FIELD STUDY, FURTHER RESEARCH AND COUNCIL RECOMMENDATIONS*

In early 2002 the MRC launched Phase II of the Data Archiving and Access Project. It convened a team consisting of representatives from MRC Head Office, from the UKDA (consultants), and from IBM to conduct a series of seven visits to sites where MRC-sponsored population-based research takes place. The purpose was to fill out the sketchy picture provided by the Phase I questionnaire, and to investigate more fully both the range and variability of MRC-sponsored population research, the current provision for DAA "at the coalface" and researchers' attitudes towards the issues raised by the establishment of an MRC- DAA policy.

In essence, the brief for the case study site visits was to:

- identify and draw on existing good data management and archiving practice;
- illustrate how data are created and used - and the needs of data creators in sharing and preserving datasets;
- identify technical, cultural and financial challenges.

3.1. *Site selection*

The seven sites were chosen specifically to represent not only a range of population datasets, methodologies, and data types; but also a range of institutional settings, from freestanding MRC units to MRC-funded research conducted within academic departments. The Project also sought to capture datasets at different points in their "lifecycle" – from projects just beginning to collect data to "orphan" datasets whose principal investigators had retired or died. The teams investigated 18 different studies across the seven sites. Appendix 1 contains site report summaries for each of the seven sites visited, and Appendix 2 contains standardised dataset descriptions for each of the 18 datasets.

Table 1: Sites and Studies

	Host Unit / Programme		**Studies**
I	MRC Social, Genetic and Developmental Psychiatry Research Centre (SGDP), Institute of Psychiatry, London	1	Twins' Early Development Study (TEDS)
		2	Depression Case Control Study (DeCC)
		3	Isle of Wight Studies (IoW)
II	MRC Environmental Epidemiology Unit, Southampton	4	Southampton Women's Survey (SWS)
		5	Wessex Fracture Prevention Study (Wessex Fracture)
		6	Study of the effect of formaldehyde on the mortality of workers in the UK chemical industry (Formaldehyde)
III	MRC Social and Public Health Sciences Unit, Glasgow	7	SHARE: Does Teacher Led Sex Education Reduce Sexual Risk Taking? (SHARE)
		8	Masculinity and Health: The Social Factors Affecting Men's Health (Masculinity)
		9	Racist & Sectarian Graffiti in Glasgow – A Pilot Study (Graffiti)
		10	The West of Scotland Twenty-07 Study (Twenty-07)
IV	Department of Community Health Sciences,	11	MRC Scottish Colorectal Cancer Study

	University of Edinburgh Medical School		(SOCCS)
V	University of Birmingham, Clinical Trials Unit (BCTU)	12	Parkinson's Disease DNA Bank (PDGEN)
		13	Parkinson's Disease Drugs Assessment Randomised Trial (PDMED)
		14	Parkinson's Disease Surgery Assessment Randomised Trial (PDSURG)
VI	MRC National Survey of Health and Development, University College London	15	MRC National Survey of Health and Development (NSHD or 1946 Birth Cohort)
VII	MRC Clinical Trials Unit (CTU), London	16	MRC/INSERM trial of zidovudine in HIV infection (CONCORDE)
		17	Evaluation of Subcutaneous Proluekin in a Randomised International Trial (ESPRIT)
		18	An open randomised trial to evaluate different therapeutic strategies of combination therapy for HIV-1 infection (INITIO)

3.2 *Methods*

The project team's first task was the construction of a case study tool or questionnaire to be used to structure the interviews undertaken at the sites. The tool was then piloted at the first site visit, the MRC Centre for Social, Genetic and Developmental Psychiatry (SGDP) at the Institute of Psychiatry in London. The tool underwent subsequent revision based on the experience of the pilot visit. A copy of the final version appears in Appendix 5.

The method undertaken was semi-structured interview. The site visit would begin with a general introductory session with all staff where the project and its aims were presented and discussed. Then followed parallel sessions where individual principal investigators, project managers, statisticians, laboratory managers and data managers were interviewed. Generally, two project team members would interview the site visit participants, with one team member asking questions whilst the other took notes. Whenever practical, interviews were audio recorded as an aide mémoire. The day would end with another group session where any other concerns could be raised and discussed.

Every effort was made to ensure that site visit participants did not feel as if the project team were evaluating them or sitting in judgement on their dataset practices. Every attempt was made to engender frank and open discussions of problems and challenges, as well as positive achievements.

Whenever possible, documents associated with the study, such as participant briefing notes, consent forms, ethics committee applications, annual reports, etc were collected for comparative analysis.

Site report summaries were created with the following subject headings:

- **Background**: description, scientific objectives and funding history.
- **Scientific value and potential for new research**: profile and perceived value of study, secondary use potential and issues recognised by PI and team.
- **Access**: terms of consent, nature of requests and procedures for current access, issues concerning future access.
- **Custodianship and ownership**: perceived ownership, formal and local contracts concerning ownership, IP issues.

- **Resources**: staffing profile, funding, overview of data collection/processing operations, storage space, digital storage.
- **Dataset technical details**: dataset size, data formats, platforms and software, security, metadata, relationships between data types (e.g. human biological samples).
- **Views towards future data sharing/possible MRC DAA service provision**: research registry, advice centre, best practice guidelines, etc.

These site summaries appear in Appendix 1. Standardised dataset study descriptions were also created for each of the eighteen studies investigated to aid analytical comparisons. These study descriptions appear in Appendix 2.

3.3. *Comparative models research*

A second strand of Phase II was to look to other domains, both within and without MRC science, for models of data sharing and data preservation. To that end, project members visited and corresponded with a number of organisations, including the Economic and Social Research Council (ESRC), Natural and Environmental Research Council (NERC), the Human Genome Mapping Project (HGMP), the Netherlands Institute for Scientific Information Services (NIWI), the Arts and Humanities Research Board (AHRB), and others. These visits served as examples not only of data sharing practice, but also policy and structures of service provision.

This comparative research will be furthered by the upcoming Horizons workshop, which will situate the current discourse into a wider context of e-science, and provide linkages to other areas of MRC investment and policy, and related enterprises in other scientific domains.

3.4 *Council recommendations*

The final strand of Phase II will be the submission to the Medical Research Council of specific policy recommendations growing out of the case studies, this consultant's report, and the Horizons workshop, for its consideration in December 2002.

4. *SITE VISIT RESULTS*

The site visits proved to be an extremely rich resource for illustrating the range and complexity of MRC investments in population-based research. They were also extremely illustrative of the variation in awareness of, and views about, issues central to the theme of data archiving and access.

4.1. *Varied study types*

One of the lessons of the case studies was that even under the rubric "population data" a very broad church was represented. Of the eighteen studies investigated, there was a lot of variation in study types and methods.

Eight of the studies could be called trials, six of the studies were longitudinal in nature, two were focused primarily on the collection of DNA for banking (although many more of the studies had a DNA element to them) and two fell under the slightly vague rubric of "medical sociology", owing more of their methodology to that social science, although their subject matter was medical or health related.

Most of the studies had survey elements, many also had clinical measurements and assessments, some also had associated physical or genetic samples. Many used standard clinical scales, for example the Rutter scales, DSM-IV, SADS, etc. Some had qualitative elements, such as focus groups, or participant-produced art. A number had audio- or video-taped interviews.

4.2. *Varied data types*

The types and forms of data collected and produced by these studies also varied tremendously.

4.2.1. Paper

All of the studies involved a certain amount of paper, if only in the form of consent forms. Paper-based data also included original questionnaires, interviewer notes, clinical measurements, drawings, photographs and maps.

4.2.2. Digital data

Electronic data included coded questionnaire responses, interview transcripts, clinical measurements, and digital images. Survey data in some studies were "born digital", whilst the majority were coded from paper originals. The software formats for these digital data were many and varied, including text or word-processed files, various databases, statistical software, and a variety of proprietary image formats (in the case of DNA, gels linked largely to the laboratory machines producing them).

4.2.3. Physical samples and DNA

Physical samples were another data type collected by a number of the studies, including epithelial cells, blood and plasma, and tissue samples (tumours, for example). A number of the studies included extraction of DNA from these samples, and went on to produce genetic sequence data.

4.3. *Current data management provision*

Among the different sites visited, data management provision varied fairly widely. Interestingly, research lifecycle seemed to be correlated with the effectiveness of the data management for the study. Studies which had just begun and studies which had been in operation for a very long time seemed to have the most difficulties with data management protocols – the former due to lack of experience, and the latter due to being "locked in" to systems and protocols established in a much earlier era, which might be bordering on obsolete. Studies in "mid-life" often seemed to have the most effective data management structures.

Effectiveness also varied by study type. The clinical trials unit, for whom data preservation is required for legal, audit trail reasons, had well-established, effective data management protocols. Studies under the rubric "medical sociology" were probably the least concerned with long-term data management, as they tended to be smaller, one-of-a-kind studies. Longitudinal studies faced particular challenges (alluded to under lifecycle above) since data collection, storage, and retrieval methods may have changed dramatically during the lifetime of the study.

It was also interesting to note that different studies conducted within the same unit could have differing levels of expertise and effectiveness of data management. Some of the effect of "lifecycle" noted above could be mitigated by more data management expertise sharing within units. This may be another reason why the clinical trials unit was so effective; because the senior statistician was involved in all the studies from the earliest stage and can ensure adherence to standards in data collection protocols, etc. Sharing of data management expertise, and the establishment of published guidelines for good practice in data management, would be extremely helpful.

Indeed, the units which had a central data person supporting more than one project often had superior data management. This was probably because the person was forced to think more generically and in a standards-based way about data management, rather than being free to tailor procedures to the needs of one particular data creation activity. This may be part of the reason why the labs that handled physical materials also had superior data management systems on the whole, since ordinarily these labs were serving multiple studies simultaneously. This finding would tend to support the notion that centralisation of service and expertise can promote quality in data handling.

4.3.1. Physical infrastructure

The physical infrastructure (facilities, computers and networks, paper storage) ranged from excellent to less than optimal. Paper storage was mentioned most often as the most pressing and most expensive problem, and many sites had recourse to an external contractor for paper storage (largely Iron Mountain). This was not, however, seen as the optimal solution, as referral back to stored paper was costly and difficult. Nearly all sites did require local storage of some paper materials which were referred to regularly. Many felt that their facilities were inadequate in terms of archival issues like fire safe and environmental control, although physical security and theft prevention were of primary concern. Preservation of audio and videotapes was particularly poor, with little understanding of (or where there was understanding, insufficient resources for) archival practice for such media.

Most sites were content with their hardware and software provision for operational purposes, but for many, issues such as media backup and network security were left to the host institution (university or hospital). Those studies dealing with genetic materials were generally most aware of electronic security issues.

4.3.2. Awareness and training in data management

Whilst many data managers were aware of some of the principals of archiving and records management, not one of the sites visited had any staff formally trained in these areas. Data managers were nearly unanimous in their desire for published guidelines to good practice, and there was a high level of good will and desire to conform to good practice expressed. However, it is also clear that preservation and documentation for secondary analysis was not in the forefront of data managers' minds on a daily basis, as most of them were busy serving immediate operational and research needs.

Whilst most data managers had significant hands-on experience in managing digital data, there was particular concern around paper-based data and what could and should be done in terms of its preservation. Almost none of the sites had considered scanning or digitisation of documents such as consent forms, and there were questions about the admissibility and authenticity of such electronic versions for which guidance is sought. Many sites wished to maintain paper questionnaires because of verbatim responses and interviewers notes which might not have been coded, but again scanning as image files or documents was not widely considered as a viable alternative to storing paper. Exceptionally, the NSHD had invested, in an earlier era, in microfilm/fiche, and was now faced with the question of whether to continue subsequent waves in this analogue format or invest anew in digital imagery, which is a lesson in the costs of shifting technology.

4.3.3. Standards awareness and use

With the exception of clinical trials where the standards are fairly well established and motivated by audit trail concerns, there was little awareness of archiving and metadata standards among the sites visited. Metadata is a particular area of concern. Whilst the content of digital databases tended to be fairly well documented, there was a distinct insufficiency of internal documentation of other kinds of data: registers of the content of paper files, audiotapes, etc. The genetic and human samples labs were a noticeable exception to this generalisation.

Nonetheless, those data managers with multiple responsibilities for multiple studies, and those who were responsible for making orphan datasets usable again, often had the best awareness of what kinds of documentation standards are useful. Those serving solely the operational needs of a particular study were the least likely to think in terms of standardisation and documentation.

4.3.4. Summary

Current data management provision varies across the sites investigated. Based on the experience of the site visits, we would offer the following observations:

- There was a desire to conform to good practice in data management, and a keen need felt for published guidelines on this topic.
- Concern was voiced, however, that if guidelines are produced, adequate resources are provided to meet them.
- Units with a central core of expertise which serviced many studies seemed to have superior data management procedures and protocols, as did studies in "mid-life".

- Expertise was often not well shared across studies within the same unit which were not centrally serviced.
- Whilst metadata surrounding digital data files was often fairly good, metadata about, and internal tracking of, ancillary non-digital data was patchy (with the exception of physical samples).
- Management of paper was viewed as a particular problem, and digital solutions were not well investigated or used.
- There were very few complaints regarding computer software and hardware; the largest complaint regarding physical infrastructure had to do with paper storage.

4.4. *Researcher Opinions 1: long term value of data*

During the interviews, principal investigators were asked about the scientific value and potential for new research of their study data. Issues covered were:

- the extent of evidence of external knowledge or value of the study/dataset;
- the unique features of the study or dataset;
- whether the study could be compared or pooled with other similar studies;
- what they considered to be the expected 'lifetime' of research for the dataset.

4.4.1. Evidence of value or external knowledge of the study/dataset

As expected, the studies that were judged to be the most renowned or with a perceived high profile were those that were unique, pioneering, expensive, with large samples, followed over time (e.g. cohort studies) or addressed significant public health issues. Such pioneering or significant funding investments such as the NSHD, the SWS, the ESPRIT trial and the PDGEN DNA bank naturally attract media and policy attention because of the hopes they offer in helping to understand the aetiology of certain diseases and to develop new treatments. In addition to web publicity, publications and the conference circuit, knowledge is transferred rapidly across the academic community as many of the experts in the same fields sit on various committees to advise on study funding; study design and analytic questions.

The NSHD, having been established for over half a century, is obviously truly unique, as is the team's expertise in formulating and running the study. The SWS is unique in the western world primarily as its approach, which is reliant on recruiting using the UK GP network, would be very difficult, if not impossible, to emulate in other countries such as the USA.

The Formaldehyde study is the biggest study of its kind with 13,500 participants and within the field, is considered to be the most informative of all studies on formaldehyde.

The Isle of Wight Study conducted by Rutter in the 1960s was unique at the time due to both the richness of data collected, that was far greater than for comparable studies, and its methods. It was the first study to involve direct assessment of children and depended on a partnership with social and medical services, which could not take place today, for example gaining direct access to their record systems.

In the area of clinical trials, the CTU has a history of co-ordinating groundbreaking HIV/AIDS trials that have both influenced therapy and contributed to understanding of the disease mechanisms. Its HIV study, ESPRIT is the largest cross-national study of its kind, funded in the region of 3 million dollars, and as a consequence, its results are much awaited. PDGEN is also very well known because of its size, rigour and large case-control design, considered to be the key factor looking at genetic factors in complex disease. Finally the Wessex Fracture Study, a big public health interest of the international osteoporosis research community, is unique as no one else is undertaking this kind of intervention.

For some of the older groundbreaking studies, where new methods of measurement had been developed (e.g. Michael Rutter, Mike Wadsworth, George Brown) the tools themselves are seen as being equally valuable future assets for the scientific community. The Dietary Assessment Tool from the SWS has also been published as a validation tool, and is tied to well known questions. To promote institutional expertise, some units also run training programmes in study design, and data collection and coding, thereby promoting their studies further.

However, research can also become well known when studies give unexpected results. One such case amongst the studies we examined, was the clinical trial, CONCORDE, referred to as the prototype HIV study in the field and known to policy makers and practitioners across the world. This was a big HIV trial investigating early or deferred treatment with a single drug with a clinical end point, measured at the time (late 1980s) by the number that died and the number that got AIDS. The trial results, which were contrary to other findings, showed that there were significantly higher differences in mortality for the early treatment group - a shock to the company, the research and clinical community. This trial sparked an international controversy that affected future clinical practice.

Finally, many investigators stressed that studies were of greatest value when they had a long-term follow-up design, for example, trials that were supported by a core funded unit with continuity.

4.4.2. Studies with which own data could be pooled or compared

Investigators were asked to consider how their data related to other studies, and to what extent it would be possible to compare or pool data. Comparison or pooling offers greater power to answer scientific questions and as a consequence can be seen as adding value to the original study.

For the pioneering or unique studies, as discussed above, the theories and methods of the study have often driven other research agendas and study designs across the world, thus making realistic comparison a reality. The NSHD, the Twenty –07 are examples of this. Other purposive cross-national data collection strategies to which the teams were contributing members, offered further comparative data sources. An example is the multi-national INITIO clinical trial. Some of the newer trials, such as PDmed and PDsurg were expected to encourage new large-scale data trials to follow in other countries to provide comparative data.

Finally, some teams recounted smaller more focused studies, from within their own units, funded by MRC, or from individual European countries, with which their data could be pooled. Additionally, new value and power could be added if a dataset could be combined with data beyond its own geographical limitations.

The idea of 'pooling' data was felt by others to be untenable, as the methodology and instrumentation between studies is often so different. Data were more likely to be compared with other studies with similar styles of data collection. In this respect The NSHD was comparable with other UK based birth cohort studies, the NCDS, 1970 Birth Cohort and ALSPAC, and may have similar psychiatric data to some of Rutter's studies, e.g. the Isle of Wight.

In the field of clinical trials, comparison appeared to be more evident, for example with the existence of trialist collaborative groups set up by the CTU. The ESPRIT study could be compared with MRC commercial studies, e.g. SILCAAT which, while addressing a different population, could provide good comparison. Equally, CONCORDE data had been pooled with a Wellcome company sponsored study that investigated early HIV treatment.

Trials investigators use meta-analysis as a key methodology in their work, and as such the CTU has its own dedicated meta-analysis team. As the results from trials are published to high standards, so these published data can be utilised for meta-analysis. However investigators emphasised how expensive meta-analysis was to undertake, and as such the relative value of such an exercise always had to be weighed up. Some of the trialists further insisted that in order to undertake a reliable and robust integration of data across trials, it was essential to return to the source data (the individual patient data), as tables were generally not published in the same format (e.g. tabulated in different ways or different outcome measures).

The other primary use of data within the sphere of comparative use was the contribution to systematic reviews.

4.4.3. Lifetime of research on datasets

Leading on from the comparative potential of data, investigators were invited to consider the expected 'lifetime of active research' on their unit's dataset, both by their own team and in terms of the utility for secondary research.

Opinions were divided as to the potential research longevity of study data, which were driven primarily by the kind of data.

Cohort studies have obvious longevity and the degree of prior investment by MRC in its largest longitudinal studies might suggest that continued investment would also be a sensible strategy. The longitudinal studies, such as NSHD, Twenty-07 and SWS already had secure funding for set periods, but it was expected that the investment would be likely to continue. Even after data collection ceased, a lengthy period of analyses within units was anticipated. Furthermore, cohort studies' tendency to address broader sets of questions and variables provided good opportunities to ask new questions.

Studies that were unique were also viewed as having a lot of lifetime, where the longer view of use should be taken at the outset. The potential existed to go back to original schedules to reformulate new analyses. However, the complexity of some data, for example in the way in which questionnaire responses may have been rated for psychiatric measures, may be a barrier to longer term use by new research teams (discussed later on).

Studies that would be generating genetic data were felt to be of the greatest long-term value for new analyses. They offered the opportunity for a dataset to have utility long after the study has finished, and, in the case of some clinical trials, had long-term commercial value. The PDGEN study and its two associated trials were thought of as long-term sources of data which presented long ranging scope for analysis. Investigators were optimistic that they could utilise the data until retirement, but the utility of the data was regarded as stretching further than this - possibly fifty years or more, when new genetic tools could be applied to the data. Indeed, DNA banks could enable prevalence studies to be performed on genes not yet identified.

Future follow-up studies also had the potential to yield sufficiently rich data to answer tomorrow's research questions. In the case of the retrospective panel study, Formaldehyde, data would be accrued until all the sample are dead, suggesting the potential to keep going for some years. Equally, for the clinical trials there may be several years' worth of analysis looking at survival rates.

For ongoing studies, such as clinical trials, the chance to add on various components to the study, for example, specific measurements, tests or questions, was viewed as extending the lifetime of the data. Outputs of the SWS were considered to be likely to contribute to development of further, independent, studies such as systematic reviews and intervention studies. Also, studies that had already opened up to external collaborators were positive about an extended lifetime.

For narrowly focused or smaller studies on the other hand, there may not be much scope to answer new questions. Clinical trials are often very focused, both in the questions they ask and the breadth and number of variables. While the published data are used for meta-analyses, there is relatively limited value in mining the data or attempting new analyses.

For other types of studies, investigators considered that medical measures were likely to become outdated and that those with social science data were likely to be superseded in a relatively short time.

Finally, the expected future utility of a dataset was also seen to depend on a number of more practical criteria. First was the state of the data, for example whether a team's data management could enable a reliable and usable dataset, and whether all essential documents, which might be necessary to undertake future work, still existed. Second was whether active collaboration of the research team would be required to make long-term use a reality – in cases where investigators wish to move on to undertake new science, the future utility of a dataset might be restricted. Third was the question of investment in the purposive construction a long-life dataset – how much value should be attached to prospective archiving versus the pursuit of the planned and immediate scientific research agenda, where funds may be in competition? These data sharing issues are addressed in the following section.

4.4.4. Conclusion

The varying views expressed by investigators regarding the future value of their datasets highlighted the belief that only some datasets have comparative potential or longevity. Datasets that could be used in comparative analyses included those that were: cross-national in scope; had influenced the design of other studies; could be used for meta-analyses; could be combined with data beyond the study's own geographical focus. Datasets with longevity were identified as those that: are longitudinal or unique in nature; have the potential for long-term funded follow-up of subjects; are large-scale unique clinical trials; are linked to large sample DNA banks; are of high quality and well documented. For other, smaller or more narrowly focused, studies there appeared to be little value in going back to old data, when compared with the opportunity to address new science. Finally additional secondary analytic value was viewed to require: focus and breadth of investigation; complexity; and high data quality.

4.5. *Researcher Opinions 2: attitudes towards sharing data and accessing others' data*

Investigators were asked a number of questions about sharing their 'own' data and re-using other data not under their control. They were asked about:

- ownership of data and whether they considered that they had responsibilities to ensure wider scientific use;
- barriers that might exist to further exploitation of their dataset by a secondary analyst;
- motivations to accept or decline access to their own data;
- whether they required access to datasets not under their control.

4.5.1. Ownership of data and the perceived responsibilities to ensure wide use

All the investigators held clear views on ownership of their datasets. MRC funded study data were considered to be owned by and copyright of the MRC unit and/or the university, depending on primary sponsorship. Often universities had provided senior staff costs, space and other costs (e.g. underwriting a study). In the case of cross-national clinical trials, the steering committee, seen to represent the national centres, was also mentioned as have a claim to ownership. For trials sponsored by pharmaceutical companies there may be valid questions concerning their potential interest in ownership of data. Finally, two researchers interviewed pointed to the patients (in trials) as having some claim to ownership of data.

Various rights of the investigators and sometimes the unit as whole were asserted pointing to:

- collectively having the right to allow others to use data for research in a manner consistent with the original ethics approval;
- exploiting the data first because of the intellectual investment made by the original team;
- making the final decision on access to primary data or DNA samples.

and having responsibilities for:

- protecting the interests of the research subjects from unnecessary harm or contact;
- optimising the research potential of the dataset so that it could be shared with collaborators/used in the future;
- publishing study protocols to avoid unnecessary duplication and to encourage collaboration;
- ensuring that data could be used maximally, but only in proper, responsible, high-quality new use, interpretation and publication;
- being receptive to sharing data;
- ensuring that the dataset is properly stored and does not become an orphan;
- formally archiving the data.

The question of how teams viewed their responsibilities towards data sharing revealed divergent views. While some teams were keen to take a role in actively encouraging and supporting secondary use, others were adamant that soliciting requests for re-use was beyond their current duty. Principal investigators generally tended to take a more active approach in promoting data sharing - via their roles as chairs of national working parties and of trials' management committees. Many studies had not considered archiving data at the time the study was set up, and as such believed that their responsibilities did not stretch beyond in-house storage and data management. Creating a fully documented archivable dataset in their opinion required recourse to dedicated funding.

When asked further about who should take prime responsibility for preserving and facilitating secondary use of the dataset, investigators pointed to the MRC, who should develop the provision of systematic approaches and enable appropriate resourcing. For high profile clinical trials, researchers felt that the steering committee should take responsibility to promote the study and its use.

On the question of intellectual property rights, in an era where research excellence is judged on publications output, some investigators were concerned that they may lose out on recognition for the original creation and management of data in publications arising from data obtained through an independent archive. In the case of sample banks and clinical trials, where provision for long-term access or international collaboration was the principal aim, these kinds of IP issues were typically addressed up front, so were less of a worry to researchers. However, two researchers expressed the need for the academic/funding world to better recognise the value of the work of trialists, including the statisticians and co-ordinators, in publications.

Regarding the question of rights and responsibilities to make decisions about secondary access to primary data or samples, the following section addresses investigators' perceived barriers to secondary use of, and motivations to accept or decline access to, data.

4.5.2. Barriers to secondary analysis of dataset

The question of what, if any, barriers existed to further exploitation of data by a secondary analyst revealed in some cases a number of key concerns, and in contrast, in others a more nonchalant response.

The following issues were cited as being distinct barriers:

- the lesser value of data without the active collaboration of the primary investigating team;
- for complex data, the degree of investigator involvement that would be required to support users;
- loss of control over data and intellectual property rights;
- the potential to misinterpret or misuse data;
- the huge task required to document data to high standards to render it usable;
- ethical and consent issues;
- concerns that data would not be used to the extent that justified investment in archiving;
- the mechanism or infrastructure in place to enable data sharing.

4.5.3. Complexity and intensive user support

The complexity, quirks, or lack of adequate documentation of data were seen, in particular by researchers in the older and larger established studies, to be a major barrier to re-using data properly, and particularly without the input of the investigating team. A high level of investigator involvement that was viewed to be essential by many teams in supporting new users. For those with experience of providing guidance on complex data, the time expended on such exercises had proved to be significant. For some study teams, the lack of staff available to support every potential collaborative study meant that data sharing was highly selective. For some studies, only the part of the dataset which was of particular interest was even cleaned, making the dataset as a whole currently unusable, even by the in-house team. In contrast, for clinical trials, this was never the case, primarily as established practice, dictated by the SOP and Clinical Trial Protocol, sought to ensure that data were frequently monitored, reliable and well documented.

4.5.4. Control and threat to intellectual property rights

Concerns were voiced about the loss of control over data or intellectual property rights when data were publicly archived. There was a perception that investigators would not get their names on secondary research papers as collaborators if the datasets were accessed through an independent archive. Moreover some teams expressed the need for ensuring that the study team could work on the data first. The Phase I DAA survey further revealed a view that secondary users could be likened to 'scavengers'. Keeping data in-house and restricting access was seen to be one way of preserving intellectual capital.

4.5.5. Misinterpretation

Worries about misinterpretation of the data arose from previous experiences of selective and opportunistic interpretation by new analysts. Interpreting complex data requires considerable expertise and an understanding of the limits of the data and how they were collected, which in turn requires considerable documentation. Genetic data was considered to be most problematic, where the huge analytic potential in linking it with phenotypic data or outcome measures data, made it a target for scientists not versed in the art of population data analysis. For one trial, the statistician considered that only geneticists with the right clinical, medical expertise would be able to analyse the data in conjunction with a team with a broad base of skills. One team noted the possibility of the negative impact of misuse on sample attrition for longitudinal studies, and another expressed concern regarding the misuse of politically-sensitive occupational or environmental data by pressure groups or industry.

Changes in current practice towards more transparent research, have seen scientific journals requesting contributors to send (or supply via the web) copies of data for verification/ auditing purposes, especially where findings may be controversial. In this respect some investigators considered that the scientific community ought to bear the consequences of mis-interpretation and expose it through academic debate, rather than prevent it from occurring in the first place.

4.5.6. Documentation of data

Study teams running ongoing studies were daunted by the prospect of transforming their data into a widely usable resource. Documenting data to the high standard required to render it usable was seen as a huge task, and one that they considered they were not explicitly funded to undertake. In the case of some studies, documents such as contact information for the sample were not kept, suggesting that there would be difficulties in conducting a future follow-up study.

4.5.7. Ethical considerations

For some studies, researchers expressed concern that ethical considerations with respect to the nature of the original informed consent agreement might impose restrictions on re-use.

These are discussed in the section in this report on Informed Consent and Implications for Sharing Population Data.

4.5.8. Re-use not justifying investment in archiving

The question of public investment in archiving was viewed by some to be approached with caution. Examples of publicly-available datasets that were not well used were cited, such as the NHANES data and the Framlington Heart Study. A significant tension between the a priori level of investment in data preparation versus new data collection and primary research, in a world in which scientific budgets are limited, should be recognised. Indeed where a data policy exists, resources for archiving are liable to compete with those for the 'science'.

Investigators saw the need for a policy to determine relative levels of investment in data preparation and documentation according to different types of datasets and for different end uses. For example, for cohort studies, or studies that have the potential for follow-up, the costs of good project housekeeping, high-quality data documentation, anonymisation, sample maintenance and sufficient document retention facilities (including paper storage, or digitisation costs) would need to be budgeted.

4.5.9. Mechanism for data sharing - access control

Finally, when discussing mechanisms or infrastructure that could support and enable data sharing, the issue of access control proved to the top priority. For those who did allow access, the efforts and personnel required to establish a mechanism for access, to appraise new proposals or check secondary reports prior to publication, should not be under-estimated. Models of access control are discussed further in the section on Researcher Opinions 3: Access Control.

4.5.10. Motivations for accepting or declining collaboration/access

In addition to the general perceived barriers to secondary use, investigators were asked to consider specific motivations for accepting or declining collaboration or access to their team's data or samples.

Criteria mentioned in support of accepting collaboration or access were:

1. For proposals:
    - having a research interest in common with the Unit;
    - asking relevant and useful questions - not easily answerable from the research literature;
    - taking the field forward not just replicating existing findings;
    - usefulness to the progress of the study;
    - asking questions that only the dataset could address – not as just a resource of samples/DNA;
    - having appropriate ethical approval where necessary;
    - having the infrastructure in place – funding, staffing and laboratory facilities;
    - multidisciplinary in nature and having scientific opportunity.

2. For researchers:
    - experience/standing/evidence of research quality of the secondary analyst;
    - those already known to and trusted by the Unit;

- competent, substantive expertise which the Unit does not have in-house;
- willingness to work with in-house statisticians.

For declining collaborations or access points the key factors were:

1. For proposals:
   - loss of focus on the study aims and the primary hypotheses;
   - not overloading the study participants – danger of attrition;
   - a directly competing project;
   - not of scientific judgement or value;
   - where the data cannot answer the question or are obviously frivolous;
   - which had no respect for longitudinal nature of the data;
   - where purely for cross-sectional analysis (for cohort studies).

2. For researchers:
   - insufficient skills to analyse particular kinds of data;
   - insufficient technical competence to handle/interpret data due to their complexity.

As expected there were different levels of motivations depending on the nature of the request and the sensitivity of the data. Those asking for access to anonymised data and samples were more easily accommodated than those investigators wishing to utilise primary data or personal information, for example to conduct follow-up studies. Finally, all investigators expressed the need to have a suitable infrastructure in place to be able to manage request to access data or collaborate. Models of service provision for archiving and data access are discussed later on, while options for controlling access to data, as suggested by investigators, are described in the section Researcher Opinions 3: Access Control.

4.5.11. Gaining access to other datasets not under own control

On the whole, investigators did not convey any major barriers to obtaining access to other data sources. For those who said they did use data other than their own team's, current access to other research team's study data tended to be straightforward and on a collaborative basis with the study investigators. This was felt by most to be an important and preferable model, rather than just gaining access to the data per se. Datasets mentioned were EPIC, NSHD and NATSAL.

However, it is evident that access via the collaborative model depends to a large extent on established networks, and that the views of the principal investigators, who are all at the forefront of their disciplines, may not reflect the desire of other younger researcher's to use existing unarchived data. One trials investigator felt that 'inner circles' often had advantageous access to information about the study and funding calls to re-use data, and the MRC should attempt to make this process more transparent.

Other researchers mentioned using datasets like the BHPS and the other birth cohort studies, via the UK Data Archive. Investigators also purchased data, for example, the Southampton and UCL teams bought the ALSPAC, in order to conduct comparative analyses.

The clinical trials teams were more likely to want to get hold of other primary data that were not accessible, particularly for cases of meta-analyses which were concerned with looking at long term outcomes (or adverse effects), and where locating and flagging the patients was essential. These data were often in the commercial domain, deriving from considerable financial investment from pharmaceutical companies.

For others there appeared to be little interest in accessing other MRC datasets, partly due to the perceived difficulties in interpreting other people's data.

4.5.12. Conclusion

On balance, study teams were mostly happy for their unit's data to be widely re-used but with caveats. Access to almost all of the studies would require some degree of vetting and many saw the need to conduct new analyses and publish on a collaborative basis. Vetting was proposed for reasons of safeguarding data, preventing misuse or mis-interpretation and helping users navigate complex data, rather than through selfish reasons.

Other points raised concerned loss of intellectual property rights or a threat to one's own research career. A further concern was the amount of work to make a dataset widely usable, and connected to this was the uncertainty of the value of MRC investing in an a priori archiving strategy in comparison to supporting new data collection efforts.

There were few problems encountered in gaining access to other researchers' data, although this did rely heavily on "who one knew" and being part of established formal and informal networks. Clinical trials units expressed a desire to utilise pharmaceutical data, but thought this possibility to be unlikely.

In quoting from the MRC Phase I DAA report, investigators tended to favour a ' system that promotes and supports responsible secondary research while rewarding primary researchers for their intellectual investment in the original study objectives and design and in managing primary data collection and analysis. There is a fear amongst investigators of loss of control over their data and of increased bureaucracy'.

4.6. *Researcher Opinions 3: access control*

The preceding section discussed the perceived barriers to both sharing and re-using data. This section sets out some of the solutions raised by investigators for dealing with some of the concerns raised – namely: data confidentiality; intellectual property rights; fear of mis-use; and the complexity of data.

In order to elicit more precise responses study teams were asked to explain or consider the following questions regarding access to data and biological samples:

- current (or likely) scale and type of requests for secondary access;
- current formal or informal access criteria or processes for collaboration/secondary use;
- nature of agreement/policy with collaborators/secondary users on ownership of data or publications from secondary analyses;
- methods of access control advocated for the future.

4.6.1. Secondary access: scale and type of request

The scale of requests varied a great deal between studies, from more than twenty per annum to none. For the well-established longitudinal studies, there existed a definite community of users, whereas for studies in progress, or completed more focused studies, no access as yet had been requested.

NSHD reported in the region of twenty requests per annum, of which ninety percent were fulfilled. Requests are only refused when the data can't answer the question or the request is obviously frivolous. Currently twenty researchers were using NSHD data and for some areas of the study, like mental health, the collaborators were responsible for driving the design and analysis.

While it was too early for requests for many of the ongoing clinical trials, many requests were expected once a reasonable amount of data/samples had been collected

In general, units' policies on secondary access to the sample population (primary research) tended to be highly restrictive, although NSHD would, in certain cases, allow access to more sensitive data on site only. For recent studies, requests for access to the sample population or biological samples, rather than to the data per se (secondary research), were certainly more common, and typically were for the purpose of undertaking new data collection. For at least two of these studies, collaborations had already been set up, for example in adding new questions on self-reported outcomes or physical measurements to the data collection. These collaborations on data collection were only considered if the proposed additions were seen to add value to the study and, crucially, were not overly intrusive to the study instrument. Other requests mentioned had been for permission to use the questionnaire or, in one instance, syntax file for derived variables. Finally, publications always tended to generate a large number of requests for information.

In addition to the unique value or size of studies, the scale of requests to access data depended to a large extent on the degree of publicity and proactivity in seeking users. The majority of units visited did not operate a policy of soliciting collaboration or secondary access.

Managing requests, or anticipating how to manage future requests, from collaborators/ new users can become a considerable operation as the volume increases. For every request judgements must be made on the suitability of the proposal and the proposers, and the study team's capacity to handle the collaboration/secondary use.

4.6.2. Current formal or informal access criteria/processes and nature of agreements with collaborators/secondary users, including IPR

Investigators were asked to describe any current formal or informal criteria or processes they used to decide whether or not to collaborate/give access to data or samples. They were also asked about the nature of agreements with collaborators or secondary users.

Only two of the units conducting population studies had formalised access criteria. For one, decisions on who to allow access were judged in-house on an ad-hoc basis by the team of senior scientists. For the other, the study steering group would consider a one-page proposal to determine access, and on acceptance would identify a liaison person to be collaborator and co-author. The preference in both cases was for 'guided access' due to the complexity of the longitudinal data. In contrast, trials teams, who were more likely to collaborate, had formalised mechanisms for sharing data built into protocols and via steering/management or data monitoring and ethics committees, but still relied on a subjective assessment of the quality of the request. The criteria used to determine access for all studies reflect those highlighted in the previous section, on motivations for accepting or declining collaboration/access. The main factors for acceptance are described in the section Researcher Opinions 2: Attitudes Towards Sharing Data and Accessing Others' Data.

Three of the units who had re-use facilities set up, or who were involved in multi-country studies, had established pro forma written agreements. The NSHD's end-user licence requires a number of undertakings, some of which are similar to those set out in the UK Data Archive Individual Access Agreement: promising confidentiality of data; not passing on personal information, the study team checking drafts of papers for publication; acknowledgement of the NHSD team; not keeping data and data descriptions on same computer; destroying data after use; and using sensitive data on-site only. The Glasgow Twenty-07 team further ask for any derived variables to be given back to the team, which is a valuable method of adding value to datasets. All of the clinical trials had established written procedures, or were in the process of drafting them.

Others were currently using informal procedures. For known and trusted researchers, agreements were likely to be contained in correspondence, whereas for 'unknown' researchers some saw the need for a tighter, signed agreement. Newer study teams were working on putting collaborative agreements in place where future requests for re-use were anticipated.

In the case of genetic information, strict procedures were often in place – for example the local genetic advisory committees would consider issues of collaboration and access.

Regardless of the formality of access procedure, users were always asked to sign a confidentiality agreement before accessing data. It was very rare that identifiable data were ever released, but for the unit that did allow access to sensitive data, those data had to be accessed on site only. One unit's collaborative access policy required data to be used only on site.

Most teams agreed, for collaboration, the need to establish an agreed approach to authorship, whereby principal investigators are given a share in authorship. One of the units had a User Manual that covered publication policy. For clinical trials, IPR procedures are typically formalised and described up front in the protocols, and the management/steering committee take responsibility for appointing and approving writing groups for trial publications. One team mentioned that sponsoring companies would always be informed of the release of data.

4.6.3. Methods of access control: future

While some units already operated formalised access control procedures, all welcomed future plans for this. Many had not considered plans for enabling access down the track, although clinical trials units traditionally did have mechanisms in place whereby a virtual committee was kept in place, made up of ex-trial investigators or committee members, who would continue to be involved in decision making some 8-10 years after the closure of the trial.

On the whole investigators advocated some degree of vetting by either the study team or by an appropriate independent body/committee, to apply criteria systematically and consistently. Equally some teams mentioned the MRC or a centralised archive resource centre to help evaluate and facilitate requests for collaboration. The Phase I DAA Working Group report found that "45% of the 40 studies analysed would want independent input into decisions on providing access". Written agreements were thought to be good practice

Some investigators argued further that they would only allow secondary usage if their own unit's investigators and statisticians are involved, while others advocated access via invitation only. Particularly for complex data, and without recourse to full documentation, 'guided access' for new users with the help of the research team was advocated. In the majority of cases, new proposals would also require ethics committee approval. Many investigators agreed that re-deposit of new data (e.g. derived data) should be a condition of re-use.

In addition, while many investigators saw the need to conduct new analyses and publish on a collaborative basis, they also welcomed the idea of an agreed rights management framework and commercial exploitation plans for MRC studies.

4.7. *Researcher Opinions 4: services desired*

Opinions were sought on the utility of a range of possible services which could be provided in the Data Archiving and Access realm.

4.7.1. Research instruments' registry

In general, there was positive support for the idea of a central repository of research instruments. This support was strongest among the project managers and data managers, and weakest among principal investigators. Some PIs asserted that they were already familiar with the study instruments relevant to their particular area of interest, so that they would be unlikely to use such a resource, but could see how it might be useful as a training tool for young researchers. Support for the registry tended to be higher within those projects which were in the early stages of research (probably because they had just been through the arduous process of constructing study tools for their own projects).

4.7.2. Published standards and guidelines for data preservation and documentation

This proposed service received widespread support from nearly all personnel at all sites, particularly those with direct responsibility for data management. There is quite clearly a desire to "do the right thing," but also quite clearly a lack of training in these areas. There is a concern that if new standards and guidelines are produced, that adequate resources be provided to meet them. For studies which have been in existence for some time, this might require considerable extra resources, to bring data and documentation produced in a different era up to modern preservation and documentation standards.

4.7.3. Advisory service for preservation/dissemination issues

This likewise received fairly widespread support, although less than for published guidelines. In order to ask for help one needs to know what questions to ask, therefore an advisory service would be most useful in the context of accompanying guidelines and expectations. A number of those interviewed remarked that simply seeing the case study tool and the kinds of questions the DAA project team was asking had already raised their awareness of particular issues to do with data archiving.

4.7.4. Data preservation/dissemination service

Some could see the usefulness of a "data warehouse" kind of facility, analogous to Iron Mountain (but for safekeeping digital data), but many were hesitant about the implications of depositing data with a third party. PIs clearly still wanted a role in determining criteria by which other researchers received their data, but some (particularly those for whom there were already significant requests for access) could see the value of having help in the mechanics of data dissemination.

5. *INFORMED CONSENT AND IMPLICATIONS FOR SHARING POPULATION DATA*

5.1. *Introduction*

This section focuses on the use of patient information sheet and consent forms by the projects examined in the case studies and how these relate to secondary or follow-up uses of data (phenotypic and genetic), samples or subjects themselves. Further, the case study organisations' ability to comply with guidelines or regulations for retaining essential documents relating to consent is discussed.

The historical origin of current ethical principles for conducting research with human subjects arises from the Nuremberg Code which sets out statements of the moral, ethical and legal principles. In 1964 the World Medical Assembly adopted the Declaration of Helsinki to provide guidance for physicians in biomedical research with human subjects. This was most recently amended in 1996. The principles for conducting research contained in the Declaration of Helsinki require that adequate information must be provided to the research participants, participation in the research must be freely volunteered, with the understanding that the research subject can withdraw at any time and, in addition, informed consent should be obtained, preferably in writing.

In the UK, responsibility for ensuring the dignity, rights, safety and well-being of all actual or potential research participants in the fields of health and social care is set out in the Central Government's *Health and Social Care Act 2001* and the Department of Health's *Research Governance Framework for Health and Social Care*. The Research Governance Framework defines the broad principles of good research governance and is key to ensuring that health and social care research is conducted to high scientific and ethical standards. As a general principle, all research proposals that involve recent or past NHS patients, records, premises or facilities must seek approval from a national Research Ethics Committee (REC). These committees are convened to provide the independent advice to participants, researchers, funders, sponsors, employers, care organisations and professionals on the extent to which proposals for research studies comply with recognised ethical standards.

5.2. *Guidelines for good ethical conduct*

The MRC and the Wellcome Trust require approval from the appropriate Research Ethics Committees, for all funded research involving human participants or biological samples. Approval from other regulatory bodies such as the Human Fertilisation and Embryology Authority or the Gene Therapy Advisory Committee in the UK should also be sought where necessary. Researchers are advised to ensure the confidentiality of personal information relating to the participants in research, and that the research fulfils any legal requirements such as those of the Data Protection Act of 1998 and the Human Rights Act of 1998.

Over the past five years a number of funding bodies have published excellent and very detailed sets of guidelines for scientific researchers on the conduct of ethical research, particularly in setting out the moral and legal responsibilities that researchers have towards research subjects. Examples in the UK include Central Office of Research Ethics Committee's (COREC) *Guidelines for Researchers applying* to RECs or MRECS, the MRC

Good Research Practice (2000), the MRC *Personal Information in Medical Research* (2000), the MRC *Operational and Ethical Guidelines on Human Tissue and Biological Sample Use for Research* (2001), the MRC *Guidelines for Good Clinical Practice in Clinical Trials* ((1998), the Wellcome Trust *Guidelines on Good Research Practice* (2002), the BBSRC *Statement on Safeguarding Good Scientific Practice* (2000), the General Medical Council, *Role and Responsibilities of Doctors: Good Practice in Medical Research* (2002), the Royal College of Physicians, *Research Involving Patients* (1990) and the EU ICH *Guidance on Good Clinical Practice* (1997). The ethical guidelines of many other professional organisations also endorse the Declaration of Helsinki principles.

Guidance provided in these documents covers the requirements for obtaining consent for projects as specified by sponsors, research ethics committees and Clinical Trials Institutional Review Board or Independent Ethics Committee. The principles and content are of great help to researchers designing Information Sheets, for trials involving patients, patient volunteers and healthy volunteers. Essential pieces of information that should form part of the consent agreement are the **Information Sheets** and **Informed Consent Forms**. Information Sheets should contain information under particular headings and in the order specified and should be written in simple, non-technical terms and be easily understood by a lay person. The headings should cover information about: the invitation to participate and selection procedures; who is organising and funding the research; the purpose, the length and aims of the study; voluntary participation and right to withdraw; expected degree of involvement in the research, including treatment and test procedures; the benefits and possible disadvantages or risks of taking part; access to medical records; storage of personal information gathered; what will happen to the results

5.3. *Information Sheets and Informed Consent Forms from Case Studies*

As part of the site visits, the nature and content of the documentation relating to the studies' Information Sheets and Informed Consent Forms for each study were examined. The principal focus was to establish the degree of restriction for future access to data, subjects and biological samples, and secondly to examine the compliance with the guidelines or regulations for storage and retention of these documents.

Regarding the procedures used for obtaining consent, the case studies fell into four categories:

- First, the clinical trials (*ESPRIT, INITIO, CONCORDE, PDMED* and *PDSURG, Wessex Fracture Prevention Study*) which have their own guidelines for running studies, set out in standardised formal documents.

- Second, the sample bank collection projects (*Parkinson's Disease DNA Bank, the Depression Case Control DNA Collection* and the *MRC Scottish Colorectal Cancer Study*) which are gaining consent specifically for generalised re-use of the information collection.

- Third, the survey-based and cohort studies (*Southampton Women's Survey, 1946 National Survey of Health and Development, The West of Scotland Twenty-07 Study, Twins' Early Development Study* and *Isle of Wight Studies*) and qualitative methods studies

(*Masculinity and Health*) which use a variety of informed consent forms and differing promises about data access.

- Finally, older studies for which follow-up were underway, (such as the retrospective cohort study *Formaldehyde* and the *Isle of Wight Studies*) had less straightforward positions on consent.

5.3.1. Clinical trials

In Europe the running of medicinal drug-based clinical trials is guided by the EU International Conference on Harmonisation Guidelines for Good Clinical Practice (ICH GCP). MRC's own *Guidelines for Good Clinical Practice in Clinical Trials* also sets out ethical and scientific standards for designing, conducting and reporting trials that involve the participation of human subjects. These documents set out to ensure that the rights, safety and well being of the trial subjects are protected, consistent with the principles that have their origin in the Declaration of Helsinki, and that the clinical trials data are credible.

Institutional Review Boards and independent ethics committees are established for every trial to review all proposed clinical trials, and demand a standard set of documents: a Clinical Trial Protocol; a set of Standards Operations Procedures (SOP); written informed consent forms (and updates); subject recruitment procedures; written information for subject and an Investigator's Brochure.

In all cases examined, the trials had detailed Protocols and SOPs, had complied fully with the applicable regulatory requirements for obtaining and documenting informed consent. Prior to the beginning of a trial, investigators had all gained IRB/IEC written approval of the proposed informed consent form and any additional written information to be provided to subjects. These documents provide information about the trial and clarification regarding the subject's involvement and rights, and whilst promising confidentiality, also clarified that information would be used for medical (or in other cases, research) purposes only. Wording of this kind implies the will to enable a more liberal use of data within the community, yet is placed alongside the reassurance of a promise of anonymity.

5.3.2. Sample banks

The MRC is increasingly supporting major collections of blood, tissue and DNA samples from groups of individuals, for example those with specific disease types. These studies typically collect both samples and phenotypic data from the subjects, which together provide important and powerful resources for research into the aetiology of specific diseases.

The consent forms from the studies of this type followed REC recommendations for gaining informed consent, and all asked for consent for the samples to be made available for future research relating to health that were ethically approved (including the DNA collection from the recent wave of the *NHSD*). Some studies specified that other research teams would have controlled access to the sample collection, as prescribed in the MRC *Operational and Ethical Guidelines on Human Tissue and Biological Sample Use for Research.* Controlled access to the collection was taken to mean vetting by: the project's steering or management committee, or MRC where appropriate; approval though a new Research Ethics Committee submission; and often external peer review of the proposal.

5.3.3. Surveys and cohort studies

Cohort studies require a form of consent that will allow for prospective follow-up of the sample. As the usefulness of the data relies upon high response rates at each wave, investigators have traditionally been concerned to protect their research subjects.

A variety of terminology was used by the survey investigators to provide reassurance to subjects about the confidentiality of the personal information they provided. The 1946 cohort study (the *NSHD)* and the *Southampton Women's Survey (SWS)* ask for explicit consent for each aspect of data collection, and permission for the team to return in the future. In these two cases, consent to view medical records or sample information was limited to members of the research team only, rendering longer-term follow-up by other researchers problematic.

Other studies use phrasing to reassure subjects such as 'information given is kept and used only in the form of numbers, so that it is entirely confidential' or 'information will be kept strictly confidential'. In the former case, the wording precludes any alternative use of any qualitative information collected. Similarly the terms of the REC review for the Isle of Wight study specified that access should be restricted to the principal investigator only.

A further concern, reported by at least one of the investigators, was that the workings of the Data Protection Act at the local level had provided serious impediments to the conduct of the project. Additionally, the only qualitative study we visited had severely restricting terms of consent for future access, whereby the team had promised to destroy the tape and transcripts after the project had been written up. In our view this practice is not necessary and is not something that is required under the law or recommended in guidelines, but nevertheless is sometimes demanded by local ethics committees. These same questions have also arisen through the ESRC's Qualitative Data Service's work, whereby some RECs or local Data Protection Officers have taken unnecessarily over-restrictive views towards consent for research projects, particularly those collecting longitudinal or qualitative data.

5.3.4. Follow-up of early studies

For studies conducted decades ago, the position on consent to re-use material or follow-up subjects is often a little hazy. For example, for the *Isle of Wight Studies* and the earlier waves of the *NHSD***,** consent to participate was not governed by any legal principles and, consequently, was assumed by virtue of the questionnaire having been returned. In this instance follow-up or re-use of data may be problematic, but the teams we interviewed saw this as the responsibility of RECs to determine the ethical position. Similarly, the *Formaldehyde* study, a retrospective cohort that is following up cancer or death registration of employees from companies' records, could not gain individual's consent, but instead relied on approval from local safety committees who represented the welfare of the employees. The same issue arises with orphan studies where consent forms have gone missing. Investigators appeared to be unsure about the precise legal framework operating in these cases, and we would recommend that the MRC provide some guidelines in this area.

5.4. *Storage and retention of study documents: legal and consent considerations*

There are a number of legal conventions that require personal information to be kept safely. First, the Data Protection Act 1998 contains enforceable eight principles of good practice applying to anyone processing or using personal data. Regarding the storage and retention of documents, data must be: processed for limited purposes, in accordance with the data subject's rights and within the terms of consent given; and kept securely and for only as long as necessary. The length of time for which documents must be kept for human-subject-focused research depends on the regulatory body, but as far we are aware there are no legal requirements on the length of time that consent forms per se must be kept. Nevertheless, there is a range of relevant guidelines or documents on good practice for study document retention.

Guidance to the NHS is provided in the 1999 'For the Record' HSc1999/053, which sets out minimum retention times for records. For patients involved in clinical trials this period is fifteen years after the conclusion of the treatment, including consent forms. For therapeutic research, the ICH GCP recommend retaining "Essential Documents" for the conduct of a clinical trial until at least two years after the last approval of a marketing application or the formal discontinuation of clinical development of the investigational product, unless specified otherwise by sponsor or appropriate regulation.

Sponsors of medical research, the Wellcome Trust and the MRC further set out guidelines for data security and storage. For MRC studies, identifying information (e.g. consent forms) must be stored in a separate place from data and research teams must maintain written procedures for keeping electronic and written information secure (whether being processed or archived), that must be enforced and reviewed at regular intervals. The MRC also expects research records relating to clinical or public health studies to be maintained for twenty years, to 'allow adequate time for review, reappraisal, or further research, and to allow any concerns about the conduct or consequences of the work to be resolved'. For all studies for which consent was obtained, the protocol, a sample of records and the consent procedures should be retained for thirty years. Further, it is recommended that the full records for historically important studies, novel clinical interventions or controversial studies are kept for longer periods.

The Wellcome Trust considers a minimum of ten years to be an appropriate retention period, however, research based on clinical samples or relating to public health might require longer storage to allow for long-term follow-up to occur.

The majority of the research teams we visited were aware of, and compliant with, the recommendations on secure storage of personal information and on document retention. Some units were extremely diligent in their operational procedures for compliance with the Data Protection Act requirements and MRC guidelines.

The majority of teams stored personal information separately from data, although two of the studies we visited were found to be filing signed consent forms together with paper data during the data processing stages. All of the teams used locked filing cabinets to hold paper documents, with controlled access to these, but very few had fireproof safes or proper archival storage facilities. Teams reported that cost was the main reason these facilities were not available to them, and that previous requests to MRC for professional storage equipment had usually been turned down. Another study had had the misfortune to have most of its key unique documents destroyed in a storeroom fire. Across the board, teams stored blood and tissue samples without identifying information other than ID. Information relating to consent and data were also stored securely.

Many of the project managers told us that they would be grateful for advice or guidelines on the storage of study materials. For those without good on-site paper storage facilities, professional storage equipment would enable more efficient document retrieval.

We identified a number of circumstances that worked against the practicality of complying with longer-term retention conventions. First, the lack of long-term adequate storage space for paper, and the high cost of professional off-site storage, meant that for about half of the centres the future retention of consent forms might be problematic. Indeed, as we point out above, adequate paper storage is a real issue for the larger studies. Many longitudinal studies were using companies such as Iron Mountain for retention of older data and associated key documents. In a couple of cases, study documents were reported to be housed in any spare place found to be available on site (basements, attics, cupboards etc) and, without an inventory.

For clinical trials, independent data-monitoring committees are responsible for assessing the progress of the trial, including the safety of data, and critical efficacy endpoints. Electronic data processing systems for trial are required to comply with established requirements for completeness, accuracy, reliability and consistent validation. Great emphasis is therefore placed on the investigators maintaining an audit trail, data trail and edit trail. All three of the trials units examined had established data monitoring procedures and audit trails and the clinical trials managers were far more likely to be fully briefed about legal requirements, namely because they had the advantage of local guidelines (SOPs) to consult. For trials, the consent forms were always stored separately from Case Report Forms at the local clinical site, rather than at the central trial units, with a further copy retained with patients' records.

Those teams who needed to consult personal information on a regular basis, such as consent forms, were keen to explore the idea of scanning these documents to enable quick and easy access. Clinical trials teams, however, were not convinced as to the utility of digitising these kinds of study materials.

5.4.1. Document format

The format in which key primary documents should be stored is less prescriptive than retention times across various guidelines, although there is separate guidance on storing paper and electronic data or document files. In the course of our research, we did not find any clear guidance on whether informed consent documents could be image-scanned and retained as legal electronic documents.

Neither the MRC nor the DoH provide any guidance on the legality of conversion, but the 1999 BSI standard (PD 0008:1999): *A Code of Practice for Legal Admissibility and Evidential Weight of Information Stored on Document Management Systems (EDM)* does offer current interpretation of best practice on legal admissibility of documents stored on EDM systems, although is not intended to guarantee legal admissibility. The Code advises on scanned images, documents created by a computer system and documents generated by third parties. The Code requires a documented procedure for scanning operations, where each document should have a unique identity number, which cannot be altered, and the date and time of scanning and the identity of the scanner operator must be recorded. The Code also allows for the scanning of photocopies of originals, which should be indicated on the electronic copy. The usual precautions regarding security of electronic data also apply.

A new European Directive on Clinical Trials currently under consultation is proposing 'Detailed Guidelines on the Trial Master File and Archiving'. This directive will provide advice on what essential documents must be retained (archived) for sufficient periods to allow for audit and inspection and the guidelines aim to give details on: the minimum set of documents to be retained; the quality of documents to be archived; minimum standards for storage conditions; media transfer and certified copies; and retention times. Much of the information is drawn from the ICH GCP, and does not propose any new standards.

MRC should follow the current debate in this area and, in particular for studies other than clinical trials, raise the issue of document retention of NHS-related documents (e.g. patient consent forms) with bodies like COREC so that advice can be incorporated into their own guidelines.

5.5. *Summary*

Our investigation into terms of consent for the range of MRC studies selected has concluded that, in general, most of the recommendations proposed by the regulatory bodies for the ethical conduct of research are followed. Indeed many of the teams' procedures demonstrated that they were acutely aware of the Data Protection regulations on the security of personal information, although less so on document retention procedures.

Regarding terms of consent, there are significant differences in the explanation and form of wording on how data will be used in the studies' information sheets and consent forms. The content is driven by a number of factors:

- the legal requirements or recommendations of regulatory bodies (e.g. Data Protection, DoH, NHS, COREC, RCP, GMC);
- the requirements of the sponsors (MRC, Wellcome Trust, etc.);
- the demands of local ethics committees;
- the concerns of the principal investigating team to keep data 'safe' in their hands.

We recommend that if MRC do adopt a data sharing policy that they recommend approved forms of wordings for consent forms and information sheets for the studies they fund (as they advise for research collecting biological samples in the *Operational and Ethics Guidelines* booklet). This could be achieved by enhancing the information in section 7 of the *Guidelines on Personal Information in Medical Research,* but also needs to be done in conjunction with a programme of education of the scientific community towards the benefits of data sharing, to prevent overly restrictive consent clauses being built in by investigators. At the same time the MRC need to inform the governing and regulating bodies that often drive the procedures for the conduct of population studies (especially local Research Ethics Committees and local Data Protection Officers) to make them fully aware of the motivations of MRC, and the implications of terms of consent in studies for operating a fruitful data sharing policy. Recommendations from COREC to researchers on less restrictive wording for consent clauses would be particularly beneficial.

The clinical trials community is perhaps the most advanced in terms of consideration for data sharing. In the interests of auditing and enabling greater access to data, the MRC might wish to follow this direction in recommending formalised basic protocols and SOPs, and data monitoring committees to be established for all projects, or clusters of projects, they support.

Finally, in order to ensure that teams comply with the relevant legal requirements on how study documents and data should be stored for MRC studies, we would advise MRC to provide clearer practical guidance, to undertake periodic checks, and to follow closely current debates on these matters.

5.6. *References and organisations' URLs*

BBSRC (2000) Statement on Safeguarding Good Scientific Practice
http://www.bbsrc.ac.uk/funding/overview/good_practice.pdf

BSI DISC PD 0008:1999 Code of practice for legal admissibility and evidential weight of information stored electronically

Central Office for Research Ethics Committees (COREC), Guidelines for Researchers: Patient Information Sheet & Consent Form
http://www.corec.org.uk/wordDocs/pis.doc

Data Protection Act (1988)
http://www.legislation.hmso.gov.uk/acts/acts1998/19980029.htm

DOH (2001) Research Governance Framework for Health and Social Care.
http://www.doh.gov.uk/research/rd3/nhsrandd/researchgovernance.htm

EC (1997) International Conference on Harmonization Guidelines for Good Clinical Practice (ICH GCP)

EC (2002) Proposed Detailed Guidelines On The Trial Master File And Archiving
http://pharmacos.eudra.org/F2/pharmacos/docs/Doc2002/june/tmf_06_2002.pdf

General Medical Council (2002) The Role and Responsibilities of Doctors: Good Practice in Research, February 2002
http://www.gmc-uk.org/standards/research.htm

HMSO (2001) 'The Health and Social Care Act 2001, Section 60: 'Control of patient information'
http://www.hmso.gov.uk/acts/acts2001/10015--g.htm#60

Human Genetics Commission: http://www.hgc.gov.uk

Human Rights Act (1998)
http://www.legislation.hmso.gov.uk/acts/acts1998/19980042.htm

Institute for Clinical Excellence
http://www.nice.org.uk

MRC (1998) Guidelines for Good Clinical Practice in Clinical Trials. London: Medical Research Council
http://www.mrc.ac.uk/pdf-ctg.pdf

MRC (2000) MRC Good Research Practice. London: Medical Research Council. http://www.mrc.ac.uk/index/publications/publications-ethics_and_best_practice

MRC (1995) Principles in the Assessment and Conduct of Medical Research and Publicising Results. London: Medical Research Council

NHS (1999) Health Service Circular 'For the Record: Managing records in NHS Trusts and Health Authorities', HSC1999/053, NHS

Royal College of Physicians, Research Involving Patients (1990)

The Wellcome Trust, Guidelines on Good Research Practice http://www.wellcome.ac.uk/en/1/awtvispolgrpgid.html

6. *EXISTING METADATA AND INTEROPERABILITY STANDARDS*

Any attempt to develop a data archiving and access policy must be mindful of potentially relevant standards. Many models for the architecture of data sharing and service provision rely heavily on adherence to standards.

6.1. *Introduction*

6.1.1. Metadata

Metadata are data that describe other data. Most commonly, the term "metadata" is used to mean information about data which can be used for searching or cataloguing data holdings. For example, if the metadata are being used to describe an online journal article then they could include the author, title, publisher, date and URL.

6.1.2. Metadata standards

Metadata standards relate to the elements that are chosen to describe data, the terminology or controlled vocabularies that are used within those elements and the formats the metadata are stored in.

Different metadata standards have different elements (for example, author, title, publisher) to describe data and sometimes identify the same element by different names. Although a large number of different standards exist, there are some widely accepted and used international metadata standards, a major one being the Dublin Core Metadata Initiative (http://dublincore.org/), a set of metadata descriptions about resources on the Internet, containing 15 data elements including title, creator, subject, description, data, type and format. Within the field of teaching and learning materials, the IMS Global Learning Consortium (http://www.imsproject.org/) defines and delivers interoperable, XML-based specifications for exchanging learning content and information about learners among learning system components. There are also accepted national standards such as the National Geospatial Data Framework (http://www.ngdf.org.uk/Metadata/met_guid.htm).

In addition, there are National and International Standards Organisations. The International Organization for Standardization (http://www.iso.ch/iso/en/ISOOnline.openerpage) has set more than 13,000 International Standards for business, government and society, including standards relevant to metadata and interoperability. For example, the Dublin Core conforms to ISO11179 for the description of data elements.

Most of the development of metadata standards proceeded from the viewpoint of library and other archive cataloguing. Therefore, there are a number of well-established library metadata standards such as classification systems and cataloguing formats, most notably MARC (MAchine Readable Catalogue Format).

6.1.3. Controlled vocabularies

Within different subject areas, there are controlled vocabularies for describing subject matter and assigning keywords to aid in cataloguing and searching. Use of a controlled vocabulary can add immensely to the power of a finding tool. If different data creators at different sites use the same limited set of keywords to describe the same subject content, then someone attempting to locate data on a particular topic can search far more precisely, and be sure that the resources discovered, even across widely different locations, correspond to the same desired concept. Use of controlled vocabularies and electronic thesauri help to mitigate some of the pitfalls of literal string or free-text searching such as use of multiple terms for a single concept ('Holland', 'the Netherlands', 'the Low Countries'), or multiple alternative spellings (labor and labour) and can even serve to map concepts across different languages and disciplines (e.g. the European Language Social Science Thesaurus ELSST: http://www.limber.rl.ac.uk/Internal/Deliverables/D4_2_final_V2.doc).

6.1.4. Interoperability and formats

Interoperability may be defined as the cooperation, interaction, or sharing among different archives, publishers, text formats, or information sectors, permitted by the use of common metadata standards, e.g. for storage, access, cataloguing, or communication.

In order to make an electronic collection automatically searchable by a wide audience, standards in terms of machine readability/understandability are also important. If the metadata elements can be mapped on to a standard computer-readable format/language then it enables computers to search across many databases located at different sites and controlled by different organisations. The system which seems to be potentially the most valuable in terms of coding concepts as computer readable descriptions is XML – eXtensible Mark-up Language. Whereas HTML (HyperText Mark-up Language) defines how elements are displayed (as in a web page), XML defines what those elements contain and allows tags to be defined by the developer. XML is a subset of SGML (Standard Generalized Mark-up Language) which is widely used in libraries. SGML and XML use a separate Document Type Definition (DTD) file that defines the format codes or tags embedded within it.

Interoperability is facilitated by standard query and communication standards such as SQL, Z39.50, CORBA and Microsoft COM. This subject, however, is beyond the remit of this report, as it has more to do with the development of software tools to locate and manipulate data resources, and less to do with what the creators of these data resources must do to ensure their data are locatable and manipulable by those software systems.

6.2. *The Biomedical Sciences*

It should be emphasised that there are no universally accepted metadata standards specific to the Biomedical Sciences, less still specific to population-based research, although some controlled vocabularies are more widely used and accepted than others. Often databases are set up that use their own controlled vocabularies and do not follow any predefined metadata standards. Although this report does not focus on metadata standards for genetic data, there is a separate section on XML and Biotechnology in Appendix 3.

6.2.1. Controlled vocabularies

The most widely used subject classification scheme for the medical sciences appears to be Medical Subject Headings (MeSH) from the National Library of Medicine (**http://www.nlm.nih.gov/mesh/meshhome.html**)

MeSH consists of a set of terms or subject headings that are arranged in both an alphabetical and a hierarchical structure. At the most general level are very broad headings such as 'Anatomy'. At more narrow levels are found more specific headings such as 'Ankle'. There is also a separate chemical thesaurus and thousands of cross-references to assist in finding the most appropriate MeSH heading.

Examples of the application of MeSH in the UK include the **Bristol Biomedical Image Archive** (http://www.brisbio.ac.uk/) and **OMNI** (Organising Medical Networked Information http://omni.ac.uk/) . The primary objective of Bristol BioMed was to make a shared resource of digital images available for reuse in the development of medical, dental and veterinary electronic learning and teaching materials. However, the usefulness of metadata originally attached to each image was frustrated by semantic and syntactic inconsistencies, use of multiple terms for a single concept and alternative spelling forms. To overcome this, MeSH was selected to fully exploit the metadata as an information retrieval resource.

UK-based OMNI, one of BIOME's (http://biome.ac.uk/) subject-specific gateways, offers free access to a searchable catalogue of Internet sites covering health and medicine and uses MeSH to index its records.

Additional controlled vocabularies and classification schemes in the areas of medical related research can be found in Appendix 3.

6.2.2. Emerging metadata standards

One potentially relevant emerging metadata standard which might be particularly suitable for some types of population-based datasets is the DDI (Dataset Documentation Initiative) standard (http://www.icpsr.umich.edu/ddi). This is an XML-based standard which is the outgrowth of an international collaboration of social science data professionals, and one upon which a number of medical-related developments have been based. Health Canada's Web-DAIS system (http://www.hc-sc.gc.ca/) is based around it, as is the US Bureau of Census DataFerrett system (http://dataferrett.census.gov/TheDataWeb/index.html) which delivers a number of US government-produced medical-related datasets. The UKDA catalogues its datasets to this standard, and provides datasets such as the Health Survey for England via the Nesstar online data browsing tools which are based around it (http://www.nesstar.org).

6.2.3. Metadata elements and formats

Some examples are given in Appendix 3 of sites specifying certain metadata elements and/or formats for metadata storage for different kinds of data in the medical domain. It can be seen that most of these are using XML. Biotechnology and genomic metadata are also being increasingly specified in XML.

6.2.4. Additional standards

Additional standards, such as those used for interchange of information among clinical health care providers, appear in Appendix 3.

7. *DATA SHARING POLICIES*

7.1. *Introduction*

There are a number of drivers for establishing data sharing policies. The key driver is that there is the growing world-view that 'data' are the primary building blocks of science. Second, legal requirements and public funding arguments are convincing motivations for research funders to establish mechanisms for enabling access to data. Third, demand from the research communities to gain access to expensive already collected data and the willingness to share their own data helps to get the issue onto the policy agenda. Finally, the dramatic advancements in the conduct of scientific research that collect massive amounts of data, that are often distributed and require expensive storage and analysis facilities, require suitable infrastructures to be in place. Opposing the drivers are barriers that can complicate data sharing policies - those of property rights and public privacy – although neither of these are insurmountable.

Enabling meaningful access to reliable scientific data merits attention to the preservation, archiving and sharing of scientific data.

Many funding bodies now recognise that there are a number of convincing reasons for investing in data sharing. The National Institute of Health (NIH, 2002) summarises these in a concise way,

> *"Sharing data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies on data collection methods and measurement, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new data sets when data from multiple sources are combined. By avoiding the duplication of expensive data collection activities, the NIH is able to support more investigators than it could if similar data had to be collected afresh de novo by each applicant.... However, NIH recognises that sharing data about human research subjects presents special challenges. The rights and privacy of people who participate in NIH-sponsored research must be protected at all times. Thus, data intended for broader use should be free of identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of individual subjects. Similarly, NIH recognises the need to protect patentable and other proprietary data and the restriction on data sharing that may be imposed by agreements with third parties."*

However, the take-up of investment in data sharing is, as yet, highly skewed across disciplines. The social sciences and humanities have led the way in implementing and promoting data policies, in some cases boasting a 30-year investment profile.

Researchers of the natural environment have found formalised ways of archiving and sharing research data, largely driven by various Information Acts and international treaties that demand access to environmental information. Dissemination of data from the Antarctic may be covered by the Antarctic Treaty and research data which constitute 'environmental information' within the meaning of the Environmental Information Regulations 1992, are often subject to the EC Directive on the Freedom of Access to Environmental Information.

In the field of physics and astronomy, scientists are also highly successful in sharing and accessing very large amounts of data via modern grid technology and this is being extended to other areas of 'Big Science' including genomics and medical images. The latest e-science initiatives are exploring ways to create a national infrastructure and the tools for data storage and analysis facilities for the natural sciences.

This section examines a number of data sharing practices across the world, drawing on practices from within and beyond the natural sciences, and provides short case studies as exemplars.

7.2. *The UK scene*

In the UK many research funders operate data sharing policies. Guiding principles are most evident in the social sciences and humanities, whilst amongst the natural sciences only the Natural Environmental Research Council (NERC) has a formal data policy. These policies vary in: how mandatory they are; how involved the recipient organisations are in appraising research applications and associated data management plans; the degree to which a budget line should be costed in for data preparation and documentation for archiving; and in rules on allowing researchers to place embargoes upon data.

7.2.1. Case 1: The Natural Environmental Research Council Data Policy

The Natural Environmental Research Council's Data Policy was created in 1996 to be consistent with legal frameworks such as the Environmental Information Regulations, the Antarctic Treaty and contractual arrangements with other bodies where, for example, NERC holds their data on a confidential basis without owning the Intellectual Property Rights.

NERC has a formal infrastructure of NERC Data Centres which work with NERC supported scientists to preserve data they create. NERC realises the value of its data by using it to further scientific understanding, create wealth and improve the quality of life. The Council recognises that this can be done in a variety of ways including: using datasets within NERC's own research and collaborative centres; giving, exchanging or licensing/selling them to other scientific researchers; licensing/selling them to commercial organisations which will themselves create wealth. The fact that data may be seen as a tradable asset is recognised in the data charging policy, which has been driven by the Freedom of Information Act.

The NERC Data Policy model is based on a distributed network of data centres with the NERC administration keeping the distributed system running smoothly. NERC data holdings are dispersed amongst seven key specialist NERC data research and resource centres across the UK. NERC's science-based archaeology community is further encouraged to deposit data with the Archaeology Data Service, part of the Arts and Humanities Data Service supported by the AHRB and JISC (see Case 3). Formal data management plans are required for all programme grants, and the data and research centres get involved in both peer review and planning.

NERC 'owns' all the data created via programmes and undertaken at NERC centres, but license the use of smaller datasets created via response mode grants. The NERC data charging strategy ensure that teams and individuals who have collected datasets are allowed a reasonable period (set out as a maximum of two years) of exclusive use during which to analyse them and publish results.

NERC has recently employed a Data Manager, located at one of the data centres, to provide the much needed interface between the Council, NERC supported researchers and the data centres. The NERC Data Policy is available on the web and as a published well-designed small A5 booklet.

7.2.2. Case 2: The Economic and Research Council Datasets Policy

The Economic and Research Council (ESRC) Datasets Policy was established in 1995 and reinforces and emphasises the ESRC's stated position relating to the acquisition and use of datasets, the requirements of which are now a condition of ESRC research funding. The ESRC requires all award-holders to offer for deposit copies of both machine-readable quantitative data, and machine- and non-machine-readable qualitative data, within three months of the end of the award. This relates not only to datasets arising as a result of primary data collection, but also to derived datasets resulting from ESRC-funded work.

In order to operate the Datasets Policy, the ESRC supports two Resource Centres with responsibilities for the cataloguing and archiving of data. The UK Data Archive (UKDA) based at the University of Essex is responsible for acquiring, documenting, disseminating and preserving digital data created during the course of ESRC research grants. The Qualitative Data Service (Qualidata), also at Essex, has special responsibility for qualitative in both digital and non-digital form. From January 2003, these will be combined within a single data archiving and dissemination service.

UKDA/Qualidata has a co-ordinated quantitative/qualitative acquisitions strategy which encourages the stream of qualitative data destined for archiving. Both centres have long-standing experience dealing with all aspects of acquisition and data collections management, including licensing agreements, working with academic award holders in the process of depositing data, and established relationships with other data producers, such as other research funders, and are well placed to operate the Datasets Policy.

The Datasets Policy requires that datasets must be deposited to a standard which would enable the data to be used by a third party, including the provision of adequate documentation. Depositors are advised to contact the two Resource Centres at the earliest opportunity should the nature of the data be such that it may be difficult to lodge the data. The earlier in the research process these discussions occur, the more likely researchers are to create datasets which are well-documented, free of confidentiality or licence constraints, and useable for secondary analysis. Support for award holders and potential depositors is generally provided through web-based guidelines and notes on preparing data for deposit. Support extends to adopting a more proactive role, working to promote the importance of sharing and preserving data within the social sciences and actively alerting award holders to their obligations.

Copyright in data deposited with the UKDA is retained by the copyright holder(s). Terms for access to the data are agreed with the copyright holder(s) and deposits are accompanied by a signed licence form. Usage of data is also subject to the acceptance by the user of a formalised access agreement, wherein the user undertakes to comply with the terms and conditions of deposit.

The UKDA holds a number of datasets from population studies and other socio-medical surveys, as outlined in Appendix 4.

UKDA/Qualidata were instrumental in helping to set the ESRC Datasets Policy, and in October 2000 proposed a set of changes to the operational procedures that would create a more robust, systematic and accountable policy. One of the central concerns of the current policy is that improved three-way communication channels between the ESRC, the award holders and the data archiving and dissemination services would be highly beneficial to the Resource Centres. The first suggestion was that the archiving and dissemination services should be involved across the life cycle of data generation, and, in particular enabling Resource Centre to have input at the grant application selection stage. Second, and in order to put the first proposal into place, the ESRC need to establish a fully co-ordinated strategy in-house, with dedicated staff to ensure the smooth running and auditing of the policy. Finally the Policy would, like the NERC principles outlined above, benefit from a requirement by data creators to produce a formalised data management plan at the application or short-listing stage, particularly for expensive research programmes; and a more stringent view towards the length of time allowed for data embargo.

The ESRC Datasets Policy is currently under review but a summary of the recent Policy can be found in section 17 of the ESRC *Guide to Research Funding*.

7.2.3. Case 3: The Arts and Humanities Research Board Information and Communications Technology (ICT) Policy

Through their research grants the Arts and Humanities Research Board (AHRB) seek to encourage the development of high-quality scholarly data resources in the arts and humanities. They support applied research in areas which promise to enhance or extend the use of ICT in the development and use of scholarly information resources and encourage the development of training and other related materials which seek to encourage effective scholarly use of digital resources and ICT. As such, a draft joint ICT Policy between the AHRB and the Arts and Humanities Data Service (AHDS) was established in 1999. This includes guidelines for AHRB grant applicants and award holders who include ICT in their projects.

The AHDS was initially established by the Joint Information Systems Committee (JISC) in 1994 as a distributed service comprising a managing Executive and five subject-based Service Providers offering data archiving and data resource provision to archaeology, history, the performing arts, textual studies, and the visual arts. Subsequently, the AHDS has received funding from the AHRB, who is now its major stakeholder.

The AHDS has a major role in advising the AHRB on matters involving ICT which seek to extend and enhance scholarly uses of digital resources and information technologies in the arts and humanities. In pursuit of this main aim they advise the AHRB in the technical assessment of those grant applications which seek to produce electronic materials, to conduct applied research, or to develop training and other materials, and provide advice on standards and best practices. Finally, the AHDS also contributes to selected AHRB-funded projects, and on a cost-recovery basis, a range of value-added expertise and services.

The AHRB normally gives preference to applications which promise to make data, applied research, and training and related materials available to the arts and humanities research, teaching, and learning communities within the UK. Applicants proposing data creation projects must include a completed technical appendix describing the value added to the project by ICT and the project's proposed methods.

As with the ESRC, under the Policy, data are licensed from the copyright holder(s) for dissemination by the AHDS by way of a non-exclusive licence agreement, which defines the terms and conditions for access by secondary analysts.

7.2.4. Others UK

The Joseph Rowntree Foundation (JRF) has a Project Funding Agreement that expects award holders to offer all machine readable data collected as part of the Project to the UKDA for deposit and, if required, to lodge the data and documentation, in an appropriate format, within six months of the completion of the project. The Humanities Research Board of the British Academy also operates a formal data policy for large award holders, where data should be offered for deposit to the AHDS or UKDA within a reasonable time after the completion of the project.

Finally, other social science and humanities funders have informal data sharing policies that both encourage and recommend grant holders to deposit data with the UKDA, Qualidata or the AHDS. The Leverhulme Trust, The Wellcome History of Medicine Programme and the Society in Medicine Programme and The Carnegie Trust for Scottish Universities all operate such polices.

At this time, the UK government has begun to take a closer interest in data sharing. The report *Privacy and data-sharing: The way forward for public services*, published by the Performance and Innovation Unit of the Cabinet Office in April 2002 looks at the issues of privacy and data-sharing in delivering public services, and charts the way forward. One of the report's recommendations is:

> *"To encourage widespread adoption of such standards, the Lord Chancellor's Department, working in conjunction with the Public Record Office, should facilitate the development and dissemination of model data-sharing protocols and codes of practice as a resource to public sector organisations. This work will need to draw on a wider understanding of the overall information architecture of government, which maps the creation, flows and uses of information sets, establishes criteria for its sharing, retention and disposal, and allocates responsibilities for sustaining access, quality, reliability and safe-keeping."*

7.3. *The United States scene*

A number of federal agencies in the US have formal data sharing polices. This is notably a result of US federal government law and policy, under which publicly funded information, including research data, should be in the public domain. National scientific organisations have made a commitment to the sharing and archiving of data through their ethical codes and publication policies. Over 15 years ago, the National Academy of Sciences described the benefits of sharing data (Fienberg 1985). For many years, the National Science Foundation (NSF) Economics Program, has required data underlying an article arising from an NSF grant to be placed in a public archive. Other Departments are committed to investment in providing public data access systems for data they collect. Moreover, many scientific journals require that authors make available the data included in their publications. In the biological sciences, protein and DNA sequences are made available to researchers through data archives, such as GenBank.

Finally, if data are cited in a Federal regulation or administrative order, then they may be accessible through the Freedom of Information Act (FOIA). Wouters' report (2000) further details the main principles and regulations in the US.

7.3.1. Case 1: The National Science Foundation

The National Science Foundation (NSF) advocates and encourages open scientific communication, and is committed to the principle that the various forms of data collected with public funds belong in the public domain. The NSF expects significant findings from research and educational activities it supports to be promptly submitted for publication, with authorship that accurately reflects the contributions of those involved. It expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections, and other supporting materials created or gathered in the course of the work. It also encourages awardees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable.

The Division of Social and Economic Sciences (SES) has formulated a policy to facilitate the process of making data that has been collected with NSF support available to other researchers.

As SES supports a wide range of disciplines, the nature of the data, the way they are collected, analysed, and stored, and the pace at which this reasonably occurs vary widely. Grant holders from all fields are expected to develop and submit specific plans to share materials collected with NSF support, except where this is inappropriate or impossible. These plans should cover how and where these materials will be stored at reasonable cost, and how access will be provided to other researchers, generally at their cost. Data include quantitative, qualitative and experimental research data, and mathematical and computer models.

For appropriate datasets, researchers should be prepared to place their data in fully cleaned and documented form in a data archive or library within one year after the expiration of an award. Before an award is made, investigators will be asked to specify in writing where they plan to deposit their dataset(s). This may be the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan, but other public archives are also available. Investigators are invited to consult with NSF program staff about the most appropriate archive for any particular dataset.

The Division of Earth Sciences (EAR) at the NSF, is responsible for the implementation of the Foundation's Data Sharing Policy on Earth Science data. The overall purpose and fundamental objective of the policy is to ensure and facilitate full and open access to quality data for research and education in the Earth Sciences. The Policy guidelines are considered to be a binding condition on all EAR-supported projects.

7.3.2. Case 2: The National Institute of Health

Since 1996 the National Institute of Health (NIH) has required data sharing in several areas, such as DNA sequences, mapping information, and crystallographic co-ordinates.

In March 2002, the NIH further announced that it was developing a statement on data sharing that expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers. The statement on data sharing is an extension of NIH policy regarding sharing research resources, which expects that recipients of NIH support will provide prompt and effective access to research tools. Furthermore, it is consistent with the policies of many scientific journals publishing the findings of NIH-supported research.

The NIH expects investigators supported by NIH funding to make their research data available to the scientific community for subsequent analyses where possible. Consequently, the NIH will require that data sharing be addressed in grant applications and in the review of applications. Funds for sharing or archiving data may be requested in the original grant application or as a supplement to an existing grant. Investigators who incorporate data sharing in the initial design of the study can more readily and economically establish adequate procedures for protecting the identities of participants and provide a useful dataset with appropriate documentation. Applicants whose research will produce data that are not amenable to sharing should include in the application reasons for not making the data available. NIH encourages investigators to consult with an NIH Program Administrator prior to submitting an application to determine the appropriateness of data sharing and a suitable mechanism to disseminate the data.

The policy is still in its consultation phase.

7.3.3. Case 3: The National Center for Health Statistics

The National Center for Health Statistics (NCHS) is the US's principal health statistics agency, and is responsible for collecting accurate, relevant, and timely data. NCHS' mission, and those of its counterparts in the Federal statistics system, focuses on the collection, analysis, and dissemination of information that is of use to a broad range of users. NCHS surveys and data collection systems include some high profile and much used population studies: the National Health and Nutrition Examination Survey (NHANES); the National Health Care Surveys (NHCS); the National Health Interview Survey (NHIS); the National Immunization Survey (NIS); the Longitudinal Studies of Aging (LSOA); and the National Vital Statistics System (NVSS).

All data collected by NCHS are collected under the authority of the Public Health Service Act that demands the timely release of data and mandates that data be made available on as wide a basis as is practicable. Public-use data files are released via web download as soon as they have been prepared and the necessary reviews and approvals have been obtained, including a full disclosure review. Release, to a collaborator, of files that are not yet ready for public release is permissible under NCHS confidentiality policy, but must be consistent with NCHS legislative authority, informed consent, and submissions for human subjects reviews. Such releases are normally carried out under an agreement specifying how appropriate confidentiality protections are to be provided by the collaborator.

NCHS does not "embargo" data that are otherwise ready for public release and does not provide collaborators with preferential early access to data files or tabulations that are otherwise ready for public release.

7.3.4. Case 4: The Robert Wood Johnson Foundation,

The Robert Wood Johnson Foundation, the largest health care philanthropic organisation in the United States, sponsors the Health and Medical Care Archive (HMCA) housed at ICPSR as its official data archive of the Foundation. The Foundation is devoted to preserving and making available research data that have significant secondary-analytic value for expanding knowledge on, and ultimately contributing to the improvement of, the health of people in the United States. Included in this archive are surveys of health care professionals, investigations of access to and financing of medical care, evaluations of innovative programs for the delivery of health care, and surveys of substance abuse.

From among the many research projects that it funds, the Foundation designates a selection of projects that are required to submit their data to HMCA. The timely submission of these data collections is closely monitored by the Foundation in collaboration with the archive. HMCA provides advice to grant holders on the preparation of their data collections for submission to the archive in a manner optimally useful for secondary analysis.

7.3.5. Other US

In the case of US clinical trial data, for example those relating to AIDS sponsored by the National Institute of Allergy and Infectious Diseases (NIAID), data are submitted to the National Technical Information Service (NTIS) for public use. Examples of shared data from clinical trials include the Asymptomatic Cardiac Ischemia Pilot, the Intermittent Positive Pressure Breathing Study, and the Safety and Efficacy Trial of Zidovudine for Asymptomatic HIV Infected Individuals.

The National Institute on Aging and the Substance Abuse and Mental Health Services Administration of (SAMHSA) the United States Department of Health and Human Services both support National Archives of Computerized Data managed and hosted by ICPSR. Their missions are to advance research in these areas by helping researchers to profit from the under-exploited potential of a broad range of datasets. The Archives promote the sharing of these data among academics, policymakers and service providers to enable greater understanding of the issues. SAMHSA provides public use files from its major data collection systems for on-line analysis to provide ready access to substance abuse and mental health research data.

The NYS GIS Data Sharing Cooperative is a group of governmental entities and not-for-profit organisations that have executed Data Sharing Agreements for the purpose of improving access to GIS data among members. The Data Sharing Cooperative was primarily developed to encourage public agencies in New York to share in the creation, use, and maintenance of GIS data sets at the least possible cost. Two key features of the agreement are that data creators (primary custodians) retain ownership of their GIS datasets, but agree to share it with other Cooperative members for free; and that secondary users of the GIS data pass updates, corrections, and revisions back to the creators of the data set, resulting in improved data quality.

7.4. *The international picture*

Since 2000 there has been a growing number of international collaborative efforts and projects that are committed to action on data sharing in the sciences. The major initiatives in this area are discussed below, but it should be noted that much of the work is still very much in progress.

7.4.1. CODATA (Committee on Data for Science and Technology) Working Group on Archiving Scientific Data

CODATA is an interdisciplinary Scientific Committee of the International Council for Science (ICSU) and has been in existence for over 30 years. Over 20 countries from across the world are members, but UK does not appear to be listed. In 2001, CODATA established a working group to focus on the special issues of archiving scientific data. In October 2000, a planning meeting was held to review issues related to the archiving and preserving of scientific data, organised during the 17th International CODATA Conference.

CODATA define scientific data as "*numerical quantities or other factual attributes generated by scientists and derived during the research process (through observations, experiments, calculations and analysis)*".

The working group aims are to propose a position paper on archiving scientific data that:

- develops a comprehensive annotated bibliography on the subject;
- documents the diversity of best practices and identifies the 'best' ones in the area of data archiving and preservation across the science domain;
- describes the role and function of CODATA in preserving and archiving scientific data;
- describes the major issues to be considered in archiving scientific and technology data;
- identifies other efforts in data archiving and preservation (e.g. efforts by ICSTI, ISO, NARA, NRC etc.).

The initial group was not representative of all scientific communities and other communities, especially from genomics, geo-sciences, astronomy, chemistry, high-energy physics, statistics, psychology, and museum collections have been approached. In addition, various panels are to be constituted to give attention to the different science disciplines. This progress of this group should certainly be followed up by MRC, and it is advised that the MRC contribute to the ongoing programme of work.

7.4.2. OECD Working Group on Issues of Access to and Sharing of Research Data from Public Funding

The third CSTP/OECD Global Research Village Conference held in 2000 addressed policy implications of the use of ICT for the global science system (OECD Working Group 2002). The key recommendation coming out of the Conference was to focus upon and draw up commonly agreed principles to guide access to publicly financed research (Access to and

Sharing of Research Data from Public Funding). Over the summer of 2001, experts from the Netherlands, Denmark, Poland, US and the ESF initiated the programme of work. Other member countries include Australia, Canada, Finland, Germany, Poland, but not UK. The group works in close liaison with the ESF, NSF and CODATA (see above). The working Group aim to produce a report that will include a science policy section, and a section on the impact of the sharing of data on the quality of the research process.

Activities of the group include a range of data gathering exercises and expert meetings. First in 2002 Paul Wouters from the Netherlands Institute for Scientific Information Services (NIWI) produced an elementary study on the State of Affairs. Second, US participants gained support from the NSF for a project combining scientific research and policy research into data sharing (to be finalised by Spring 2003). Third, a funded study on national legislation relevant to access to and sharing of research data was conducted by a professor of Law at Leiden University.

Wouters' report (2002) found that over half of the countries who responded to his mini survey (21) considered data sharing to be an issue of science policy, and currently on the policy agenda. Very few countries had national legislation for data sharing, whilst others had various policies addressing data sharing.

Following these activities an expert meeting was held on the economics and management of digital research data for the Global Science System held in Maastrict in September 2002. MRC would be advised to follow the progress of this Working Group and volunteer to participate where appropriate.

7.4.3. Bioinformatics initiatives

The bioinformatics community has well-established data sharing practices, at the national, European and wider levels. A good example is the European Molecular Biology Laboratory (EMBL) which is supported by 16 countries including nearly all of Western Europe and Israel, and consists of five research/data service facilities across Europe. The UK facility, the European Bioinformatics Institute (EBI) is a non-profit academic organisation that forms part of this network. The mission of the EBI is to ensure that the growing body of information, including databases from molecular biology and genome research is placed in the public domain and is freely accessible to all facets of the scientific community in ways that promote scientific progress. The Institute manages databases of biological data including nucleic acid, protein sequences and macromolecular structures. On the same site is the Wellcome Trust sponsored Sanger Institute for genome research and the UK Medical Research Council Human Genome Mapping Project Resource Centre (HGMP). Together, the three institutes provide one of the world's largest concentrations of expertise in genomics and bioinformatics.

An example of a high quality truly international product of the advanced collaborative activities in this field is the EMBL Nucleotide Sequence Database which constitutes Europe's primary nucleotide sequence resource. This comprehensive database of DNA and RNA sequences is collected from the scientific literature and patent applications and directly submitted from researchers and genome sequencing groups. Data collection is done in collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ).

In terms of a data sharing policy, it is a commonly accepted principle that European researchers wishing to publish in the field of genomics and protein data research should deposit their sequences and analyses with one of the three genetic data bank sites.

7.4.4. EU DataGrid Project

DataGrid is a project funded by European Union with the main objective of building the next generation computing infrastructure providing intensive computation and analysis of shared large-scale databases, at the PetaBytes level, across widely distributed scientific communities. The DataGrid project brings together scientists from Biological Sciences, Earth Observation and High-Energy Physics where large-scale, data-intensive computing is essential.

In April 2002, The UK government officially opened the National e-Science Centre (NeSC), a centre that is part of an international effort to enable scientific researchers to take advantage of the computing power distributed across the globe. In March 2002, the project released a version of its Testbed middleware software, that has proved capable of reliable job distribution over the five main sites, including RAL and CERN. The projects Biology testbed aims to provide a platform upon which to store, share and analyse databases of genomic data and medical images.

7.5. *Conclusion*

This report has highlighted some of the key national and international data-sharing activities. The UK and the US provide notable examples of the will to share publicly funded data for research purposes, and have established data sharing procedures and regulations.

The bio-medical sciences, with the exception of the fields of genomics and medical imaging, are perhaps the least advanced in this respect, but efforts to examine the benefits of data sharing policies are taking place at the national level (MRC) and at the international level.

In summary, and drawing on Wouter's report (Wouters' 2000) data sharing principles rest upon a number of key issues:

- whether public access to data is stated as a basic policy principle;
- what the motivation for data sharing rules are;
- whether data sharing is a condition of research funding;
- who is responsible for providing access to data;
- whether data types are distinguished;
- how issues of property rights are treated;
- how the limits of data sharing are recognised.

The timing of these new developments is opportune for MRC in considering a data sharing policy. MRC should draw on work in progress arising from the rising tide of activity seeking to explore and implement data-sharing policies in the sciences.

7.6. *References and organisation URLs*

AHRB ICT Policy *http://www.ahrb.ac.uk/strategy/c_it_policy.htm*

Committee on Data for Science and Technology (CODATA)
http://www.codata.org/about.pdf
CODATA Working Group on Archiving Scientific Data *http://www.nrf.ac.za/codata/*

NYS GIS Data Sharing Cooperative *http://www.nysgis.state.ny.us/coop_gis.htm*

DATAGRID *http://eu-datagrid.web.cern.ch/eu-datagrid/*

European Molecular Biology Laboratory (EMBL) *http://www.embl-heidelberg.de/ExternalInfo/GeneralInfo/*
EMBL-EBI *http://www.ebi.ac.uk/Information/index.html*

ESRC Datasets Policy *http://www.esrc.ac.uk/esrccontent/researchfunding/sec17.asp*

Us Freedom of Information Act *http://www.usdoj.gov/04foia/*

Inter-University Consortium for Political and Social Research (ICPSR) *http://www.icpsr.umich.edu/*

Joseph Rowntree Foundation (JRF) Project Funding Agreement *http://www.jrf.org.uk/funding/applyforfunding/agreement.asp*

Fienberg, S. *Sharing Research Data*, Committee on National Statistics, National Research Council)

The National Center for Health Statistics (NCHS) *http://www.cdc.gov/nchs/datawh/ftpserv/ftpdata/ftpdata.htm*

The US National Institute in Aging *http://www.nih.gov/nia*

National Technical Information Service (NTIS), US Department of Commerce *www.ntis.gov*

NSF Data Sharing Policy *http://www.nsf.gov/search97cgi/vtopic*

NSF Data Sharing Policy for Earth Sciences *http://www.geo.nsf.gov/ear/EAR_data_policy_204.doc*

NERC Data Policy http://www.nerc.ac.uk/data/policy.shtml

OECD (2002) *Public Domain of Digital Research Data*, Report from the Follow up Group on Issues of Access to Research Data from Public Funding, 2002

Performance and Innovation Unit of the Cabinet Office (2002) Report on 'Privacy and data-sharing: The way forward for public services', *http://www.cabinet-office.gov.uk/innovation/2002/privacy/report/*

Qualitative Data Service *http://www.qualidata.essex.ac.uk/*
Qualitative Data Service Guidelines for Data Creators *http://www.qualidata.essex.ac.uk/creatingData/introduction.asp*

The Robert Wood Johnson Foundation *http://www.rwjf.org/*

The US Substance Abuse and Mental Health Services Administration
http://www.samhsa.gov/

UK Data Archive *http://www.data-archive.ac.uk/*
UK Data Archive Guidelines for Data Creators *http://www.data-archive.ac.uk/creatingData/advice.asp*

Wouters, Paul (2002), 'Data Sharing Policies', Report prepared for the OECD Follow up Group on Issues of Access to Research Data from Public Funding, NIWI-KNAW

8. *MODELS OF SERVICE PROVISION*

Drawing on the lessons of the site visits and the examples of other data organisations and research funding bodies, it is now possible to draw a series of possible architectural models for the provision of data archiving and access services for MRC population-based research. These models fall in a continuum from centralised and integrated to de-centralised and distributed.

8.1. *Fully centralised: the national archive model*

8.1.1. Description

Under a fully centralised model, data archiving and dissemination all occur within a single centralised facility. Researchers would be required to pass over data and documentation to the centre, likely under a rights management framework which either cedes copyright to the centre, or sets out the conditions under which the centre is permitted to disseminate the data. This model was the basis of the original ESRC Data Archive.

8.1.2. Advantages

- Centralised services are often the most cost effective, since infrastructural costs need not be replicated across multiple sites, and often there are economies of scale.
- A centralised MRC archive would presumably be staffed by trained data archivists, and therefore issues such as development of and adherence to standards could be more easily managed and enforced.
- By centralising expertise, extensive training in data preservation and management within data producing units is not necessary, also a cost saving.
- From the secondary analyst's perspective, there is a simple 'one-stop-shop' for locating and accessing data and for supporting the use of those data.
- A unified, standardised rights management framework is also easier for users to navigate - a single user license could cover the entire collection.
- The service would remove the burden of data preservation and dissemination from research staff, who may not be equipped, skilled, or resourced to deal with it.

8.1.3. Disadvantages

- This is the least popular solution among Principal Investigators, largely because too much control would be relinquished.
- A single centralised service cannot possibly provide expert content support for each and every dataset in its collection ('jack of all trades, master of none' syndrome).
- Not all datasets can be adequately anonymised, and therefore to protect fully respondent confidentiality, onsite collaboration with PIs rather than a centralised service would be the appropriate dissemination mechanism.

8.2. *Centralised infrastructure/distributed expertise: the AHDS/ESRC model*

8.2.1. Description

This model is a compromise which attempts to harness the advantages of both the centralised and distributed models. In it, the basic infrastructure and mechanics of acquiring, validating, preserving, and disseminating data are undertaken by a central service. Supporting users in the use of particular datasets, however, remains the responsibility of centres of expertise; most likely the researchers who created the data themselves, or those with particular expertise in a certain type of data, methodological approach, or subject matter.

Examples of this model include the Arts and Humanities Data Service (AHDS), and the new ESRC data archiving and dissemination service. There is a core management and coordination function which oversees the preservation of data, the machine infrastructure, and the development of the service; whilst the support and promotion of the use of the data falls to specialist services on a subject basis (AHDS) or a data type/methodology basis (ESRC).

8.2.2. Advantages

- This model maintains most of the centralised advantage of reducing infrastructure costs.
- Centralised control of preservation ensures enforcement of standards.
- If the coordination is well-managed, the service still appears to be a seamless 'one-stop-shop' from the user's point of view.
- Unified rights management framework, as above.
- Lifting the burden of preservation and dissemination off the researcher.
- User support is high quality, better focused, and more complete than a purely centralised model.
- Researchers may maintain a measure of involvement in the use of their data.

8.2.3. Disadvantages

- PIs may still object to loss of control over data.
- Issues of anonymisation still apply.
- Will be more expensive than a centralised model.

8.3 *Subject-focused distributed service: the NERC model*

8.3.1. Description

Under this model, the data world is divided into a number of subject areas, and a series of data centres are established with responsibility for data within each area. Each data centre replicates the functions of data acquisition, validation, preservation, dissemination, and user support. An example of this model is the Natural Environmental Research Council's Data Centres.

8.3.2. Advantages

- Since data within certain subject areas may have certain commonalities, data management practices can be better tailored to the specific data types.
- Likewise preferred methods of accessing and using data may share commonalities within subject areas, so dissemination services can also be better tailored.
- User support is likely to be more focused and of higher quality, particularly if (as is the case with NERC) these data centres are also research centres.

8.3.3. Disadvantages

- Much more costly, as services, infrastructure and expertise are replicated across sites.
- Still requires PIs to give up control; and may in fact be potentially more contentious, as control might be ceded to researchers who are actually viewed as competitors in their field.
- Without some fairly strong overarching management, standards and rights management frameworks may be difficult to agree and enforce.
- Whilst satisfying the mainstream disciplinary researcher as an appropriate "one stop shop" it may discourage cross-disciplinary and multi-disciplinary research, since multiple "shops" would need to be navigated.
- Potential confusion amongst both users and depositors, as disciplinary boundaries are not always clear.

8.4. *Virtual integration only: the GRID model*

8.4.1. Description

This model is a more purely distributed model, where local data producers make their data available via common standard interoperability protocols, and the integration occurs on the user's desktop. This is a very flexible model, in that it allows for the inclusion of a tremendous range of data held anywhere in the world, the only caveat being that the data must be presented in such a way that the virtual tools can read them. Whilst harnessing the widest range of data resources and taking fullest advantage of the incredibly rich existing information landscape, it does not in itself mandate or control the preservation of those resources – this is left to the individual sites.

8.4.2. Advantages

- Control over the data is left in the hands of those who know and understand it best.
- Potential for innovative cross-disciplinary, cross-method research increased.
- Cost of the service essentially falls to the local data producer (but has to be provided for somehow).
- Provenance of the data is immaterial, since the tools locate and present them as virtually integrated.
- Takes best advantage of the existing investment of other organisations worldwide.

8.4.3. Disadvantages

- Preservation is left to the data producer. If adequate preservation is mandated, then each data producer is required to acquire the hardware and skills to undertake it him/herself, or outsource it.
- Rights management is a potential stumbling block in this model, as either only freely available data are included, or the user must jump through multiple authorisation hoops to gain access to multiple resources.
- User support is left to the data producer – no one stop shop for support in using the data (although presumably what content support there is would be highly focused and very expert).

8.5. *The portal approach: the RDN model*

8.5.1. Description

This is not strictly a data archiving and access model at all. In it, data access is enabled by the funding not of an archiving or dissemination service, but of a service which establishes and maintains a portal which links through to existing resources elsewhere. The service consists in the identification and creation of searchable, browsable metadata for resources which actually reside anywhere in the world. It is a useful service for users, a one-stop shop for the *location and identification* of resources, but it does not in itself provide for accessing or acquiring resources. Preserving data is not in the service's brief, although it would of course maintain and preserve its own metadata repository. The Resource Discovery Network (RDN http://www.rdn.ac.uk/) is an example of this model.

8.5.2. Advantages

- A portal is much less costly than a full-blown archive.
- Would assist in locating data resources outside of MRC science.
- Can bring together research data resources with other resources (articles, publications, web sites, etc) in one metadata framework and one search/browse environment.

8.5.3. Disadvantages

- Does not provide for either short or long term data preservation.
- Does not ensure data access, only access to information about research data; thereafter researchers are left to negotiate access to data resources themselves.

9. *RECOMMENDATIONS*

It is beyond the scope of this report to present specific recommendations to the MRC concerning the specific form and content of its data archiving and access policy and service provision. However, this report offers some general observations and lessons learned from the course of the site visits and accompanying research.

9.1. *Encouraging the culture of data sharing*

It is clear that most researchers support the idea of data sharing. However, in discussion of the mechanisms of how access would operate, there is a reluctance to relinquish control. There are significant cultural barriers that need to be overcome, and researchers need to be encouraged to think in terms of sharing data as a normative activity, as it is in other areas of MRC science. Formalising access control procedures and protocols might serve to make researchers more comfortable with data sharing. It is important that a policy which mandates data sharing not be confused with a requirement to open all data to free and uncontrolled access. Researchers will understandably feel far more comfortable if there is a sense that secondary users have been in some measure "vetted" before access is allowed to potentially sensitive or particularly complex data. Depending upon the access control procedures implemented, such protocols, however, may come to seem unnecessarily onerous to the secondary analyst, who may quite reasonably feel that the peer review process already in place for research proposals and resulting publications would obviate such a requirement. Enforcing the sharing of data may require both "carrots" and "sticks" in the form of additional funds to researchers to prepare their data for secondary analyses, and perhaps making the incorporation of data sharing into the research plan a condition of the grant award.

9.2. *Standards and guidelines*

It is clear that there is a demand for the re-use of population data in the medical research field, as well as a good scientific and economic rationale for its preservation and dissemination. It is also equally clear that in order for data to be reasonably re-usable and preservable, certain standards, particularly of dataset documentation, must be met, and that researchers would welcome assistance in meeting them, particularly in the form of published guidelines and advice.

9.3. *Rights management framework*

A good rights management framework is essential for the establishment of a data archiving and access service. Rights management in its broadest sense covers not only the relationship between the data creators and secondary data users, but also protects the rights of institutions, funders of research and even research subjects. A unified rights management framework for MRC data would greatly improve a data access service, no matter what the particular model of service provision.

9.4. *Consent*

Consent arrangements need not preclude the use of data for secondary analysis, and it is recommended that the MRC draft appropriate guidelines for the wording of consent agreements which would allow secondary analysis. It is important when issuing guidelines on consent and data sharing that the MRC educate both researchers and other bodies who are concerned with the ethics of research with human subjects.

9.5. *Selection of data for DAA*

All data are not created equal, and in a world of limited resources, priorities for DAA must be drawn. Investigators' views should be of great value in helping MRC to consider a strategy for both assessing and prioritising datasets for archiving. In summary, the following criteria may be most appropriate for making these decisions, some of which were previously suggested from the MRC DAA Phase I survey:

- the degree of uniqueness or the size/longevity of the study/dataset;
- the reputability and quality of the study methods and dataset;
- the potential to answer new important research questions cost-effectively;
- the likelihood that a dataset can be pooled or combined with other data to provide power that individual datasets cannot provide;
- the degree to which the study fulfils ethical and legal requirements to enable the re-use of personal information;
- the degree of MRC sponsorship;
- the anticipated cost of preparing a dataset to professional archival standards;
- studies that were conducted recently;
- building in prospective data preparation and documentation plans that enable secondary access for all new studies .

A set of criteria such as these would provide the framework in which to apply independent evaluation of MRC datasets for archiving.

9.6. *Desired services*

There was nearly unanimous support among researchers for published guidelines for good practice in the management of data and their preservation. Many also supported the idea of a data archiving and access advisory service. A central registry of research instruments also received strong support, particularly among younger researchers and project managers. Support for a freestanding data archiving and dissemination service was mixed, as many researchers feared losing control altogether; support was stronger for a preservation service.

9.7. *Service provision*

There are a number of different possible models for the architecture of service provision for data archiving and access, ranging along a continuum from completely centralised to completely virtual and distributed. Each model has advantages and disadvantages which must be weighed, including cost effectiveness, quality and focus of service and support, and ease of use and navigation for both data creators and data users, and support among the research community. Centralised services are often most cost effective because of lack of replication of infrastructure and expertise, and may be an easy focal point or "one stop shop" for secondary users and data creators, but may also offer the least focused user support, and may face the greatest resistance from researchers concerned about access control. Different models also require, and enable, different degrees of data and metadata standardisation. The

MRC must determine the relative priority of these competing factors in choosing a model for service provision.

9.8. *Policy and resources*

Regardless of the model of service provision, funding is required for long-term preservation and to facilitate data access and user support. Any data archiving and access policy needs to be adequately resourced, and may also require some contractual "teeth" to overcome cultural barriers and ensure compliance.

APPENDIX 1: SITE VISIT SUMMARIES

N.B. These are not yet available for distribution

I. MRC SOCIAL, GENETIC AND DEVELOPMENTAL PSYCHIATRY RESEARCH CENTRE, INSTITUTE OF PSYCHIATRY, LONDON

II. MRC ENVIRONMENTAL EPIDEMIOLOGY UNIT, SOUTHAMPTON

III. MRC SOCIAL AND PUBLIC HEALTH SCIENCES UNIT, GLASGOW

IV. DEPARTMENT OF COMMUNITY HEALTH SCIENCES, UNIVERSITY OF EDINBURGH MEDICAL SCHOOL

V. PARKINSON'S DISEASE DNA BANK, DEPT. OF NEUROLOGY, DUDLEY ROAD CITY HOSPITAL, BIRMINGHAM

VI. MRC NATIONAL SURVEY OF HEALTH AND DEVELOPMENT, UNIVERSITY COLLEGE LONDON

VII. MRC CLINICAL TRIALS UNIT, LONDON

APPENDIX 2: DATASET DESCRIPTIONS

1. Twins' Early Development Study (TEDS)

Study Title	Twins' Early Development Study
Study Acronym	TEDS
Unit/Centre Name	MRC Social, Genetic and Developmental Psychiatry Research Centre (SGDP)
Unit/Centre Address	Institute of Psychiatry, London http://www.iop.kcl.ac.uk/main/ResRep/Centre.htm http://www.iop.kcl.ac.uk/IoP/Departments/SGDPsy/research/TEDS.stm
Unit/Centre Director	Prof Peter McGuffin
Date of site visit	15/05/2002
PI	Prof Robert Plomin
Summary of major aims	Genetic and environmental investigation of the early emergence and persistence of language difficulties and other problems and their links to externalising behaviour problems. About 16,000 pairs of twins born in 1994-96 in England and Wales have been recruited. Twins showing persistent difficulties through to 4 years of age (and an unselected control group) are selected for intensive study in their homes. Quantitative genetic analyses of the total sample and the selected subsamples will be performed in order to elucidate the interplay between genetics and environment in the early development of these mild mental disorders and their co-occurrence with behaviour problems.
Study Location	England and Wales
Duration	1995-2000 2000-2005
Start Date	1995
Principal funding source	MRC
Study type	Longitudinal
Variables	Approx. 1500
Cases	33,000
Variable types	99% coded, a little verbatim text, a few dates, heights/weights

2. Depression Case Control (DeCC) Study

Study Title	Depression Case Control Study - A DNA collection for Case-Control studies of unipolar depression
Study Acronym	DeCC
Unit/Centre Name	MRC Social, Genetic and Developmental Psychiatry Research Centre (SGDP)
Unit/Centre	Institute of Psychiatry, London

Address	http://www.iop.kcl.ac.uk/main/ResRep/Centre.htm
Unit/Centre Director	Prof Peter McGuffin
Date of site visit	15/5/2002
PI	Prof Peter McGuffin
Summary of major aims	Family, twin and adoption data provide a compelling case that major depression is strongly influenced by genes, but suggest that this is likely to be, as with other common conditions, the result of several, perhaps many genes of small effect. The genetic contribution is complicated and almost certainly involves an interplay between multiple genes and environmental factors. This project aims to establish a resource for one of the main methods of locating and identifying genes, allelic association. The DNA collection is comprised of samples from patients suffering from severe recurrent unipolar depressive disorder and ethnically matched controls screened for absence of psychiatric disorder. The depressed participants are over the age of eighteen and have had two or more episodes of depression as defined by the research criteria of ICD10 and by DSMIV. The sample size will allow sufficient power to detect or replicate associations with susceptibility loci of modest or small effect and will be suitable both for candidate gene studies and linkage disequilibrium mapping in genomic regions of interest identified by linkage studies.
Study Location	SGP Research Centre/South London & Maudsley NHS Trust, Cardiff, Birmingham
Duration	2000-2003
Start Date	Grant Awarded 2000
Principal funding source	MRC
Study type	Genetic, trial
Variables	Approx. 1800
Cases	365 current, 800 target
Variable types	most coded, some ordinal, few dates

3. Isle of Wight Studies

Study Title	Isle of Wight Studies
Study Acronym	IoW
Unit/Centre Name	MRC Social, Genetic and Developmental Psychiatry Research Centre (SGDP)
Unit/Centre Address	Institute of Psychiatry, London Http://www.iop.kcl.ac.uk/main/ResRep/Centre.htm
Unit/Centre Director	Prof Peter McGuffin
Date of site	15/05/2002

visit	
PI	Prof Sir Michael Rutter Dr Barbara Maughan Jack Tizzard
Summary of major aims	Wave A: To assess the prevalence and correlates of psychiatric disorder and reading difficulties in childhood and adolescence, and the implications for the provision of services. Wave B: To examine continuity and discontinuity of disorders/difficulties over time (changes between childhood, adolescence and adulthood).
Study Location	Isle of Wight
Duration	1964/65 1968/69 Ongoing followup
Start Date	Start of grant: 1963 Start of data collection 1964
Principal funding source	Current funders: MRC. Previous funders: DoE, Nuffield Foundation, SSRC (ESRC)
Study type	Longitudinal
Variables	Approx. 5000
Cases	2300 adults with detailed assessment; intensive interviews for 615 children and 350 adults.
Variable types	95% coded, 5% scores, real numbers (weights)

4. Southampton Women's Survey (SWS)

Study Title	Southampton Women's Survey
Study Acronym	SWS
Unit/Centre Name	MRC Environmental Epidemiology Unit
Unit/Centre Address	University of Southampton, Southampton General Hospital, Southampton, SO16 6YD http://www.mrc.soton.ac.uk/project.asp?proj=37, http://www.swsurvey.soton.ac.uk/sws/default.htm
Unit/Centre Director	Prof David Barker
Date of site visit	11/06/2002
PI	Dr Hazel Inskip
Summary of major aims	The Southampton Women's Survey (SWS), a large-scale epidemiological study, is looking at 20,000 20-34 year old women in the Southampton area. It looks at their health and nutrition before they conceive and during their pregnancy and relates these factors to the subsequent health of their babies. Its aim is to understand how issues such as a woman's diet and body composition before and during pregnancy, and her own growth in the womb, affect the development of her foetus.
Study Location	Southampton
Duration	1998-2002/3

Start Date	Dunhill Grant : 1998, later funding from MRC
Principal funding source	MRC, Dunhill Medical Trust, 5 yr funding to continue project sought from British Heart Foundation
Study type	Longitudinal
Variables	Approx. 5200
Cases	20,000 women, target 3000 babies
Variable types	85% coded, 10-15% physical measurements, some dates

5. Wessex Fracture Prevention Study

Study Title	Wessex Fracture Prevention Study
Study Acronym	Wessex Fracture
Unit/Centre Name	MRC Environmental Epidemiology Unit
Unit/Centre Address	University of Southampton, Southampton General Hospital, Southampton, SO16 6YD http://www.mrc.soton.ac.uk/project.asp?proj=21
Unit/Centre Director	Prof David Barker
Date of site visit	11/06/2002
PI	Prof Cyrus Cooper
Summary of major aims	Do annual intramuscular injections of vitamin D, given with the well-established influenza immunisation, reduce the risk of non-vertebral fractures sustained by men and women aged over 75 years old in general practice?
Study Location	Wessex, Isle of Wight, West Sussex and Somerset
Duration	Ongoing. Report expected 2004/5
Start Date	Data Collection 1998
Principal funding source	MRC, NHS R&D
Study type	Trial
Variables	550
Cases	10,000
Variable types	95% coded, some dates, real numbers

6. Study of the effect of formaldehyde on the mortality of workers in the UK chemical industry

Study Title	Study of the effect of formaldehyde on the mortality of workers in the UK chemical industry
Study Acronym	Formaldehyde
Unit/Centre	MRC Environmental Epidemiology Unit

Name	
Unit/Centre Address	University of Southampton, Southampton General Hospital, Southampton, SO16 6YD http://www.mrc.soton.ac.uk http://www.mrc.soton.ac.uk/project.asp?proj=52
Unit/Centre Director	
Date of site visit	11/06/2002
PI	Prof David Coggon
Summary of major aims	To determine the mortality attributed to cancer and to other causes of death in workers exposed to formaldehyde in comparison with men of the same age in the same geographical region of England. Cohort design based on registers of all men ever employed by companies in which exposure to formaldehyde began before 1960 and which retained records identifying all previous members of the workforce at least since 1960. Identification and exposure details.
Study Location	UK
Duration	Ongoing with follow-up every 10 years
Start Date	Original Project 19xx, Follow Up : 2001
Principal funding source	MRC (funding follow-up), original funders Chemical Industry Association and Colt Foundation
Study type	Longitudinal
Variables	38
Cases	13,500
Variable types	Nearly all coded, couple dates

7. SHARE: Does Teacher Led Sex Education Reduce Sexual Risk Taking?

Study Title	SHARE: Does Teacher Led Sex Education Reduce Sexual Risk Taking?
Study Acronym	SHARE
Unit/Centre Name	MRC Social and Public Health Sciences Unit
Unit/Centre Address	University of Glasgow, 4 Lilybank Gardens, Glasgow, G12 8RZ http://www.msoc-mrc.gla.ac.uk/
Unit/Centre Director	Prof Sally Macintyre
Date of site visit	19/06/2002
PI	Dr Danny Wight Ms Marion Henderson
Summary of major aims	A randomised controlled trial (RCT) of a previously piloted, teacher-led sex education programme in schools in Lothian and Tayside, Scotland. Half the schools randomised to receiving training in and to deliver the SHARE programme, and half delivering their usual sex education programme. The

	outcome measures are attitudes and behaviours of pupils, assessed through baseline and post-programme questionnaires of the whole sample. The process of delivering the SHARE programme is being investigated through qualitative research involving teacher questionnaires and interviews with senior schools managers, supplemented with detailed studies of a selection of programme and comparison schools.
Study Location	Lothian and Tayside, Scotland
Duration	1996-2000
Start Date	1996
Principal funding source	MRC
Study type	Socio, trial
Variables	100-200
Cases	110
Variable types	94% coded, 5% real numbers, 1% text

8. Masculinity and Health: The Social Factors Affecting Men's Health

Study Title	Masculinity and Health: The Social Factors Affecting Men's Health
Study Acronym	Masculinity
Unit/Centre Name	MRC Social and Public Health Sciences Unit
Unit/Centre Address	University of Glasgow, 4 Lilybank Gardens, Glasgow, G12 8RZ http://www.msoc-mrc.gla.ac.uk/
Unit/Centre Director	Prof Sally Macintyre
Date of site visit	19/06/2002
PI	Dr Ros O'Brien
Summary of major aims	Examines the social factors that affect men's health. It aims to discover how men understand their own health and the meanings that they invest in health practices – and how these are related to masculinity. There are two stages. The first comprises twenty focus groups to explore men's health, using purposive sampling to structure the groups (e.g. to create particular age groups) and achieve diversity (e.g. of occupational environment). The output will be information about the wider social factors that shape men's health. The second stage comprises tape-recorded in-depth interviews designed to add specific histories of masculinity and health to the focus group data.
Study Location	Glasgow & London (some groups only)
Duration	2000-2001
Start Date	Fieldwork : Feb 2000
Principal funding source	MRC

Study type	Socio
Variables	n.a.
Cases	9
Variable types	NVIVO coding on themes

9. Racist and Sectarian Graffiti in Glasgow – A Pilot Study

Study Title	Racist and Sectarian Graffiti in Glasgow – A Pilot Study
Study Acronym	Graffiti
Unit/Centre Name	MRC Social and Public Health Sciences Unit
Unit/Centre Address	University of Glasgow, 4 Lilybank Gardens, Glasgow, G12 8RZ http://www.msoc-mrc.gla.ac.uk/
Unit/Centre Director	Prof Sally Macintyre
Date of site visit	19/06/2002
PI	Ms Anne Ellaway
Summary of major aims	A systematic exploration of the location and content of sectarian and racist graffiti in Glasgow, Scotland. The project stems from perceptions and observations that graffiti and discrimination based on race in the west of Scotland is associated with poor health. It was not the aim of this project to add immediately to the understanding of mechanisms that might link neighbourhood graffiti and health. Instead, it addressed issues such as the use of graffiti as a public statement of prejudice, to represent affiliation with a cause, or to mark territory. In this pilot project , the content and location of graffiti in a selection of locations in Glasgow were recorded by photography and mapped to social and physical features of the environment. No research data were recorded relating to individuals.
Study Location	Glasgow, Scotland
Duration	2000
Start Date	2000
Principal funding source	MRC
Study type	Socio
Variables	n.a.
Cases	n.a.
Variable types	Maps and photographs

10. The West of Scotland Twenty-07 Study

Study Title	The West of Scotland Twenty-07 Study
Study Acronym	Twenty-07

Unit/Centre Name	MRC Social and Public Health Sciences Unit
Unit/Centre Address	University of Glasgow, 4 Lilybank Gardens, Glasgow, G12 8RZ http://www.msoc-mrc.gla.ac.uk/
Unit/Centre Director	Prof Sally Macintyre
Date of site visit	19/06/2002
PI	Prof Sally Macintyre
Summary of major aims	A longitudinal survey based in the Central Clydeside Conurbation, Scotland. The study seeks to explain the social patterning in a number of health measures by social class, gender, marital status, age, area of residence and ethnicity. The study comprises 3 cohorts, who were aged 15, 25 and 55 when first studies in 1987/88. The study is currently in its third sweep. It examines the relative importance of factors that may cause variations in physical and mental health and wellbeing. These factors include lifestyle behaviours, such as diet, smoking and exercise; selection by health into social positions, e.g. unemployment, marriage; financial and material resources; and exposure to physical and social risk.
Study Location	Central Clydeside Conurbation, Scotland
Duration	1987-2007
Start Date	1987
Principal funding source	MRC
Study type	Longitudinal
Variables	
Cases	Approx. 3,000
Variable types	Most coded, some measurements

11. MRC Scottish Colorectal Cancer Study (SOCCS)

Study Title	MRC Scottish Colorectal Cancer Study
Study Acronym	SOCCS
Unit/Centre Name	Dept. of Community Health Sciences
Unit/Centre Address	Public Health Sciences, University of Edinburgh Medical School, Teviot Place, EH8 9AG
Unit/Centre Director	Dr Harry Campbell
Date of site visit	20/06/2002
PI	Dr Harry Campbell
Summary of major aims	Population based DNA sample collections from colorectal cancer patients' close relatives and matched controls.
Study	Scotland

Location	
Duration	2001-2004
Start Date	March 2001
Principal funding source	MRC
Study type	Genetic
Variables	Approx. 300
Cases	8,500 (4000 cases, 4000 controls, 500 family members)
Variable types	75% coded, some dates, nutrient values, measurements

12. Parkinson's Disease DNA Bank

Study Title	Parkinson's Disease DNA Bank
Study Acronym	PDGEN
Unit/Centre Name	University of Birmingham Clinical Trials Unit (BCTU)
Unit/Centre Address	Park Grange, 1 Somerset Road, Edgbaston, Birmingham. B15 2RR http://www.bctu.bham.ac.uk/
Unit/Centre Director	Dr Carl E Clarke
Date of site visit	24/06/2002
PI	Dr Carl E Clarke
Summary of major aims	To develop a Birmingham-based DNA bank from patients with PD and controls using samples from large pragmatic randomised controlled trials (RCT) such as PD MED and PD SURG and to distribute these samples to approved researchers working in the field.
Study Location	UK
Duration	2000 onwards
Start Date	August 2000
Principal funding source	MRC, BCTU
Study type	Genetic
Variables	
Cases	8,000
Variable types	Genetic samples

13. PDMED Parkinson's Disease Drugs Assessment Randomised Trial

Study Title	A large randomised assessment of the relative cost-effectiveness of classes of drugs for Parkinson's disease
Study Acronym	PDMED

Unit/Centre Name	University of Birmingham Clinical Trials Unit (BCTU)
Unit/Centre Address	Park Grange, 1 Somerset Road, Edgbaston, Birmingham. B15 2RR http://www.bctu.bham.ac.uk/ http://www.bctu.bham.ac.uk/PDMED/PDMEDintro.htm
Unit/Centre Director	Dr Carl E Clarke
Date of site visit	24/06/2002
PI	Dr Carl E Clarke
Summary of major aims	PDMED is a large, simple, "real-life" open-label randomised trial to evaluate the roles of different classes of drugs as therapy for both early and later PD.
Study Location	UK
Duration	2000-2005
Start Date	August 2000
Principal funding source	NHS R&D, BCTU
Study type	Trial
Variables	
Cases	5,000 (3000 early PD and 2000 later PD)
Variable types	90% coded, 10% real numbers/uncoded text, some dates

14. PDSURG Parkinson's Disease Surgery Assessment Randomised Trial

Study Title	A large randomised assessment of the relative cost-effectiveness of surgery for Parkinson's disease
Study Acronym	PDSURG
Unit/Centre Name	University of Birmingham Clinical Trials Unit (BCTU)
Unit/Centre Address	Park Grange, 1 Somerset Road, Edgbaston, Birmingham. B15 2RR http://www.bctu.bham.ac.uk/ http://www.bctu.bham.ac.uk/PDSURG/PDSURGintro.htm
Unit/Centre Director	Dr Carl E Clarke
Date of site visit	24/06/2002
PI	Dr Carl Clarke
Summary of major aims	PDSURG is a large, simple, "real-life" randomised trial to evaluate the role of surgery as therapy for PD.
Study Location	UK
Duration	2000-2010
Start Date	2000
Principal funding	MRC, BCTU, Parkinson's Disease Society

source	
Study type	Trial
Variables	
Cases	28 to date, 400-600 target
Variable types	90% coded, 10% real numbers/uncoded text, some dates

15. MRC National Survey of Health and Development

Study Title	MRC National Survey of Health and Development
Study Acronym	NSHD aka 1946 Birth Cohort
Unit/Centre Name	MRC National Survey of Health and Development Unit
Unit/Centre Address	University College London http://www.ucl.ac.uk/epidemiology/mrc/mrc.html
Unit/Centre Director	Prof Mike Wadsworth
Date of site visit	25/06/2002
PI	Prof Mike Wadsworth
Summary of major aims	A long term follow-up of a national cohort of 5,362 births from 1946 to the present day. Information has been collected regularly from birth so far to the age of 43 years on this study population, and current concerns are with subjects' mid-life physical and mental health, and their rates of change of health with age. Data on social circumstances and health in childhood, adolescence and adult life are used to investigate the precursors of change in health, health risk factors and ill health in middle life.
Study Location	GB
Duration	1946-ongoing
Start Date	1946
Principal funding source	MRC currently
Study type	Longitudinal
Variables	Appx 13,000
Cases	Originally 4695; 3035 in last wave
Variable types	90% coded, rest real numbers, very little free text

16. CONCORDE - MRC/INSERM trial of zidovudine in HIV infection

Study Title	CONCORDE - MRC/INSERM trial of zidovudine in HIV infection
Study Acronym	CONCORDE
Unit/Centre Name	MRC Clinical Trials Unit (CTU)
Unit/Centre	222 Euston Road, London, NW1 2DA

Address	http://www.ctu.mrc.ac.uk/hiv_division/hiv_home.asp
Unit/Centre Director	Prof Janet Darbyshire
Date of site visit	02/07/2002
PI	Prof Ian Weller
Summary of major aims	To compare immediate treatment with zidovudine with treatment deferred until the onset of symptomatic disease in terms of survival and disease progression and of toxicity in asymptomatic HIV infected individuals.
Study Location	UK & France. Follow-up ongoing in UK.
Duration	1988-1993
Start Date	01/01/1988
Principal funding source	MRC, Wellcome, ANRS, INSERM
Study type	Trial
Variables	
Cases	1749 (half U.K.)
Variable types	

17. ESPRIT - A study of subcutaneous recombinant IL-2 (Proleukin®) in patients with HIV

Study Title	Evaluation of Subcutaneous Proleukin® in a Randomised International Trial
Study Acronym	ESPRIT
Unit/Centre Name	MRC Clinical Trials Unit (CTU)
Unit/Centre Address	222 Euston Road, London, NW1 2DA http://www.ctu.mrc.ac.uk/hiv_division/hiv_home.asp http://www.espritstudy.org/
Unit/Centre Director	Prof Janet Darbyshire
Date of site visit	02/07/2002
PI	Prof Brian Gazzard
Summary of major aims	A large randomised, open-label, phase III, international study of subcutaneous recombinant IL-2 (Proleukin®) in patients with HIV-1 infection and CD4+ cell counts of 300/mm^3 or greater. The purpose of the trial is to compare the effects of IL-2 or no IL-2 on progression of HIV disease and mortality over a 5-year period in patients taking combination anti-retroviral therapy.
Study Location	24 UK sites & 22 other countries
Duration	1998-2004
Start Date	1998
Principal funding	MRC, NIH

source	
Study type	Trial
Variables	
Cases	400 (300 U.K.)
Variable types	

18. INITIO - comparing different combination anti-retroviral treatment in HIV infected individuals

Study Title	INITIO - An open randomised trial to evaluate different therapeutic strategies of combination therapy for HIV-1 infection
Study Acronym	INITIO
Unit/Centre Name	MRC Clinical Trials Unit (CTU)
Unit/Centre Address	222 Euston Road, London, NW1 2DA http://www.ctu.mrc.ac.uk/hiv_division/hiv_home.asp http://www.ctu.mrc.ac.uk/initio
Unit/Centre Director	Prof Janet Darbyshire
Date of site visit	02/07/2002
PI	Prof Jonathan Weber
Summary of major aims	International trial involving 17 countries, comparing different strategic approaches to combination anti-retroviral therapy that includes both the first and subsequent regimens (if a change is necessary for therapeutic failure or intolerance) in individuals with HIV infection who wish to start treatment.
Study Location	UK, and 16 other countries
Duration	1998-2004
Start Date	1998
Principal funding source	Roche, SmithKline Beecham, Bristol Myers, and Verco. Plus MRC Core.
Study type	Trial
Variables	23 Files with variables
Cases	950 (77 U.K.)
Variable types	Approx. 90%, 70% real numbers

APPENDIX 3: METADATA STANDARDS AND INTEROPERABILITY IN THE BIOMEDICAL SCIENCES

THE BIOMEDICAL SCIENCES

It should be emphasised that there are no universally accepted metadata standards specific to the Biomedical Sciences, although some controlled vocabularies are more widely used and accepted than others. Often databases are set up that use their own controlled vocabularies and do not follow any predefined metadata standards. Although this report does not focus on metadata standards for genetic data, there is a separate section on XML and Biotechnology in the Metadata Elements and Formats section below.

CONTROLLED VOCABULARIES

Medical Subject Headings (MeSH)

(www.nlm.nih.gov/mesh/meshhome.html)

(MeSH) is the National Library of Medicine's (NLM www.nlm.nih.gov/nlmhome.html) controlled vocabulary thesaurus. Keywords from MeSH are assigned to MEDLINE, NLM's database of more than 11 million bibliographic citations and abstracts covering the fields of medicine, nursing, dentistry, veterinary medicine, health care systems and preclinical sciences. MeSH is one of the world's pre-eminent controlled vocabularies and is widely used internationally.

MeSH consists of a set of terms or subject headings that are arranged in both an alphabetical and a hierarchical structure. At the most general level are very broad headings such as 'Anatomy'. At more narrow levels are found more specific headings such as 'Ankle'. There is also a separate chemical thesaurus and thousands of cross-references to assist in finding the most appropriate MeSH heading.

Examples of the application of MeSH in the UK include the Bristol Biomedical Image Archive (www.brisbio.ac.uk/) and OMNI (Organising Medical Networked Information omni.ac.uk/). The primary objective of Bristol BioMed was to make a shared resource of digital images available for reuse in the development of medical, dental and veterinary electronic learning and teaching materials. However, the usefulness of metadata originally attached to each image was frustrated by semantic and syntactic inconsistencies, use of multiple terms for a single concept and alternative spelling forms. To overcome this, MeSH was selected to fully exploit the metadata as an information retrieval resource.

UK-based OMNI, one of BIOME's (biome.ac.uk/) subject-specific gateways, offers free access to a searchable catalogue of Internet sites covering health and medicine and uses MeSH to index its records.

Unified Medical Language System (UMLS)

(www.nlm.nih.gov/research/umls/)

The NLM's Unified Medical Language System (UMLS) project develops and distributes multi-purpose, electronic "Knowledge Sources" and associated lexical programs. The purpose is to aid the development of systems that help health professionals and researchers retrieve and integrate electronic biomedical information from a variety of sources and to make it easy to link disparate information systems, including computer-based patient records, bibliographic databases, factual databases and expert systems. There are 3 UMLS knowledge sources: UMLS Metathesaurus, Specialist Lexicon and UMLS Semantic Network. The Metathesaurus provides a uniform, integrated distribution format from over 60 biomedical vocabularies and classifications, and links many different names for the same concepts. The Lexicon contains syntactic information for many terms, component words and English words, including verbs, that do not appear in the Metathesaurus. The Semantic Network contains information about the types or categories to which all Metathesaurus concepts have been assigned and the permissible relationships among these types. NLM also distributes associated lexical programs and software helpful in producing customised versions of the UMLS Metathesaurus.

Additional controlled vocabularies and classification schemes (not an exhaustive list)

Systematized Nomenclature of Medicine (SNOMED)
(www.snomed.org/)
Selected as the standard for the UK NHS Electronic Patient Record (EPR).

Multilingual Glossary of technical and popular medical terms in nine European Languages (EUGLOSS)
(allserv.rug.ac.be/~rvdstich/eugloss/welcome.html)

The European multilingual thesaurus on health promotion in 12 languages
(www.hpmulti.net/)
Relationships between medical terminologies used within Europe.

Bioethics Thesaurus
(www.georgetown.edu/research/nrcbl/ir/BT99about.htm)
Published by the Kennedy Institute of Ethics. Each element is assigned a two-letter qualifier e.g. KW – Keyword, PT – Publication Type.

Life Sciences Thesaurus – Cambridge Scientific Abstracts (CSA)
(www.csa.com/helpV3/lscthes.html)
CSA is a privately owned information company located in Bethesda, Maryland. Life Sciences Thesaurus is used to aid searching of various CSA databases.

Dewey Decimal, Library of Congress and NLM Classification Schemes
Traditional generic and specific library classification schemes.

Royal College of Nursing (UK) thesaurus.

International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10)
(www.who.int/msa/mnh/ems/icd10/icd10.htm)
Classification of diseases and related health problems for the collation of medical statistics. Widely used in Europe, it has not yet supplanted the part of the 9th Edition commonly referred to as ICD-9-CM in the US.

METADATA ELEMENTS AND FORMATS

Some examples are given below of sites specifying certain metadata elements and/or formats for metadata storage. It can be seen that most of these are using eXtensible Mark-up Language (XML). There is a separate section on 'XML and biotechnology'.

MEDLINE and PubMed
Publishers whose journals are indexed in MEDLINE can submit citation and abstract data electronically for inclusion in PubMed. Submissions are required in a standard tagged XML format (with required and optional tags) and resources are available to assist with this including a PubMed DTD and XML File Validator (www.ncbi.nlm.nih.gov/entrez/query/static/spec.html). Where possible, ISO standards are followed. For example, the optional language tag requires the user to choose from the language codes in ISO 639. MeSH can also be downloaded in XML, ASCII and MARC (Machine Readable Catalogue Format) formats.

Bristol BioMedical Image Archive (UK)
Control and standardisation of catalogue records was achieved by use of the Dublin Core Metadata Element Set, extended to accommodate the range and variety of biomedical subject matter.

Health*Insite* (Australia)
(www.healthinsite.gov.au)
Health*Insite* is a Commonwealth Government of Australia initiative which aims to improve the health of Australians by providing easy access to quality information about human health. Standards are compliant with Australian Government Locator Service specification which is based on the Dublin Core standard. There is also extended syntax provided to enable external search engines to recognise keywords since Dublin Core syntax is not recognised by many external search engines. Metadata is recorded in HTML.

National electronic Library for Health (NeLH)
(www.nelh.nhs.uk/)
Conforms to the UK e-Government Interoperability Framework (e-GIF) recommendations to adopt the Internet and World Wide Web standards for all government systems. There is a strategic decision to adopt XML as the core standard for data integration and management of presentational data. The NeLH will take the lead in some areas of schema development, for example, in national guidelines publication.

Health Information Disclosure, Description and Evaluation Language (HIDDEL)
(www.medcertain.org/english/metadata/index.htm)

This metadata standard will allow webmasters to describe their privacy, ethics, advertising, content and data quality policies using XML.

The vocabulary is expected to be used to achieve interoperability between third-party rating/evaluation services such as OMNI. The vocabulary will allow users to express their own preferences and needs (e.g. the desired target audience of a site) in a standard language (using an add-on to their browser). Tag usage is explained at two levels, suitable for use by healthcare professionals and non-professionals.

Clinical Data Interchange Standards Consortium Operational Data Model (CDISC ODM)
(www.cdisc.org/models/odm/v1.1/odm1-1-0.html)
The Operational Data Model (ODM) provides a format for representing the study metadata, study data and administrative data associated with a clinical trial. It represents only the data that would be transferred among different software systems during a trial, or archived after a trial. Metadata is used to version study data. The data structure is defined as an XML Document Type Definition.

***meta*Register of Controlled Trials and International Standard Randomised Controlled Trial Number**
(www.controlled-trials.com/)
The *meta*Register of Controlled Trials (*m*RCT) contains more than 10,000 records and is a major international searchable database of ongoing randomised controlled trials in all areas of healthcare, built by combining registers held by public, charitable and commercial sponsors of trials. At the moment the *m*RCT also contains some completed trials. The *m*RCT is a free service that allows users to search all participating registers, all of which are asked to submit trial records including specified essential data items. The content of all the trial records in the *m*RCT has been indexed so that they can be searched efficiently.

Essential data items come under the headings of identification, trial details, funding and contact. Identification details include an International Standard Randomised Controlled Trial Number (ISRCTN). The ISRCTN is a simple numeric system for the identification of randomised controlled clinical trials worldwide. It will simplify the identification of trials and provide a unique number that can be used to track all publications and reports resulting from each trial. The ISRCTN Register is a database of trials with ISRCTNs. It is still being developed and is not yet available online.

Current Controlled Trials (CCT)
(www.controlled-trials.com/links/)
CCT Links give access to more than 200 other online registers of controlled trials, some of which are in languages other than English.

The Open Healthcare Group
(www.openhealth.org/)

The Open Healthcare Group is an organisation devoted to the promotion and distribution of an open source health record, XChart using XML Internet technology

XML AND BIOTECHNOLOGY

XML has become the foundation of several mark-up languages for storing biological data, a selection of which are listed below:

BIOpolymer Mark-up Language (BIOML)
BIOML was designed to be used to describe experimental information about proteins, genes and other biopolymers).

Protein Sequence Database Mark-up Language (PSDML)
(pir.georgetown.edu/)
An open-standard mark-up language used to store protein information in the Protein Information Resource (PIR) database.

Bioinformatic Sequence Mark-up Language (BSML)
(www.ncbi.nlm.nih.gov/ & www.labbook.com)
An open-standard protocol for the encoding and display of graphic genomic displays of DNA, RNA, and protein sequence information. The web-based Basic Browser can import gene sequences from local or remote repositories such as GenBank (the US National Institutes of Health's genetic sequence database, an annotated collection of all publicly available DNA sequences.

Genome Annotation Markup Elements (GAME)
(www.bioxml.org/Projects/game)
A mark-up language used in molecular biology for annotation of biological sequences.

Gene Expression Mark-up Language (GeneXML)
(www.ncgr.org/genex/)
An open-standard mark-up language for DNA microarray and gene expression data. GeneXML was recently renamed from GEML.

Chemical Markup Language (CML)
(www.xml-cml.org)

ADDITIONAL STANDARDS

ISO TC215 & CEN/TC251 Health Informatics.

(www.centc251.org)Used for the interchange of information between health care providers within Europe, and elsewhere around the world.

Health Level Seven HL7
(www.hl7.org/about/)

Health Level Seven, the basis of an ISO standard, aims to provide standards for the exchange, management and integration of data that support clinical patient care and the management, delivery and evaluation of healthcare services. It is widely used in US hospitals. HL7-sanctioned national groups also exist in Australia, Germany, Japan, the Netherlands and New Zealand.

APPENDIX 4: SECONDARY USAGE OF MEDICAL-RELATED DATA

MEDICAL-RELATED DATA AT THE UK DATA ARCHIVE

At present, datasets with medical-related content account for 319 (or around 7 per cent) of the UK Data Archive's 4,500-strong collection. Of these 319 datasets, more than four-out-of-five (268) contain data on the health, lifestyle, behaviour or beliefs of respondents; around a third (97) contain epidemiological data; one-in-ten (34) deal with patient satisfaction; and just 3 per cent (10) focus on trial data. In addition, around one-in-seven (44) are 'general' surveys (e.g. the General Household Survey) that include questions on health-related topics.

Table 1: Medical-related Datasets and Usage (1993*-present)

Type of data	N studies	N orders	Ratio – orders: studies
TRIAL DATA	10	3	0.3
Epidemiological	97	1781	18.4
HALS/behaviour/beliefs	268	3657	13.6
Patient satisfaction	34	17	0.5
General (inc. health)	44	1408	32.0
Totals	**319**	**3876**	

* Usage can only be measured from 1993 onwards.
Note: Columns do not add up due to some studies being allocated to more than one category (e.g. National Surveys of NHS Patients are categorised as both 'HALS/behaviour/beliefs' and 'patient satisfaction').

Medical-related data and its usage

Despite their relatively small representation in the Archive's holdings, the 319 medical-related datasets have, between them, been ordered more than 3,800 times over the past decade.

In terms of the number of orders, usage has been concentrated on the HALS/behaviour/beliefs datasets (3,657 orders have included a dataset with HALS content), with epidemiological and general surveys also proving popular. In contrast, datasets with a patient satisfaction or trial data focus have been ordered only very occasionally (only 20 orders in total).

It is not clear whether the lack of interest in patient satisfaction and trial data indicates that there is little demand for this kind of data among researchers who use the UKDA, or whether low interest is a reflection of the fact that the Archive's holdings are not extensive in these areas. In support of the former explanation, the ratio of orders to studies is very low for patient satisfaction and trial data – certainly in comparison to the same ratios for epidemiological, HALS or general data. However, in support of the latter explanation, it should be noted that much of the patient satisfaction and trial data dates back to the 1970s and is unlikely to be relevant to current researchers.

A contributing factor may also be the level of promotion and ease of accessibility of the datasets. Relatively few of these datasets are supported by specialist 'major studies' UKDA web pages, or are available via the Nesstar online analysis tools or Direct Download service. We know that when lesser-used datasets are made more readily available their usage does increase (e.g. comparing the nine months before and after the mounting of the ONS Omnibus surveys on Nesstar and via Download, usage increased by over 400%).

MRC-funded data and its usage

Datasets funded/sponsored by the MRC account for 8 per cent (25/319) of the Archive's medical-related data holdings. The type of data deposited and the pattern of usage are very similar to the wider medical-related holdings. As is the case generally, almost all the MRC-funded datasets have some HALS/behaviour/beliefs content (24/25). Much lower proportions of MRC-funded datasets contain epidemiological data (7/25), trial data (4/25) or patient satisfaction scores (1/25).

Table 2: MRC-Funded Datasets and Usage (1993-present)

Type of data	N studies	N orders	Ratio – orders: studies
TRIAL DATA	4	0	0.0
Epidemiological	7	231	33.0
HALS/behaviour/beliefs	24	336	14.0
Patient satisfaction	1	1	1.0
General (inc. health)	0	0	n/a
Totals	**25**	**336**	

Note: Columns do not add up due to some studies being allocated to more than one category.

Table 3: Full Listing of Medical-Related Datasets and Usage

UKDA SN/GN	Title	N studies	MRC-funded	Trial data	Epidemiological	HALS/behaviour/beliefs	Patient satisfaction	General (inc. health)	N available Nesstar	N available Download	Major studies pages	N orders
4476	*Project SIGMA : Gay Men's Panel Study, 1987-1994*	**1**	**Yes**			1						**0**
33326	*National Surveys of NHS Patients, 1998-*	**2**				2	2					**7**
4449	*Quality of Life Among People Aged 75 and Over in Great Britain, 1994-1998*	**1**	**Yes**	1		1						**0**
4442	*Alcohol and Suicide, Jews and Protestants, 1999-2000*	**1**				1						**0**
4350	*1918-1919 Influenza Pandemic Mortality in England and Wales*	**1**			1							**2**
4127	*Decline of Infant Mortality in England and Wales, 1871-1948 : A Medical Conundrum; Vaccination Registers, 1871-1913*	**1**			1							**1**
33090	*General Household Survey, 1971-*	**28**				28		28	*5*	*12*	*Yes*	**1342**
33071	*National Food Survey, 1974-*	**26**				26						**56**
33241	*Health and Lifestyle Survey*	**3**				3						**201**
33261	*Health Survey for England, 1991-*	**10**			10	10			*2*	*9*	*Yes*	**445**
4351	*Scottish Household Survey, 1999 and 2000*	**1**				1		1		*1*		**37**
33267	*Road Accident Data*	**10**				10						**58**
33260	*Continuous Household Survey*	**15**				15		15				**28**
33282	*OPCS/ONS Surveys of Psychiatric Morbidity*	**5**			5	5						**55**
33004	*National Child Development Study, 1958-*	**6**			6	6			*1*	*6*	*Yes*	**586**
33275	*National Diet and Nutrition Surveys*	**3**	**Yes**			3						**101**
4226	*Adult Dental Health Survey (ADH), 1998*	**1**			1	1						**2**
33308	*Scottish Migration and Emigration, 1861-1911*	**3**			3							**2**
33263	*Smoking, Drinking and Drug Use Among Young Teenagers*	**9**				9						**23**
33302	*National Study of Health and Growth*	**3**			3							**1**
4175	*Dietary Survey of Vegetarians in Great Britain, 1994-1995*	**1**				1						**5**
33310	*Vital Statistics for England and Wales*	**18**			18							**78**
4176	*Welsh Health Survey, 1998*	**1**			1	1			*1*			**6**
33294	*Health Education Monitoring Survey (HEMS)*	**4**				4						**19**
33307	*Teenage Smoking Attitudes Surveys*	**3**				3						**5**
4118	*Southampton Ageing Project, 1977-1998*	**1**				1						**0**
33305	*Great Britain Historical Database Online, 1841-1939*	**10**			10							**27**

4109	*TAPS Study of Long-term Non-demented Patients, 1985-1998*	**1**		1		1	1					**1**
4090	*Disability Follow-up to the 1996/97 Family Resources Survey*	**1**				1						**20**
4093	*Digest of Welsh Historical Statistics: Population, 1570-1974*	**1**			1							**2**
4005	*Management of Back Pain, 1996*	**1**		1		1						**0**
3976	*Social Variations in Health in Early Old Age : Investigation of Precursors in a 60 Year Follow-Up Study, 1998*	**1**				1						**0**
3831	*Nottingham Study of Food Choice in Later Life, 1994-1996*	**1**				1						**1**
33320	*Scottish Health Surveys, 1995-*	**2**			2	2				*2*		**41**
3903	*General Household Survey, 1994 : Follow-Up Survey of the Health of People Aged 65 and Over*	**1**				1				*1*		**19**
33251	*Infant Feeding Survey*	**3**				3						**5**
3569	*Income, Expenditure and Disability, 1993*	**1**				1						**2**
3808	*Young People's Involvement in Sport, March 1993 - October 1994*	**1**				1						**3**
3779	*Physical Health of Prisoners, 1994*	**1**				1						**0**
33229	*1970 British Cohort Study (BCS70)*	**6**	**Yes**		6	6			*1*	*7*	*Yes*	**231**
3759	*Infant Feeding in Asian Families, 1994-1996; Waves 1-5*	**1**				1						**4**
3641	*PPRU Surveys of Disability, 1989-1990*	**1**			1	1						**3**
3546	*Heights and Weights of British Schoolchildren, 1908-1950*	**1**			1							**3**
3552	*Causes of Death in England and Wales, 1851-60 to 1891-1900 : The Decennial Supplements*	**1**			1							**17**
3625	*Local Mortality Datapack : Population and Deaths by Cause, 1979-1992*	**1**			1							**12**
3554	*Young People and Sport in England, 1994*	**1**				1						**9**
3469	*Northern Ireland Health and Activity Survey, 1994*	**1**			1	1						**0**
33270	*Trent Health Lifestyle Survey*	**2**				2						**2**
33280	*Population Based Computer Assisted Telephone Interviewing Survey of Lifestyles and Health*	**2**				2						**2**
33272	*Cambridge Prenatal Screening Study*	**2**				2	2					**1**
3464	*Expectations and Experiences of Childbirth, 1987*	**1**				1	1					**0**
3434	*National Survey of Sexual Attitudes and Lifestyles, 1990*	**1**				1						**45**
3303	*Allied Dunbar National Fitness Survey, 1990*	**1**				1						**27**
3304	*Health Education Authority National Survey of Activity and Health, 1991*	**1**				1						**8**

3212	*Families of Teenagers with Down's Syndrome; Parent, Child and Sibling Adaptation, 1991*	**1**				1						**0**
3150	*Scottish Heart Health Study, 1984-1986 and the First Scottish MONICA Survey, 1986*	**1**			1	1						**3**
3149	*Jewish Depression, 1991-1993*	**1**				1						**0**
3108	*Anthropometric Data Relating to Working-Class Children, 1841*	**1**			1							**4**
2987	*Living with AIDS : The Experience of Homosexual Men with HIV Infection Or AIDS, 1988-1989*	**1**				1						**0**
2984	*Outpatients and their Doctors : A Study of Patients, Potential Patients, General Practitioners and Hospital Doctors, 1989*	**1**				1	1					**0**
2985	*Life Before Death, 1987*	**1**	**Yes**			1	1					**1**
2982	*Contraceptive Services and Recent Mothers, 1989*	**1**				1						**1**
2943	*Family Expenditure Survey Follow-up Survey of Disabled Adults, 1986-1987*	**1**				1						**8**
2902	*Historic Mortality and Population Data, 1901-1992*	**1**			1							**9**
2861	*Coale Indices of Fertility and Nuptiality in Scotland, 1881-1911*	**1**			1							**0**
2836	*Dietary and Nutritional Survey of British Adults, 1986-1987*	**1**			1	1						**123**
2834	*Adult Dental Health, 1988*	**1**			1	1						**6**
2738	*Substance Abuse and Perceptions of Risk : Young People's Attitudes to Personal Health, 1990*	**1**				1						**1**
2713	*Young People's Leisure and Lifestyles in Modern Scotland, 1987*	**1**				1						**2**
2708	*Census Enumerators' Books : Four Rural Areas, 1851-1881*	**1**			1							**8**
2693	*Survey of Family Planning Services in Scotland, April - July, 1982*	**1**				1	1					**2**
2658	*Care for Elderly People at Home, 1989*	**1**				1						**1**
2657	*Schoolchildren's Dietary Survey, 1983*	**1**				1						**14**
33214	*OPCS Surveys of Disability in Great Britain*	**4**			4	4						**56**
2592	*Census Enumerators' Books, Keighley, West Yorkshire, 1851-1881*	**1**			1							**6**
2552	*More Trouble with Feet : a Survey of the Feet Problems and Chiropody Needs of the Elderly, 1985*	**1**				1	1					**0**
2529	*Impact of Life Events on Heroin, Alcohol and Tobacco Use, 1985-1986*	**1**				1						**2**
2503	*Focus on Health Care : Surveying the Public in Four Health Districts, 1987*	**1**				1	1					**3**

2445	*Diet in the Home, 1969*	**1**				1						**0**
2316	*Transition to Parenthood, 1979-1981*	**1**				1						**2**
2310	*Survey of Smoking Attitudes and Behaviour, 1981*	**1**				1						**1**
2312	*Wessex Survey of Marathon Runners, 1984-1985*	**1**				1						**1**
2308	*AIDS Advertising Evaluation, 1986-1987*	**1**				1						**1**
2264	*Disability in Adolescence, 1977-1979 : the Psychological and Social Problems of Teenagers with Cerebral Palsy and Spina Bifida with Hydrocephalus*	**1**				1						**0**
2174	*Elderly and Their Medicines, 1984*	**1**				1						**1**
2216	*Health Evaluations, 1984-1985*	**1**				1						**6**
2175	*Teenage Mothers and Their Partners, 1979-1980*	**1**				1						**1**
2126	*Handicapped and Impaired in Great Britain, 1968-1969*	**1**			1	1						**0**
2097	*Factors Mediating the Effects of Unemployment on Health, 1982-1984*	**1**				1						**0**
33178	*Long-Term Changes in Nutrition, Welfare and Productivity in Britain*	**4**			4	4						**12**
2100	*Richmond Fellowship of Australia, 1983*	**1**				1	1					**0**
2137	*Perinatal Mortality Survey, 1958*	**1**			1							**22**
2087	*Alternative Medicine, 1984*	**1**				1	1					**2**
2032	*Visually Handicapped in the City of Nottingham*	**1**	**Yes**			1						**0**
2026	*Social and Psychological Consequences of Unemployment in Young People, 1982-1983*	**1**	**Yes**			1						**0**
2046	*National Heights and Weights Survey, 1980*	**1**			1							**20**
2053	*Caring for the Health of Young Children, 1982*	**1**				1						**0**
1900	*Family Formation Survey, 1976*	**1**			1	1						**3**
1903	*Hearing and Vision Screening in Pre-School Children*	**1**		1		1						**0**
1943	*Pedestrian Accidents*	**1**				1						**0**
1786	*Life After a Death : a Study of the Elderly Widowed*	**1**				1						**1**
1818	*Survey of Residents of Two Therapeutic Communities for Former Drug Users, 1973*	**1**				1						**1**
1824	*Towards the Rationalisation of Pre-Employment Health Assessments, 1971-1978*	**1**				1						**0**
1587	*Norfolk Health Care Survey, 1980*	**1**				1	1					**0**
1687	*Rethinking General Practice : Dilemmas in Primary Medical Care*	**1**		1		1						**0**
1169	*Life Stress, Symptomatology and First Year Examination Performance in Overseas Students*	**1**				1						**1**

1171	*Studies Using the Nottingham Health Index, 1976-1977*	**1**				1						**0**
1263	*Experiences of Childbearing : the Dignity of Labour?*	**1**				1	1					**0**
1266	*Psychological Adjustment of Immigrants*	**1**		1		1						**1**
1278	*Changes in the Structure of General Practice : the Patient's Viewpoint*	**1**				1	1					**0**
1317	*Young People's Knowledge of Sex and Birth Control*	**1**				1						**1**
1410	*Aircraft Noise and Prevalence of Psychiatric Disorders, 1977*	**1**	**Yes**	1		1						**0**
1427	*WHO/ICS Medical Care Utilization Study Data, 1968-1969*	**1**			1	1						**0**
1493	*Hospitalised Children in Swansea*	**1**				1	1					**0**
1650	*Attitudes to the National Health Service*	**1**					1					**0**
33127	*Socio-Psychiatric Survey on Distribution and Aetiology of Psychiatric Disorder, 1969-1976*	**2**	**Yes**	2		2						**0**
33162	*Surveys on Smoking, 1976*	**5**				5						**0**
961	*School Child Chest Health Survey, 1966*	**1**			1	1						**0**
962	*Patients and Their Doctors in 1977*	**1**				1	1					**0**
965	*Measurement of Psychological Disturbance in Asian Immigrants*	**1**				1						**0**
1005	*Transport Services in General Practice*	**1**		1		1	1					**0**
71007	*Consumers' Association Survey: Maternity Services, 1981*	**1**					1					**0**
33053	*Patients and Their Doctors, 1964*	**10**				10	10					**0**
33038	*Child Development Research Unit Longitudinal Study*	**6**	**Yes**			6						**3**
808	*Longitudinal Study of Child Development*	**1**	**Yes**			1						**0**
526	*Family Planning in Trinidad : the Problem of Discontinuation*	**1**				1						**0**
362	*Survey of Old People in Telford*	**1**				1						**0**
392	*Trouble with Feet*	**1**				1	1					**0**
396	*Study of Family Size and Family Spacing, 1973*	**1**				1						**0**
399	*Problems of Hospital Communication : An Experimental Study, 1964; Stage I*	**1**				1	1					**0**
1470	*Personal Injury Survey, 1966-1973*	**1**			1							**0**
33085	*Birth Control Services, 1970*	**5**				5						**0**
718	*Southampton Health Centre Study, 1973*	**1**				1	1					**0**
393	*Life Before Death, 1969*	**1**				1						**1**
114	*MRC 83 County Boroughs Study, 1948-1973*	**1**	**Yes**		1							**0**
33054	*Survey of Abortion Patients for the Committee on the Working of the Abortion Act, 1972*	**3**				3						**0**

	Totals (studies/orders)	**319**		10/3	97/1781	268/3657	34/17	44/1408				**3876**

Table 4: MRC-Funded Datasets and Usage

UKDA SN/GN	Title	N studies	MRC-funded	Trial data	Epidemiological	HALS/ behaviour/ beliefs	Patient satis-faction	General (inc. health)	N available Nesstar	N available Download	Major studies pages	N orders
4476	*Project SIGMA : Gay Men's Panel Study, 1987-1994*	**1**	**Yes**			1						**0**
4449	*Quality of Life Among People Aged 75 and Over in Great Britain, 1994-1998*	**1**	**Yes**	1		1						**0**
33275	*National Diet and Nutrition Surveys*	**3**	**Yes**			3						**101**
33229	*1970 British Cohort Study (BCS70)*	**6**	**Yes**		6	6			*1*	*7*	*Yes*	**231**
2985	*Life Before Death, 1987*	**1**	**Yes**			1	1					**1**
2032	*Visually Handicapped in the City of Nottingham*	**1**	**Yes**			1						**0**
2026	*Social and Psychological Consequences of Unemployment in Young People, 1982-1983*	**1**	**Yes**			1						**0**
1410	*Aircraft Noise and Prevalence of Psychiatric Disorders, 1977*	**1**	**Yes**	1		1						**0**
33127	*Socio-Psychiatric Survey on Distribution and Aetiology of Psychiatric Disorder, 1969-1976*	**2**	**Yes**	2		2						**0**
33038	*Child Development Research Unit Longitudinal Study*	**6**	**Yes**			6						**3**
808	*Longitudinal Study of Child Development*	**1**	**Yes**			1						**0**
114	*MRC 83 County Boroughs Study, 1948-1973*	**1**	**Yes**		1							**0**
	Totals (studies/orders)	**25**		4/0	7/231	24/336	1/1	0/0				**336**

APPENDIX 5: DATA ARCHIVING & ACCESS: CASE STUDY TOOL

ADMINISTRATIVE INFORMATION

Contact:

Address:

Tel:

Email

website:
etc.

QUESTIONS FOR THE PRINCIPAL INVESTIGATOR(S)

OVERALL ACTIVITY

1. How many data sets is the Unit / project managing in total?
2. How many are closed to new accrual of data?

RESOURCES

3. Brief funding history

SCIENTIFIC VALUE AND POTENTIAL FOR NEW RESEARCH

4. What are the overall scientific objectives for which this dataset has been assembled?
5. What other evidence have you of the value put on the study / dataset by external users e.g. peer review of the research or publications? Have your research techniques/instruments/scales that relate to this study been adopted elsewhere/published?
6. In what way is the study unique? PROBE: its objectives, methodology, scale, study population, measures…?)
7. What other studies are there with which this dataset could be compared or pooled (e.g. through formal meta-analysis)?
8. What is the expected "lifetime" of active research by the PI and his/her team on these data? What is the expected "lifetime" of the data in terms of their utility for secondary research?
9. Do you see any particular issues concerning re-use after a PI has retired or funding has come to an end? Have you a successor custodian in mind?

10. What valuable research questions might such pooling / mining address, either in terms of improving the questions the data-set is currently designed to address, or – potentially – new questions?

ACCESS

11. What were the terms of the consent with study participants? PROBE: Study team require copies of up-to-date sample consent forms and patient information

12. What would be involved in anonymising (unlinked) these data? If all identifying characteristics were removed, would the remaining data be useful for secondary analysis?

13. What, if any, would be any barriers be to further exploitation of the dataset by a secondary analyst? PROBE: cost, data completeness, data ownership/intellectual property/copyright, data quality, consent/confidentiality issues

14. How widely known is your dataset? PROBE: What is the audience for information about the research – own or wider research communities, policy makers, practitioners… and how do people find out about your dataset? PROBE: Research literature, Web site, the media, active promotion e.g. through a Data Archive

15. What scale is the access by new collaborators / others to the data-set (how often / how much…?) What is the nature of their requests? How many requests do you receive? And how many do you accept [standardise – over the past 12 months / 2 years?].

16. What motivates you to accept new collaborations / new uses? PROBE how much do you consider their proficiency/ability to use the data?

17. What motivates you *not* to accept new collaborations / new uses?

18. What methods of access control would you favour/do you see as necessary for this collection? PROBE: vetting by PI; 'guided' access by team; ethics committee review

19. Do you have *formal* criteria and processes to decide whether or not to collaborate / give access? If so, what are they? If no, what are your *informal* criteria and processes to decide whether or not to collaborate / give access? PROBE: Is there independent involvement? Is there representation from the participants?

OWNERSHIP

20. Who, in your opinion, owns the dataset – and what do you mean by ownership? What rights and responsibilities do you consider that you (as the PI) have over the use of the data now and it the future? Particularly with respect to:

    - Archiving the data (to what kinds of level / quality)

    - Ensuring wider use by the scientific / policy community?

    - Ensuring commercial exploitation opportunities are identified and managed?

21. What rights and responsibilities rights do you consider your employer or funder or have?

22. Is this codified as an explicit institutional agreement or policy? (If so, what?)

23. Are any other funders likely to have a proprietorial interest? If so, what particular issues might this raise with respect to archiving /re-use? PROBE: ownership/copyright/ability to agree

24. Do you make any formal written agreements with your collaborators and any independent new users about ownership, further use and new use?

25. Do these data contain or incorporate data created elsewhere (e.g. quality of life or nutrition scales, international classifications, look-up tables) for which intellectual property resides outside the research team?

26. What about derived data created by your researchers in the course of their analyses – where does the copyright and intellectual property in these derived data resources lie?

27. For this project, what is your Centre/Unit's ability to meet the demands of the MRC Good Clinical Practice Guidelines (20 years for data; 10 years for lab data)?

GENERAL

28. How are the case studies typical of the whole range of Unit datasets?
29. What advantages could you see an MRC data archive/data dissemination service offering you as an active researcher? PROBE: good practice, usability for successors. relevance of a dissemination service

30. Are there any datasets out there not under your control that you would like to be able to access for your own research?

31. How useful would you consider a national register of research instruments to be?

32. Do you feel you would welcome assistance/advice on data preservation standards and practices? Would you use an advisory service or published guidelines to enable you to preserve your own data? Would you use a service that actually preserved data for you?

33. Are you aware of any international developments in the field of preservation/sharing of biomedical/health/social sciences data?

QUESTIONS FOR THE PROJECT MANAGER(S)

QUALITATIVE INFORMATION – SCIENCE

34. Study Title

35. Acronym

36. Summary of study aims

37. Duration of study

38. Start date [1st accrual of data]

39. Methodology (we have a typology from the Phase 1).

40. Study Location

SAMPLE

41. Study population:

42. Sampling frame

43. Number of cohorts

44. Sample size of cohort

45. Major topics covered

46. Data sources utilised? PROBE: medical records, school records, other ancillary or related information in addition to data directly created by research team

HUMAN BIOLOGICAL SAMPLES

47. Have you established/are you establishing a collection of human biological samples in this study?

48. If yes, what are: a) the current numbers sampled? and b) the anticipated final numbers?

49. Where is the collection held?

50. What data are held on the samples? How are they held and maintained? (What information is known about the sample donors, including relevant personal and family data, or health or treatment information (including confirmation of diagnosis)? Include intellectual property - i.e. data you may have derived from your work with these samples.

51. Can genetic information be derived from the samples?

52. Where is this genetic information held?

53. Which organisation do you consider is the custodian of the collection?

The custodian organisation has responsibility for safekeeping of samples and control of their use and eventual disposal in accordance with the terms of consent given by the donor.

54. Is this organisation also considered to be the 'owner' of the data held on the samples?
55. Have you had any applications to access the collection? If so, how are these requests dealt with?

RESOURCES

56. Staff (FTEs) directly employed by the project in what kinds of capacities? PROBE: scientists, research assistants, IT people, data people. If Unit has closed, how many staff were employed?

57. How long have these people been involved with the project? With the institution? If closed, how long were the staff involved with the project?

58. Staff directly or indirectly supporting the work of the project in what kinds of capacities? PROBE: computer services, IT consultants, University infrastructure, warehouses

59. Any staff with formal experience or training in data preservation / archiving / records management?

PHYSICAL STORAGE

60. How are the data currently stored? PROBE: Paper, electronic data; audio-visual materials
61. Volume measure
62. Rate of increase.
63. Location and Contractors
64. Cost measure, of storage and retrieval costs for data stored offsite pa?
65. What security measures are in place for protecting the integrity of digital data? PROBE: back-up media and procedures, version control, access control for multi-researcher teams, system security? PROBE: Physical security : firesafes, climate control, offsite backups

66. What security measures are in place for non-digital (e.g. paper and analogue image or sound) data PROBE: fireproof cabinetry, acid-free paper, climate control, locks and access control, etc?

67. In terms of the current facilities you have for paper storage, how adequate do you think they are?PROBE: have you ever applied for specific funding for storage?

68. If data were originally collected in structured or unstructured paper-based formats, how often is recourse made to the paper/audio materials? For what purposes?

69. How useful would you find it to have paper-based data held in electronic format? e.g. PDF

70. How are interviewer manuals/training guides/rating notes stored and in what format? Are these used for training /promotional purposes beyond the scope of this project? Are they formally published at all?

GENERAL

71. What advantages could you see an MRC data archive/data dissemination service offering you as a project manager?

72. Are there any datasets out there not under your control that you would like to be able to access for the project?

73. How useful would you consider a national register of research instruments to be?

74. Do you feel you would welcome assistance/advice on data preservation standards and practices? Would you use an advisory service or published guidelines to enable you to preserve your own data? Would you use a service that actually preserved data for you?

75. Are you aware of any international developments in the field of preservation/sharing of biomedical/health/social sciences data?

QUESTIONS FOR DATA MANAGER(S)

DATA ARCHIVING

Data digitisation issues:

76. If data were collected in structured paper-based formats, what proportion have been made digital (e.g. data entered into a computer file)? What software was used to enter the data, and what formats/platforms/media are they stored in? How often is recourse made to the paper/audio materials? For what purposes?

77. If data were collected in unstructured paper-based formats, what proportion have been made digital and by what method (e.g. coded after the fact, verbatim copy-typing, scanning of papers)? What software was used to enter the data, and what formats/platforms/media are they stored in? How often is recourse made to the paper? For what purposes?

78. If data are audio, visual, or moving images, what proportion have been made digital and by what method? What software was used to digitise the data, and what formats/platforms/media are they stored in? How often is recourse made to the original tape/images? For what purposes?

79. If the data were collected in digital formats originally, what software was used to collect the data, and what formats/platforms/media are they stored in?

80. What software formats and platforms are used to analyse the data in house?

Version control issues:

81. For open datasets, how are subsequent data accruals added to the existing (digital) dataset?

82. For multi-researcher teams: is there a single definitive digital version of the data, or do separate researchers on the teams have their own "massaged" versions? Are these versions saved? Are the relationships among the different versions documented?

83. In the course of data analyses by the research team, are derived data created, and are these saved? Are the methods of their derivation recorded in a fashion useful to a secondary analyst?

Metadata: information about data

84. Is there a catalogue or register of the contents of paper files or analogue audio/image collections? What form does this take and how is it maintained?

85. Is there a catalogue or register of the contents of digital files? What form does this take and how is it maintained? PROBE: Are there formal 'codebooks' or 'data dictionaries' kept?

86. Are study instruments (e.g. questionnaire pro-formas, interviewers guides/manuals, etc) kept for data, and if so in what form? Are coding frames described?

87. Are protocols and methods for clinical or laboratory measurements recorded, and if so how?

88. Are these publicly available, and if so in what form?

89. Are there publications in the public domain based on the data? What form do these take? Is there a publicly available catalogue or register of these publications; how is it maintained and accessed?

90. How accessible are your study instruments and data collections tools to the wider research communities? PROBE: published in manuals; widely circulated, well known; patented

91. Would you see value in producing a web based register of study instruments?

Personnel

92. Who is responsible for maintaining the original (non digital) data? Who is responsible for maintaining the digital data? What is their experience or training, if any, in preservation / migration / archiving / records management?

QUANTITATIVE INFORMATION – INDIVIDUAL *DATA-SET*

93. Numbers of variables*?*

94. What proportion of the digital data are coded (labelled) variables (e.g. survey responses, etc)?

95. What proportion of the digital data are real numbers / measurements?

96. What proportion of the digital data is uncoded textual (free text)?

97. Are there any date/time variables or physical measurements?

GENERAL

98. What advantages could you see an MRC data archive/data dissemination service offering you as a data manager?

99. Do you feel you would welcome assistance/advice on data preservation standards and practices? Would you use an advisory service or published guidelines to enable you to preserve your own data? Would you use a service that actually preserved data for you?

100. How useful would you consider a national register of research instruments to be?

101. Are you aware of any international developments in the field of preservation/sharing of biomedical/health/social sciences data?

QUESTIONS FOR LAB MANAGER(S)

HUMAN BIOLOGICAL SAMPLES

102. Have you established/are you establishing a collection of human biological samples in this study? (e.g. DNA, Blood samples)

103. If yes, what are: a) the current numbers sampled? and b) the anticipated final numbers?

104. Cost measure, pa.

105. In terms of the lab storage facilities, how adequate do you think they are? PROBE: have you ever applied for specific funding for storage?

106. What data are held on the samples? How are they held and maintained? What information is known about the sample donors, including relevant personal and family data, or health or treatment information (including confirmation of diagnosis)?

Include intellectual property - i.e. data you may have derived from your work with these samples. Where is the linking information between the samples and data stored?

107. Can genetic information be derived from the samples? Where is this genetic information held?

108. What security measures are in place for data held on the samples? PROBE: access control etc?

109. What security measures are in place for protecting the integrity of digital data? PROBE: back-up media and procedures, version control, access control for multi-researcher teams, system security? PROBE: Physical security : firesafes, climate control, offsite backups

110. Which organisation do you consider is the custodian of the collection? (The custodian organisation has responsibility for safekeeping of samples and control of their use and eventual disposal in accordance with the terms of consent given by the donor).

111. Is this organisation also considered to be the 'owner' of the data held on the samples?

112. Have you had any applications to access the collection? If so, how are these requests dealt with?

113. Are you aware of any international developments in the field of preservation/sharing of biomedical/health/social sciences data?

APPENDIX 6: GOOD PRACTICE IN MANAGING DATA FOR SHARING AND PRESERVATION

LIFECYCLE APPROACH TO DATASET MANAGEMENT

Sharing and Preserving Datasets

Access to the information contained in a digital dataset is dependent on a constantly shifting landscape of hardware and software. Datasets can only be used when appropriate hardware and software are available to read and present the data. Equally, datasets will only remain useful if the expertise and knowledge needed to interpret and manipulate the dataset is available. Thus it is important not to view sharing and preserving datasets as a purely technical challenge.

To share or preserve a dataset, it must contain adequate documentation, be technically compatible, and have appropriate provisions for ensuring the authenticity of the data and managing access to the data. A comprehensive data management strategy must consider these issues at all stages in the lifecycle of a dataset (Figure 1), from its creation right through to its eventual long-term preservation and re-use, or destruction (for legitimate reasons). Choices made during the creation of a dataset are particularly important, as poor decisions made at this stage can limit the usefulness of a dataset throughout its life.

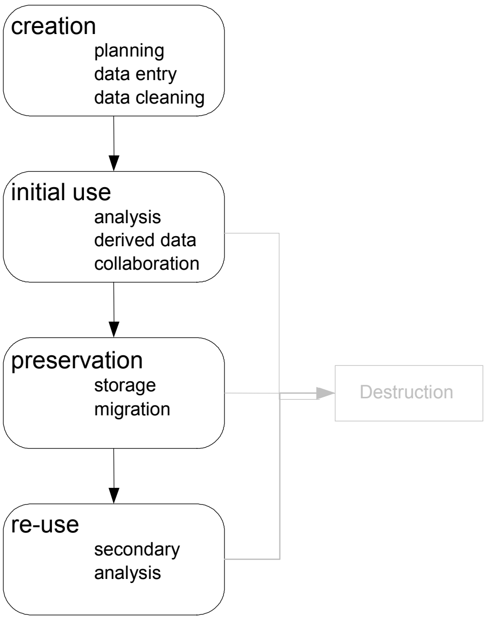

Fig 1. Lifecycle of a Dataset

Documentation
Comprehensive documentation is vital if datasets are to be shared and preserved as useable resources, but many datasets are not properly documented. Publications arising from a dataset may not provide all the necessary documentation to fully understand a dataset. Documentation includes items such as field notes, codebooks, research reports, data entry instructions, notes on data manipulation (scaling, weighting, aggregation and other

operations), survey instructions, questionnaires, interview prompts and any other material which informs the use of the data.

Documentation should be written as the dataset is created and developed. The departure of a key member of staff can mean the loss of vital knowledge about a dataset if the task of documenting is left until after the work is completed. Indeed, each time an individual or group ceases to be involved with a dataset, there is a danger that information critical to making informed use of the dataset may be lost. For example, information about problems with the wording of questions that may affect the trustworthiness of the answers may be lost if those conducting survey/interviews or collating information are not involved in later stages of the project.

Technical compatibility
Ensuring technical compatibility or, more pragmatically, minimising technical incompatibility should be given a particularly high priority during the creation of a dataset. Technical incompatibility is often caused by the unnecessary use of multiple data formats, or through the use of unusual software functionality. When more than one data format is used, the data must be frequently exported from one software package to another, an operation that often generates subtle changes/errors in the data. Similarly, the use of incompatible operating systems and hardware creates the need to constantly move data between different computer systems, which can also introduce errors into the data.

More significant is the problem posed by complex datasets that make use of software - or possibly hardware - specific functionality. These datasets can be difficult to share or preserve because they are dependent on a specific piece of software that may be expensive to purchase, require unusual expertise to operate, or will not run on all common operating systems.

Authenticity and access
The great ease with which digital data can be altered *without trace* lies behind the importance of implementing policies that can guarantee the authenticity of a dataset. A variety of technical procedures, such as digital signatures or recording unique file characteristics can be used to assist in demonstrating the authenticity of a dataset, but the basis for demonstrating authenticity is maintaining a clear record of how a dataset has been created and altered, who did the work and why.

Digital preservation strategies
Many issues involved in preserving data are similar to the issues involved in sharing data. However, the volatile nature of digital technologies generates additional problems for the long-term preservation of datasets. The conversion of digital *data* into meaningful *information* is dependent on the continuing availability of the digital technologies - the computer hardware, software, and the format specifications that hardware and software rely on - used to create the digital data. The rapid pace of change in computer technology means that these technologies can go from new to obsolete in a matter of years, quickly rendering datasets inaccessible.

Any digital preservation strategy must therefore focus primarily on avoiding technological obsolescence. Two main approaches to avoiding technological obsolescence have been proposed: migration and emulation.

Migration is the process of converting information held in an obsolescent data format into a newer format. Data is preserved into the future by migrating it to a new data format with each new generation of software. Selecting open and well-used data formats should reduce the frequency of migration.

Alternatively, instead of changing the data to fit the computer, it may be possible to change the computer to fit the data by emulating the original hardware environment on a modern computer.

Migration is the most commonly adopted strategy for digital preservation at the moment. This is primarily because, unlike emulation, migration can be conducted without full access to the specifications of files, software and hardware, and it does not require the same level of software development work. The main danger of migration is that information may be lost during each migration exercise. Emulation has only been tested in very limited contexts. It relies on the creation of complex and accurate emulators. It is accepted as likely to be appropriate in some situations where data is especially dependent on specific hardware or software, but its wider feasibility is still a matter of debate.

RECOMMENDATIONS FOR TECHNICAL COMPATIBILITY

Backup

The media used to store digital data are fragile but easy to copy. To reduce the risk of damage or loss of data, all digital data should therefore be backed up. A good backup procedure will protect against a range of mishaps such as: accidental changes to data; accidental deletion of data; loss of data due to media or software faults; virus infections; hackers; catastrophic events (such as fire or flood).

Frequency of backup
The more frequently data are changed, the more frequently they should be backed up. Data that are changing significantly every day should be backed up every day. If frequent data backup is required, this process should be automated to reduce the burden on resources.

Rolling backup copies
Earlier backup copies should not be immediately discarded when a new backup is made. Backup copies made at several points of time, over a period of at least three months, should be kept as this will reduce the risk of incorporating recent, but as yet undetected, problems in all backup copies.

Off-site backup copies
At least one backup copy should be held off-site to ensure that a disaster, such as a fire in the office, will not destroy all copies of the data. Off-site is a relative term and the location of off-site copies should reflect the level of protection needed. Datasets of short-term, project-specific significance might be stored in another building at the same institution. Backups of datasets of national and long-term significance should be stored elsewhere in the country

It is also useful to keep a backup copy on-site. An on-site backup copy can be quickly retrieved and work recommenced if there is a minor mishap, such as the accidental deletion of an important file.

Institutional backup policy
When a project relies in part, or in whole on an institutional backup system, then project staff should ensure that the institution's policies will be appropriate to their needs:

- institutions may maintain backups for a limited period;
- institutions may, intentionally or unintentionally, not backup all data on their network;
- institutions may not restore individual users' files.

Independent backups of critical data should always be maintained.

Validate backup copies!
A backup that does not actually work is of no use at all. Backup copies should be tested to ensure all data has been successfully saved and can be retrieved.

Recommended choice of backup media
All copies of data should be held on new media. Do not continue to use media once they start to develop faults. The primary storage medium for the preservation of digital data is tape, but CD-R is often used for short-to-medium term backup and as an ease-of-use alternative to the primary storage media. Floppy disks should not be used for backup or preservation of data. DVD is not yet recommended for preservation, as its viable lifespan is still uncertain.

For additional risk reduction, use media from different batches and write media using different drives in order to avoid replicating faults caused by a faulty production batch or drive.

Refresh media
Because the media used to store digital data are fragile and can degrade quickly it is good practice to 'refresh' data regularly by transferring it from old media to new media.

Storage conditions for physical media
Media should be stored according to the manufacturer's recommendations for temperature, humidity, light levels and other factors, to ensure the recommended conditions are constantly maintained.

Checksums
Data backup and media refreshment both involve copying data. Checksums[1] should be used to ensure that data have been copied successfully. When media are placed in storage, checksums should be stored with the data (as a separate file), and another copy of the checksums should be kept separately.

Data Formats

To make sharing and preserving data easier, data should, ideally, be stored in formats that are software independent, non-proprietary and widely supported.

Use non-proprietary standards
Data formats can be grouped into three broad categories relevant to the issues of sharing and preserving datasets.

- Proprietary standards - formats such as Microsoft Word that are owned by a company and not made generally available. Data held in these formats can only be accessed through software that has been licensed to read the format.
- Available standards - the specification of some proprietary standards, such as Microsoft Rich Text Format (RTF) and Adobe Portable Document Format (PDF) are made available to other software developers and the general public. Available standards are still proprietary and restrictions may be placed on their use in the future.
- Open standards - standards created by a co-operative group that are then made freely available to anybody to use without restriction.

---

[1] A checksum is a numerical calculation, based on the bits in a file, that is calculated before and after data is copied. A difference between the checksums indicates that an error has occurred. Checksums are a way of verifying that data has been copied successfully.

Open formats should be used whenever possible. Available formats, such as Microsoft RTF and Adobe PDF should be considered next. Proprietary formats should only be considered as a last resort.

Use widely supported formats
Preference should be given to widely supported data formats because these are more easily shared and less likely to become obsolete unexpectedly.

Careful consideration should be given to the adoption of open standards, some of which may not be adequately supported by available software.

Avoid specialist proprietary formats
Reliance on specialist proprietary formats should be avoided as this commits a project to using one set of software products, even if better options emerge, or the software company ceases to support the software.

Select software independent formats
Select data formats that can be read by more than one software package, and preferably more than one type of software package (word processor, statistical package, spreadsheet, database). Simple plain text, delimited text and, increasingly, marked-up text formats often meet these requirements. For example, many different software packages can import and export delimited text, making it a useful format for storing and exchanging data.

Record file format details
Details of the version of software used to create or modify data should be recorded. It is also useful to record details of versions of software used to read files.

Validate data export between formats
Data may be exported for use in different software, to be shared with researchers using different hardware and software, or to store the data in more stable formats suitable for long-term preservation (these are not necessarily the formats most convenient for analysis).

Whenever data is exported from one format to another, checks should be made to verify that the export has not altered the information content of the dataset. Examples of potential problems to check are:

- rounding of numeric values;
- truncation of textual values;
- numeric values that exceed allowable range;
- mismatched colour palettes in images;
- non-supported code pages or glyphs[2] in text.

Data should be exported by data managers or other project staff familiar with the data as they are best placed to spot any errors or inadvertent changes to the data that may occur in the export process.

---

[2] A glyph is the symbol used to indicate a character in an alphabet. For example, the glyph for the character DOLLAR SIGN is "$"

Suggested data formats

Data Type	Suggested Preservation Format	Suggested Format for Sharing
tabular data	tab delimited text	text-based formats such as SPSS portable
image data	PNG or TIFF	PNG or TIFF except for large images delivered online where JPEG is appropriate
qualitative data	XML marked-up text according to an appropriate DTD or schema	software specific formats such as NUD*IST or plain text
documentation	plain text or XML marked-up text according to an appropriate DTD or schema (e.g. XHMTL 1.0)	RTF or PDF

All text should be encoded as ASCII or UNICODE. When data may contain non-ASCII characters (generally, any non-Latin characters) it should always be encoded as UNICODE. Note that newer versions of software are likely to use UNICODE by default. Note that XML requires the use of UNICODE.

Data Capture from Hardcopy

When data is captured from hardcopy originals, such as survey forms, it is important to minimise the introduction of errors into the digital data due to the data capture process.

Scanning
Hardcopy documentation should be scanned at a resolution and colour depth that creates a legible image. Black and white (1 bit) page images at a resolution of 200dpi are adequate for plain typed documents. Higher resolution grey scale images are likely to be more suitable for handwritten material. Full colour (24 bit or higher) images at 600dpi should be considered for documents that include fine detail, such as photographs or diagrams.

Optical Character Recognition
Output from Optical Character Recognition (OCR) software of less than about 95% accuracy should not be accepted. Output should be proof-read against the hardcopy.

Keyboard data entry
Ideally, keyboard data entry should be double-keyed (all data entered by two individuals and then compared for differences). As a minimum, a random sample of data entry should be proof-read against the hardcopy.

RECOMMENDATIONS FOR AUTHENTICITY AND ACCESS

Digital data can be copied, altered or deleted very easily, and this makes it very important to be able to demonstrate the authenticity of data, and to prevent unauthorised access to data for ethical, legal and quality reasons

Master Files

Assign responsibility for master files
Responsibility for maintaining the master versions of all material (digital and hardcopy) should be clearly assigned to individual members of the project team.

Restrict write access to master versions
Because digital data is very easy to duplicate or alter, it is particularly important to ensure that access to the data is restricted to authorised individuals. Write access (i.e. the ability to alter the data) to the master files for a dataset should be restricted to the member(s) of the project team with responsibility for the master files.

Formalise destroying master files
A formal procedure should be drawn up to ensure that master files are not accidentally or prematurely deleted. The procedure should establish that the file is not in current use, has not been referred to in a published work, does not have potential for re-use, is not used to generate other files that are still in use and does not need to be maintained for administrative, ethical or legal reasons.

Record changes to master files
A log recording all changes to the master files should be maintained. The log should include, at a minimum, the time, date, and person responsible for all changes to master files.

Maintain old master files
To guard against accidental alteration or deletion of master files, old versions of master files should be maintained after they have been superseded, unless this is legally or ethically impossible.

Version Control

Uniquely identify files
Each file within a dataset should be uniquely identified. A formal procedure should be written to govern the assignment of names to ensure that they remain unique.

Files that are made available outside the project team (e.g. as references in publications, or as files sent to collaborators) should be identified using a persistent name, such as a URN (Uniform Resource Name), to ensure that the information can be found irrespective of its current or future location.

Record version and status

The status (e.g. 'draft', 'interim', 'final', 'internal' etc.) and version, either as part of their unique name, or within their content, should be recorded.

Record relationships between items

In many cases the information contained in a single file is supported by information held in other files.

- A delimited text data file may be used in conjunction with an SPSS setup file, specifying variable formats, labels etc., and may also be supported by a descriptive user guide which explains how variables were collected and collated.
- The paper questionnaire that provided the raw data for each record in the dataset should be easily traceable from the data file.

Whenever a change is made to a file (either its content or its file format), a check should be made to ensure no other files are affected.

Track the location of all items

If digital or hardcopy items are kept in more than one location then a list, index or other finding aid should be maintained to ensure that all relevant material can be easily located.

Legal Rights and Responsibilities

Rights of respondents

Survey, interview, and other forms of information collected about individuals must be accompanied by clear documentation indicating the conditions under which respondents agreed to allow the data to be used.

User licences

Formal agreements should be made with any collaborators or other users outside the project team specifying the data they may use, the purposes they may use it for and the period they may keep the data for.

Legal mandate

At the extreme, there may be a need to ensure the legal admissibility and evidential weight of information stored digitally. The British Standards Institute has produced a code of good practice on these matters *(Legal Admissibility and Evidential Weight of Information Stored Electronically*, DISC PD 0009:1999).

Computer Security

Network security

Access to project files should be restricted to approved individuals through the use of user accounts and passwords. Additional restrictions can be applied, if necessary, by assigning directory and file level access restrictions.

A member of the project staff should be responsible for ensuring that the list of individuals allowed access to the dataset remains up-to-date.

MRC units that store data on networks managed by other organisations should be aware that many networks are insecure. Confidential data should not be stored on servers that host internet services (web or email). Especially sensitive material should be stored on computers that are not connected to a network.

Upgrades and patches
To prevent unauthorised access to data, it is important to apply all relevant security-related upgrades and patches to operating systems and applications as quickly as possible.

A formal procedure should be adopted to ensure that new patches and updates are applied to all project computers.

Viruses
Project staff should be given training to recognise suspicious files and emails.

All project computers should have regularly updated (preferably daily through a contract with an anti-virus software vendor) virus detection software.

Email should not be received on computers storing data.

Physical security of systems
Prevention of unauthorised access to computer hardware and hardcopy material is very important. Sensible physical security precautions should be taken, such as locking rooms when staff are absent, limiting access to rooms where computers or media are held to a few individuals, logging computer media or hardcopy material that are removed from store rooms, recording who holds keys, etc.

RECOMMENDATIONS FOR DOCUMENTATION

Creating and Managing Documentation

Documentation is an integral part of a research project. It should be written as the dataset is created and developed. The documentation should be comprehensive and provide all the necessary information to enable informed use of the dataset.

Guidelines for documenting a dataset
Guidelines should be developed to ensure that all project staff are aware of what information is needed to adequately document a dataset.

Review documentation
Documentation should be reviewed by members of the project (or other appropriate individuals outside the project) to ensure that it is comprehensive and understandable.

Indicate the relationship between documentation and data
Clearly indicate in the documentation how it relates to the data (which file, which variable, which coding scheme).

Hardcopy documents should be clearly marked with all the details needed to find other related hardcopy or digital material (see Version Control).

A periodic check should be made to ensure that documentation for a dataset remains up-to-date.

Documentation in both hardcopy and digital forms
When documentation is partially digitised a clear statement should be included with the digitised documentation explaining how it differs from the hardcopy original.

When documentation exists in both digital and hardcopy form then one version should be clearly identified as the master version that should be used as the basis for any future revision of the documentation.

Documentation Content

The documentation for a dataset should provide all the information needed to make informed use of the data. Members of the project will be best placed to make detailed decisions on what and how to documented, although as a rule, it is better to be inclusive and record information that is not vital, rather than create sparse documentation that does not provide all necessary information.

Documentation checklist

Provenance

- History of the originating project
    - the purpose of the project
    - topic(s) of research
    - geographic and temporal limits
    - funders
    - principal investigators
    - information about methods
- Methods used to create the dataset
    - consistency checks
    - error corrections
    - sampling strategies employed
- Details of existing material used to create the dataset
    - existing sources of data used
    - procedures for updating, combining, or enhancing existing source data
    - coding and classification schemes used
    - description of any known copyrights held on existing source material

Technical Details

- Data characteristics

  data model (flat data file, relational database) and relevant details
  variable data types
  text encoding
  file formats
- Ancillary information

  list of filenames and description of contents
  description of identification numbers assigned
  description of any known problems
  details of derived data
  details of codebooks and data dictionaries
  history of format changes to dataset
- Software

  software used for creation (including specialized software such as OCR)
  software used for analysis
  software used for conversion, import and export of data
  operating system(s) and platforms that software ran on
  specialized hardware used (e.g. digital cameras and scanners)

Access and Use

- terms of consent
- bibliographic references to any publications, or web sites, about the project
- access conditions (free/restricted)
- intellectual property rights statements
- rights held by third parties in the dataset
- history of how the dataset has been used
- indication of how long archive is to be retained (indefinitely or fixed period)

GLOSSARY

ADH	Adult Dental Health Survey
AHRB	Arts and Humanities Research Board
AIDS	Acquired immune deficiency syndrome
ALSPAC	Avon Longitudinal Study of Parents and Children
ASCII	American Standard Code for Information Interchange
BBSRC	Biotechnology and Biological Sciences Research Council
BHPS	British Household Panel Survey
BIOME	BIOME is a collection of gateways which provide access to evaluated, quality Internet resources in the health and life sciences, aimed at students, researchers, academics and practitioners.
BIOML	BIOpolymer Mark-up Language
BSI	British Standards Institute
BSML	Bioinformatic Sequence Mark-up Language
CCT	Current Controlled Trials
CDISC ODM	Clinical Data Interchange Standards Consortium Operational Data Model
CEN	Comit é Européen de Normalisation European Committee for Standardization
CERN	European Organization for Nuclear Research
CML	Chemical Markup Language
CODATA	Committee on Data for Science and Technology
CONCORDE	MRC/INSERM trial of zidovudine in HIV infection
CORBA	Common Object Request Broker Architecture
COREC	Central Office for Research Ethics Committees
CSA	Cambridge Scientific Abstracts
CSTP	Committee for Scientific and Technological Policy
CTU	Clinical Trials Unit
DAA	MRC Data Archiving and Access Project
DAIS	Data and Information System
DDBJ	DNA Data Bank of Japan
DDI	Dataset Documentation Initiative
DeCC	Depression Case Control Study
DNA	deoxyribonucleic acid
DoH	Department of Health
DSM-IV	Diagnostic and Statistical Manual (Fourth edition)
DTD	Document Type Definition
EAR	Division of Earth Sciences at the National Science Foundation (NSF)
EBI	European Bioinformatics Institute
EDM	electronic document management
EMBL	European Molecular Biology Laboratory
EPR	Electronic Patient Record
ESF	European Science Foundation
ESPRIT	Evaluation of Subcutaneous Proleukin® in a Randomised International Trial
ESRC	Economic and Social Research Council
EUGLOSS	Multilingual Glossary of technical and popular medical terms in nine

	European Languages
FOIA	Freedom of Information Act
Formaldehyde	Study of the effect of formaldehyde on the mortality of workers in the UK chemical industry
FTE	Fixed-term employment
GAME	Genome Annotation Markup Elements
GCP	Good Clinical Practice
GeneXML	Gene Expression Mark-up Language
GIS	Geographic Information Systems
GMC	General Medical Council
GP	General Practitioner
Graffiti	Racist and Sectarian Graffiti in Glasgow – A Pilot Study
HALS	Health and Lifestyle Survey
HEMS	Health Education Monitoring Survey
HGMP	Human Genome Mapping Project
HIDDEL	Health Information Disclosure, Description and Evaluation Language
HIV	Human immunodeficiency virus
HL7	Health Level Seven
HMCA	Health and Medical Care Archive
HSC	Health Service Circular
HTML	HyperText Mark-up Language
IBM	International Business Machines
ICD-10	International Statistical Classification of Diseases and Related Health Problems (10th Revision)
ICH	International Conference on Harmonisation
ICPSR	Inter-University Consortium for Political and Social Research
ICS	International Collaborative Study
ICSTI	International Council for Scientific and Technical Information
ICSU	International Council for Science
ICT	Information and Communications Technology
IEC	Independent Ethics Committee
INITIO	An open randomised trial to evaluate different therapeutic strategies of combination therapy for HIV-1 infection
INSERM	L'Institut National de la Santè et de la Recherche Medicale/French National Institute for Health and Medical Research
IoW	Isle of Wight Studies
IPR	Intellectual Property Rights
IRB	Institutional Review Board
ISO	International Organization for Standardization
ISRCTN	an International Standard Randomised Controlled Trial Number
JRF	Joseph Rowntree Foundation
KNAW	Royal Netherlands Academy of Arts and Sciences
LSOA	Longitudinal Studies of Ageing
MARC	MAchine Readable Catalogue Format
Masculinity	Masculinity and Health: The Social Factors Affecting Men's Health
MeSH	Medical Subject Headings
Microsoft COM	Microsoft Component Object Model
MRC	Medical Research Council
mRCT	metaRegister of Controlled Trials
NARA	National Archives and Records Administration

NATSAL	National Survey of Sexual Attitudes and Lifestyles
NCDS	National Child Development Study
NCHS	National Center for Health Statistics
NeLH	National electronic Library for Health
NERC	Natural Environment Research Council
NeSC	National e-Science Centre
NESSTAR	Networked Social Science Tools and Resources
NHANES	National Health and Nutrition Examination Survey
NHCS	National Health Care Survey
NHIS	National Health Interview Survey
NHS	National Health Service
NIAID	National Institute of Allergy and Infectious Diseases
NIH	National Institute of Health
NIS	National Immunization Survey
NIWI	Netherlands Institute for Scientific Information Services
NLM	National Library of Medicine
NRC	National Research Council
NSF	National Science Foundation
NSHD	National Survey of Health and Development
NTIS	National Technical Information Service
NVSS	National Vital Statistics System
NYS	New York State
OECD	Organisation for Economic Co-operation and Development
OMNI	Organising Medical Networked Information
ONS	Office for National Statistics
OPCS	Office of Population Censuses and Surveys
PDF	Portable document format
PDGEN	Parkinson's Disease DNA Bank
PDMED	Parkinson's Disease Drugs Assessment Randomised Trial
PDSURG	Parkinson's Disease Surgery Assessment Randomised Trial
PI	Principal investigator
PSDML	Protein Sequence Database Mark-up Language
RAL	Rutherford Appleton Laboratory
RCP	Royal College of Physicians
RDN	Resource Discovery Network
REC	Research Ethics Committee
RNA	ribonucleic acid
SAMHSA	Substance Abuse and Mental Health Services Administration
SES	Social and Economic Sciences
SGDP	MRC Centre for Social, Genetic and Developmental Psychiatry
SGML	Standard Generalized Mark-up Language
SHARE	SHARE: Does Teacher Led Sex Education Reduce Sexual Risk Taking?
SILCAAT	A Study of Interleukin-2 (IL-2) in People with Low CD4+ T-Cell Counts on Active Anti-HIV Therapy
SNOMED	Systematized Nomenclature of Medicine
SOCCS	MRC Scottish Colorectal Cancer Study
SOP	Standards Operations Procedures
SQL	(pronounced "ess-que-el") stands for Structured Query Language
SWS	Southampton Women's Survey
TAPS	Team for the Assessment of Psychiatric Services

TEDS	Twins' Early Development Study
Twenty-07	The West of Scotland Twenty-07 Study
UCL	University College London
UKDA	United Kingdom Data Archive
UMLS	Unified Medical Language System
Wessex Fracture	Wessex Fracture Prevention Study
WHO	World Health Organization
XML	eXtensible Mark-up Language
Z39.50	The OSI client/server-based protocol established as a standard by the National Information Standards Organisation (NISO) which allows computer users to query a remote information retrieval system (server) using the software of a different system, and display results in the interface of the system used for input (client).