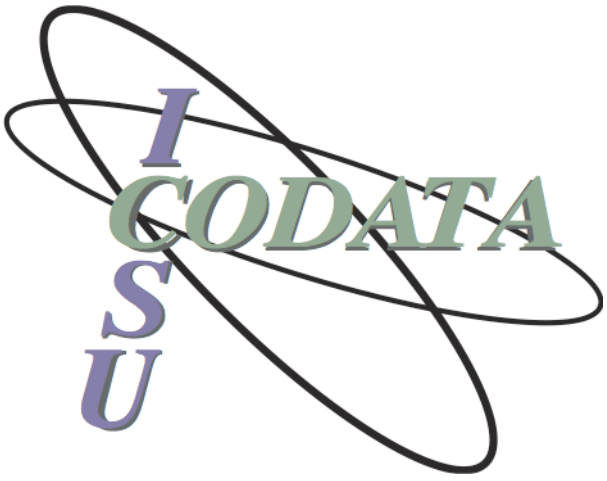
Current Best Practice for Research Data Management Policies

A Memo for the Danish e-Infrastructure Cooperation and the Danish Digital Library

APPENDICES

Simon Hodson and Laura Molloy



Appendices

Appendix 1: Benefits of Data Sharing in the Health Sciences, Public Health and Epidemiology

Sharing research data to improve public health: full joint statement by funders of health research

Sharing data in the health sciences has particularly strong human benefits. There have been a number of initiatives to add impetus internationally to data sharing for public health. This includes the so-called Foggy Bottom Statement or, to give it its proper title *Sharing research data to improve public health: full joint statement by funders of health research*: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm

The statement makes clear the benefits of data sharing for public health and the intention of health research funders internationally to promote this:

> The importance of data sharing in advancing health is becoming increasingly widely recognised, and has been strongly endorsed by the H8 group of global health organisations. … Much of the data collection that could improve public health research is expensive and time-consuming. As public and charitable funders of this research, we believe that making research data sets available to investigators beyond the original research team in a timely and responsible manner, subject to appropriate safeguards, will generate three key benefits:
>
> - faster progress in improving health
> - better value for money
> - higher quality science.
>
> Each funding institution will work within its own legal and operational framework, and we are committed to working towards these goals together. We intend to establish joint working groups where appropriate. We call on governments and other actors that generate routine health service statistics and other types of public health data to adopt a similar approach.
>
> This Statement establishes guiding principles and desired goals. It recognises that flexibility and a variety of approaches will be needed in order to balance the rights of the individuals and communities that contribute data, the investigators that design research and collect and analyse data, and the wider scientific community that might productively use data for further research.

The *Joint Statement of Purpose* reads as follows:

> **Vision**
>
> We, as funders of health research, intend to work together to increase the availability to the scientific community of the research data we fund that is collected from populations for the purpose of health research, and to promote the efficient use of those data to accelerate improvements in public health.
>
> **Principles**
>
> Funders agree to promote greater access to and use of data in ways that are:
>
> Equitable: Any approach to the sharing of data should recognise and balance the needs of researchers who generate and use data, other analysts who might want to reuse those data, and communities and funders who expect health benefits to arise from research.
>
> Ethical: All data sharing should protect the privacy of individuals and the dignity of communities, while simultaneously respecting the imperative to improve public health through the most productive use of data.

Efficient: Any approach to data sharing should improve the quality and value of research and increase its contribution to improving public health. Approaches should be proportionate and build on existing practice and reduce unnecessary duplication and competition.

Goals

While we recognise that progress may be gradual as we develop mechanisms and resources consistent with these principles, we aim to work in concert to achieve the following.

Immediate goals

Data management standards support data sharing

Standards of data management are developed, promoted and entrenched so that research data can be shared routinely, and re-used effectively.

Data sharing is recognized as a professional achievement

Funders and employers of researchers recognize data management and sharing of well-managed datasets as an important professional indicator of success in research.

Secondary data users respect the rights of producers and add value to the data they use

Researchers creating data sets for secondary analysis from shared primary data are expected to share those data sets and act with integrity and in line with good practice - giving due acknowledgement to the generators of the original data.

Longer-term aspirations

Well documented data sets are available for secondary analysis

Data collected for health research are made available to the scientific community for analysis which adds value to existing knowledge and which leads to improvements in health.

Capacity to manage and analyse data is strengthened

The research community, particularly those collecting data in developing countries, develop the capacity to manage and analyse those data locally, as well as contributing to international analysis efforts.

Published work and data are linked and archived

To the extent possible, datasets underpinning research papers in peer-reviewed journals are archived and made available to other researchers in a clear and transparent manner.

Data sharing is sustainably resourced for the long term

The human and technical resources and infrastructures needed to support data management, archiving and access are developed and supported for long-term sustainability.

Appendix 2: The Influence of Intelligent Openness

Definition of Intelligent Openness from the Royal Society report on *Science as an Open Enterprise*

The *Science as an Open Enterprise* report provides the following definition of the components of 'intelligent openness':

a. Accessible. Data must be located in such a manner that it can readily be found. This has implications both for the custodianship of data and the processes by which access is granted to data and information.

b. Intelligible. Data must provide an account of the results of scientific work that is intelligible to those wishing to understand or scrutinise them. Data communication must therefore be differentiated for different audiences. What is intelligible to a specialist in one field may not be intelligible to one in another field. Effective communication to the non-scientific wider public is more difficult, necessitating a deeper understanding of what the audience needs in order to understand the data and dialogue about priorities for such communication.

c. Assessable. Recipients need to be able to make some judgment or assessment of what is communicated. They will, for example, need to judge the nature of the claims that are made. Are the claims speculations or evidence based? They should be able to judge the competence and reliability of those making the claims. Are they from a scientifically competent source?7 What was the purpose of the research project and who funded it? Is the communication influenced by extraneous considerations and are these possible sources of influence identified? Assessability also includes the disclosure of attendant factors that might influence trust in the research. For example, medical journals increasingly require a statement of interests from authors.

d. Usable. Data should be able to be reused, often for different purposes. The usability of data will also depend on the suitability of background material and metadata for those who wish to use the data. They should, at a minimum, be reusable by other scientists.

The Royal Society (2012), *Science as an Open Enterprise*, p. 14-15;
https://royalsociety.org/policy/projects/science-public-enterprise/Report/

EC Guidelines on Data Management in Horizon 2020

The guidance from the EC's *Guidelines on Data Management in Horizon 2020* shows the influence of the Royal Society *Science as an Open Enterprise* report's requirements for intelligent openness and the G8 Science Ministers' Statement which expands on the criteria:

> Scientific research data should be easily
>
> **1. Discoverable**
>
> a. DMP question: are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier)?
>
> **2. Accessible**
>
> a. DMP question: are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses (e.g. licencing framework for research and education, embargo periods, commercial exploitation, etc.)?
>
> **3. Assessable and intelligible**

a. DMP question: are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review (e.g. are the minimal datasets handled together with scientific papers for the purpose of peer review, are data is provided in a way that judgments can be made about their reliability and the competence of those who created them)?

4. Useable beyond the original purpose for which it was collected

a. DMP question: are the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data (e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists)?

5. Interoperable to specific quality standards

a. DMP question: are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc. (e.g. adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing re- combinations with different datasets from different origins)?

Appendix 3: EC Horizon 2020 Model Grant Agreement, Limitations on Openness

Horizon 2020 Annotated Model Grant Agreements, Version 1.6 2 May 2014,
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf

ARTICLE 27 — PROTECTION OF RESULTS — VISIBILITY OF EU FUNDING

27.1 General obligation to protect the results [p. 182-4]

Each beneficiary must examine the possibility of protecting its results and must adequately protect them — for an appropriate period and with appropriate territorial coverage — if:

(a) the results can reasonably be expected to be commercially or industrially exploited and

(b) protecting them is possible, reasonable and justified (given the circumstances).

When deciding on protection, the beneficiary must consider its own legitimate interests and the legitimate interests (especially commercial) of the other beneficiaries.

[…]

[Annotation: Protection may be sought through patent, trademark, industrial design, trade-secret or confidentiality.]

27.3 Information on EU funding

Applications for protection of results (including patent applications) filed by or on behalf of a beneficiary must — unless the [Commission][Agency] requests or agrees otherwise or unless it is impossible — include the following:

'The project leading to this application has received funding from the [European Union's Horizon 2020 research and innovation programme][Euratom research and training programme 2014-2018] under grant agreement No [number]'.

ARTICLE 36 — CONFIDENTIALITY

36.1 General obligation to maintain confidentiality [p.212]

During implementation of the action and for four years after the period set out in Article 3, the parties must keep confidential any data, documents or other material (in any form) that is identified as confidential at the time it is disclosed ('confidential information').

If a beneficiary requests, the [Commission][Agency] may agree to keep such information confidential for an additional period beyond the initial four years.

If information has been identified as confidential only orally, it will be considered to be confidential only if this is confirmed in writing within 15 days of the oral disclosure.

Unless otherwise agreed between the parties, they may use confidential information only to implement the Agreement.

The beneficiaries may disclose confidential information to their personnel or third parties involved in the action only if they:

(a) need to know to implement the Agreement and

(b) are bound by an obligation of confidentiality.

This does not change the security obligations in Article 37, which still apply.

The [Commission][Agency] may disclose confidential information to its staff, other EU institutions and bodies or third parties, if:

(a) this is necessary to implement the Agreement or safeguard the EU's financial interests and

(b) the recipients of the information are bound by an obligation of confidentiality.

Under the conditions set out in Article 4 of the Rules for Participation Regulation No 1290/201343, the Commission must moreover make available information on the results to other EU institutions, bodies, offices or agencies as well as Member States or associated countries.

The confidentiality obligations no longer apply if:

(a) the disclosing party agrees to release the other party;

(b) the information was already known by the recipient or is given to him without obligation of confidentiality by a third party that was not bound by any obligation of confidentiality;

(c) the recipient proves that the information was developed without the use of confidential information;

(d) the information becomes generally and publicly available, without breaching any confidentiality obligation, or

(e) the disclosure of the information is required by EU or national law.

ARTICLE 37 — SECURITY-RELATED OBLIGATIONS

[p.214-5]

Annotation:

'If the security scrutiny (carried out by the Commission/Agency during the selection procedure) finds that the action raises security issues because it deals with information that is 'EU-classified' (under the Commission internal Rules of Procedure), the Commission/Agency will make it subject to 'security requirements' (set out in a 'Security Aspect Letter (SAL)' and 'Security Classification Guide (SCG)' in Annex 1 of the GA).'

In this case, the beneficiaries must comply with the security requirements.

These requirements include restrictions on disclosure and subcontracting.

ARTICLE 39 — PROCESSING OF PERSONAL DATA

[…]

39.2 Processing of personal data by the beneficiaries [p.222-3]

The beneficiaries must process personal data under the Agreement in compliance with applicable EU and national law on data protection (including authorisations or notification requirements).

The beneficiaries may grant their personnel access only to data that is strictly necessary for implementing, managing and monitoring the Agreement.

The beneficiaries must inform the personnel whose personal data are collected and processed by the [Commission][Agency]. For this purpose, they must provide them with the service specific privacy statement (SSPS) (see above), before transmitting their data to the [Commission][Agency].

Appendix 4: Analyses of Data Policy Elements

A number of analyses of research data policies exist, primarily constructed to guide researchers to funders' expectations. However, we have considered in more helpful to present the elements of data policies in the structure given here and outlined in the Executive Summary.

UK Digital Curation Centre

The UK Digital Curation Centre's guidance regarding funder research data policies provides information about the following components.

Policy Stipulations	
Time limits	Covers both the time allowed before data should be made available (period of privileged access) and the time for which the data should be retained in an archive.
Data plan	Covers under what circumstances the funder requires a data management plan.
Access/data sharing	Data in scope and limitations on data sharing
Long term curation	Means (and sometimes period) for long term data stewardship.
Monitoring	Intention to monitor compliance.
Support Services	
Guidance	Whether the funder, or a related agency, provides guidance.
Data Centre	Whether the funder provides a data centre and if not what recourse is expected.
Costs	Statement on responsibility for the costs of data management and long term stewardship.

DCC Overviews of UK Research Funder Data Policies: http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

Analysis of Funder Requirements for London School of Hygiene and Tropical Medicine

Gareth Knight, Funder Requirements for Data Management and Sharing (for projects at the London School of Hygiene and Tropical Medicine): http://researchonline.lshtm.ac.uk/208596/

This survey for a research data management support project at the London School of Hygiene and Tropical Medicine covers largely health related funders. It uses a useful set of criteria to communicate high level information about the funder requirements.

1. Data outputs covered
2. Data management plan required
3. Funding arrangement for data management and sharing
4. Documentation requirements
5. Retention requirements
6. Data standards
7. Publication requirements
8. Data sharing rights
9. Designated data centre
10. Monitoring
11. References
12. Institutional requirements

ICPSR Guidelines for funding agencies in responding to the OSTP Memo 'Increasing Access to the Results of Federally Funded Scientific Research'

In response to the OSTP memo, ICPSR (the US Inter-university Consortium for Political and Social Research) has produced guidelines for agencies in responding to the memo. This provides 'Elements of a Public Access Plan for Scientific Data' which is for the most part a framework for a funder research data policy. Although this is designed to respond to a specific US government policy, it offers a useful outline to research data policy content and associated guidance.

Elements of a Public Access Plan for Scientific Data

The most important elements are:

Maximize access

Protect confidentiality and privacy

Preserve intellectual property rights and commercial interests

Balance demands of long-term preservation and access

Use of data management plans

Include cost of data management in funding proposals

Evaluate data management plans

Ensure researcher compliance with DMPs

Promote public deposit of data

Private-sector cooperation to improve access

Mechanisms for identification & attribution of data

Data stewardship workforce development

Long-term support for repository development

ICPSR Guidelines for OSTP Data Access Plan:
http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/ostp.html

De-Mystifying the Data Management Requirements of Research Funders

Analysis from Dianne Dietrich, Trisha Adamus, Alison Miner, Gail Steinhard, De-Mystifying the Data Management Requirements of Research Funders, Issues in Science and Technology Librarianship, DOI:10.5062/F44M92G2

Element	Element Description	Element Category
Organizational data Policy	Funder recommends or requires that a project have a policy for management of research data.	General policy
Data Plan for Proposal	Funder recommends or requires a data management plan as part of all research proposals.	General policy
Funder recommends or requires a particular timeframe for the data	Data Timeframe management plan to be implemented.	General policy
Compliance	Funder monitors or enforces compliance.	General policy
Funding	Funder specifies if funding for any aspect of data management can be written into a proposal.	General policy
Scope	Whether a funder's data management policy applies to all research data, or only to certain types of data.	General policy
Guidance	Whether funder provides guidance for meeting data management requirements.	General policy
Data Standards	Funder recommends or requires specific file formats or other standardization of data.	Standards
Metadata Standards	Funder recommends or requires use of particular metadata standards.	Standards
Data Access	Funder recommends or requires access to data.	Access and preservation
Data Embargo	Funder addresses embargo periods for data.	Access and preservation
Data Preservation	Funder recommends or requires preservation of data.	Access and preservation
Data Center	Specified Funder specifies that data be deposited in a particular data center.	Access and preservation
Data Center Supported	Funder supports or maintains a data center for use by funding recipients.	Access and preservation
Open Access to Publications	Funder recommends or requires open access to resulting publications.	Publications
Publication Repository Specified	Funder specifies that publications are to be deposited in a particular repository.	Publications
Publication Repository Supported	Funder supports or maintains a publication repository	Publications
Date of policy	Any date of issue or posting date on the policy.	-

Appendix 5: Definitions of Research Data

OECD Principles

> In the context of these Principles and Guidelines, "research data" are defined as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated.
>
> This term does not cover the following: laboratory notebooks, pre- liminary analyses, and drafts of scientific papers, plans for future research, peer reviews, or personal communications with colleagues or physical objects (e.g. laboratory samples, strains of bacteria and test animals such as mice). Access to all of these products or outcomes of research is governed by different considerations than those dealt with here.
>
> These Principles and Guidelines are principally aimed at research data in digital, computer-readable format. It is indeed in this format that the greatest potential lies for improvements in the efficient distribution of data and their application to research because the marginal costs of transmitting data through the Internet are close to zero. These Principles and Guidelines could also apply to analogue research data in situations where the marginal costs of giving access to such data can be kept reasonably low. *OECD Principles, p. 13-14.*

EPSRC on analogue data

In some cases, there is specific clarification that physical resources such as notebooks and other objects are subject to the data policy, e.g. EPSRC Expectation iv: 'Publicly-funded research data that is not generated in digital format will be stored in a manner to facilitate it being shared in the event of a valid request for access to the data being received (this expectation could be satisfied by implementing a policy to convert and store such data in digital format in a timely manner).'
http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx

BBSRC on biological samples

BBSRC 'expects that any biological resources accompanying the data should also be made available … for a minimum period of two years following the publication of any paper describing those organisms.' BBSRC Policy, p9; http://www.bbsrc.ac.uk/web/FILES/Policies/data-sharing-policy.pdf

NIH

> For this purpose, "data" means recorded information, regardless of the form or media on which it may be recorded, and includes writings, films, sound recordings, pictorial reproductions, drawings, designs, or other graphic representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing or computer programs (software), statistical records, and other research data. NIH Policy 8.2.1
> http://grants.nih.gov/grants/policy/nihgps_2013/nihgps_ch8.htm#_Toc271264947

NSF

The general policy of the NSF and the specific policies of some NSF divisions sometimes contain an outright contradiction in the definition of research data.

The NSF general policy states 'Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants.'
http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/aag_6.jsp#VID4

The policies of particular divisions, however, for no specific reason, often refer to a definition contained in OMB (US Office of Management and Budget) Circular A-110 http://www.whitehouse.gov/omb/circulars/a110/a110.html which contradicts this definition by *excluding* 'preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues [and] … physical objects (e.g., laboratory samples).'

Furthermore, those divisional policies which use the OMB Circular use it to exclude sensitive data from the definition of research data.

Many of the US / NSF policies use a definition of research data which is taken from a US policy document OMB Circular A-110 http://www.whitehouse.gov/omb/circulars/a110/a110.html

> *Research data also do not include:*
>
> *(A) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and*
>
> *(B) Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study.*

The authors of this briefing are not convinced that this is a helpful definition and think that it has the potential to confuse matters. These things can very well be 'research data'; often they are. But they are research data that for various reasons should not be disclosed and which need to be treated in a secure way.

A clear data policy should govern the best practice management of all data produced by research projects. Therefore, it should have a coherent description of research data and be equally clear about what types of data are necessarily not subject to the Open data or data sharing aspects of the policy.

Definitions in Institutional Policies

Taken from **http://ands.org.au/guides/what-is-research-data.html**

The **Queensland University of Technology** Management of research data policy states:

> Research data means data in the form of facts, observations, images, computer program results, recordings, measurements or experiences on which an argument, theory, test or hypothesis, or another research output is based. Data may be numerical, descriptive, visual or tactile. It may be raw, cleaned or processed, and may be held in any format or media.

The **University of Melbourne** draft policy on the Management of Research Data and Records states:

> Research Data: Data are facts, observations or experiences on which an argument, theory or test is based. Data may be numerical, descriptive or visual. Data may be raw or analysed, experimental or observational. Data includes: laboratory notebooks; field notebooks; primary research data (including research data in hardcopy or in computer readable form); questionnaires; audiotapes; videotapes; models; photographs; films; test responses. Research collections may include slides; artefacts; specimens; samples. Provenance information about the data might also be included: the how, when, where it was collected and with what (for example, instrument). The software code used to generate, annotate or analyse the data may also be included.
>
> The University of Melbourne makes no functional distinction between physical research products, digital research data and records of research, which can include items such as correspondence, application documents, reports and consent forms.

The **Monash University** Research Data Policy provides the following definition:

> Research data: The data, records, files or other evidence, irrespective of their content or form (e.g. in print, digital, physical or other forms), that comprise research observations, findings or outcomes, including primary materials and analysed data.

Appendix 6: Categories of Data in Scope

The Guidance on the Horizon 2020 Open Data Pilot uses a helpful two-part classification of the research data which is in scope and should be made Open.

> The Open Research Data Pilot applies to two types of data:
>
> 1) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;
> 2) other data [e.g. curated data not directly attributable to a publication, or raw data], including associated metadata, as specified and within the deadlines laid down in the data management plan.

Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, p. 10; http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

This distinction also features in the key Article 29.3 of the Model Grant Agreement.

> **29.3 Open access to research data**
>
> [OPTION for actions participating in the open Research Data Pilot: Regarding the digital research data generated in the action ('data'), the beneficiaries must:
>
> (a) deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:
>
> (i) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;
>
> (ii) other data, including associated metadata, as specified and within the deadlines laid down in the data management plan (see Annex I);
>
> (b) provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves).
>
> This does not change the obligation to protect results in Article 27, the confidentiality obligations in Article 36, the security obligations in Article 37 or the obligations to protect personal data in Article 39, all of which still apply.
>
> As an exception, the beneficiaries do not have to ensure open access to specific parts of their research data if the achievement of the action's main objective, as described in Annex I, would be jeopardised by making those specific parts of the research data openly accessible. In this case, the data management plan must contain the reasons for not giving access.]
>
> [OPTION: not applicable]

Horizon 2020 Annotated Model Grant Agreements, Version 1.6 2 May 2014, p. 188; http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf

Appendix 7: Simple typology of different research data types in relation to data policies.

The authors have provided a simple typology of different data types in relation to data policies.

All research data	Should be subject to data policies requiring good management of data during and after a research project.
Open Data, data to be shared	There are some very good reasons why certain data should not be shared openly. However, the presumption or default is that data arising from publicly funded research should be shared unless these reasons apply.
Data underpinning published research conclusions	Should be made freely available concurrently with the publication for reasons of verification and research integrity.
Substantial datasets produced by research projects (particularly any unrepeatable, longitudinal or expensively created data)	Should be made available on completion of the project or after a specific period of privileged use. It is helpful if policies link both to general guidelines about what data is likely to be in scope and to more details information about selection of research data. Observational data that cannot be repeated and data which would be very expensive to recreate are good examples.
Sensitive data containing human subject information	Must be managed according to good practice during the course of the project. Great care must be taken not to release any data that may contain personal information. When appropriate anonymisation is done, it may be possible for such data to be made available. An alternative is for such data to be accessible under controlled conditions. Specialised social science data centres have great experience in these matters and in many cases will be the appropriate home for such data.
Sensitive data containing commercially privileged information	Partnership, knowledge-transfer or enterprise/innovation related research projects may have legitimate reasons for not releasing data. There may be legitimate grounds for delaying data release if it is intended to support patent applications.
Sensitive data which must be restricted for other reasons	Policies must allow for additional sound reasons for restricting data release (for example national security, information about endangered species etc).

Appendix 8: Data Selection *NERC Data Value Checklist*

General guidance on the selection of data for long-term preservation:
http://www.nerc.ac.uk/research/sites/data/policy/data-value-checklist.pdf

Selection of data should be based on integrity, originality and geographic coverage. Data retained must tender a contribution to the scientific knowledgebase. The data may be used to inform national policy making or in an international context.

1. RELEVANCE TO MISSION

Are the data aligned with the NERC current strategy and fall in the environmental data remit of the NERC Data Centres? Consideration should be given to any legal or legislative requirements to retain data, for example, compliance with the Environmental Information Regulations and any contractual obligations which exist relating to the long term management and storage of data.

2. SCIENTIFIC OR HISTORICAL VALUE

Is there, or could there potentially in the future be, a use for the data? Could the material be scientifically or communally important? Are the data exemplary or do they set a precedent? Predicting future trends in research is difficult but consideration should be given to current trends in research awards and the scientific direction of research institutes alongside any educational value which might be obtained.

3. UNIQUENESS

Is this the primary and most complete unprocessed version of the data, to which no irreversible transformations have been applied? There are areas covered by NERC research where this approach may not be appropriate for example with 3D and 4D seismic data the volumes of raw data are so large that post-stack or pre-stack are more suitable for long term retention and storage, however, these will be exceptions rather than the norm. The NERC Data Centre will hold the principal copy of the data. Is the information new and unique or a re-working of previous material? If other copies existed would they be at risk or will they be preserved, and if so, are they the most complete and up to date?

4. NON-REPLICABILITY

Where it is not realistic to reproduce data, this is usually constrained by the costs incurred in the creation of the original dataset. Observations and sampling are seen as non-repeatable. Simulations could potentially be run again and experiments are repeatable, subject to cost implications.

5. POTENTIAL FOR REDISTRIBUTION

How reusable will the data be? Will they be stored in a format which will enable future re-use? Will the data be tied to a specific type of software? Will there be technical issues in reusing the data? Are there precautions which can be taken i.e. storing software alongside the data to future proof their reuse. Are there any Intellectual Property Rights issues associated with the data and their reuse? Are there constraints from the original funder which would restrict future use of the data or are there any contractual or licensing terms which affect future use of the data? Is this an unaltered dataset which has not been changed in any way with its original integrity retained?

6. ECONOMIC CASE

When considering the preservation of data the cost of retention (identified not simply as storage but including managing, sharing, accessing, backing up and long term maintenance of data,) should be balanced against evidence of potential reuse of the data. A full economic case for retention will need to be made once a grant application is accepted. The Data Centre will need to consider the likely cost of preservation.

7. FULL DOCUMENTATION

Is there information i.e. completed metadata which will support the sharing, access and re-use of the data? In preparing this information considerations must be given to the fact that the individuals preparing the material for retention may not be involved in later projects which re-use the material.

Appendix 9: Sample Institutional Data Policy

Where they exist, the data policies of research performing organisations tend to be relatively straightforward and high-level and serve to indicate the responsibility of the institution and researcher to comply with funder policies. Typically they include the following statements:

- that research data in the organisation will be managed to agreed standards and in accordance with funder requirements;
- that PIs have a responsibility to ensure compliance with RDM policy in funded projects and to ensure the execution of the data management plans;
- that the institution has a responsibility to provide guidance and support;
- that data management plans should be produced in accordance with good practice;
- that research data should be offered for deposit in an appropriate data service, whether this is an international or national data centre or a data repository run by the research performing organisation;

Institutional policies generally also include a statement of who is responsible for overseeing and updating the policy. Often an indication of the likely institutional and general benefits is included.

The example below is taken from the University of Leeds in the UK. It is simple and straightforward, but has the notable feature of in-line linking to explanatory information and guidance.

University of Leeds Research Data Management Policy (July 2012)

http://library.leeds.ac.uk/research-data-management-policy

The management of Research Data reflects our:

- *commitment to research excellence*
- *recognition of our duty to our funders*
- *appreciation of the value of our data - to us and to others*

1. Research data will be managed to agreed standards throughout the research data lifecycle and according to funder requirements funder requirements.
2. Responsibility for research data management during any research project or programme lies with responsible owners such as Principal Investigators (PIs).
3. The University is responsible for the provision of training support and advice on research data management
4. A data management plan that explicitly addresses the capture, management, integrity, confidentiality, preservation, sharing and publication of research data must be created for each proposed research project or funding application. Sufficient metadata shall also be created and stored to aid discovery and re-use. Data management plans should take account of and ensure compliance with relevant legislative frameworks which may limit public access to the data (for example, in the areas of data protection, intellectual property and human rights).
5. All research data should be offered and assessed for deposit and preservation in anappropriate University, national or international data service or domain repository, unless specified otherwise in the data management plan.
6. Data should not be deposited with any organisation that does not commit to itsaccess and availability for re-use, unless this is a condition of the project funding or arising from other requirements.
7. At the completion of each research project, the PI should ensure that all relevant research data are made available, subject to meeting appropriate requirements, in the location specified in the data management plan.

8. Research and Innovation Board will be responsible for reviewing and updating the policy.

The University recognises the following benefits of implementing this policy:

a. support for the re-use of data

b. benefit future generations

c. improved data integrity, security and access management

d. opportunities for further research collaboration

e. improved research reproducibility and validation

f. further development of research skills

g. the ability to cite data as a publication

h. improved institutional research reputation

i. improved relationship with research funders

Appendix 10: Questions to be addressed in Data Management Plans

ESRC

The ESRC Research Data Policy communicates the expectation that the DMP to be submitted with the grant application should cover the following points:

1. an explanation of the existing data sources that will be used by the research project with references
2. an analysis of the gaps identified between the currently available and required data for the research
3. information on the data that will be produced by the research project, including:
    a. data volume
    b. data type, eg qualitative or quantitative data
    c. data quality, formats, standards documentation and metadata
    d. methodologies for data collection
4. planned quality assurance and back-up procedures [security/storage]
5. plans for management and archiving of collected data
6. expected difficulties in data sharing, along with and causes and possible measures to overcome these difficulties
7. explicit mention of consent, confidentiality, anonymisation and other ethical considerations
8. copyright and intellectual property ownership of the data
9. responsibilities for data management and curation within research teams at all participating institutions.

ESRC Research Data Policy, p.5; http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

NSF

The NSF general policy lists the following five questions which should be addressed in a two-page data management plan to be submitted with grant applications.

> 1. The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
>
> 2. The standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);
>
> 3. Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
>
> 4. Policies and provisions for re-use, re-distribution, and the production of derivatives; and
>
> 5. Plans for archiving data, samples, and other research products, and for preservation of access to them.

Five NSF Questions: http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#dmp

Appendix 11: Data Management Plan Templates

DCC Checklist for a Data Management Plan

Policy makers and support services considering developing guidance on data management planning and DMP templates should first consult the DCC Checklist for a Data Management Plan
http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP_Checklist_2013.pdf

The checklist contains the following sections:

1. Administrative information (details of the project etc)
2. Data collection: what data will you collect or create? how will the data be collected or created?
3. Documentation and metadata: what documentation and metadata will accompany the data?
4. Ethics and legal compliance: how will you manage any ethical issues? How will you manage copyright and IPR issues?
5. Storage and backup: how ail the data be stored and backed up during the research? How will you manage access and security?
6. Selection and preservation: which data should be retained, shared and/or preserved? What is the long-term preservation plan for the dataset?
7. Data sharing: How will you share the data? Are any restrictions on data sharing required?
8. Responsibilities and resources: who will be responsible for data management? What resources will you require to deliver your plan?

Horizon 2020 Guidance Data Management Plan (DMP) template

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

The purpose of the Data Management Plan (DMP) is to provide an analysis of the main elements of the data management policy that will be used by the applicants with regard to all the datasets that will be generated by the project.

The DMP is not a fixed document, but evolves during the lifespan of the project.

The DMP should address the points below on a dataset by dataset basis and should reflect the current status of reflection within the consortium about the data that will be produced.

1. Data set reference and name

Identifier for the data set to be produced.

2. Data set description

Description of the data that will be generated or collected, its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.

3. Standards and metadata

Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.

4. Data sharing

Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the

repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).

In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).

5. Archiving and preservation (including storage and backup)

Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.