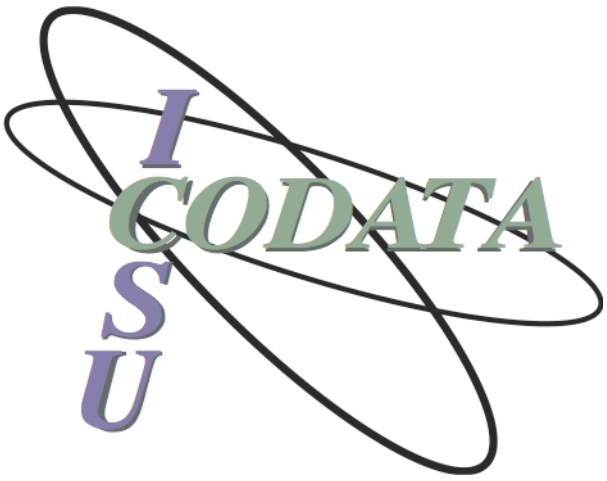
Current Best Practice for Research Data Management Policies

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

Simon Hodson and Laura Molloy

May 2014



Executive Summary

This memo responds to the request by the Danish e-Infrastructure Cooperation (DeIC), in partnership with Denmark's Electronic Research Library (DEFF), for a document 'to provide an overview of current best practices for research data management (RDM) policies within a number of subject areas, and as such inspire the development of a Danish national strategy on the area of RDM policies.'

In accordance with the instruction, the memo considers current best practices in an international context, highlights what are considered the 'pivotal points' or key elements of the policies mentioned and explores particular variations and distinctive features of those policies relating to particular subject areas (specifically: humanities, social sciences, health sciences, natural sciences and technical sciences).

Policy Drivers and General Principles

It is important for a survey of good practice to start by considering the drivers and principles underlying data policies which are currently mostly published by funders of research. Institutional research data management policy-making at the current time is mostly in response to these funder policies. The OECD's *Principles and Guidelines for Access to Research Data from Public Funding* (2007) have, since their publication, been of particular influence on research funders across countries and research disciplines. Very recently, there has been great emphasis on the idea of Open Data, the principle of 'open by default' and the criteria for data reuse (particularly as expressed in the concept of 'intelligent openness'). All funder research data policies underline the necessary limits on openness and particularly those relating to personal information and commercial considerations.

Accordingly, good practice in research data policies might be said to start with the following considerations:

1. **An account of the general drivers and principles:** these include the validation of research results, research opportunities for data reuse, the principle of open access by default to the outputs of publicly-funded research, and broader societal and economic benefits.
2. **A discussion of the requirements for effective data sharing:** e.g. 'intelligent openness' and the need for data to be 'discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards.'[1]
3. **A statement of the necessary limits of openness:** these are imposed, in particular, by the need to protect personal information, by the requirement to respect commercial considerations and by security concerns.

These general principles outlined, the policy can then move to more specific elements required for implementation.

Key Policy Elements

We provide a summary of what emerge from our survey as the key elements of current good practice in research data policies. This has the potential to be used as a framework for developing and assessing research data policies. In addition to the principles above, data policies generally contain the following key elements:

4. **A definition of research data:** many policies provide a definition of research data.
5. **An overview of the data within the scope of the policy:** which generally includes two definitions:
    a. The data that directly underpin or substantiate published research findings (i.e. those that are required for validation). Such data should be made available concurrently with the research publication.

[1] G8 Science Ministers Statement, 13 June 2013 https://www.gov.uk/government/news/g8-science-ministers-statement

b. The data assets that are created by the research project, but which may not directly underpin the published research findings. Such data should generally be made available within a specified time after creation or project end.

6. **An indication of general criteria for the selection of research data:** it is helpful for policies to indicate which data are likely to be the most important to select for sharing. Policies should also underline the need to make software and tools available.

7. **A summary of responsibilities:** in Table 1, we provide an overview of stakeholder responsibilities, as expressed in the research data policies surveyed.

Table 1: Summary of High-Level Stakeholder Responsibilities

Stakeholder	Responsibility
Funder	Develop and communicate RDM policy; provide advice directly or through data services; review implementation.
Researcher	Conform to policy in grant proposals,during the lifetime of the project; some responsibilities may remain after the completion of the project.
Research performing organisations	Ensure that the execution of the policy requirements by grant holding lead researchers is adequately supported; do this through institutional policies and provision of support and guidance, particularly for creation and execution of data management plans; depending on national data infrastructure provision the research performing organisation may also need to provide long-term stewardship for some data.
Research data services/centres	Provide long-term stewardship for specific data in accordance with funder policies; provide guidance and support according to role designated by funder.

8. **An indication of the availability of infrastructure and responsibility for costs:** the policy should indicate what expectations there are for the provision of research data infrastructure and the associated costs of data management and stewardship. This applies to data management during the lifetime of a project, but is particularly important with regard to the long-term stewardship of the data. International or national data centres serving the research area should be used, if available. Where these are not available, the policy should indicate where responsibility lies. This is generally with the research performing institution. What costs may be covered from the research grant and is it, for example, acceptable for deposit charges to be paid? How is the cost of long-term stewardship of research data in the host institution to be resourced?

9. **An overview of data management planning requirements:** it is good practice to require researchers to prepare data management plans (DMPs) and most data policies surveyed make this stipulation. The policy should lay out the procedure around the DMP. Many funder policies require either a brief statement or a more detailed plan to be presented as part of a project proposal. Ideally, the policy should indicate how the DMP will be assessed. It is good practice to provide guidance for those reviewing the DMP.

   Some policies require a more detailed DMP to be prepared once the project is underway. Again, the policy should lay out any procedures for reviewing and reporting against the plan.

   Above all, the policy should indicate, broadly at least, the issues that should be addressed in a DMP and should link to more detailed guidance and support if available. A number of examples of these guidelines are included in the memo and appendices.

   Inter alia, the DMP should address who will take responsibility for the long-term stewardship of data created. For cases where this falls to the research performing organisation, the policy may indicate general expectations of how long data should be retained. There is some variation in what the policies surveyed say on this matter and practice is still in the process of evolving.

10. **Recommendations on enabling discovery and reuse:** policies generally make a number of specific recommendations or requirements that relate to the objective of enabling discovery and reuse. These typically include the requirement for published research to state how the supporting data may be

accessed as well as recommendations relating to the use of appropriate metadata, permanent identifiers and licenses which enable reuse.

11. **Stipulations to encourage recognition and reward for data providers**: policies generally acknowledge that moving to an open data regime requires for many research areas a shift in practice and culture. For this reason policies often include, alongside the requirement that publicly funded data should be made open, some statement of the need for appropriate recognition and reward for those researchers who make data open.

    There are two notable policy implications of this principle. Firstly, many policies require acknowledgement of data reuse and the citation of data where it underpins further research findings. Secondly, some policies allow periods of privileged access. In relation to the latter, there is considerable diversity. Certain policies, while stressing the time should be limited, uphold it as an important principle. Others reject the principle and require deposit within a limited timescale of project completion or data generation.

12. **A summary of reporting requirements, compliance monitoring and any possible sanctions:** policies should indicate how compliance will be monitored, what reporting is required and what sanctions may be imposed. The policies surveyed are usually specific on the reporting procedures but do not generally indicate which precise sanctions may be imposed.

Subject Area Considerations

The third part of this memo provides an overview with examples of variations in research data policies in different subject areas. There are some limitations that should be mentioned here.

Many variations between the policies of funders in different research areas appear to be contingent and procedural rather than based on specific and intrinsic issues to do with data sharing in the given research area.

Many research data policies present broad, high-level principles. Nevertheless, some funders in a given research area highlight particular issues which are characteristic for that discipline. However, it is often only at the level of accompanying or more detailed guidance and support that a lot of such issues are really addressed, rather than in the research data policy itself.

With policy development and - above all - data practice still hardly mature, we would hesitate to draw particularly strong conclusions about subject-specific variations from the range of examples surveyed. Nevertheless, the most significant variations between the policy concerns in given subject areas relate to:

Legal and ethical requirements specific to the type of research being conducted. For example, privacy issues are particularly important in the health and social sciences.

Existence of more established data infrastructure and practice. For research disciplines where international or national data centres have been established, policies more often provide (in appendices or guidance) lists of appropriate data centres or databases and allude to any existing technical standards or practices widely used in that discipline.

Accepted technical approaches in a specific research area. In some areas of the social sciences, life sciences and natural or technical sciences, specific data format or metadata standards have emerged and become common practice in that community. Where this is the case, these standards may be mentioned directly or indirectly in policy documents and recommendations.

Introduction

The memo that follows responds to the request by the Danish e-Infrastructure Cooperation (DeIC), in partnership with Denmark's Electronic Research Library (DEFF), for a document 'to provide an overview of current best practices for research data management (RDM) policies within a number of subject areas, and as such inspire the development of a Danish national strategy on the area of RDM policies.'

In accordance with the instruction, this memo considers current best practices in an international context, highlights what are considered the 'pivotal points' or key elements of the policies mentioned and explores particular variations and distinctive features of those policies relating to particular subject areas (specifically: humanities, social sciences, health sciences, natural sciences and technical sciences).

Accordingly, this memo comprises sections addressing the following issues:

1. RDM policy drivers and general principles featuring in research data policies;
2. Key elements of research data policies;
3. Particular variations and distinctive features of those policies relating to particular subject areas.

In order to provide supporting material, the memo comes with a number of Appendices that we hope will be useful to Danish stakeholders' discussions around research data policies.

In order to provide a framework that may be useful for developing research data policies, this memo primarily summarises research funder data policies. Funder policies are considerably more detailed and further developed that those produced so far by institutions. And arguably, the primary purpose of data policies for research performing institutions so far is to demonstrate compliance with funder policy. Where journal editorial boards have released data policies, these generally amount to a requirement to state how the data may be accessed. Research funder policies cover the major general considerations which can form a framework for the development of research data policies, as presented below.

Section 1: RDM Policy - Drivers and Principles

It is important for a survey of good practice to start by considering the drivers and principles underlying data policies which are currently mostly published by funders of research. Institutional research data management policy-making at the current time is mostly in response to these funder policies. The OECD's *Principles and Guidelines for Access to Research Data from Public Funding* (2007)[2] have, since their publication, been of particular influence on research funders across countries and research disciplines. Very recently, there has been a great emphasis on the idea of Open Data, the principle of 'open by default' and the criteria for data reuse (particularly as expressed in the concept of 'intelligent openness'). All funder research data policies underline the necessary limits on openness and particularly those relating to personal information and commercial considerations.

1.1 Data are a Fundamental Asset of Research

Data are fundamental to the research lifecycle in all subject areas, both as an essential raw material and as a reusable output of research. The UK Economic and Social Research Council (ESRC) policy makes a particularly strong statement for the fundamental role of data in social and economic research: 'Data are the main asset of economic and social research. They are the basis for research and also the ultimate product of research.'[3] Increasingly, government and funder policies have focused on the societal benefits of research: benefits that cannot be achieved without adequate communication of research findings and access to the

[2] OECD's Principles and Guidelines for Access to Research Data from Public Funding (2007), http://www.oecd.org/sti/sci-tech/38500813.pdf

[3] ESRC Research Data Policy, September 2010, Revised March 2013, p.1; http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

underlying data. Researchers have a fundamental responsibility to communicate findings and outputs, including data. The ESRC Policy also stresses that data management and sharing is, therefore, an essential 'indivisible part of the research project' - not an administrative and bureaucratic add-on.[4] The aim of many funder policies has been to encourage practice in which care for data, during and after a research project, is an essential part of the activity. There are a number of drivers for this.

1.2 Research Methods and the Verification of Results

A widely stated principle in research data policies is to enable the verification of results. Fundamental to good research practice and the scientific method are verification, reproducibility and collective self-correction. For this reason research data policies generally attach particular importance to the availability of the data that underpin and substantiate the findings in research publications: as the UK Natural Environment Research Council (NERC) data policy proposes, access to such data 'supports the fundamental scientific requirement of allowing others to confirm or challenge research results'.[5]

1.3 Broader Policy Objectives: Benefits of Data Availability and Reuse

The ESRC policy preamble also mentions two significant and relatively recent developments which have influenced data policies. These are the increasingly 'data intensive' characteristics of some research areas and the broader governmental moves in the UK towards Open Data[6] and open public data.[7] Data-intensive research has shown in some disciplines the benefits of access to data for further analysis and repurposing. The open data movement advances the principles of transparency and the social and economic benefits of broader access both to governmental public data and data produced by research. Key recent policy documents summarising these tendencies are the *G8 Science Ministers Statement* (2013),[8] the *G8 Open Data Charter and Technical Annex* (2013),[9] the *US OSTP Memorandum Increasing Access to the Results of Federally Funded Scientific Research* (2013),[10] and the related *US OSTP Executive Order - Making Open and Machine Readable the New Default for Government Information* (2013).[11] Common to these policy statements is the view that Open Data will promote transparency, encourage innovation and have beneficial economic effects. The influence of these policy statements on the Open Data pilot in the EC's Horizon 2020 programme is clear. Specifically in the research sphere it is argued that Open Data will:

- allow easier and quicker validation of results and will reduce duplication of effort, leading to more reliable, transparent and efficient research.
- increase collaboration around data and their reuse will lead to accelerated innovation.
- encourage reuse by citizens, private and third sector organisations leading to greater transparency and engagement in research.[12]

Particular benefits internationally are also expected from more open data in the Health Sciences and have led to a joint statement by funders of health research to coordinate promotion of data sharing (see Appendix 1).

---

[4] ESRC Research Data Policy, September 2010, Revised March 2013, p.5; http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

[5] NERC Data Policy http://www.nerc.ac.uk/research/sites/data/policy/data-policy.pdf

[6] Described at http://data.gov.uk/

[7] ESRC Research Data Policy, September 2010, Revised March 2013, p.1; http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

[8] G8 Science Ministers Statement, 13 June 2013 https://www.gov.uk/government/news/g8-science-ministers-statement

[9] G8 Open Data Charter and Technical Annex, 18 June 2013, https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex

[10] US OSTP Memorandum Increasing Access to the Results of Federally Funded Scientific Research, 22 February 2013 http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[11] US OSTP Executive Order -- Making Open and Machine Readable the New Default for Government Information, http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

[12] See the documents mentioned and the Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf and Guidelines on Data Management in Horizon 2020 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

1.4 Open Access to Assets Produced by Public Funding

The forerunner to these recent policy statements was the OECD's *Principles and Guidelines for Access to Research Data from Public Funding* (2007).[13] This document has had a considerable influence and many funders directly reference these principles as part of their research data policy's justification and foundation.[14] Emerging themselves from an earlier *Declaration* (2004),[15] the *Principles and Guidelines* expressed an increasing prevalent view of the importance of appropriate access to the data outputs of publicly funded research. Data sharing in the genomics, earth observation and social science communities were cited as important examples.

The principles state:

- Publicly-funded research data are a public good, produced in the public interest.
- Publicly-funded research data should be openly available to the maximum extent possible.

The 'public good' can be understood as, firstly, a commodity or service provided without profit to all members of a society, and, secondly, the benefit or well-being of the public. Access to research data, then, is part of a larger ambition by many funders to contribute to public wellbeing (including but not limited to the research community) without requiring the support of a profit-making business model, as it is recognised that data is often an output of research which has already consumed public funds. This is confirmed by the OECD *Principles* which conclude the sharing of research data will achieve scientific, economic and social benefits: 'access to research data increases the returns from public investment in this area; reinforces open scientific inquiry; encourages diversity of studies and opinion; promotes new areas of work and enables the exploration of topics not envisioned by the initial investigators.'[16]

As noted, a large number of data policies explicitly reference these principles and arguments. The RCUK *Common Principles on Data Policy*, drawing on the OECD *Principles*, state that 'Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property.'[17]

1.4.1 Open By Default

All research data policies and policy statements surveyed recognise that there are some legitimate limits to the open sharing of research data. And it is generally agreed that these limits are set when there are 'concerns in relation to privacy, safety, security, commercial interests and the legitimate concerns of private partners.'[18] These necessary limits will be discussed in more detail below.

Building on the OECD *Principles*, the more recent policy statements have stressed the principle of 'Open by Default' as a way of minimising *unnecessary* restrictions on the availability of research data. The presumption should be towards the Open availability of the research data for further use. Although there are some variations, definitions of Open often comprise the following (e.g. as in the G8 Open Data Charter):

- Free of charge: 'open data should be available free of charge in order to encourage their most widespread use';
- Open Formats: encourage the use of open, non-proprietary, formats;

[13] OECD's Principles and Guidelines for Access to Research Data from Public Funding (2007), http://www.oecd.org/sti/sci-tech/38500813.pdf

[14] E.g. RCUK Common Principles on Data Policy http://www.rcuk.ac.uk/research/datapolicy/; referenced also in a number of the specific UK Research Council policies, e.g. ESRC http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx; NERC http://www.nerc.ac.uk/research/sites/data/policy/data-policy.pdf; and the EPSRC Principles http://www.epsrc.ac.uk/about/standards/researchdata/Pages/principles.aspx.

[15] OECD Declaration on Access to Research Data from Public Funding (2004) [30 January 2004 - C(2004)31/REV1, http://webnet.oecd.org/OECDACTS/Instruments/ShowInstrumentView.aspx?InstrumentID=157&InstrumentPID=153&Lang=en&Book

[16] OECD's Principles and Guidelines for Access to Research Data from Public Funding (2007), http://www.oecd.org/sti/sci-tech/38500813.pdf, p.3.

[17] RCUK Common Principles on Data Policy http://www.rcuk.ac.uk/research/datapolicy/

[18] G8 Science Ministers Statement, 13 June 2013 https://www.gov.uk/government/news/g8-science-ministers-statement

- Open Licenses: encourage use of open licences where possible 'so that no restrictions or charges are placed on the re-use of the information for non-commercial or commercial purposes, save for exceptional circumstances'.[19]

Using the same criteria, the Horizon 2020 Open Data pilot asserts that Open means that it is 'possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user'.[20] The EC's position is summarised effectively, as follows:

> The European Commission's vision is that information already paid for by the public purse should not be paid for again each time it is accessed or used, and that it should benefit European companies and citizens to the full. This means making publicly-funded scientific information available online, at no extra cost, to European researchers, innovative industries and citizens, while ensuring long-term preservation.'[21]

1.5 Intelligent Openness and Principles for Reuse

To achieve these benefits it is not enough for research data simply to be made available. Additional information must be provided so that the data may be discovered and understood. As the RCUK *Common Principles* state: 'To enable research data to be discoverable and effectively re-used by others, sufficient metadata should be recorded and made openly available to enable other researchers to understand the research and re-use potential of the data. Published results should always include information on how to access the supporting data.' Similar statements appear in many research data policies.

As a way of emphasising this point and exploring the issues in more detail, the influential Royal Society report *Science as an Open Enterprise* (2012)[22] coined the term 'intelligent openness' to describe the preconditions for the effective communication of data. 'Intelligent openness' imposes four requirements: 'data must be accessible and readily located; they must be intelligible to those who wish to scrutinise them; data must be assessable so that judgments can be made about their reliability and the competence of those who created them; and they must be usable by others.'[23]

These requirements have been influential and very similar criteria for making scientific data fully open and reusable are presented in the *G8 Science Ministers' Statement*: 'Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards.'[24]

The terms used in the *G8 Science Ministers' Statement* are reproduced in the 'Additional Guidance for Data Management Plans' as an Annex to the EC's *Guidelines on Data Management in Horizon 2020* [See Appendix 2].[25]

1.6 Necessary Limits of Open: Legitimate Limitations on Research Data Availability

All funder research data policies surveyed are clear that necessary constraints exist on the appropriate availability of research data. Policies generally agree that these constraints include the following, as taken from the EC *Guidelines on Data Management in Horizon 2020*:

- Protection of personal data;
- Security concerns;
- Protection of intellectual property;
- Protection of commercial interests of project partners.[26]

---

[19] G8 Open Data Charter and Technical Annex, 18 June 2013, https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex
[20] Guidelines on Data Management in Horizon 2020, p.3.
[21] Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, p.4.
[22] Royal Society, Science as an Open Enterprise (2012) https://royalsociety.org/policy/projects/science-public-enterprise/Report/
[23] Royal Society, Science as an Open Enterprise (2012), pp. 7, 12, 14-15.
[24] G8 Science Ministers Statement, 13 June 2013 https://www.gov.uk/government/news/g8-science-ministers-statement
[25] Guidelines on Data Management in Horizon 2020, p.6;
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

The guidance stresses that recipients of EC funding under Horizon 2020 will be under obligation:

- to protect results and possible commercialisation through appropriate measures (Article 27);
- to maintain confidentiality of 'any data, documents or other material (in any form) that is identified as confidential at the time it is disclosed ('confidential information')' (Article 36);
- to protect any data or information in relation to which the Commission, during the selection procedure may have identified as raising security issues (Article 37);
- to protect personal data 'in compliance with applicable EU and national law on data protection (including authorisations or notification requirements)' (Article 39.2).[27]

Interestingly, the EC Horizon 2020 policy allows an additional reason, related to the conduct of research, for not sharing data: if it can be reasonably shown that 'the main objective of their research be compromised by making data openly accessible.'[28] Such a consideration is also contained in the UK Medical Research Council (MRC) Data Policy which is concerned that the ongoing and increasing value of longitudinal cohort studies in particular might be damaged by premature data release: 'Ongoing research contributing to the completion of datasets must not be compromised by premature or opportunistic sharing and analysis. Sharing should always take account of enhancing the long-term value of the data.'[29]

1.6.1 Protection of Personal Data

The processing and communication of personal data is governed by data protection legislation, notably the EC Data Protection Directive[30] and its various transpositions into individual member states' law (in the UK this is the Data Protection Act 1998[31]). Accordingly, research data policies surveyed universally stress the importance of protecting personally identifiable data as a pre-eminent restriction on the sharing of research data unless adequately anonymised.

The US National Institute of Health (NIH) *Data Sharing Workbook*, accessible from information pages about the data policy, makes the important observation that:

> There are two basic tools to protect from disclosure of sensitive data and subjects' identities: restricting information in the dataset, and restricting access to the data. Thus, data intended for broader use should be free of identifiers that would permit linkages to the research participants and free of content that would create unacceptably high risks of subject identification.[32]

Concern for the protection of personal data is particularly heavily emphasised in policies relating to the social and health sciences where the involvement of human subjects is necessarily greater. These research areas have longstanding ethical guidelines for the conduct of research, and research projects involving human subjects must go through an ethical review and clearance process, generally administered according to guidelines by committees in the research performing organisations.

The ESRC Data Policy contains an important message, however, which is less forcefully expressed elsewhere. It is the responsibility of the grant receiving researcher to take into account necessary restrictions on data sharing and to plan and conduct data gathering and data management *in such a way as to maximise the potential for data sharing.*[33] This is also a key message in the ESRC-related guidance provided by the UK Data Archive / Data Service and will be mentioned below.

---

[26] Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, pp.9, 14.
[27] See 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 and relevent excerpts in Appendix 2.
[28] Guidelines on Data Management in Horizon 2020, p.4.
[29] MRC Data Policy http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/datasharing/Policy/index.htm
[30] EC Data Protection Directive (Directive 95/46/EC) http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML
[31] UK Data Protection Act 1998 www.legislation.gov.uk/ukpga/1998/29/contents
[32] US NIH (National Institute of Health) Data Sharing Workbook, p.2; http://grants.nih.gov/grants/policy/data_sharing/data_sharing_workbook.pdf
[33] ESRC Data Policy http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx

A similar message is contained in the MRC Data Policy. The policy stresses that 'risks such as inappropriate disclosure of personal information must be managed in a proportionate yet robust manner' and that 'for medical research involving personal data, the appropriate regulatory permissions – ethical, legal and institutional – must be in place before the data can be shared.' However, the policy also communicates the message that 'Researchers, research participants and research regulators *must ensure that within the regulatory requirements of the law, opportunities for new uses are maximised*'[34] (our emphasis). In some circumstances, with data from health studies - particularly medical records – where sufficient anonymisation is not possible, only controlled reuse by other research groups will be appropriate.

1.6.2 Commercial Interests

It is generally the hope and intention of research funders that there will be positive economic outcomes from research and in order to achieve this, many research programmes encourage research institutions to partner with commercial organisations; in such cases, most funders allow the commercial exploitation of intellectual property created by the project. Two considerations for data policies arise from this: firstly, the need to protect the IPR and commercial interests of prospective research partners. It is this concern that led to certain parts of Horizon 2020 not being included in the Open Data pilot. Such considerations may also provide grounds for delaying data release. Secondly, to encourage commercial exploitation, some funders such as the MRC, recognise 'that it may be necessary on occasion to delay [data] publication for a short period to allow time for [patent] applications to be drafted.'[35] Similar provisions are contained in the Model Grant Agreement for Horizon 2020 (Article 27) [see Appendix 3 for the limitations in the EC Model Grant Agreement] and in US policies. For example, the US NIH strongly encourages the free sharing of data, innovations and research tools arising from grant funding but at the same time recognises 'the rights of grantees and contractors to elect and retain title to subject inventions developed with federal funding'.[36]

1.7 Acceptable Limits of Open

While protection of personal data, commercial interests and security concerns may be regarded as universally recognised *necessary* limits on data availability, policies mention other limits which are more *contingent*. The most common example is the concern to allow the data creators a privileged period of exclusive use.

1.7.1 Privileged Period of Exclusive Use

Data created by research is regarded as an asset by researchers for exploitation. Many researchers consider it unfair to be obliged to share data before it has been fully exploited. In respect of this, some research data policies allow an embargo period. This might be regarded alternatively as a necessary accommodation of research culture and its requirements or as a pragmatic concession. There is little doubt that perception of the appropriateness (or not) of such a period of exclusive use varies somewhat across research areas. Nevertheless, it is generally argued that any exclusive period should be limited in line with normal practice in that subject area. It is also commonly stated that the option of an embargo should not apply to those data that directly underpin published findings and which therefore are necessary for validation.

Section 2: Key Policy Elements

This section details the key elements that are included in most funder research data policies. The authors consider it good practice to address these elements in a general research data policy and to provide links to information resources and guidance to support implementation of the policy.

[34] MRC Research Data Policy http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/datasharing/Policy/index.htm
[35] MRC Research Data Policy http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/datasharing/Policy/index.htm
[36] NIH Data Policy, in Grants Policy Statement, 8. Administrative Requirement; particularly 8.2.1 Rights in Data and 8.2.3 Sharing Research Resources: http://grants.nih.gov/grants/policy/nihgps_2013/nihgps_ch8.htm#_Toc271264947

In this section, we move from the general principles and drivers described above, to the specific details for implementation of the policy. Accordingly, good practice in research data policies might be said to start with the following considerations:

1. An account of the general drivers and principles (validation of research results, research opportunities for data reuse, the principle of open access by default to the outputs of publicly-funded research, broader societal and economic benefits).
2. A discussion of the requirements for effective data sharing, e.g. 'intelligent openness' and the need for data to be 'discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards.'
3. A statement of the necessary limits of openness imposed: in particular by the need to protect personal information, by the requirement to respect commercial considerations and by security concerns.

These general principles outlined, the policy can then move to more specific elements required for implementation. A number of analyses of research data policies exist, primarily constructed to guide researchers to funders' expectations. However, we have considered it more helpful to present the elements of data policies in the structure given here and outlined in the Executive Summary (see Appendix 4 for analyses of RDM policy elements).

2.1 Definition of Research Data

It is helpful for a data policy to include a definition of research data. Many definitions of research data exist. The definition provided by the OECD *Principles and Guidelines* serves as a useful starting point:

> '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.[37]

The OECD definition goes on to stress that it is primarily concerned with digital data, but acknowledges that analogue representations of data could equally well be in scope if the cost of sharing is not excessive, a point echoed in the UK by the Research Council UK (RCUK) *Common Principles on Data Policy*[38]. However, the OECD *Principles* explicitly exclude 'laboratory notebooks, preliminary 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)', observing that '[a]ccess to all of these products or outcomes of research is governed by different considerations than those dealt with here.' By contrast, many policies explicitly include analogue representations of data as well as 'grey literature', notes and sources analogous to laboratory notebooks.[39] Data policies often also make reference to the procedures which should govern physical and biological samples[40] (see further examples of research data definitions in Appendix 5).

It can be argued that good practice policies strive to achieve a definition that makes clear the range of things which *could* be included in the definition, while at some point in the policy helping to limit the scope and prioritise so that it is understood that the expectation does not unreasonably encompass 'everything'.

The discussion around the recent update to the Public Library of Science data availability policy is instructive in this regard. The updated policy used the following definition:

> *What do we mean by data?*
>
> "Data are any and all of the digital materials that are collected and analyzed in the pursuit of scientific advances." Examples could include spreadsheets of original measurements (of cells, of fluorescent intensity, of respiratory volume), large datasets such as next-generation sequence reads, verbatim responses from qualitative studies, software code, or even image files used to create

[37] OECD Principles, p.13.
[38] RCUK Common Principles on Data Policy, http://www.rcuk.ac.uk/research/datapolicy/
[39] E.g. EPSRC Research Data Expectations http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx
[40] E.g. BBSRC Research Data Policy, p. 9; http://www.bbsrc.ac.uk/web/FILES/Policies/data-sharing-policy.pdf

> figures. Data should be in the form in which it was originally collected, before summarizing, analyzing or reporting.[41]

However, in response to contributors' concern that they were being asked to deposit all the data, PLoS clarified this part of the policy thus:

> *Correction*
>
> We have struck out the paragraph in the original PLOS ONE blog post headed "What do we mean by data?", as we think it led to much of the confusion. Instead we offer this guidance to authors planning to submit to a PLOS journal.
>
> *What data do I need to make available?*
>
> We ask you to make available the data underlying the findings in the paper, which would be needed by someone wishing to understand, validate or replicate the work. Our policy has not changed in this regard. What has changed is that we now ask you to say where the data can be found.[42]

2.2 Data in Scope

A clear data policy should govern the best practice management of *all* data produced by research projects, including sensitive data. 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.

In terms of best practice it is also important to provide a clear definition of the research data central to the scope of the policy and which is most important to be made Open. Policies governing the data produced by publicly funded research projects generally divide data into two categories:

1. The data that directly underpin or substantiate published research findings (i.e. those that are required for validation). Such data should be made available concurrently with the research publication.
2. The data assets that are created by the research project, but which may not directly underpin the published research findings. For example, a research project may generate a large dataset or data collection of which only a part or subset is used to obtain the findings of a given publication.

This categorisation features in the Horizon 2020 Guidance and consequently in Article 29.3 of the Model Grant Agreement. (See Appendix 6; combining this with the necessary limits on Open data, the authors have developed a 'Simple typology of different research data types in relation to data policies', see Appendix 7).

The second category inevitably raises issues of selection of research data as it is recognised that not all the data generated by a project will be worth sharing. Policies generally ask that the data management plan should describe the most significant and potentially valuable data assets to be created by the project.

2.2.1 Selection of Research Data

Methodologies for the evaluation and selection of research data exist and are beyond the scope of a data policy. Nevertheless, it is helpful to indicate in the policy or supporting guidance the characteristics of data that should be retained. As well as the data that underpins published research results, this includes significant and substantial data created by the project and particularly if those data are: unrepeatable observations, longitudinal studies of human or natural events, or experimental results that would be impossible or expensive to reproduce.

The UK Biotechnology and Biological Sciences Research Council (BBSRC) data policy, for example, has identified three scientific areas, described by type of data, where the case for data sharing is particularly strong, and expects data sharing to take place: data arising from high volume experimentation; low

[41] 'PLoS' New Data Policy: Public Access to Data' http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/
[42] 'PLOS' New Data Policy: Part Two', http://blogs.plos.org/everyone/2014/03/08/plos-new-data-policy-public-access-data/

throughput data arising from long time series or cumulative approaches; and models generated using systems approaches. In other areas, data sharing is 'encouraged' rather than 'expected'.[43]

Similarly, the NIH *Data Sharing Workbook* has a useful and straightforward summary of those data that 'constitute' unique resources that are most important to share:

> In summary, all data should be considered for sharing. Data that constitute "unique resources" especially should be shared unless there is a strong reason not to. Such data are difficult if not impossible to replicate because of cost (e.g. large national longitudinal surveys), special circumstances (e.g. health effects associated with a natural disaster), or rare population (e.g. a sample of centenarians). Less likely candidates for sharing are data from small studies involving research procedures that are easily replicated or data from human subjects that might identify them.[44]

The NERC *Data Value Checklist* also represents straightforward good practice in this area[45] (see Appendix 8 for criteria for data selection in the NERC Data Value Checklist). Supporting documentation and guidance should summarise such principles and link to methodologies such as that presented by the Digital Curation Centre.[46]

2.2.2 Software and Tools

It is widely regarded as important to stress that software and tools necessary to generate (and replicate) the data and research results should also be made available. This might cover software, code, algorithms, protocols, analysis tools, etc. Although excluded in some definitions, it is now widely regarded as essential and unproblematic for these resources to also be made available along with the research data.[47]

2.3 Responsibilities

An important purpose of a policy is to lay out the responsibilities of various stakeholders towards ensuring the long-term preservation, availability and reuse of data assets. This is nicely stated in the covering document for the ESRC policy: 'This policy aims to support grant holders who collect, produce and re-use data, by defining the roles and responsibilities of researchers, ESRC and its data service providers.'[48] As indicated here, the stakeholders will typically include funders, researchers, research performing institutions and data centres/services.

The RCUK *Common Principles* assign responsibilities to both the research performing organisation and researcher: they specify the need for data management policies at the institutional level and for specific research projects (the latter generally in the form of data management plans). Research organisations are responsible for raising the awareness of their research (and research support) populations of relevant funder policies, and for supporting compliance with funder policy. In turn, in the UK Engineering and Physical Sciences Research Council (EPSRC) policy, the individual researcher or research student is 'required to comply with research organisation policies in this area or, in exceptional circumstances, to provide justification of why this is not possible.'[49]

Although present, as indicated in the *RCUK Common Principles*, the emphasis on the role of the research performing institution in providing an excellent environment for research, including the management of research data, is greater in the EPSRC *Policy and Expectations*. The EPSRC's expectation that each institution in receipt of its funds should have a roadmap to ensure compliance by May 2015 has stimulated a lot of activity in the UK and the development of institutional RDM policies. UK institutional policies have been listed by the Digital Curation Centre: they are relatively straightforward and high-level and serve to indicate

[43] BBSRC Research Data Policy, p. 9; http://www.bbsrc.ac.uk/web/FILES/Policies/data-sharing-policy.pdf
[44] NIH Data Sharing Handbook, p.1; http://grants.nih.gov/grants/policy/data_sharing/data_sharing_workbook.pdf
[45] NERC Data Value Checklist http://www.nerc.ac.uk/research/sites/data/policy/data-value-checklist.pdf
[46] Whyte, A. & Wilson, A. (2010). "How to Appraise and Select Research Data for Curation". DCC How-to Guides. Edinburgh: Digital Curation Centre: http://www.dcc.ac.uk/resources/how-guides/appraise-select-data
[47] E.g. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, pp. 11.
[48] ESRC Research Data Policy, p.2; http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf
[49] EPSRC Expectation, iii: http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx

the responsibility of the institution and researcher to comply with funder policies.[50] The data policies of research performing organisations should include the same elements discussed here but from the point of view of the organisation rather than the funder. (An outline of a typical high-level institutional data policy is contained in Appendix 9.)

In contrast to generally more holistic UK approaches, US funder policies have generally simply specified that responsibility for compliance rests with the lead researcher in receipt of the grant. Nevertheless, it is understood that research performing institutions have broader responsibilities to support the good execution of grant funded research and that this includes the obligations of good research data management.

The high-level stakeholder responsibilities generally characterised in RDM policies are summarised in the table below.

Table 1: Summary of High-Level Stakeholder Responsibilities

Stakeholder	Responsibility
Funder	Develop and communicate RDM policy; provide advice directly or through data services; review implementation.
Researcher	Conform to policy in grant proposals, during the lifetime of the project; some responsibilities may remain after the completion of the project.
Research performing organisations	Ensure that the execution of the policy requirements by grant holding lead researchers is adequately supported; do this through institutional policies, and provision of support and guidance, particularly for data management planning and the execution of those plans; depending on national data infrastructure provision the research performing organisation may also need to provide long-term stewardship for some data.
Research data services/centres	Provide long-term stewardship for specific data in accordance with funder policies; provide guidance and support according to role designated by funder.

2.4 Availability of Infrastructure and Responsibility for Costs

Research data management and sharing is an integral, indivisible part of research projects and essential to the realisation of research data policy objectives. It is a necessary consequence of this principle that the human and technical infrastructure required to achieve these policy objectives must be considered. Ideally, there should be some indication of how the necessary infrastructure will be provided and how the costs of this infrastructure will be shared.

2.4.1 Provision of RDM Infrastructure

One of the key elements of a policy, dependent upon infrastructure provision, will be to indicate where responsibility for the long-term stewardship of the data will fall. This is most commonly either with a national or international data centre, or with the research-performing institution hosting the research project. Not all policies are as clear as they might be on this matter. US National Science Foundation (NSF) policies, by and large, simply state that the responsibility for ensuring long-term preservation lies with the grant-holding lead researcher. Some - e.g. that of the NSF Division of Ocean Sciences or those relating to social sciences - mention national data centres. For some US sub-programmes there is an explicit responsibility on the grant recipient to deposit data in such facilities.[51]

By and large, the UK funder policies are clearer in this regard. For those funders like ESRC and NERC which have provided data services, the responsibility lies with the researcher to offer data outputs to these facilities. In the biosciences, represented by BBSRC, there has been a proliferation of community-led and often international databases and resources: these therefore recommend practice in accordance with

[50] DCC list of UK Institutional Data Policies http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies
[51] Data policy of the NSF Ocean Sciences Division http://www.nsf.gov/pubs/2011/nsf11060/nsf11060.pdf

community norms. The EPSRC policy, on the other hand, states that in the absence of any appropriate national or international data centre it is the responsibility of the research performing institution to see to the long-term curation of the data.

2.4.2 Costs of RDM Infrastructure

Research data management requires time and effort. It is a part of research activity and it has a cost. Funder policies often attempt to make clear that it is legitimate to use public funds to support RDM effort.[52] Different systems may have different ways of allocating responsibility for these costs and principally for governing the way they are divided between the funding body and the research performing organisation. It is important that there should be clarity on how the 'in-project' costs of RDM should be funded and what provision there is for funding long-term stewardship of data.

2.4.3 In-Project Costs of RDM

UK research funders are increasingly encouraging applicants to ensure that RDM is adequately and proportionately funded. The ESRC, for example, is clear that the costs of in-project research data management should be included in the grant proposal. The ESRC will 'review the data management and sharing plan, including any costs for its implementation, as an integral part of the funding decision and based on this decision provide appropriate funding for data management'.[53]

Similarly, the Horizon 2020 Guidelines state that participating projects will receive dedicated support for RDM. 'In particular, any costs relating to the implementation of the pilot will be reimbursed and specific technical and professional support services will be provided.'[54]

Various NSF Divisional RDM policies also state that the costs of in-project RDM should be included in the general grant proposal. The NSF Atmospheric and Geospatial Sciences Division includes, as an example, the cost of creating a website if that is to be the means of data dissemination. However, it should be stressed that unless the research performing institution takes explicit responsibility for the long-term maintenance and curation of such a website, this is a far from optimal solution for providing data preservation and access.[55]

The EC plans to provide research data management support for Horizon 2020, both during projects and for long-term stewardship, by means of projects themselves funded under the e-Infrastructures section of the *Research Infrastructures Work Programme 2014-15*.[56] 'Exact plans for service delivery will be available in the Participants Portal by end of 2014. Full support services [are] expected to be available at the latest in 2015 and coordinated with similar national initiatives. These services will be available only to research projects funded under Horizon 2020, with preference to those participating in the Open Research Data Pilot.'[57] This is an admirable attempt at a joined-up approach. However, although these services will build on existing initiatives, it will be interesting to observe how effectively they can support projects in the Open Data Pilot given the timescales involved.

2.4.4 Costs for Long-term Preservation and Access

Research infrastructure is funded in a variety of ways: international collaborations, joint funding agreements and support from national research funders are the most common. Where a given research funder provides a data archive service there is an expectation that appropriate data should be offered to that archive (this is the case for NERC and ESRC in the UK). The Archaeology Data Service in the UK now charges for deposit and holders of archaeology-related AHRC grants must budget for data deposit fees. It is possible, even likely,

[52] RCUK Common Principles on Data Policy http://www.rcuk.ac.uk/research/datapolicy/
[53] ESRC Research Data Management Policy, p.3; http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf
[54] Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, p.12; http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
[55] NSF Atmospheric and Geospatial Sciences Division Research Data Policy http://www.nsf.gov/geo/geo-data-policies/ags/index.jsp
[56] Research Infrastructures Work Programme 2014-15, pp. 25+; http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-infrastructures_en.pdf
[57] Guidelines on Data Management in Horizon 2020, p.4: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

that – on the model of Article Processing Charges for (so-called) Gold Open Access – data repository services will increasingly charge for deposit. This is the sustainability model, for example, of the not-for-profit Dryad Digital Repository which charges a small 'Data Publishing Charge' (generally between US$25 and US$80, depending on the pricing plan) for ensuring the long-term stewardship and access to the package of data associated with a given peer-reviewed research article.[58] The necessary policy corollary of such an approach is that research funders make a clear statement on the responsibility for data deposit charges and whether it is appropriate to pay them with funds from a research grant. In the UK, a recent RCUK 'Clarification on Costs' indicates that it *is* appropriate to pay such charges from the grant, so long as this is done within the lifetime of the project and that the repository does not also receive central or direct funding that might be considered to cover the data deposit in question.[59]

2.4.5 Recommendations on Data Repositories

Ideally data repositories should be 'trusted' or 'accredited' to provide long-term stewardship of research data. A number of initiatives exist to promote repository accreditation and standards.[60] However, most policy makers recognise that practice remains relatively immature and the infrastructure landscape is incomplete. Good practice would be to recommend the use of trusted or accredited repositories, while recognising that where these do not exist institutional or other infrastructure should be used.

The task of helping researchers find an appropriate repository has been simplified by Re3Data and DataBib (which have announced a merger, as Re3Data, which will be sustained in the long term by DataCite). Guidance documents supporting research data policies should mention the Re3Data service.[61] However, there is still a place for policies to help researchers make more specific recommendations about appropriate repositories. In this context, as practice develops it will be potentially important to track the BioSharing initiative, which proposes a pilot to link recommendations or requirements of policies with available standards and databases or data centres. There is the intention to extend this beyond the sphere of life sciences.[62]

As noted, certain research areas have a better-developed and more established data infrastructure than others. Where this is the case, the archiving systems and the communities and funders around them have generally developed more rigorous data practices and also more broadly accepted data formats and metadata standards. Funder policies in those disciplines often specify the requirement to deposit data in certain national data centres or international public data facilities (particularly genome sequencing).

It is particularly – but not exclusively – in the social sciences and the environmental sciences that national data archives have been established and therefore that national funder policies require deposit. In these fields the data policy or supporting information will often provide a list of and information about the appropriate national data centres.

As we have seen, where such resources do not exist, policies make the responsibility for ensuring long-term preservation and access to the data fall to the host research performing institution. In the UK this is most notably the case with the EPSRC policy. The MRC policy also observes that the 'data custodians', who have responsibility for full lifecycle care of the data including release and long-term preservation, will 'often' be 'those individuals or organisations that received MRC funding to create or collect the data.'[63]

[58] Dryad Digital Repository, Pricing Plan http://datadryad.org/pages/pricing
[59] See 'RCUK Responses' to questions relating to RDM funding http://blogs.rcuk.ac.uk/files/2013/07/RCUK-Responses-to-DCC-RDMF-Funder-Questions-.pdf and the associated blog post 'Supporting research data management costs through grant funding' http://blogs.rcuk.ac.uk/2013/07/09/supporting-research-data-management-costs-through-grant-funding/
[60] See the overview of a three-layer, 'basic, extended and formal' certification model at http://www.trusteddigitalrepository.eu/Site/Trusted%20Digital%20Repository.html
[61] See http://www.re3data.org/ and DataCite, re3data.org, and Databib Announce Collaboration http://www.re3data.org/2014/03/datacite-re3data-org-databib-collaboration/
[62] See http://biosharing.org/
[63] MRC Data Policy http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/datasharing/Policy/index.htm

2.5 Data Management Plans

One of the most important features of funder research data policies is the requirement that researchers in receipt of funding should manage data in accordance with good practice and to do so, therefore, by means of a data management plan (DMP).

The DMP requirement serves an important purpose for policy makers by obliging grant holding lead researchers, research groups and their host institutions to think seriously about their practice relating to data and to plan accordingly. It has been seen by policy makers and data experts in the UK and US as an important means of inculcating good practice. Additionally, requiring DMPs allows funders to avoid being too prescriptive about (for example) the data or metadata standards to be used or the process of selection of data – it is precisely these things which should be detailed in the plan.

Almost all funders with research data policies make some reference to data management plans, if only as a recommendation of good practice. Many funders require a data management plan or a short data management and sharing statement to be submitted for assessment with grant applications. Many funders, additionally or alternatively, require a data management plan to be completed in the early stages of the project. For some, it is expected that the project DMP should be reviewed as a deliverable; for others, e.g. the EPSRC, the expectation is that the host institution should take responsibility for quality assuring and overseeing the execution of the data management plan.

2.5.1 DMPs as Part of Project Proposals

Many funder policies require a statement relating to data to be submitted as part of a funding application. This is the case for the NSF, for most of the UK Research Councils and, more recently for the EC for those parts of the Horizon 2020 programme in the Open Data pilot.[64]

In Horizon 2020, for participating areas an outline data management plan should be included in the project proposal under the section on impact. There does not currently appear to be an indication of the expected length of this plan in the available documentation. The Horizon 2020 Data Management Guidelines provide an outline of the information to be provided and the questions to be addressed:

> Where relevant, applicants must provide a short, general outline of their policy for data management, including the following issues:
>
> - What types of data will the project generate/collect?
> - What standards will be used?
> - How will this data be exploited and/or shared/made accessible for verification and reuse? If data cannot be made available, explain why.
> - How will this data be curated and preserved?

Although funders are understandably concerned not to burden researchers at the application stage there has been a perceptible shift towards requiring increasing levels of detail and a shift from a data sharing statement to a more complete data management plan.

NSF requires a two page data management plan to be submitted with the research proposal. Notwithstanding some very minor inconsistencies, the various subject-oriented divisions of NSF require the data management plan to be structured around the same five questions (see Appendix 10).[65]

The ESRC has increased the level of detail requested, and requires a general data sharing statement to be completed in the online submission system *and* for a 'data management and sharing plan' to be prepared 'as an integral part of the application'.[66] The ESRC guidance contains nine areas in which information should be provided and data management issues addressed (see Appendix 10). Likewise, the MRC has recently

[64] The decision only to include certain areas of the Horizon 2020 programme (listed in the Guidelines, p.9) is interesting. In consultations, in which the authors participated, a very strong case was made that developing a data management plan is good practice in all research projects and particularly those with sensitive data that cannot be shared.
[65] For the NSF DMP requirement see http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#dmp
[66] ESRC Research Data Policy, p.5; http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

updated its DMP guidance and has provided a template which (uncompleted) runs to three pages.[67] The NERC, however, requires only a brief, outline DMP at the point of application; successful applicants are then required to complete a full DMP[68].

2.5.2 Assessment of DMPs Submitted with Project Proposals

Many research data policies rightly stress that proposals will be judged on their scientific merit. Nevertheless, the ability to demonstrate appropriate provision for research data management is part of the project delivery process, and so it is reasonable for it to be one of the criteria against which a lead researcher, research team or institution can be assessed. Policies which require the submission of a data management and sharing statement or a DMP with the grant proposal should be clear how this statement or DMP will be assessed as part of the proposal, but this is not always the case. One of the better practice examples is the data policy of the ESRC:

> [The ESRC] will seek an assessment of data management plans via its peer review and assessment processes. Although the application will first and foremost be assessed on grounds of its scientific merit, nonetheless, an assessment of the data management and sharing plan will be included in the general assessment of the application. Hence a poorly prepared data management and sharing plan may have a detrimental effect on an otherwise strong application.'[69]

It is notable as well that the ESRC provides guidance for peer reviewers.[70] Increasing awareness among peer review panels of the importance of data management planning is an important ongoing task of awareness-raising.

2.5.3 DMP in the Early Stages of the Project

As mentioned above, NERC provides an example of a two stage process in requiring a very simple data statement at application stage and a more detailed plan 'within three to six months of the start date of the grant'.[71] This approach is also taken by the EC in the Horizon 2020 Open Data pilot.

> Projects taking part in the Pilot on Open Research Data are required to provide a first version of the DMP as an early deliverable within the first six months of the project. … Since DMPs are expected to mature during the project, more developed versions of the plan can be included as additional deliverables at later stages. The purpose of the DMP is to support the data management life cycle for all data that will be collected, processed or generated by the project.[72]

2.5.4 DMP Templates and Tools

A template for the DMP to be produced as a project deliverable is included in the EC Horizon 2020 Guidelines. It covers the following areas:

1. Data set reference and name
2. Data set description
3. Standards and metadata
4. Data sharing
5. Archiving and preservation (including storage and backup)

Some funders in the UK provide DMP templates (e.g. NERC, MRC).[73] The UK DCC has developed an online tool (DMPonline) to help researchers create data management plans.[74] This tool includes *de facto* templates

[67] MRC Guidance on Data Management Plans http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/datasharing/DMPs/index.htm
[68] NERC guidance on data management planning http://www.nerc.ac.uk/research/sites/data/dmp/
[69] ESRC Research Data Policy, p.6; http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf
[70] ESRC 'Data management plan: guidance for peer reviewers': http://www.esrc.ac.uk/_images/Data-Management-Plan-Guidance-for-peer-reviewers_tcm8-15569.pdf
[71] NERC guidance on data management planning http://www.nerc.ac.uk/research/sites/data/dmp/
[72] Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, p. 3; http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
[73] NERC Data management planning http://www.nerc.ac.uk/research/sites/data/dmp/; MRC Data Management Plans http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/datasharing/DMPs/index.htm

designed by the DCC on the basis of funders' requirements. Some funders (e.g. MRC) recommend that researchers use DMPonline. In addition, the DCC has created generic checklists for information to be included in a DMP. This should be the starting point for any policy maker and support service considering what information is necessary and what guidance on data management planning should be communicated in a research data policies.[75] DMPTool provides similar DMP support for US researchers[76] (see Appendix 10 for the DCC's DMP Checklist and the EC DMP Template).

2.5.5 Timescales for Data Preservation

Some policies indicate an expectation relating to the length of time for which data should be retained. For data which will be deposited in an international or national data centres, the responsibility for any such decision will lie with the data centre. Where such infrastructure is not available, the existence of policy guidelines may help the research performing institution that will likely then have responsibility for stewardship of the data. Many policies surveyed do not address this issue and where they do there is significant variation (ranging from a minimum of three years in some cases, through to ten years from the last data access, and up to thirty years in others). For data underpinning published research findings, there is often a presumption that this should remain indefinitely available as part of the 'record of science'. There are a number of factors involved, and any organisation involved with data stewardship needs to have appropriate policies and procedures in place. In the light of this, some funders ask for data management plans to consider the likely necessary retention times and indicate what provision will be made.

2.6 Enabling Discovery and Reuse

Research data policy drivers, statements such as the OECD Guidelines and the principle of 'intelligent openness' advanced by the Royal Society *Science as an Open Enterprise* report make clear that ***the value of data lies in reuse***. This means that shared research data should be 'easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards.'[77] Research data policies increasingly articulate this as a general principle. It is useful also to provide some indications of how these attributes may be achieved.

2.6.1 Links from Published Articles

We have seen that research data policies increasingly make a distinction between data that directly underpin or substantiate published research findings and the broader data assets that are created by the research project. The former should be made available concurrently with the research publication and the information should be provided on how the data can be accessed. Ideally a link should be provided to a web-based location and a permanent identifier. The RCUK *Common Principles* state that 'Published results should always include information on how to access the supporting data.' By way of implementation, since 1 April 2013 the *RCUK Policy on Open Access and Supporting Guidance* requires RCUK grant-holders to include a statement on data availability in published academic papers. Where necessary, this statement should be used to adumbrate the reasons why the supporting data cannot not be made accessible.[78]

2.6.2 Standards, Formats and Metadata

It is a widely articulated principle in research data policies that to enable research data to be discoverable and effectively reused by others, sufficient metadata should be recorded and made openly available to enable other researchers to understand the research and reuse potential of the data.

Funders' data policies rarely provide specific recommendations on the specific standards to be used (whether referring to data formats or metadata standards) and generally restrict themselves to a statement that the metadata should be appropriate, rich, high-quality, and comply with community norms and good

[74] https://dmponline.dcc.ac.uk/
[75] http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP_Checklist_2013.pdf
[76] DMPTool is available at https://dmp.cdlib.org
[77] G8 Science Ministers Statement, 13 June 2013 https://www.gov.uk/government/news/g8-science-ministers-statement
[78] RCUK Policy on Open Access and Supporting Guidance, pp.3-4; http://www.rcuk.ac.uk/RCUK-prod/assets/documents/documents/RCUKOpenAccessPolicy.pdf

practice by using a recognised standard where such exist. The RCUK *Common Principles* state: 'Data sharing should be done in accordance with relevant standards and community best practice.' To a large extent the general absence of specific recommendations is understandable – in many research areas there remains considerable debate and volatility around metadata standards. Some subject specific policies recommend particular standards and some of these are considered in Section 3. Some examples are the use of DDI (Data Documentation Initiative)[79] in the social sciences, the use of CIF (the Crystallographic Information Framework)[80] in crystallography, and the use of the NetCDF and HDF5 data formats in atmospheric and meteorological science.

2.6.3 Identifiers

For data to be reusable it should be uniquely and persistently identified. The Horizon 2020 *Guidance*, in common with many policies, recommends identifiers that are 'persistent, non-proprietary, open and interoperable'.[81] The recently formulated *Joint Declaration of Data Citation Principles*, the collaborative work of a number of organisations and experts,[82] state: 'A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community'.[83] DOIs (Digital Object Identifiers), administered by the DataCite organisation, are emerging as a widely-used standard for data and many policies either recommend the use of DataCite DOIs or cite it as an example.[84] There is some concern that DOIs may not be optimal identifiers for activities involving machine-aided analysis at scale. Nevertheless, it is fair to say that DataCite DOIs are becoming established as a widely used means to attach a permanent identifier to data.

2.6.4 Licence

It is widely recognised that an important precondition to reuse of research data is clarity and openness around licensing. Accordingly, research data policies increasingly require that researchers should make data available using the least restrictive licence possible.

The existence of the *sui generis* European Union database protection makes it particularly important to include a waiver, licence or public domain dedication in order to remove any uncertainty over possible restrictions on reuse. Version 4 of the Creative Commons (CC) waiver and licence suite has been specifically adapted in order to ensure that *sui generis* database rights fall 'squarely within the scope of the license unless explicitly excluded by the licensor'.[85] It also allows database providers to use CC licenses to explicitly license those rights.'[86] The Horizon 2020 Guidance, therefore, recommends the use of the new Creative Common licences (v.4) CC-BY or CC0.[87]

A Knowledge Exchange Report examines in detail the legal status of research data in Denmark, and provides useful guidance, particularly from the point of view of the reuser of data.[88] It is important to stress that a lot of restrictions and 'legal interoperability' issues can be overcome through the application of an open licence or waiver. Note also that policies which envisage the greatest reuse possible of assets created through

[79] See http://www.ddialliance.org/
[80] See http://www.iucr.org/resources/cif
[81] Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, p.8: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
[82] On the background to this declaration see: http://codata.org/blog/2013/11/25/data-citation-synthesis-group-draft-declaration-of-data-citation-principles/
[83] Joint Declaration of Data Citation Principles: https://www.force11.org/datacitation
[84] See EPSRC Expectation v, http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx
[85] See http://creativecommons.org/version4
[86] See in particular the advice on using CC licenses for databases http://wiki.creativecommons.org/Data#Can_databases_be_released_under_CC_licenses.3F
[87] Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, p.11: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
[88] The legal status of research data in Denmark, Annex 2 to the Knowledge Exchange report 'The legal status of research data in the Knowledge Exchange partner countries': http://www.knowledge-exchange.info/Default.aspx?ID=461

public funding argue that the data should be reusable free of charge by *any user* and therefore argue against the 'non-commercial' restriction included in some licences.[89]

2.7 Recognition and Reward

Policies generally acknowledge that moving to an open data regime requires for many research areas a shift in practice and culture. For this reason policies often include, alongside the requirement that publicly funded data should be made open, some statement of the need for appropriate recognition and reward for those researchers that make data open. This is neatly expressed as a general principle in the G8 Science Ministers' Statement: 'To ensure successful adoption by scientific communities, open scientific research data principles will need to be underpinned by an appropriate policy environment, including recognition of researchers fulfilling these principles, and appropriate digital infrastructure.'[90] There are two notable policy implications of this principle: 1) to require acknowledgement of data reuse and the citation of data where it underpins further research findings, and 2) to allow periods of privileged access for the data creators.

2.7.1 Policy Requirements on the (Re)Users of Data: Referencing and Citation

Credit and recognition are important aspects of research culture and the need to accord due credit is underlined in many research data policies. As the RCUK *Common Principles* puts it:

> In order to recognise the intellectual contributions of researchers who generate, preserve and share key research datasets, all users of research data should acknowledge the sources of their data and abide by the terms and conditions under which they are accessed.[91]

The principle means of according credit in scholarly communication is through citation. Recognition of this has led to calls to propagate the practice of data citation. The objective is eloquently put in the *Joint Declaration of Data Citation Principles*:

> Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse.[92]

Such a principle is explicitly reflected in the ESRC Data Policy: 'We emphasise the responsibilities that data sharing places upon those who plan to reuse existing data for research purposes. Where such data sharing leads to publication of related research findings in any format, full and appropriate acknowledgement, via citation, should be made of data sources.'[93] In support of this, ESRC makes a guide to data citation available from the policy information pages.[94]

Like some other data policies, that of the MRC stresses that 'data sharers should receive full and appropriate recognition by funders, their academic institutions and new users for promoting secondary research'. However, perhaps reflecting the distinct culture of medical research, the MRC adds the distinctive recommendation that research that results from data-sharing 'is often most fruitful when it is a collaboration between the new user and the original data creators or curators, with the responsibilities and rights of all parties agreed at the outset.'[95]

---

[89] E.g. see the G8 Open Data Charter https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex, the Open Data License Examples provided by the US Open Data Project http://project-open-data.github.io/license-examples/ and Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, p.11: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[90] G8 Science Ministers' Statement https://www.gov.uk/government/publications/g8-science-ministers-statement-london-12-june-2013

[91] RCUK Common Principles http://www.rcuk.ac.uk/research/datapolicy/

[92] *Joint Declaration of Data Citation Principles* (2014) https://www.force11.org/datacitation

[93] ESRC Data Policy p.4; http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

[94] Data Citation: what you need to know http://www.esrc.ac.uk/funding-and-guidance/grant-holders/data-citation.aspx

[95] MRC Research Data Policy http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/datasharing/Policy/index.htm

2.7.2 Periods of Privileged Access

A number of funder research data policies conceive that it may be appropriate to allow a limited period of privileged use for data creators. The reasoning behind this provision is clearly expressed in the RCUK Common Principles:

> To ensure that research teams get appropriate recognition for the effort involved in collecting and analysing data, those who undertake Research Council funded work may be entitled to a limited period of privileged use of the data they have collected to enable them to publish the results of their research. The length of this period varies by research discipline and, where appropriate, is discussed further in the published policies of individual Research Councils.[96]

Similarly, NERC, in order 'to protect the research process' allow a period of privileged access for the data creators: 'this period will normally be a maximum of two years from the end of data collection'. EPSRC allows that the length of privileged access that is appropriate will vary between scientific disciplines.[97] However, EPSRC expects that the metadata describing the data should normally be made public within 12 months of the data being generated.[98] NSF policies encourage timely release of data and suggest that 'a reasonable standard of timeliness is to make the data accessible immediately after publication, where submission for publication is also expected to be timely.'[99] The AHRC also requires prompt action: archaeology projects need to offer data to the Archaeology Data Service for reuse within 3 months of the project end[100]. Conversely, NSF Ocean Studies Division allows a period of privileged access for two years subsequent to data collection.[101]

The Horizon 2020 Guidance follows the distinction between those data directly underpinning published research findings and other datasets produced by projects. For the former, the data should be made available as soon as possible. For the latter, the Guidance makes no recommendation, but says that data release should follow a timescale laid out in the Data Management Plan.[102]

Some general policy statements contain a presumption that it will be in the humanities and social sciences that a greater period of privileged access will be appropriate because of the longer timescales involved in research and the cultural norms in those research communities. It is not clear to what extent this is factually based and it would seem to be related to the generally longer embargo times which have been allowed in some areas of the humanities and social sciences for scholarly publications before release to Open Access. A counter-example, along with the case of AHRC policy for archaeology, above, is the ESRC data policy which requires submission of data to the ESRC data service within three months of the completion of the project.

It should be noted that while many policies hold as reasonable that a period of privileged access should be accorded to the researcher or team creating the data, the genomics community has had, since 1991, a succession of policies and agreements that favoured *pre-publication data release* by requiring deposit to a public nucleotide sequence database (GenBank, EMBL or DDBJ) within a short time of the sequence data being created. In 2003, the Fort Lauderdale agreement reduced this period to 24 hours or 1 week depending on criteria relating to the sequence and the project approach.[103]

2.8 Reporting, Compliance Monitoring and Sanctions

As stated in the research data policies surveyed, the regimes for reporting on data management and sharing vary considerably, during and even after the lifetime of a funded project.

Possibly the most stringent existing policy in this regard is that of the ESRC. Like many funders ESRC requires reporting against the DMP as part of the research project's annual report. Notably, however, the ESRC

[96] RCUK Common Principles
[97] EPSRC Research Data Principles http://www.epsrc.ac.uk/about/standards/researchdata/Pages/principles.aspx
[98] EPSRC Research Data Expectations, v; http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx
[99] E.g. NSF Engineering Division Data Policy http://nsf.gov/eng/general/ENG_DMP_Policy.pdf
[100] http://archaeologydataservice.ac.uk/advice/adviceForAHRCGrantApplicants
[101] NSF Ocean Science Division, Data Management Policy, http://www.nsf.gov/pubs/2011/nsf11060/nsf11060.pdf
[102] Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, 3; http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
[103] Summary of genomics data policies http://www.genome.gov/page.cfm?pageID=10506537

requires 'that the data must be made available for preparation for re-use and/or archiving with the ESRC data service providers within three months of the end of the award' and will 'withhold the final payment of an award if data have not been offered for archiving to the required standard within three months of the end of the award'.[104]

The standard NSF approach, like many funders, is to require reporting against the plan as part of the annual and final reporting process. The final report should describe the implementation of the DMP including any changes from the original DMP and contain the following information:

- The data produced during the award period
- The data that will be retained after the award expires
- How the data will be disseminated and verification that it will be available for sharing
- The format (including community standards) that will be used to make the data – including any metadata – available to others
- Where the data generated by the project has been deposited/is being stored for long-term public access[105]

The NSF Ocean Sciences policy, which conditionally allows a period of two years' privileged access, specifies that 'In cases where the Final Report is due before the required date of sample or data submission, the PI [Principal Investigator] must report submission of metadata and plans for final submission. The PI should notify the cognizant Program Officer by e-mail after final data and/or sample submission has occurred, even if this is after the expiration date of the award.'[106]

Beyond the ESRC's statement that it will withhold the final award payment pending data deposit, there is little evidence of precisely formulated sanctions in the policies surveyed. The NSF notes that in the future applicants will be required to report on data deposit relating to any previous grant. The BBSRC assesses the adherence to those activities outlined in the DMP as part of project final reporting assessment, and that performance in data sharing activities may be considered as part of the selection process for future funding proposals from that applicant. The EPSRC, meanwhile, has said that it will perform light-touch monitoring of RDM capacity in research performing institutions as part of its general oversight processes. It is not explicitly stated what sanctions may be applied, and it is likely hoped that the implication that a future proposal would be weakened will encourage compliance.

Section 3: Subject Area Considerations

The third part of this memo provides an overview with examples of variations in research data policies in different subject areas. There are some limitations which need to be mentioned here.

Some variations between the policies of funders in different research areas appear to be contingent rather than intrinsic to data sharing in the given research area. For example, most of the UK Research Councils require funded projects to produce a DMP of some sort that will be submitted alongside applications for funding and assessed in some way. In contrast, the EPSRC expects a DMP to be created and used but does not require review of it at bidding stage. This difference does not appear to reflect any consideration of the nature of research in EPSRC-funded fields: it is simply a procedural decision about whether it is the role of the funder and peer-review process to assess the DMP.

Many funder policies and most institutional data policies are deliberately high-level and general. The policy aims to lay down general principles and expectations and acknowledges that the application of those principles may vary according to the research area or, more precisely, in relation to the characteristics of the data in question. As noted above, this makes the role of the DMP particularly important. At the policy level it is generally simply stated that a DMP should be produced to provide details about what data will be created, how it will be stored, described, shared and preserved. The specificities, which may relate to the

[104] ESRC Research Data Policy, p.3; http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf
[105] NSF Directorate of Biological Sciences Research Data Policy http://www.nsf.gov/bio/pubs/BIODMP061511.pdf
[106] NSF Ocean Sciences Division p.5; http://www.nsf.gov/pubs/2011/nsf11060/nsf11060.pdf

data in question (volume or security of storage, metadata standards, potential or not for sharing), are then to be addressed in the DMP.

The policy of a funder in a given research area *may* highlight some issues which are characteristic for that discipline. However, it is often only at the level of accompanying or more detailed guidance and support that a lot of such issues are really addressed, rather than in the research data policy itself.

We provide a number of examples of variations in data policies between research areas. Some of these examples are taken from supporting documents which add detail to the general statement in the data policy. This was deemed to be helpful for the purpose of this survey as providing a fuller picture. It should be noted, however, that policies themselves will rarely go into a great level of detail.

With policy development and data practice still hardly mature, we would hesitate to draw particularly strong conclusions from this range of examples. Nevertheless, the most significant variations between the policy concerns in given subject areas relate to:

Legal and ethical requirements specific to the type of research being conducted. For example, privacy issues are particularly important in the health and social sciences.

Existence of more established data infrastructure and practice. For research disciplines where international or national data centres have been established, policies more often provide (in appendices or guidance) lists of appropriate data centres or databases and allude to any existing technical standards or practices widely used in that discipline.

Accepted technical approaches in a specific research area. In some areas of the social sciences, life sciences and natural or technical sciences, specific data format or metadata standards have emerged and become common practice in that community. Where this is the case, these standards may be mentioned directly or indirectly in policy documents and recommendations.

Some research initiatives or programmes have had specific requirements and data management policies and sometimes data centres or offices to provide bespoke support for data management. An example in the UK was the interdisciplinary *Rural Economy and Land Use* programme.[107]

3.1 Humanities

Research datasets in the humanities and arts may often be small, highly qualitative, unfunded, consist of a mixture of digital and analogue materials, often gathered rather than created, and held by lone scholars with little motivation to share.[108] These characteristics are in contrast with the longstanding collaborative research practices found in many of the sciences based on the generation and systematic sharing of large datasets.

Archaeology, however, stands as the exception to much of the above: as a form of research within the humanities but with many characteristics of scientific practice such as the generation of large and complex numerical datasets, archaeology benefits from more structured support for research data than many other humanities disciplines. The UK's Arts and Humanities Research Council (AHRC) contributes to the running of the Archaeology Data Service (ADS). As noted in section 2 above, it is a condition of award for AHDS-funded archaeology projects that 'any significant digital resources created as an output of [the] project' are offered for deposit with the ADS, 'accompanied by appropriate documentation … within three months of the end of the project.'[109]

Other than in the case of archaeology, however, there are some differences between the approach to data policy taken by funders in arts and humanities subject areas in comparison with the funders of the sciences. As an example, the AHRC has not published a discrete general research data policy to date, perhaps

[107] Rural Economy and Land Use Programme: http://relu.data-archive.ac.uk/data-sharing/planning

[108] As described by the Sudamih project (2010) at University of Oxford http://sudamih.oucs.ox.ac.uk/docs/Sudamih_Interface2010.pdf and at the 10th Research Data Management Forum, 3-4 September 2013, notes available at http://www.dcc.ac.uk/events/research-data-management-forum-rdmf/rdmf10-research-data-management-arts-and-humanities.

[109] ADS (2008). Advice for AHRC Grant Applicants, v2. Available at http://archaeologydataservice.ac.uk/advice/adviceForAHRCGrantApplicants.

reflecting the lower level of certainty about the nature and proper ambitions of arts and humanities research data. Alternatively it is possible that the funder wishes to avoid use of the term 'research data' in its general policy framework as many researchers in the arts and humanities may well assume this refers to the generation of numerical databases and therefore is not relevant to their work[110].

As a member of RCUK, the AHRC has discharged its obligation to perpetuate the RCUK *Common Principles* through the AHRC Research Funding Guide[111] and supporting online guidance[112]. The *Guide* outlines all requirements of the funding application process including the expectations of preservation and sustainability of all digital outputs beyond project end and includes research data amongst these. [113]

The AHRC recognises the range of possible forms this data may take: 'electronic data, including sound or images' is mentioned as one of a possible range of outputs subject to the guidance. There is little demarcation in the *Guide* between research data as material underpinning research findings, and other outputs generated by funded research activity; this is in contrast with other UK research funders.

Other than in the case of archaeology, the AHRC does not tightly specify a timescale wherein research data must be made available, and requires a minimum of three years after the end of project funding for both preservation[114] and sustainability[115] of digital outputs (including research data, as described above). This approach appears to make significantly lower demands on the researcher to publish promptly and sustain for the longer periods expected by science funders, possibly reflecting lower expectations of the urgency of release of arts and humanities research data and lower ambitions for its reuse in some research areas.

The AHRC no longer funds its former network of subject-specific data centres but instead implies the importance of the institution amongst the possible appropriate stewards for research data[116].

In the US, the National Endowment for the Humanities (NEH) takes a similar approach, although with tighter specification of some areas and a specific published statement on data management planning[117]: applications for NEH funding now (i.e. since 2013) require a DMP detailing the type of research data which is anticipated to arise from the research, and the plans for its management and dissemination. Guidance provides a useful and specific list of likely data types, but again the timescale for publication of data is not specified. The NEH guidance does not refer explicitly to open access for research data; in contrast, the AHRC is clear that open access to outputs including research data (and open source for any software developed) constitutes the default position and exceptions must be justified. The NEH, however, specifies a more detailed monitoring process which includes the requirement for a white paper at project end which must include 'a description of data management challenges' encountered by the project.[118]

3.2 Social Sciences

3.2.1 Social Sciences: data sharing infrastructure

The distinctive features of social science research data policies generally relate to the good management of data involving human subjects and related ethical issues. Social sciences have a longer tradition of data infrastructure. Social science data archives like the UK Data Archive (UKDA) and the US-based Inter University Consortium for Political and Social Research (ICPSR) have been proactive in providing guides to researchers on meeting funder requirements, data management planning and management of research

[110] See the findings of the Sudamih project (2010), described at http://sudamih.oucs.ox.ac.uk/
[111] AHRC (2014). Research Funding Guide, v2.6, Jan 2014. Available at http://www.ahrc.ac.uk/SiteCollectionDocuments/Research-Funding-Guide.pdf.
[112] AHRC: 'Technical Plan'. Webpage, available at http://www.ahrc.ac.uk/Funding-Opportunities/Research-funding/RFG/Application-guidance/Pages/Technical-Plan.aspx
[113] See AHRC (2014).
[114] Here defined as 'the storage of a project's digital outputs for a period beyond the end of funding'.
[115] Here defined as 'your plans for ensuring that digital outputs remain publicly accessible and usable for a period beyond the end of funding.'
[116] The AHRC's Arts and Humanities Data Service closed in 2008, but the Archaeology Data Service and Visual Arts Data Service both remain, with alternative funding sources including, in the case of the ADS, a charging model for research data deposit.
[117] NEH Office of Digital Humanities (2013). Data Management Plans for Proposals and Awards. Available at http://www.neh.gov/files/grants/data_management_plans_2013.pdf
[118] See NEH (2013).

data. A more developed data sharing practice and infrastructure means that the social sciences and economic sciences have 'an established international standard for the description of data', namely DDI, the Data Documentation Initiative.[119]

In the UK, the ESRC data service providers, particularly the UK Data Service, are funded both to 'ensure long-term access to data which has been placed in their care', and to provide support services to researchers.[120]

3.2.2 Social Sciences: issues of sensitive data

The support provided by social science data services relates in particular to guidance 'on issues related to confidentiality, security and ethics in data sharing, together with offering guidance on providing additional information to research participants on the purposes and benefits of data archiving.'[121] The online guidance provided by the UK Data Service covers legal issues, consent for data sharing, anonymisation, access control, research ethics and guidelines for Research Ethics Committees.[122]

Notably, UKDS support aims to help maximise the potential for research data involving information about human subjects to be archived and shared. The ESRC position is that:

> sensitive and confidential data can be shared ethically provided researchers pay attention right from the planning stages of research to the following aspects:
>
> - when gaining informed consent, include consent for data sharing
> - where needed, protect participants' identities by anonymising data
> - address access restrictions to data before commencing research in the data management and sharing plan.[123]

An important feature of good practice, which appears in supporting documentation but not in the ESRC research data policy, is to avoid gathering personal information which 'may not be strictly necessary for the research': this will make the process of anonymisation significantly easier. By allusion to the Principle of Minimum Necessary, the same point is made in guidance linked to the US NIH policy summarising the US Privacy Rule.[124]

The approach of the ESRC with regard to consent is notable: most data policies surveyed simply stress the need to avoid revealing personal information and suggest anonymisation as the only recourse. The US OSTP memo mentions the importance of very careful anonymisation because of the 'mosaic effect', whereby dataset aggregation may make it possible to identify participants. The possible recourse to informed consent for data sharing does not seem to be mentioned in other policies, though, for the MRC, consent is a precondition for the reuse even of anonymised data from health studies or patient records.[125]

3.3 Health Sciences

3.3.1 Health Sciences: Protection of Personal Information

As with the social sciences, the particular concern of data policies in the health and medical sciences relates to the protection of personal information. MRC is clear that 'appropriate regulatory permissions' must be in place before data is shared. Researchers must undertake necessary steps in order to ensure that 'opportunities for new uses are maximised'. Although the MRC data policy cites the OECD principles, it is clear that much health data cannot be 'Open' in the accepted sense but can be shared - ethically, with consent - with other research groups.

---

[119] http://www.ddialliance.org
[120] ESRC Research Data Policy, p.6; http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf
[121] ESRC Research Data Policy, p.6; http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf
[122] http://www.ukdataservice.ac.uk/manage-data/legal-ethical.aspx
[123] ESRC Research Data Policy, p.7; http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf
[124] United States Department of Health and Human Services, Office for Civil Rights, *Summary of the HIPAA Privacy Rule*, p.10; http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/privacysummary.pdf
[125] MRC, *Personal Information in Medical Research* (2000, updated 2003), p.33; http://www.mrc.ac.uk/consumption/idcplg?IdcService=GET_FILE&dID=6233&dDocName=MRC002452&allowInterrupt=1

Detailed MRC guidance covering many aspects of using personal information including research design, collection and management of data and data sharing is provided, linked to the data policy, in the document *Personal Information in Medical Research*.[126] It is good practice for such information to be linked to the data policy. The guidance as such is of too great detail to be presented as part of a good practice policy, but we reproduce here the two most salient points:

> 1. Personal information of any sort which is provided for health care, or obtained in medical research, must be regarded as confidential. Wherever possible people should know how information about them is used, and have a say in how it may be used. Research should therefore be designed to allow scope for consent, and normally researchers must ensure they have each person's explicit consent to obtain, hold, and use personal information. In most clinical research this is practicable.
>
> 3. All personal information must be coded or anonymised as far as is possible and consistent with the needs of the study, and as early as possible in the data processing. Only personal identifiers that are essential should be held.

3.3.2 Health Sciences: Anonymisation and Coding

According to the MRC guidance, there are three generally recognised levels of anonymisation: 'coded information' (where identifying information is 'concealed by a code' with the code book held by the research team); 'linked anonymised data' (where the code book is held by a third party and the data is anonymous to the research team); and unlinked anonymised data in which, in principle, 'the link to individuals [via the code] has been irreversibly broken'. The MRC guidance observes:

> 'With both linked and unlinked anonymised data, there is sometimes potential to deduce individuals' identities through combinations of information, either by the people handling research data, or by those who see the published results. The most important potential identifiers are:
>
> - rare disease or treatment, especially if an easily noticed illness/disability is involved;
> - partial post-code, or partial address;
> - place of treatment or health
> - professional responsible for care;
> - rare occupation or place of work;
> - combinations of birth date, ethnicity,
> - place of birth, and date of death.'

Judgements on the potential risk of identification through aggregation or 'the mosaic effect' must be made on a case by case basis.[127]

3.3.3 Health Sciences: Storage and Reuse of Research Data

Research in the health sciences has well developed rules about the security attached to data storage. These are detailed at a high level in MRC's supporting documentation. There is an established ISO Standard for Information Security processes (ISO-27001) which is often required of facilities holding sensitive medical and personal data (particularly in the context of long-term human cohort studies).[128] This is beyond the scope of the current survey. However, it is worth considering the MRC guidance that relates to data selection and retention schedules:

[126] MRC, *Personal Information in Medical Research* (2000, updated 2003); http://www.mrc.ac.uk/consumption/idcplg?IdcService=GET_FILE&dID=6233&dDocName=MRC002452&allowInterrupt=1
[127] MRC, *Personal Information in Medical Research*, Section 5, p.27-30.
[128] See a useful overview of Information Security from the perspective of a university secure data centre that was part of the Jisc Managing Research Data programme: http://blogs.ucl.ac.uk/dmp-ss/2013/03/26/information-security-explained/

> Research records need to be preserved for the longer-term for a number of reasons - other than for historical posterity. Firstly, records may be needed later on for scientific validation of research, or for future research and audit.[129]

The guidance suggests a minimum retention period of 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', though for studies 'which were of historical importance, where novel clinical interventions were first used, those which have proved controversial, or where research is ongoing' the data may need to be retained longer. For the MRC, it is generally the research unit which created the data, or their host institution, that is considered to have responsibility as the long-term data custodian.

3.3.4 Health Sciences: Reuse of Data

The MRC *Guidelines* on personal information make clear that consent should be obtained from patients and subjects at the time of the first study for the use *and* the reuse of data. This distinction should be made clear to the subject and consent for both obtained. Furthermore, 'researchers obtaining information with consent should, wherever possible, anticipate likely needs to archive the data, and to share data sets with other researchers, and make this clear to the people involved.'

When sharing data 'normally, only anonymised data should be passed on'. Identifiable data may be passed on to another group under controlled circumstances and with Research Ethics Committee approval. Identifiable data can only be passed onto groups with appropriate controls and in countries with appropriate data protection legislation (importantly, according to the guide this includes the European Economic Area, but not the USA). Conversely, Ethical Committee approval 'is not needed for re-use of unidentifiable data obtained directly from study participants, or for re-analysis, by any research group, of unidentifiable data from previous research'.[130]

3.3.5 Health Sciences: Cohort Management

Health and social science research conducted longitudinally on cohorts have specific requirements in relation to data management. Many of these are to do with ethics and privacy. Issues around data sharing in the health (and social) sciences may also relate to maintaining the engagement of the cohort. Considerations behind the management of longitudinal cohort studies drive the MRC's concern that 'Ongoing research contributing to the completion of datasets must not be compromised by premature or opportunistic sharing and analysis. Sharing should always take account of enhancing the long-term value of the data.'[131]

3.3.6 Health Sciences: NIH Data Sharing Guide on 'De-identification'

The NIH *Data Sharing Workbook* also provides guidance for researchers in sharing data. The guide stresses the imperative of sharing data that constitute 'unique resources', while underlining the importance of protecting information about human subjects. The guide also provides useful examples of data archives which may be used by NIH-funded researchers as well as secure 'data enclaves' for sensitive data.

The Workbook provides information from the US Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule on 'de-identification' or anonymisation of data.[132]

The workbook notes that 'Deductive disclosure of individual subjects becomes more likely when there are unusual characteristics or the joint occurrence of several unusual variables. Samples drawn from small geographic areas, rare populations, and linked datasets can present particular challenges to the protection of subjects' identities.' For this reasons, additional precautions made be necessary in some circumstances. These include restricting access to the data set, or sometimes just to those variables with 'low prevalence rates' which might allow deductive disclosure.[133]

---

[129] MRC, *Personal Information in Medical Research*, Section 7 Storage and re-use of research data, pp. 33-4.
[130] MRC, *Personal Information in Medical Research*, Section 7 Storage and re-use of research data, pp. 33-4.
[131] MRC Research Data Policy, http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/datasharing/Policy/index.htm
[132] NIH Data Sharing Workbook, p.2; http://grants.nih.gov/grants/policy/data_sharing/data_sharing_workbook.pdf
[133] NIH Data Sharing Workbook, p.2-3; http://grants.nih.gov/grants/policy/data_sharing/data_sharing_workbook.pdf

3.4 Natural (and Environmental) Sciences

For those natural sciences and environmental sciences involving earth observation of various kinds, where the data produced are typically unrepeatable observations with significant long-term reuse value (e.g. relating to climate change, natural disaster, biodiversity, etc) then there is a very strong presumption in favour of retention, curation and sharing. Accordingly, data sharing practices and the necessary data infrastructure are typically somewhat more developed in these fields.

In the UK the seven NERC Data Centres cover the following domains: polar and cryosphere science, atmospheric science, marine science, terrestrial and freshwater science, earth sciences, earth observation, solar and space physics.[134] As noted above the NERC Data Policy encourages data to be submitted to the appropriate data centre 'as soon after the end of data collection as possible' although an embargo period is allowed. NERC-funded researchers are encouraged to liaise closely with the data centre staff about their data management plan, about data formats and about appropriate metadata. The NERC Data Policy Guidance makes some general statements about metadata relating to observations and relating to model output. Beyond this, however, the Policy and Policy Guidance declares specific data formats and metadata standards to be beyond scope: detailed subject specific guidance on all aspects of metadata creation is available from the data centres.[135] Although the data policy of the NSF Atmospheric and GeoSpace Division, mentions two widely used data formats for 3D meteorological data (HDF5[136] and NetCDF[137]) it does so merely as examples, without providing any further guidance on data formats or on metadata.[138]

3.4.1 US NSF Ocean Sciences

In the US NSF the Division of Ocean Science has a significantly more detailed and developed data policy than other NSF divisions. The Ocean Science policy goes beyond general principles into guidance and rules for implementation which relate to the data infrastructure provided. This data policy existed before 2011 and was updated following the general NSF requirements for data management plans with project proposals.

The US NSF Ocean Sciences Data policy requires submission of data, physical samples and other supporting materials within two years of collection. The policy document provides a list of recommended data centres. In cases where an appropriate data centre is not available for the data in question, then the responsibility falls on the PI (and presumably her institution) to curate the data. For research conducted on board 'NSF-supported oceanographic research vessels' there is a R2R 'Rolling Deck to Repository' programme which takes care of appropriate data submission from the voyage. Additionally, oceanographic researchers involved in sea-going research voyages must comply with 'all legal requirements for submission of data and research results that are imposed by foreign governments as a condition of that government's granting research clearances'.[139]

Like some other funders, NSF Ocean Sciences allows a period of privileged access, set at two years from data collection. For data not subject to the R2R process, therefore, a period of two years is allowed before data deposit. Ocean Sciences require the appropriate programme officer to be informed by e-mail when data deposit has been completed, even if this is after the project has formally completed. On the other hand, for data submitted by ship operators through the R2R 'Rolling Deck to Repository' programme the default is that 'the underway data from NSF-funded cruises will be placed in the public domain for unrestricted open access after 60 days of the cruise end date, unless a request for restricted access is submitted through R2R. If such a request is submitted, a proprietary hold period of up to two (2) years will be maintained by R2R with approval of the cognizant NSF Program Officer.'[140]

Physical samples are to be submitted to the appropriate repository no later than 2 years post cruise; however, the metadata describing these samples 'must be submitted to the appropriate National Data

[134] See http://www.nerc.ac.uk/research/sites/data/
[135] NERC Data Policy Guidance p.6-7; http://www.nerc.ac.uk/research/sites/data/policy/datapolicy-guidance.pdf
[136] http://en.wikipedia.org/wiki/Hierarchical_Data_Format
[137] http://en.wikipedia.org/wiki/NetCDF
[138] NSF Atmospheric and GeoSpace Division http://www.nsf.gov/geo/geo-data-policies/ags/ags_data_mgt_guidance-v3.docx
[139] NSF Ocean Sciences Data Policy pp.3-4; http://www.nsf.gov/pubs/2011/nsf11060/nsf11060.pdf
[140] NSF Ocean Sciences Data Policy pp.5; http://www.nsf.gov/pubs/2011/nsf11060/nsf11060.pdf

Center and to IEDA, the Marine Geology and Geophysics Data Management Group' within 60 days of the end of the research cruise.[141]

Particular sub-programmes in NSF Ocean Sciences have offices providing data support services. These are the Biological and Chemical Oceanography Data Management Office[142] and the Marine Geology and Geophysics Data Management Office run by IEDA (Integrated Earth Data Applications).[143] Taking the role of a data management support service, IEDA provides an earth sciences-specific DMP Tool[144] and a Data Compliance Reporting Tool.[145]

3.4.2 Atmospheric Sciences: Data Release Schedules

The data policy of the NSF Atmospheric and Geospace Division indicates that there may be different schedules for release of different data types (e.g. raw data vs processed data, observations vs models, etc.)[146] We would be interested to understand the reasoning behind this distinction and precisely what timescales are envisaged, as these are not given. Generally, processed data would underpin a research publication and therefore be concurrently released. Depending on the research area 'raw' and particularly observational data may be unrepeatable and therefore have a strong case for long-term preservation. Data produced by models is in principle repeatable, but it may be very expensive to do so (and impossible unless the code and input data is available).

3.5 Technical Sciences

The NSF data policies in the technical science divisions for the most part reiterate the general NSF policy, without manifesting any major subject area-specific variations. In the UK, the EPSRC has taken a distinct approach and its data policy places greater emphasis on the responsibilities of the research performing organisation to take responsibility to support data management planning, the implementation of DMPs and potentially, if a specialist data service does not exist, the long-term stewardship of the data assets created. However, it would seem hard to maintain that this different in approach has intrinsic roots in subject area differences.

3.5.1 Technical Sciences and Commercialisation

Policies relating to engineering, materials science and related disciplines tend to place greater emphasis on the limitations to data availability which may be required by partnership arrangements or reasons of prospective commercialisation. For example, the NSF *Engineering Division Research Data Policy* lists specific sub-programmes where this may be the case, including activities in nanoscale science, engineering, knowledge transfer and collaboration programmes involving industry, small business and universities. Any data management and sharing issues related to prospective commercialisation or confidentiality agreements, 'including conditions that might prevent or delay the sharing of data', should be identified in the DMP.[147]

It is good practice to request that such issues be considered and presented in the DMP both at application stage and once the funded project is underway. A project with commercial ties will very likely produce data underpinning publications and datasets of potential reuse value to which commercial restrictions do not and need not apply.

Unusually for an NSF policy the Engineering Division specifies a minimum period of data retention. Where data supports a patent it must be retained for the life of that patent. In other circumstances, the data should be retained for a minimum of three years after the conclusion of the award, or after public release,

[141] NSF Ocean Sciences Data Policy pp.7; http://www.nsf.gov/pubs/2011/nsf11060/nsf11060.pdf
[142] http://bco-dmo.org/
[143] http://www.iedadata.org
[144] http://www.iedadata.org/compliance/plan, as distinct from that hosted by the California Digital Library, available at https://dmp.cdlib.org/
[145] http://www.iedadata.org/compliance/report.
[146] NSF Atmospheric and Geospace Division Data Policy http://www.nsf.gov/geo/geo-data-policies/ags/ags_data_mgt_form.pdf
[147] NSF Engineering Division Research Data Policy http://nsf.gov/eng/general/ENG_DMP_Policy.pdf

whichever is later.[148] The more general condition, not related to patents, appears to be another example of a contingent rather than intrinsic difference in policy as there is no apparent reason why a funder should apply a policy for data retention to engineering rather than other research areas. And by contrast, the EPSRC policy specifies a retention period of ten years since last reuse.

[148] NSF Engineering Divison Data policy, http://nsf.gov/eng/general/ENG_DMP_Policy.pdf