ENGINEERING RESEARCH DATA MANAGEMENT PLAN REQUIREMENT SPECIFICATION

ALEX BALL, MANSUR DARLINGTON, TOM HOWARD, CHRIS MCMAHON, STEVE CULLEY

erim6rep100901ab11.pdf

ISSUE DATE: 17th January 2011

Catalogue Entry

Title	Engineering Research Data Management Plan Requirement Specification
Creator	Alex Ball (author)
Contributor	Mansur Darlington, Tom Howard, Chris McMahon, Steve Culley
Subject	data management; data association; metadata; data infrastructure
Description	The research context supported by this specification is one in which academics and researchers construct a data management plan (DMP) at the point of writing a funding proposal, primarily to avoid generating duplicate data and to ensure that the data that are generated are optimized for re-use. Throughout the course of the research, the DMP is then used as a record of how the data are being managed, so that on completion of a project, the DMP brings together the information needed to deposit the data for long-term curation. The specification does not, however, cover how the DMP would need to be revised at the point of ingest to support long-term curation; it is aimed at data creators rather than data librarians. The purpose of this specification is not only to outline the information that should be recorded in a DMP for engineering research, but also the manner in which the DMP should be implemented and the functionality that the supporting data management infrastructure should provide.
Publisher	University of Bath
Date	1st September 2010 (creation)
Version	1.1
Type	Text
Format	Portable Document Format version 1.5
Resource Identifier	erim6rep100901ab11
Language	English
Rights	

Citation Guidelines

Alex Ball, Mansur Darlington, Tom Howard, Chris McMahon, Steve Culley. (2011). *Engineering Research Data Management Plan Requirement Specification* (version 1.1). ERIM Project Document erim6rep100901ab11. Bath, UK: University of Bath.

CONTENTS

1 Introduction . . . 3
2 Glossary . . . 5
3 Infrastructure and implementation issues . . . 6
3.1 Relating the DMP to other documentation . . . 6
3.2 Relating other documentation to the DMP . . . 7
3.3 Understanding the DMP . . . 8
3.4 Rôles and responsibilities . . . 8
3.5 Review of the DMP . . . 8
3.6 Revision of the DMP . . . 9
3.7 Budget . . . 9
3.8 Storage, back-up and security . . . 9
3.9 Receiving repository . . . 10
4 DMP Contents . . . 10
4.1 Summary of Research Activity . . . 10
4.2 Data re-use . . . 11
4.3 Relating new data to existing data . . . 11
4.4 Future use of the data . . . 12
4.5 Project record manifest . . . 13
4.6 Data generation and manipulation . . . 13
4.7 Data organization . . . 14
4.8 Data quality . . . 15
4.9 Data structures and formats . . . 15
4.10 Data semantics . . . 16
References . . . 16

1 INTRODUCTION

The aim of the ERIM Project is to produce data management plans (DMPs) that support three types of activity:

1. making existing research data available and fit for a future known research activity (*data re-purposing*);
2. managing existing research data such that it will be available for a future unknown research activity (*supporting data re-use*).
3. using research data for a research purpose or activity other than that for which it was intended (*data re-use*);

The DMPs produced under this specification have two functions. In the first instance, they act as a guide to researchers on re-using existing data, re-purposing their own data and supporting data re-use throughout the Research Activity. At the end of the Research Activity, their purpose is to act as a record of how the data have been re-used and re-purposed, where applicable, and how data re-use has been supported. DMPs under this specification are not intended for use in the preservation stage, though they (along with other documentation) should provide sufficient information to allow Data Librarians/ Managers to construct a suitable DMP for the long-term care of the data.

The DMPs under this specification

- identify any future known Research Activities that may make use of the research data;
- describe how researchers will make, are making or have made the Data Case available and fit for these Research Activities, if applicable;
- describe how Data Creators will manage, are managing or have managed the Data Case to make it amenable for use in a future unknown Research Activity;
- provide additional information, where needed, to allow a Data Librarian/Manager to continue to manage the Data Case, enabling its use in the identified known Research Activities and in future unknown Research Activities generally.

This data management plan specification is based on the template for DMPs drawn up by the Digital Curation Centre (DCC) [DJ10], and is informed by the Principles for Engineering Research Data Management [Dar+10] and a thematic analysis of DMP tools and exemplars [Bal10]. The conclusions of the latter report were that the most relevant pieces of information to include within a DMP focused on data re-purposing and re-use were as follows.

- Suitability or otherwise of prior data for supporting the research
- Inventory of the data
- Bodies and groups that are likely to be interested in the data
- Foreseeable contemporary or future uses for the data
- Foreseen or actual relationships between prior data and new data
- Integrability of the data with established data collections:
  - Provenance
  - Trustworthiness
  - Data quality
  - Standard formats, ontologies, conventions, methodologies
- Detailed description of data generation: methodology, technology, conventions, etc.
- Methods of data organization during research and in the final data case
- Provision of contextual data records:
  - Information or resources necessary for processing/rendering the data
  - Information necessary for understanding the data
  - Information necessary for understanding the processing history of the data
  - Manual and automated methods for capturing this information
  - Metadata standards (formats, ontologies)
  - Rationale for the above
- Quality assurance procedures and standards

Also identified were pieces of information with indirect relevance:

- Requirements and guidance that shape the data management plan:

  - Departmental and institutional policies
  - Requirements and guidance from the place of deposit
  - Requirements and guidance from the funding body
- Budget for data management activity
- Human infrastructure for data management activity

It is noted that as these lists are the result of focusing on particular aspects of data management, they are not a sufficient base for a comprehensive data management strategy and other aspects need to be addressed in order to ensure data can be re-used in practice.

2 GLOSSARY

This document uses the ERIM Data Management Terminology version 3 [see How+10], from which the following terms are extracted for ease of reference.

Research Activity The process through which research Data and context Data are accumulated and developed.

Research Data Development Process One of a set of processes that are commonly carried out during the Research Activity which changes or adds to the research Data associated with a Research Activity or project.

Data Case The set of Data Records associated with some discrete Research Activity (project, task, experiment, etc.).

Record Information in any medium, created, received and maintained as evidence of an activity.

Data Record The Data Object which contains the Data.

Associative Data Record A Record which documents the association between other Data or Data Records. The Data contained within an Associative Data Record is a special case of contextualizing Data and the Data Record a special case of a Context Data Record.

Context Data Record A Record containing Data explicitly intended to place in context other Data or abstract aspects of the Research Activity or subject.

Experimental Apparatus Data Record A Digital Object which is analogous to the physical experimental apparatus familiar in much laboratory-based research.

Research Data Record A Record containing research Data, i.e. Data that are descriptive of the research object.

Research Object Data Record A Data Object which is itself the object of research interest or which together with Research Object Data Records constitutes the object of research interest.

Data Object Either a Physical Object or a Digital Object.

Digital Object An object composed of a set of bit sequences.

This document also uses the terminology defined by Swan and Brown [SB08] for the rôles associated with managing research data (see Figure 1).

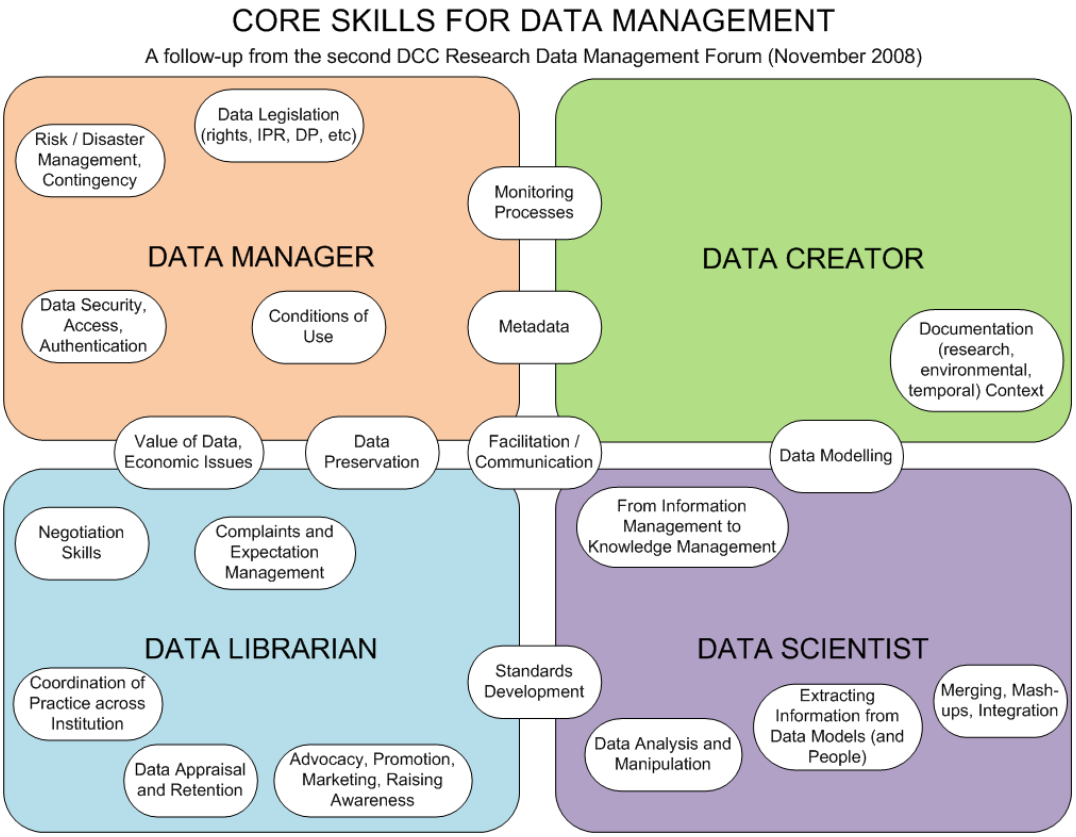

Figure 1: The four research data management rôles and their respective skills and responsibilities [PD09].

Data Creator Researchers with domain expertise who produce data. These people may have a high level of expertise in handling, manipulating and using data.

Data Scientist People who work where the research is carried out – or, in the case of data centre personnel, in close collaboration with the creators of the data – and may be involved in creative enquiry and analysis, enabling others to work with digital data, and developments in data base technology.

Data Manager Computer scientists, information technologists or information scientists and who take responsibility for computing facilities, storage, continuing access and preservation of data.

Data Librarian People originating from the library community, trained and specialising in the curation, preservation and archiving of data.

3 INFRASTRUCTURE AND IMPLEMENTATION ISSUES

3.1 Relating the DMP to other documentation

Requirement

A system must be in place to allow readers of the DMP to find the following information.

- High level project documentation relating to the Research Activity, e.g. proposal document, detailed plan.
- Confidentiality agreements, IPR statements and other documents that affect how the research data may be used.
- Requirements and guidance from receiving repository, if any, in relation to data management.
- Requirements and guidance from the funding body in relation to data management.
- University of Bath policy in relation to data management.

Guidance

Possible systems include documented directory structure conventions and use of dereferenceable URIs within the DMP for the other documents.

The University of Bath's Online Publications Store (OPuS) does not yet impose requirements or provide guidance on depositing data. The University's policy in relation to data management is currently its *Good Practice Code for Research*[1]. The EPSRC does not currently add any requirements or guidance to that provided by RCUK in its *Policy and Code of Conduct on the Governance of Good Research Conduct,*[2] 2009.

Rationale

It is important that Data Librarians/Managers of a Data Case have access to the documents that impact on how the data are to be managed. They also need as much contextual information as possible in order to produce descriptive metadata to aid the discovery of the data. As this information already exists for other purposes, there is no need to reproduce it within the DMP, but Data Librarians/Managers need to be able to find it easily.

3.2 Relating other documentation to the DMP

Requirement

A system must be in place to allow readers of high level project documentation and users of the Data Case itself to find the corresponding DMP.

Guidance

Possible systems include documented directory structure conventions and use of a dereferenceable URI for the DMP within other documents.

Rationale

It is important that Data Librarians/Managers of a Data Case have access to the documents that impact on how the data are to be managed. They also need as much contextual information as possible in order to produce descriptive metadata to aid the discovery of the data. As this information already exists for other purposes, there is no

1. http://www.bath.ac.uk/opp/resources.bho/goodpracticeresearch2010.pdf
2. http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/reviews/grc/goodresearchconductcode.pdf

need to reproduce it within the DMP, but Data Librarians/Managers need to be able to find it easily.

3.3 Understanding the DMP

Requirement

A system must be in place to allow readers of the DMP to find and read this specification.

Guidance

Possible systems include generic guidance attached to all IdMRC Data Cases, and use of a dereferenceable URI within the DMP for this document.

Rationale

Readers of the DMP should be aware of its aims, purpose and target audience. As this information is set out in this document, it does not need to be reproduced in the DMP itself, as long as readers of the DMP are made aware this document.

3.4 Rôles and responsibilities

Requirement

The responsibilities for implementing and reviewing the DMP must be enumerated and associated with named individuals or organizational rôles. This list must be kept up to date.

Guidance

While this information may form part of the DMP contents, it may be more visible as part of a wider list of responsibilities, or as part of a system that can remind individuals when action needs to be taken.

Rationale

A clear set of rôles and responsibilities must be defined in order to prevent situations where data management is always left for someone else to do.

3.5 Review of the DMP

Requirement

A system should be in place to ensure adherence to and accuracy of the DMP.

Guidance

These checks could form part of regular project meetings.

Rationale

DMPs act as both guidance and as a record of activity. Adhering to a DMP promotes good data management practice, while providing an accurate record assists both Data Librarians/Managers (when reworking the DMP for long-term archiving and preservation) and future Data Creators (enabling more realistic DMPs to be drafted for future projects).

3.6 Revision of the DMP

Requirement

The DMP should be under some form of version control. Dates and authorship should be recorded in case of future difficulty. Contact details should be available for all authors.

Guidance

It may be simplest to record this information within the DMP itself. Contact details are typically easier to keep up to date using a separate system.

Rationale

The DMP is not a static document, but should be updated throughout the project to reflect what actually happened.

3.7 Budget

Requirement

The cost of implementing this DMP and, if appropriate, future preservation activity should be estimated and factored into the total project budget.

Rationale

Data management must not be left as an afterthought, but should be a budgeted activity in the project proposal.

3.8 Storage, back-up and security

Requirement

A centre- or departmental-level Data Management Policy should enforce good practice with regard to data storage, back-up and security.

Guidance

Example conditions that may form part of such a policy include the following.

- The primary store for working data is the University of Bath's X drive, run by BUCS.
- Each Research Activity should have an allocated directory in which its associated Data Records are kept.

- Researchers should keep any additional copies of their data (on their own hard disks, on removable media) in sync with the copies on the X drive, so each may act as back-up for the other.
- Sensitive data on portable equipment should be encrypted and password-protected.

Rationale

Even in the relatively brief period of active research, disasters and systems failures can occur. A reasoned approach to storage, back-up and security can avoid serious or catastrophic data loss or leakage.

3.9 Receiving repository

Requirement

The target repository for the data of each Research Activity should be specified.

Guidance

It is expected that most research will be carried out in the same funding and organizational context, and thus there will be a default repository to which most research data will be submitted. There should therefore be, at departmental or centre level, a statement or policy that names this default repository and stipulates that, unless otherwise stated in a DMP, research data will be prepared for deposition and submitted to that repository. The requirement at section 3.1 would apply to this statement or policy.

For any Research Activity for which the default position does not apply (e.g. due to funding from a body that stipulates deposition in another repository), the DMP should include a statement specifying what will happen to the Data Case on completion of the Research Activity.

Rationale

Knowledge of the Data Librarians that will care for the data post-project, and their standards and requirements, allows researchers to optimize their data handling processes to avoid problems later on.

4 DMP CONTENTS

4.1 Summary of Research Activity

Requirement

Provide a brief summary of

- high level project documentation relating to the Research Activity;
- confidentiality agreements, IPR statements and other documents that affect how the research data may be used.

Guidance

It is expected that full details will be kept in separate documents that should be discoverable from the DMP (see section 3.1), so this summary need contain only the most notable facts, such as the project name, project dates, funder(s), and organizations involved.

Access restrictions and ownership of information should be stated here in the simplest, most concise form; where important complexities are glossed, prompts to consult the full documentation should be given.

Rationale

This is provided as a quick reference, as some data management operations require only this level of detail.

4.2 Data re-use

Requirement

Could the Research Activity's data requirements be met in whole or in part by existing data?

If not, briefly indicate how this is known.

If so, identify the data that could be used and indicate any foreseen issues with accessing the data. Explain why new data are being generated as well, if applicable.

Guidance

Typical reasons for not re-using data: searches found no similar previous research; object of research has not previously been studied; all research in the area is covered by strict confidentiality agreements.

Typical access issues: access to data is by application only; unclear data licensing.

Typical reasons for generating new data: comparison over time; extension to cover new areas.

Rationale

A Research Activity can either mine existing data for new results, add to an existing body of data (to fine tune, generalize or place limits on previous results), or create an entirely new body of data. When planning a new Research Activity, researchers should be able to justify taking one of these three approaches.

4.3 Relating new data to existing data

These requirements only apply if new data are to be generated. It may be helpful to bear Requirement 2 in mind while answering Requirement 1.

Requirement 1

Describe how the newly generated data relates to the wider landscape of existing data.

Guidance 1

This is not concerned so much with existing data that may be used in the Research Activity, but rather with the disciplinary context. A typical answer might identify a body of data with which it would be helpful to harmonize newly generated data, or from which methodologies might be drawn, e.g. ISO standard materials testing data, time/motion studies data.

Requirement 2

State the measures that will be/have been taken to ensure integrability between newly generated data and existing data.

Guidance 2

The following are possible issues to consider. Only brief answers are required here: full details should be given in corresponding sections later in the DMP.

- Method of assuring data quality.
- Method of recording provenance.
- Mechanisms for ensuring trustworthiness of data.
- Choice of standard formats, ontologies, conventions, etc. for the data.
- Choice of standard formats, ontologies, conventions, etc. for the metadata.

Rationale

A typical way in which data are re-used is in combination with similar data. This is considerably easier if compatibility issues are addressed in the planning stages of a Research Activity (Principle of Reusability [McM+09]).

4.4 Future use of the data

Requirement 1

List any bodies/groups which might be interested in the data, and the foreseeable contemporary or future uses to which they might put the data.

Guidance 1

It is acceptable to define groups based on discipline, research interest or specific research topic. It is acceptable to list bodies or groups without reference to uses, and foreseeable uses without reference to specific groups, if appropriate.

Requirement 2

State the measures that will be/have been taken to prepare the data for these bodies/ groups/uses.

Guidance 2

The following are possible issues to consider. Only brief answers are required here: full details should be given in corresponding sections later in the DMP.

- Forms of data organization.
- Choice of standard formats, ontologies, conventions, etc. for the data.
- Choice of standard formats, ontologies, conventions, etc. for the metadata.

Rationale

If the future uses for research data are known or can be predicted at the outset, special provisions can be made during the research that increase the compatibility of the data with that future use (Principle of Reusability [McM+09]). Explicitly stating where this has been done can help Data Librarians/Managers continue this work in the preservation stage.

4.5 Project record manifest

Requirement

Describe the method by which Data Records and the relationships between them will be/were recorded in a project record manifest. Present this manifest (once it exists), or otherwise indicate how it may be accessed.

Guidance

Ideally a Research Activity Information Development (RAID) diagram should be presented, alongside instructions on accessing a computer-interpretable version; in this case, the DMP should cite the RAID modelling method and specifications, alongside notes on how these will be/were implemented in this case.

A possible alternative would be to present an annotated list of Data Records. The procedure for maintaining this list must be given.

Rationale

Providing details of what Data Records are included in a Data Case, how they came about and what relationships exist between them, helps future researchers to understand the data, assess their suitability and re-use them for new research (Principle 7 for ERDM [Dar+10]). In particular, recording the relationships between Data Records (and between data) satisfies some users' requirements for provenance information.

4.6 Data generation and manipulation

Requirement

Give a detailed account of how the data will be/were generated and manipulated, including the methods, technology, conventions, coding schemes, etc. that will be/were used.

Guidance

It is expected that the level of detail provided here will be low initially, but will increase as the plans are implemented. At all stages, do bear in mind that using standard or generic tools makes it easier to revisit data at a later stage; virtual machine images can help to make specialist tools more portable (Principle 5 for ERDM [Dar+10]).

When writing a DMP in retrospect, it is acceptable to cite a journal/conference paper containing the information, provided it is detailed enough and that a pre- or post-print is available in case of access difficulties. In the normal course of events, the information should be provided here first and then adapted for use in a journal/conference paper.

It may be helpful to provide this information in the form of a commentary on a RAID diagram.

Rationale

Providing this information helps the Data Creator/Scientist to review whether a particular generation or manipulation stage really accomplishes what is intended. It helps a re-user to compare the data with other similar data, and to harmonize different data sets by reprocessing intermediate data. It also helps others (e.g. referees) reprocess data in order to verify the results.

4.7 Data organization

Requirement

Describe how the data will be/have been organized.

Guidance

This refers both to how data are organized within Data Records and how Data Records are organized within the Data Case. Example topics:

- file-naming convention
- use of individual data table files or an integrated database/spreadsheet
- version control
- convention for associating administrative metadata with Data Records
- content packaging

Rationale

Providing this information makes it easier for a Data Librarian/Manager or re-user of the data to navigate the Data Records and find specific parts. It can also help the Data Creator/Scientist to check that all the Data Records have been included.

4.8 Data quality

Requirement

Describe the quality assurance procedures and standards that will be/were used. If any data quality issues were encountered, list them and describe what was done to resolve them.

Rationale

Asking for this information prompts the Data Creator/Scientist to consider data quality issues. By adhering to quality assurance procedures and standards, Data Creators/ Scientists can raise the level of trust that re-users put in the data.

4.9 Data structures and formats

Requirement

Specify the information, tools or resources that would be needed to manipulate or render the Data Records, along with any special instructions. Provide an explanation of why a particular format has been selected for use.

Guidance

'Render' here means display to a human.

At the project planning stage, indicate the hardware and software environment in which the Research Activity will be conducted. Also indicate how this section will be completed during the course of the Research Activity.

Once Data Records have been made, start by specifying the hardware and software environment in which the data were generated or manipulated, and then consider alternative environments, tools and libraries that might support the data. If specialist tools were used, consider installing them on a virtual machine; in which case, provide details here of how to run the virtual machine. If available/known, cite here format specification documents for all data formats used.

If the choice of formats has been justified elsewhere in the DMP (e.g. section 4.3 and 4.4), readers may be directed to those sections in place of a recapitulation here.

Rationale

This information gives considerable assistance to future re-users of the data, and helps Data Librarians deal with technological obsolescence in the preservation stage. Given the impact that software/format choices can have on the longevity of data, it is important that Data Creators/Scientists make them in a reasoned manner.

4.10 Data semantics

Requirement

Provide any additional information that would be needed to understand the Data Records, once rendered. Provide justification for the conventions used.

Guidance

As an example, tabular data can have terse column headings; fuller explanations of what a column represents can be given here. Other examples of information to provide here include data dictionaries, coding schemes and ontologies. The information can be given directly in the DMP, or instructions can be given on how to look up the information for each Data Record.

At the project planning stage, provide a general statement about the conventions that will be used, and indicate how this section will be completed during the course of the Research Activity.

If the choice of conventions has been justified elsewhere in the DMP (e.g. section 4.3 and 4.4), readers may be directed to those sections in place of a recapitulation here.

Rationale

Data cannot be re-used if their meaning has not been properly recorded. Given the impact that the choice of ontologies, conventions and so on can have on the longevity of data, it is important that Data Creators/Scientists make it in a reasoned manner.

REFERENCES

[Bal10] A Ball (2010-09-08). *Thematic Analysis of Data Management Plan Tools and Exemplars*. ERIM Project Document erim6rep100701ab10. Version 1.0. University of Bath: Bath, UK. URL: http://opus.bath.ac.uk/21278 (2010-11-15).

[Dar+10] M Darlington et al. (2010). *Principles for Engineering Research Data Management*. ERIM Project Document erim6rep101028mjd10. University of Bath: Bath, UK. URL: http://opus.bath.ac.uk/22201 (2011-01-31).

[DJ10] M Donnelly & S Jones (2010-01-06). *Template for a Data Management Plan*. Version 1.2. Digital Curation Centre. URL: http://www.dcc.ac.uk/sites/default/files/DMP_template_v1.2_100106.rtf (2010-07-12).

[How+10] T Howard et al. (2010-09-17). *Understanding and Characterizing Engineering Research Data for its Better Management*. ERIM Project Document erim2rep100420mjd10. Version 1.0. University of Bath: Bath, UK. URL: http://opus.bath.ac.uk/20896/ (2010-09-21).

[McM+09] CA McMahon et al. (2009-10-01). *The Development of a Set of Principles for the Through-Life Management of Engineering Information*. KIM Project Document kim40rep007mjd10. Version 1.0. University of Bath & University of Cambridge: Bath, UK & Cambridge, UK. URL: http://www.bath.ac.uk/idmrc/themes/projects/kim/kim40rep007mjd10.doc (2010-09-22).

[PD09] G Pryor & M Donnelly (2009). 'Skilling Up to Do Data: Whose Role, Whose Responsibility, Whose Career?' *International Journal of Digital Curation* 4:2, 158–170. ISSN: 1746-8256. URL: http://www.ijdc.net/ijdc/article/view/126 (2010-09-21).

[SB08] A Swan & S Brown (2008-06). *The Skills, Role and Career Structure of Data Scientists and Curators: Assessment of Current Practice and Future Needs*. Final Report. JISC: London. URL: http://www.jisc.ac.uk/publications/reports/2008/dataskillscareersfinalreport.aspx (2010-09-21).