THEMATIC ANALYSIS OF DATA MANAGEMENT PLAN TOOLS AND EXEMPLARS

ALEX BALL

erim6rep100701ab10.pdf

ISSUE DATE: 8th September 2010

Catalogue Entry

Title	Thematic Analysis of Data Management Plan Tools and Exemplars
Creator	Alex Ball (author)
Subject	appraisal; data re-purposing; data re-use; delivery; discovery; legal compliance; preservation; shareability; supporting data re-use
Description	Data management plans (DMPs) are a useful way of ensuring that research data outputs are properly prepared for preservation and re-use. The range of issues they can address in order to achieve this end, though, is much wider than those two areas. An analysis of the guidance issued by five organizations shows that their templates and exemplars address at least nine different challenges or themes in relation to data management. Were researchers asked to follow this guidance in full, there is a danger they may regard DMPs as nothing more than added bureaucracy, instead of valuable tools for producing re-usable data. A focused subset is therefore selected for researchers served by the ERIM Project.
Publisher	University of Bath
Date	1st July 2010 (creation)
Version	1.0
Type	Text
Format	Portable Document Format version 1.4
Resource Identifier	erim6rep100701ab10
Language	English
Rights	

Citation Guidelines

Alex Ball. (2010). *Thematic Analysis of Data Management Plan Tools and Exemplars* (version 1.0). ERIM Project Document erim6rep100701ab10. Bath, UK: University of Bath.

CONTENTS

1 INTRODUCTION

While data management plans are becoming a required element for funding applications to most research funders, the EPSRC does not yet require of its applicants or funded projects a written statement on how their research data will be managed [Jon09; Jon10]. The ERIM Project – dealing with EPSRC-funded research data – therefore has more latitude than some of its peer projects in JISC's Research Data Management programme to determine what should be included in a data management plan and how it should be constructed. The aim of the ERIM Project is to produce data management plans 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*);

Note that the first two of these types are from the perspective of the data creator and data curator, whereas the latter is from the perspective of a data consumer.

In tandem with the increasing ubiquity and specificity of requirements for data management plans, organizations such as the Digital Curation Centre (DCC) have been producing guidance, in the form of templates and exemplars, to assist researchers in meeting those

requirements. Some sets of guidance aim to be comprehensive and universally applicable, while other sets are aimed at fulfilling a particular set of requirements with an emphasis different from that of the ERIM Project. As an example, some guidance places a strong emphasis on the shareability of data, an issue that is orthogonal to the issue of how re-usable they are. The task of this document is to analyse a selection of guidance documents and determine the challenges and issues that each part aims to address, and in so doing, identify those parts that are of particular relevance to ERIM.

The sets of guidance analysed in this document are those offered by the DCC, MIT Libraries, the Australian National University, the National Institutes of Health and the Rural Economy Land Use Programme. In performing this exercise, we hope to ensure that the data management planning methodology proposed by ERIM is tightly focused on supporting the three types of activity above, and feels to the researcher less like a form-filling exercise and more like a genuine commitment to address the issues.

2 THEMES

In the course of examining data management plan guidance, the following set of challenges or themes was enumerated. There are of course many ways in which the diverse issues associated with data management and curation may be grouped into themes; the following scheme focuses on potential points of failure that would prevent the data generated/collected by one researcher being re-used by another researcher. The aim is for the themes to be as independent as possible, which is not to say that full independence has been achieved or is even achievable.

Appraisal The process that determines whether a record or set of records should be curated or deleted. For the purposes of this analysis, issues of deletion are also grouped under this theme.

Data re-purposing The process of making existing research data available and fit for a future known research activity.

Data re-use The process of using research data for a research purpose or activity other than that for which it was intended.

Delivery The process by which a researcher collects existing data.

Discovery The process by which a potential re-user of data learns of their existence.

Legal compliance The extent to which data and data management processes comply with legislation, funders' requirements, contractual obligations and agreements with (human or organizational) research subjects.

Preservation The activity of maintaining data over time so they can be accessed and understood. For the purposes of this analysis, issues of storage and safeguarding data from (accidental or deliberate) damage are grouped under this theme.

Shareability The extent to which it is possible to share data outside the originating team of researchers. For the purposes of this analysis, issues of security relating to unauthorized access are grouped under this theme.

Supporting data re-use The process of managing existing research data such that it will be available for a future unknown research activity.

3 DIGITAL CURATION CENTRE

The DCC has developed a template and corresponding online tool for writing data management plans [DJ10].

3.1 Introduction and context

> Basic project information
>
> 1. Name of project
> 2. Funding body/bodies
> 3. Budget
> 4. Duration
> 5. Partner organisations

These items of information are not so much part of the data management plan as a way of tying the plan to particular set of records. The primary usefulness of specifying funding bodies is to assist automated tools or curators in looking up the requirements that those funding bodies impose relating to data management plans; conceivably it could be used to assist in monitoring those requirements for changes that apply retrospectively. The total project budget is not pertinent to data management; it may serve as a reminder to allow for data management costs in the budget, although that is covered more extensively below (see section 3.7). Specifying partner organizations may help data curators track down missing information, or warn them that there may be additional complexity in terms of methodology, wording of consent forms or agreements, etc. It may be argued, however, that such information would sit better within a Context Data Record.

> What are the aims and purpose of the research?

It is important for potential re-users of data to know the purposes for which the data were originally intended. This information is most pertinent to re-users at the investigation or experimental level, although higher-level information (such as at the project level) is may be useful at the **discovery** stage.

> Related policies
>
> 1. Funding body requirements relating to the creation of a data management plan
> 2. Institutional or research group guidelines
> 3. Other dependencies

These pieces of information are also better thought of as meta-information rather than as part of the data management plan itself. Asking for them within the template serves as a reminder of the requirements that the plan ultimately has to satisfy, and may also provide context should the plan later prove to be unclear.

Specifying the funding body requirements removes the need for later detective work to determine which requirements were in force when the research was funded; if a funding body were to adopt the practice of applying requirements retrospectively, only the information about the funding body itself (above) would be needed as the latest set of requirements would always apply. Institutional or research group guidelines may contain instructions/recommendations or substantive content (or both); any substantive content should form part of the DMP whether included directly or referenced as an annex.

Basic Data Management Plan information

1. Date of creation
2. Aims and purpose
3. Target audience for this plan
4. Statement on plan revision schedule
5. Does this version supersede an earlier plan?

These pieces of information assist with version control, concentrate the efforts of the author of the plan and provide context for the reader of the plan. They do not have a direct impact on the management of data.

Glossary of terms

This is an aid to the reader in understanding the document.

3.2 Legal and ethical issues

Ethical and privacy issues

1. Are there ethical and privacy issues?
2. If so, how will these be resolved? Examples: anonymisation of data, institutional ethical committees, formal consent agreements.
3. Is the data 'personal data' in terms of the Data Protection Act 1998 (the DPA)?
4. What have you done to comply with your obligations under the DPA?

These pieces of information are partly concerned with **legal compliance** and partly with the **shareability** of data.

Intellectual property rights

1. Is the dataset covered by copyright or the Database Right? If so, who owns the copyright and other intellectual property? Ideally, this should address the risk of movement of staff between institutions mid-project.
2. How will the dataset be licensed if rights exist? Examples: any restrictions or delays on data sharing needed to protect intellectual property, copyright or patentable data.

3. What is the dispute resolution process and/or mechanism for mediation?

Again, these pieces of information are partly concerned with (civil) **legal compliance** and partly with the **shareability** of data.

3.3 Access, data sharing and re-use

Data sharing and re-use

1. Will you share the data you capture or create?
2. Which bodies/groups are likely to be interested in the data?
3. What are the foreseeable contemporary or future uses for the data?
4. Are there any reasons not to share or re-use data? Examples: ethical, non-disclosure, quality-related.

The first of these pieces of information is clearly about the **shareability** of data. The second and third are, in ERIM terminology, concerned with **data re-purposing**. The fourth piece of information covers both issues.

Access

1. Do you have an obligation to make the data available? Examples: due to research funder policy or Freedom of Information (FoI) legislation. Note that FoI legislation differs in Scotland from England and Wales.
2. How and when will you make the data available?
3. Will any permission restrictions need to be placed on the data?
4. What is the procedure for gaining access to the data?
5. Will access be chargeable?
6. Do you plan to publish findings which rely on the data?
7. If so, do your prospective publishers place any restrictions on other avenues of publication?

The first and last of these pieces of information are concerned with **legal compliance**. Items 2, 4 and 5 deal with the **delivery** of data, while item 6 is a hint towards issues of **discovery**. Item 3 is again concerned with data **shareability**.

Timing

1. Is there a right-of-first-use agreement for the original data collector/creator/principal investigator?
2. Details of any embargo periods for political/commercial/patent reasons

These are once more issues concerned with the **shareability** of data.

3.4 Data standards and capture methods

> What does the term 'data' comprise for the research? Data description, including volume, type, content to be created, etc.
>
> What data types will you be creating or capturing? Examples: experimental measures, qualitative, raw, processed.

These two questions ask the researcher to predict what forms of data the research will produce. Providing an inventory of data associated with a research activity is a pre-requisite for most management and curation activities.

> Existing and new data
>
> 1. Have you surveyed existing data, in your own institution and from third parties?
> 2. What existing datasets could you use or build upon?
> 3. Are there any access issues?
> 4. What 'added value' will the new data you create or capture provide to existing datasets?
> 5. Why do you need to capture or create new data?
> 6. What is the relationship between new dataset(s) and existing data?
> 7. How will you manage integration between the data being gathered in the project and pre-existing data sources? This should cover provenance, trust and data quality.

Items 1, 2, 3 and 5 are concerned with the **re-use** (or otherwise) of existing data,[1] while the remaining items in this list are about a specific method of **supporting data re-use**: ensuring interoperability and integration with a wider body of data, making it easier for other researchers to re-use the new data.

> How will you capture or create the data? This should cover content selection, instrumentation, technologies and approaches chosen, methods for naming, versioning, meeting user needs, etc., and should be sensitive to the location in which data capture is taking place.

This information continues the theme of **supporting data re-use**, partly on the basis of integrating with existing data (common methods of collection, etc.) and partly through documenting the conventions in use.

> Which file formats will you use, and why? Examples: recourse to staff expertise, Open Source, accepted standards, widespread usage.

Choosing a widely supported file format can be seen as **supporting data re-use**, at least in the short term. It is also useful for data curators to be aware of the file formats in use for the purposes of **preservation**.

1. Items 1 and 3 are also about **discovery** and **delivery**, respectively, but not of the data records covered by the plan.

> Metadata
>
> 1. What contextual details are needed to make the data you capture or collect meaningful?
> 2. How will you create or capture these metadata?
> 3. What form will the metadata take?
> 4. To what extent will metadata creation be automated?
> 5. Which metadata standards will you use?

Collecting metadata that enable the research data to be understood is an act of **supporting data re-use**.

> Why have you chosen particular standards and approaches for metadata and contextual documentation? Examples: recourse to staff expertise, Open Source, accepted domain-local standards, widespread usage.

This piece of information invites the researcher to consider the extent to which the metadata choices made above serve the purpose of **supporting data re-use**.

> What criteria will you use for Quality Assurance/Management? Examples: documentation, calibration, validation, monitoring, transcription metadata, peer-review.

Quality assurance is important for data purposing, **data re-purposing** and **supporting data re-use** as higher quality data give more reliable results.

3.5 Short-term storage and data management

> Anticipated data volumes? Ballpark figures, orders of magnitude.

Data volumes are principally a storage concern, although if they are very large they impact on **shareability** and the ability of **preservation** systems to maintain their integrity and usability.

> Storage
>
> 1. Where (physically) will you store the data?
> 2. On what media will you store the data?
> 3. Whose responsibility is the storage of the data?
> 4. How will you transmit the data, if required? Should address encryption if appropriate.
>
> Back-up
>
> 1. How will you back-up the data? Should address off-site storage.
> 2. How regularly will back-ups be made?
> 3. Whose responsibility will this be?

Again, these pieces of information are of interest from a **preservation** point of view.

> Security
>
> 1. How will you manage access arrangements and data security?
> 2. How will you enforce permissions, restrictions and embargoes?
> 3. Other security issues. Should address (where relevant) sensitive data, off-network storage, storage on mobile devices (laptops, smartphones, flash drives, etc.), policy on making copies of data, etc.

These pieces of information explore how the limits of **shareability** will be enforced.

3.6 Deposit and long-term preservation

> What is the long-term strategy for maintaining, curating and archiving the data? Reminder that project can consult institutional archivist(s) and/or records managers in long-term retention plans.

This is clearly concerned with **preservation**.

> Specifics
>
> 1. On what basis will data be selected for preservation?
> 2. How long will (or should) data be kept beyond the life of the project? N.B. this may simply link to relevant institutional or funding body requirements/ policies: political, temporal, commercial, legal.
> 3. How will you dispose of/transfer sensitive data? Include justification of decisions.

These pieces of information are concerned with the **appraisal** aspect of curation, with implications for **preservation**. Item 3 is also concerned with enforcing the limits of **shareability**.

> Which archive/repository/central database/data centre have you identified as a place to deposit data?

It is important to know this information at the beginning of the research; different repositories have different requirements for the data offered to them for ingest, and these requirements should colour what contextual information is collected, what formats and conventions are used for recording the data, and so on. The choice of repository therefore has an impact on a project's response to many data management challenges.

> What transformations will be necessary to prepare data for preservation/data sharing? Example: data cleaning/anonymisation where appropriate.

This is primarily focused on **preservation** and **shareability**.

> What related (representation) information will be deposited? Example: references, reports, research papers, fonts, the original bid proposal, etc.

All these types of information contribute towards **supporting data re-use**.

> Metadata, documentation and backup
>
> 1. What metadata/documentation will be created at each stage of deposit/ transformation? Examples: descriptive, structural, administrative, preservation, etc.
> 2. How will this be created and by whom?
> 3. Will you include links to published materials and/or outcomes?
> 4. How will you address the issue of persistent citation?

The first two of these items are critical for **supporting data re-use**. Item 3 also helps in **supporting data re-use**, while any backlinks would help **discovery**. The last item helps in the **delivery** of data.

> What procedures does your intended long-term data storage facility have in place for preservation and backup? How regular, by whom, methods used (e.g. format normalisation, migration...)

This is once more squarely in **preservation** territory.

3.7 Resourcing

> Staff/organisational roles and responsibilities for implementing this plan. This should include time allocations, project management of technical aspects, training requirements, contributions of non-project staff, etc. Individuals should be named where possible. Continue in an Annex if necessary.
>
> Financial issues. This should cover (e.g.) payments to service providers within institutions, payments to external data centres for hosting data, income derived from licensing data, etc. It is also important to remember to build costs of in-project data management into the project budget.

Unsurprisingly, these pieces of information are concerned with both the human infrastructure and the financial sustainability of the curation effort. These issues underlie many of the other data management challenges.

3.8 Adherence, review and long-term management

> Adherence
>
> 1. When will adherence to this data management plan be checked or demonstrated?

> 2. Who will do this?
>
> How and when will this data management plan be reviewed?

These pieces of information are not part of the data management plan as such, but rather statements concerning its implementation. Associating tasks with named individuals helps to ensure they are completed.

> Longer-term responsibilities
>
> 1. Is there a formal process for transferring responsibility for the data?
> 2. Who will have responsibility over time for decisions about the data once the original personnel have gone? Likely to be custodians in data centres.
> 3. Who will meet the costs of long-term management and storage?

These pieces of information again concern the human infrastructure and financial sustainability of the archiving effort, with particular emphasis on **preservation**.

3.9 Annexes

> Contact details and expertise of nominated data managers/named individuals.

These pieces of information perform a practical function of providing a contact with whom the data or the plan may be discussed, but providing details of expertise is a way of demonstrating the seriousness with which the management of data is treated.

4 MIT LIBRARIES

MIT Libraries have produced a guide to data management that includes a planning checklist and a list of suggested topics for a data management plan to cover [Mit].

4.1 Data planning checklist

> What type of data will be produced? Will it be reproducible? What would happen if it got lost or became unusable later?

Providing an inventory of data associated with a research activity is a pre-requisite for most management and curation activities. The second and third questions here are useful for deciding which data take priority for data management resources.

> How much data will it be, and at what growth rate? How often will it change?

Data volumes are principally a storage concern, although if they are very large they impact on **shareability** and the ability of **preservation** systems to maintain their integrity

and usability. Open (expanding, dynamic) datasets have a different set of challenges from closed (dormant, definitive) datasets.

> Who will use it now, and later?

Consideration of known, future uses is the starting point for **data re-purposing**.

> Who controls it (PI, student, lab, MIT, funder)?

Understanding intellectual property rights helps towards **legal compliance**, and may affect the **shareability** of the data.

> How long should it be retained? e.g. 3–5 years, 10–20 years, permanently

This relates to the **appraisal** aspect of curation, with implications for **preservation**.

> Are there tools or software needed to create/process/visualize the data?

This question has implications for **supporting data re-use**. If uncommon, specialist tools or software are needed in order to handle the data, this limits the extent to which other researchers will be able to re-use them.

> Any special privacy or security requirements? e.g. personal data, high-security data
>
> Any sharing requirements? e.g. funder data sharing policy

These questions explore the drivers for and against **shareability**.

> Any other funder requirements? e.g. data management plan in proposal

Compliance with these requirements is necessary for receiving funding.

> Is there good project and data documentation?
>
> What directory and file naming convention will be used?

Good documentation of the project, the data and the conventions used in both is key to **supporting data re-use**.

> What project and data identifiers will be assigned?

Identifiers are useful when describing the structure of the data (**supporting data re-use**) and also help with data **discovery**.

> What file formats? Are they long-lived?

Using file formats that are common, well-supported by tools and likely to remain so for a long time is one way of **supporting data re-use**, and also reduces the amount of **preservation** effort required.

> Storage and backup strategy?

This strategy is clearly important for **preservation**.

> When will I publish it and where?

The place of data publication is important for **discovery**. The timing is perhaps more important from the perspective of funding the data preparation and publication process.

> Is there an ontology or other community standard for data sharing/integration?

Using community standards is another way of **supporting data re-use** as it makes it easier to integrate the data with other data, and to use standard tools with it.

4.2 Data management plans

> Name of the person responsible for data management within your research project

Assigning data management to a named person helps to ensure its completion.

> Description of data to be collected and the methodology

As mentioned above, providing an inventory of data associated with a research activity is a pre-requisite for most management and curation activities. Using a standard methodology helps in **supporting data re-use** as it makes the data more directly comparable with other similar data.

> How data will be documented throughout the research project

As mentioned above, good documentation of the project, the data and the conventions used in both is key to **supporting data re-use**.

> Data quality issues

Quality assurance is important for data purposing, **data re-purposing** and **supporting data re-use** as higher quality data give more reliable results.

> Backup procedures

Backup procedures are important for **preservation**.

> How data will be made available for public use and potential secondary uses

The primary concerns addressed by this information will be issues of **discovery** and **delivery**.

> Preservation plans

Preservation plans are, of course, important for **preservation**.

> Any exceptional arrangements that might be needed to protect participant confidentiality or intellectual property

Such arrangements enforce the limits of **shareability**.

5 AUSTRALIAN NATIONAL UNIVERSITY

The Australian National University Data Management Manual [Inf08] offers the following advice to those writing a data management plan.

> **Project description.** Write a few paragraphs about the research project to give some perspective to the remainder of the plan. Use this section to introduce any terminology that will be used in the DMP.

As the guideline states, this is provided primarily as a service to the reader rather than to address data management issues.

> **Survey of existing data.** Whilst not compulsory, it is good practice to see if there are existing data that could replace or augment the data you are planning to create. It is a condition of ESRC grants that you conduct a review of the UK Data Archive to ensure that the data you are planning to create does not already exist.
>
> - Have you searched the web and data archives for similar datasets?
> - Are there any datasets that could assist with your research?
> - How do the existing datasets fail to meet your requirements and therefore require new data to be created?

Performing such a survey may lead to the **re-use** of existing data.[2]

> **Data to be created.** You should list all the data that will be created during the project. The remainder of the DMP then deals with how each item of data will be managed.

Providing an inventory of data associated with a research activity is a pre-requisite for most management and curation activities.

> **Data organisation methods.** Data organisation methods are largely a matter of personal preference and will usually not be of interest to the recipients of the DMP. The exception would be if resources were required for IT infrastructure, software, or training.

2. Again, **discovery** is an issue here, but not in relation to the data covered by the plan.

Organizing data in an intuitive way (or at least a fully documented, explicable way) is a way of **supporting data re-use**. While, in a sense, all that matters is the way the data are organized in the final data case, getting the data into that state is made harder or easier by choices made (or neglected) about how data are organized in the course of the research.

> **Funding and legislative requirements.** List any relevant policies. Some policies (such as data archiving) are relevant to all research projects, whereas privacy will usually be associated with medical and social science projects.
>
> - Does any of your data contain personal information that must be kept confidential?
> - Does your funding agreement require data archiving?
> - Are there any other Data Management requirements in your funding agreement?

Understanding these requirements is key to **legal compliance** and to securing funding. Issues of confidentiality impact on the **shareability** of data.

> **Data owners and stakeholders.** List the owners and stakeholders of the data. Also note who will own any intellectual property created by your research.

Again, understanding intellectual property rights helps towards **legal compliance** and may effect the **shareability** of the data.

> **Access and security.** List who will have access to the research data and what Access Permissions they will have for specific data. If the data will be distributed at some point, list the Access Restrictions and any embargoes that will be used.
>
> Describe how the Access Permissions will be enforced and what IT Security practices will be used. If you have sensitive data, describe any special measures used to store and backup this data.
>
> - Is any data of a sensitive nature?
> - What are the implications of unauthorised access to this data?
> - Are any special measures warranted? (encryption, external hard-drive in locked cabinet/safe)

These pieces of information again delimit the **shareability** of the data and indicate how these limits will be enforced.

> **Backups.** List what data will be backed up and what the backup schedule is. Also mention if any data will be kept under version control and how that will be implemented.
>
> - Is there a backup service already available or will you need to do it yourself?
> - How often will backups occur?
> - Who will be responsible for performing backups?
> - How will sensitive data be backed up?

Here the concern is primarily about **preservation**. There is an overlap between this section and the previous one.

> **File formats, standards, and conventions.** List what formats, standards, and conventions will apply to each data item. Justify the use of particular formats in terms of usability, longevity, suitability for archiving.
>
> - Will other researchers be able to use this format?
> - Will this format be usable in 10 years time?
> - Does your archive accept this file format or can you easily convert to an acceptable format?

The wording here indicates several ways in which the choice of file formats, standards and conventions impacts on **supporting data re-use** and **preservation**.

> **Sharing.** List what data will be made available for other researchers to use.
>
> - What data will be shared?
> - What facilities will be used/required to distribute the data?
> - How will the data be licenced?
> - What access restrictions will be placed on each item of data?

These pieces of information are, of course, about the **shareability** and **delivery** of data.

> **Archiving and disposal.** Estimate the amount of storage space required for archiving, which archive you intend to use, and the whether or not you have discussed your project with the archive manager. If the data is sensitive, describe how you will ensure the data will be safely disposed.
>
> - Which archiving service will be used?
> - How long must you keep your data archived for?
> - When do you plan to archive each item and will they have an embargo period?
> - How much time and resources will be required for archiving?
> - What metadata will be needed for each data type.

These pieces of information cover a range of concerns. While the primary concern is storage, the embargo period is about data **shareability**, and the metadata can support a range of activities including **supporting data re-use** and the remaining aspects of **preservation**. The choice of archiving service impacts on many other data management decisions.

> **Responsibilities.** List who will be responsible for ensuring each item in the data management plan is carried out. Also note who is responsible for reviewing and modifying the data management plan.

These pieces of information are not part of the data management plan as such, but rather statements concerning its implementation. Associating tasks with named individuals

helps to ensure they are completed.

> **Budget.** Now that the data management methods and responsibilities have been established, you can estimate the costs of data management for your project. Often the time involved in documenting, writing metadata, and archiving are underestimated. Make note of any costs associated with using data management services or purchasing equipment (such as fileservers, backup media, software, etc.) used for data management.

Budgeting for data management is of course vital to ensuring it takes place and assuring its financial sustainability.

6 US NATIONAL INSTITUTES OF HEALTH

The National Institutes of Health (NIH) provide three short example data-sharing plans [Nih], one of which is reproduced below.

> The proposed research will include data from approximately 500 subjects being screened for three bacterial sexually transmitted diseases (STDs) at an inner city STD clinic. The final dataset will include self-reported demographic and behavioral data from interviews with the subjects and laboratory data from urine specimens provided. Because the STDs being studied are reportable diseases, we will be collecting identifying information. Even though the final dataset will be stripped of identifiers prior to release for sharing, we believe that there remains the possibility of deductive disclosure of subjects with unusual characteristics. Thus, we will make the data and associated documentation available to users only under a data-sharing agreement that provides for: (*1*) a commitment to using the data only for research purposes and not to identify any individual participant; (*2*) a commitment to securing the data using appropriate computer technology; and (*3*) a commitment to destroying or returning the data after analyses are completed.

As the name suggests, these plans are focused on the **shareability** of data. They describe the nature of the research and, to a lesser or greater extent, the types of data that will be collected. The plans go on to describe any barriers to sharing, and whether and how these barriers may be overcome: the example above mentions anonymization and a data-sharing agreement.

7 RURAL ECONOMY LAND USE PROGRAMME

The Rural Economy Land Use Programme Data Support Service (RELU-DSS) has produced a form to assist successful applicants to the programme in completing a communication and data management plan for their project [RD]. The section on data management includes the following.

> Please list and describe any existing datasets which will need to be acquired for the research to be carried out (third party data sources). Please also identify any specific issues relating to access to these data and how you will overcome any difficulties.

This information relates to the **re-use** of existing data.[3]

> Please list both the quantitative or qualitative data that will be collected or generated by the project, together with a brief summary of each dataset (description and methodology), its format and how it will be managed and stored.

This section asks for detailed predictions of the data to be collected. Providing an inventory of data associated with a research activity is a pre-requisite for most management and curation activities. The description and methodological summary for each dataset aid **discovery** and are part of **supporting data re-use**, as they help other researchers determine the usefulness of the data for their own research, and assess how compatible they are with other similar data. A careful choice of format can support data re-use and **preservation**. The request for information on management and storage is probably aimed at ensuring a sensible preservation regime.

> Please briefly describe the procedures for quality assurance that will be carried out on the datasets (Quality issues to be addressed could for example include: documenting the calibration of instruments, the collection of duplicate samples, data entry methods, data entry validation techniques, methods of transcription).

Quality assurance is important for data purposing, **data re-purposing** and **supporting data re-use** as higher quality data give more reliable results.

> Please describe the data back-up procedures that you will adopt to ensure the data and metadata are securely stored. For example: 'Recognising the susceptibility of hard disks to failure, collected digital data will be transferred on a weekly basis to IOMEGA Zip disks, which will be stored in the University fire safe.' Methods of version control should also be stated.

Backups are primarily a **preservation** issue.

> The Research Councils require all RELU data to be made available for long-term, post-project management within the Research Councils' data centres so they can be made available for secondary research. Do you envisage any difficulties in making any data available, for example, for access constraints or licence conditions of third party datasets used? If so, how might these difficulties be overcome?

This section is concerned with **shareability**, both the barriers to sharing and proposed solutions.

3. Again, **delivery** is an issue here, but not in relation to the data covered by the plan.

> Please state who owns the copyright/IPR of the datasets that you have collected.

Understanding intellectual property rights is important to ensure **legal compliance**.

> Please identify the first point of contact for data management issues, including metadata and quality issues. If different people are responsible for different datasets, please specify below.

Associating data management tasks with a named individual helps to ensure they are completed, both from clarity of responsibility and from archives having someone they can 'chase' if tasks have not been completed satisfactorily.

8 CONCLUSIONS

This examination of guidance on writing data management plans has shown that they relate to challenges as diverse as appraisal, shareability, legal compliance (in terms of both statutory law and contracts, agreements, etc.), preservation, discovery and delivery, as well as those challenges that are the focus for ERIM: data re-purposing, supporting data re-use and data re-use itself.

From the guidance examined, the pieces of information within data management plans that are of most significance for data re-purposing and re-use are 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

There are also some pieces of information that have 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 must of course be remembered that while the other challenges mentioned above are not the focus of ERIM, they still represent points of failure that could prevent the eventual re-use of data. If data are not shareable they cannot be re-used; if they are not preserved properly they cannot be re-used; if future researchers cannot find them they cannot be re-used; and so on. Even though ERIM focuses on a narrow set of issues, the place of the other issues must also be acknowledged, and these issues addressed outside the Project.

REFERENCES

[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).

[Inf08] Information Literacy Program (2008-08-05). *ANU Data Management Manual: Managing Digital Research Data at the Australian National University*. Version 1.1. Australian National University. URL: http://ilp.anu.edu.au/dm/ANU_DM_Manual_v1.01.pdf (2010-07-12).

[Jon09] S Jones (2009-03-30). *A Report on the Range of Policies Required for and Related to Digital Curation*. Deliverable H1.1. Version 1.2. Digital Curation Centre. URL: http://www.dcc.ac.uk/docs/reports/DCC_Curation_Policies_Report.pdf (2010-06-01).

[Jon10] S Jones (2010-01). *Summary of UK Research Funders' Expectations for the Content of Data Management and Sharing Plans*. Digital Curation Centre. URL: http://www.dcc.ac.uk/sites/default/files/documents/publications/UK%20research%20funder%20expectations%20for%20data%20plan%20coverage.pdf (2010-03-25).

[Mit] *Data Management and Publishing* (2009-07-16). MIT Libraries. URL: http://libraries.mit.edu/guides/subjects/data-management/ (2010-07-01).

[Nih] *NIH Data Sharing Policy and Implementation Guidance* (2003-03-05). National Institutes of Health. URL: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm (2010-07-12).

[RD] Rural Economy Land Use Programme Data Support Service. *Project Communication and Data Management Plan*. UK Data Archive. URL: http://www.data-archive.ac.uk/relu/DMP.doc (2010-07-13).