IJDC | *General Article*

Remediation Data Management Plans: A Tool for Recovering Research Data from Messy, Messy Projects

Clara Llebot
Oregon State University Libraries, Oregon State University

Abstract

Data Management Plans (DMPs) have been used in the last decade to encourage good data management practices among researchers. DMPs are widely used, preventive tools that encourage good data management practices. DMPs are traditionally used to manage data during the planning stage of the project, often required for grant proposals, and prior to data collection. In this paper we will use a case study to argue that Data Management Plans can be useful in improving the management of the data of research projects that have moved beyond the planning stage of the research life cycle. In particular, we focus on the case of active projects where data has already been collected and is still being analyzed. We discuss the differences and commonalities in structure between preventive Data Management Plans and remedial Data Management Plans, and describe in detail the additional considerations that are needed when writing remedial Data Management Plans: the goals and audience of the document, the data inventory, and an implementation plan.

Submitted 12 February 2018 ~ *Accepted* 12 February 2018

Correspondence should be addressed to Clara Llebot. 121 The Valley Library Corvallis OR 97331–4501 Email: clara.llebot@oregonstate.edu

An earlier version of this paper was presented at 13th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v13i1.667

Introduction

Motivation and Case Study

The Watershed Research Cooperative (WRC) is a project whose goal is 'to conduct research on the effects of current and expected forest practices on intensively managed commercial forestland on water quality, fisheries, and other water-related values'[1]. The director of the WRC contacted Research Data Services at Oregon State University Libraries and Press in the summer of 2016 because, after 15 years of data collection without a consistent data management plan, it had become 'apparent that there is a need for additional investment in data management if the benefits from the WRC studies are to be realized the project needed help' (Souder, Hatten, Ganio & Bladon, 2016). In other words, the project was in the synthesis phase and the lack of consistent data management practices during the project was making the publication of the results extremely difficult. In addition, newly hired post-docs were having difficulty integrating data across disciplines within a particular study, as well as data across studies within disciplines. The WRC is a complex and ambitious project: it has involved more than 50 researchers during the last 15 years, with funding from two federal agencies, two state agencies, two forest industry associations and three private timber companies. The WRC is structured into subprojects (three geographically distinct paired watershed studies) that have mostly independent teams of researchers; and it is a data intensive project that has generated large volumes of data. This complexity led the directors of the project to write a proposal (Souder et al., 2016) to the Fish and Wildlife Habitat in Managed Forests Research Program (FWHMF) in the College of Forestry at Oregon State University. The goal of the proposal was to 'design a WRC data stewardship and management framework', and to 'begin structuring the WRC datasets to meet data publication standards' (Souder et al., 2016). The proposal was funded and as a result the WRC was able to buy time from the Data Management Specialist of the Oregon State University Libraries (OSU Libraries), the author of this article, to take on the project.

The result of the project is a Data Management Plan, to be finalized in 2019, that the WRC will use as a guide to organize, document, share, and archive their datasets. The project has worked on the assumption that a data management plan can be useful as a remediation tool to recover data that has been generated in the absence of consistent data management practices. I will use this case study to illustrate the WRC Data Management Plans' differences and commonalities with the more traditional Data Management Plan that accompanies a grant proposal.

Context of Data Management Plans

Research groups are increasingly geographically dispersed (Wagner, Whetsell & Leydesdorff, 2017), often interdisciplinary, and employ techniques that generate immense amounts of raw data and analyses (Adams, 2012). As a consequence, managing research data is becoming a challenge for researchers. Data management is also in the spotlight because of concerns related to data reproducibility, integrity, and the reusability of science

[1] Watershed Research Cooperative: http://watershedsresearch.org/

for new research; all of these concerns are especially challenging when research data is not fully curated and openly available. Sharing research data and research results openly also holds researchers accountable for publicly funded research.

Data Management Plans (DMPs) are widely used tools that have been used in the last ten years to encourage good data management practices among researchers. The idea is that a researcher writing a grant proposal is required to have a plan for how research data will be managed if the project is funded. The plan includes the roles and responsibilities of the members of the team, an explanation of the expected data, specification of data formats and documentation, policies for data dissemination and policies for public access, and data storage and archiving.

DMPs are a preventive tool to encourage good data management practices. DMPs are traditionally used to manage data in the planning stage of a project and are usually included with the grant proposal, when data has yet to be collected. However, this proactive approach will only take care of future research. It is easy to imagine that we, at this moment, have huge amounts of data that have been produced or that are being generated right now, that are not guided by Data Management Plans and that are disorganized, undocumented, and the future storage of which may be in the garage of a retired researcher in a cardboard box. Unfortunately, there is no estimate of how much data is, or has been, produced. In this article we argue that DMPs are an effective tool to manage data in later stages of the research life cycle, not just the planning stage. We will use the case study of the Watershed Research Cooperative to show how a Data Management Plan can also be a remediation tool that can help recover already existing data from an active project. The WRC is an example of a project in an intermediate stage in the data life cycle. We define the intermediate stage as projects in which some or all of the data has already been collected; some of the data may have been published and archived; some of the data is or has been analyzed; and the interpretation of the data and writing of manuscripts is still underway. In this paper we will use the terminology pDMPs for preventive Data Management Plans, done in the planning stages of a project, and rDMPs for remedial Data Management Plans, written during the intermediate stages of a project.

In this paper we will discuss the areas of a rDMP that are fundamentally different than a pDMP (Section Differences Between rDMP and pDMP), and we will outline the commonalities between both types of plans (Section Commonalities Between rDMP and pDMP). In each section we will describe the case study and discuss general considerations for all rDMPs.

Differences Between rDMP and pDMP

Goals and Audience

Case study

The main challenge that a data management strategy needs to address for the WRC is the size of the project, both in terms of datasets generated, and number of researchers participating in the project. The goals and audience of the rDMP for the WRC are aimed at two levels:

 doi:10.2218/ijdc.v13i1.667

1. The work that will need to be done to organize, document, share and archive the WRC data is so vast that the WRC will need resources (time, money, expertise) to implement the rDMP; because of this, the rDMP is not just a tool to plan the best strategies for data organization, documentation, sharing, and preservation. The rDMP also describes the magnitude of the problem, presents priorities, and is a document that will be used to inform and convince administrators and program managers of the necessity to invest resources for the implementation of the plan.

2. The WRC research has been performed by dozens of researchers of different backgrounds. Each researcher has a different level of awareness and expertise on data management, and different priorities for their research. As in every large project, the researchers of the WRC have developed relationships, often generating fruitful collaboration but also, in some cases, conflict and disagreement. Disagreements extend to opinions about how the data should be managed. One of the goals of the WRC rDMP was to start a conversation among researchers and administrators to reach agreements on how to manage the data until the end of the project, and whether some of these data management strategies should be enforced or not. In this context, the WRC rDMP is a document that outlines recommendations and informs. Again, it is a document directed to administrators and managers, rather than to researchers.

General considerations

The case study illustrates that rDMPs are a flexible tool that can be used for many goals. The ultimate goal of a DMP is to end a research project with a dataset that is organized, documented, shared in a secure way, and preserved. The paths to accomplish this goal will vary depending on the project and the project stage. A pDMP is usually written by researchers for researchers. An rDMP could certainly be written by researchers for researchers also. However, as the case study illustrates, an rDMP may be used for other intermediate goals, and addressed to a different audience, such as funders that will decide to give resources for data management, or project managers that may decide to enforce a series of data management practices for a group of researchers.

Data Inventory

Case study

The volume of files in the WRC was so large that a data inventory based on collecting information of individual files, or even groups of files, was quickly identified as impossible. The WRC data inventory used three strategies: (1) Classification of data into groups according to how the data has been managed. This was accomplished through interviews with the people identified as data managers or as responsible for a particular dataset. (2) Analysis of number, size, format, and date of files. This task was automated with a disk space analysis tool. (3) Classification of data in thematic groups (labeled Datasets) according to the field of study that they adhered to.

Data from the WRC was summarized into five groups using this strategy. There was no sensitive data in this project, all data had already been collected, and none of the datasets had been published or made publicly available in any way at the time that the rDMP was written. Researchers responsible for each dataset were identified in the data

inventory, as well as data managers when they existed.

1. **Relational databases**: data organized and managed by a data manager, organized in three similar relational databases, one for each of the paired watershed studies.
    - Dataset: climate, hydrology, temperature, nutrients, dissolved oxygen and sediments.
    - Management strategy: the data manager receives data from individual researchers, quality controls the data, documents the data, and organizes it in a relational database. Manager also prepares data requested by researchers to conduct research.
2. **Trask tabular data**: data from the subproject Trask (one of three subprojects)
    - Dataset: macroinvertebrates, amphibians, geospatial, isotopes, primary productivity, vegetation
    - Management strategy: for tabular and geospatial data (the bulk of the data), the group adopted a data and metadata template developed for the HJ Andrews Experimental Forest (Long-Term Ecological Research project). Many of the researchers from the Trask subproject of the WRC work on the HJ Andrews Experimental Forest, and are familiar with the templates and its advantages. The templates for tabular data are in Excel format and follow the structure of the Ecological Metadata Language standard. This strategy is required for all researchers in the Trask subgroup.
3. **Fish database**: this database has been managed by a researcher acting as a data manager (covers all three subprojects).
    - Dataset: fish and fish habitat
    - Management strategy: the data manager collects data, deposits it in the database, and prepares it for researchers. Data quality control and documentation are responsibility of the researcher, not of the data manager.
4. **Other digital data**: the rest of the data generated by the WRC in a digital format.
    - Dataset: climate, hydrology, temperature, nutrients, dissolved oxygen, sediment, macroinvertebrates, fish, geospatial data
    - Management strategy: no consistent management strategy. Each researcher makes decisions about their dataset.
5. **Physical samples**: collection of more than 10,000 paper filters of water samples used to calculate suspended sediments.
    - Dataset: sediment
    - Management strategy: a data manager organized the samples, that are stored in a room at college facility. There are no standard procedures to get access to the samples, and some have been lost already due to the lack of consensus about what other researchers can or cannot do with these samples.

 doi:10.2218/ijdc.v13i1.667

The Trask tabular data, the fish database, and the other digital data categories are stored in a shared drive managed by the IT staff at Oregon State University's College of Forestry. Overall, they hold 388,167 files with 344GB of information. Not all the files are research data files, but it is difficult to distinguish these that are from other type of files (administrative, data from other projects, etc). The most common type of content is data in a tabular format, followed by images, databases, presentations and text, and geospatial files. Overall, the most common type of files are Excel files (a total of 49,549 files that take more than 73GB of space). Images are the second most common files, accumulating a total of 80GB and 52,459 .jpeg and .tif. Access databases take up more than 16GB with 334 files, and PowerPoint presentation files accumulate about 30GB of space in more than 8,000 files. Finally, geospatial file formats account for more than 22GB and 16,000 files. With so many files, the quality varies widely. Some files, especially the ones that use data and metadata templates, are well documented and organized. However, the majority of the files, specially the ones in belonging to the "Other digital data" category, lack documentation and do not follow good data organization strategies.

General considerations

The data inventory is a time consuming and essential part of the rDMP, unnecessary for a pDMP by definition (because data has not yet been collected in a pDMP). The data inventory in a rDMP substitutes, in general, the pDMP data description section, although an rDMP may include a data description section if the project is still planning on collecting data. The goal of a data inventory is to identify the data that the project has generated, and some metadata about it:

- Subject of the dataset: area of knowledge that identifies the dataset (a short description of the data in the dataset);
- Location of the dataset: where is the dataset stored, and what are the measures that have been taken to keep the data safe (e.g. backups, encryption, password protected, etc);
- Responsibilities: researcher or researchers responsible for the dataset - these researchers get to make decisions, such as who gets access to the dataset;
- Manager: researcher or researchers who perform day to day actions to manage the dataset, including collection, organization, quality control, version control;
- Versions: how many versions of the dataset exist and how they are different;
- Formats: formats of the files and data standards, if they exist;
- Documentation: whether there is documentation for the dataset, where it is, and how complete it is, and a description of metadata standards, if they exist;
- Sensitivity: whether data are classified as sensitive, protected, or require any kind of protection;
- Sharing status: whether the data has been shared internally (within the group) or externally (made publicly available) or not.

The data inventory strategy will need to adapt to the size and characteristics of the project. Collecting all this metadata about every data file would be the ideal situation,

but it is an unrealistic scenario for many medium or large projects, as the case study illustrates. The three steps taken for the data inventory in the case study (identification of data management strategies, identification of subject specific datasets, and file analysis) are general enough that they can be used for virtually any project, and are functional for writing a rDMP.

Implementation Plan

Case study

The implementation plan of the rDMP for the WRC will be a lengthy process that will take years. This section includes two parts: setting priorities of datasets that should be revised, and estimating resources needed. The rDMP sets three levels of priorities:

1. High priority: Clean, document, and preserve in ScholarsArchive@OSU, the institutional repository, quality-controlled datasets, using article publications as triggers. There are a few key datasets that are necessary for most researchers (e.g. almost everybody needs hydrology to interpret their results). These will be used as pilot datasets and will serve as an example to researchers when working on the organization, documentation, and preservation of their data. It is also high priority to maintain and archive the relational databases created for the project.

2. Mid priority: Clean, document, and preserve data associated to past publications.

3. Low priority: Triage data in shared drive folders. Eventually, the data in the shared folders will need to be addressed. By taking this as a last step we ensure that most of the important information will have already been identified and moved to a secure storage space.

The resources needed to implement the rDMP are challenging to estimate because it depends on the level of implementation that the group decides to take. This question has to be decided by the researchers and managers of the WRC, and therefore cannot be part of the plan. The rDMP includes scenarios so that it is clear what the minimum standards are (e.g. prepare well organized datasets in open formats with documentation in a readme file to preserve in ScholarsArchive@OSU), and what the ideal scenarios are (e.g. prepare datasets using a metadata standard, including keywords from a controlled vocabulary; deposit all the versions of the dataset to ScholarsArchive@OSU, including the raw and quality controlled datasets and also the data and code used to generate plots in every published article).

The rDMP suggests the possibility of hiring post-docs or faculty research assistants to work on specific datasets. It also talks about the incentives for the researchers to work on the data following the recommendations of the rDMP (increased efficiency for the researcher, increased visibility and impact for the researcher, reproducibility and acceleration of scientific breakthrough). The rDMP recommends that a decision will be made at a management level, taking the opinion of the researchers into consideration, so that the WRC can take a consistent approach to data management. In this sense the rDMP is acting as an educational tool so that administrators, managers, and researchers will understand the importance of the measures outlined in the rDMP.

 doi:10.2218/ijdc.v13i1.667

General considerations

The implementation part of the rDMP is one of the most challenging parts, but necessary to make sure that action will be taken. This is necessary in a rDMP, but it is not in a pDMP because when data management is considered during the intermediate stages of a project, something needs to change. After all, the least disruptive option is to continue with business as usual. One of the advantages of rDMP over pDMP is that the researchers have actually seen the negative effects of not managing data, and are more motivated to act.

The plan implementation section discusses how the rDMP will be implemented by setting up priorities or stages of implementation; it estimates the resources that are necessary (in terms of money, personnel, time); it identifies the individuals who will take on new data management responsibilities. When rDMP are written by researchers for their own project, this section can be very specific. In other cases, such as the case study, this section speaks directly to administrators and managers because there are choices to be made, and the responsibility of motivating researchers to act, via carrots or sticks, is theirs.

Commonalities Between rDMP and pDMP

Case study

The bulk of the rDMP for the WRC can be summarized as follows.

Organization

To organize the data the rDMP defines four data quality levels: a level zero (L0) of raw data downloaded directly from an instrument or model; a level one (L1) of raw data in a human readable format; a level two (L2) of verified data that have undergone quality control, including but not limited to detecting sensor malfunctioning, assessment of outliers, calibration, and corrections for sensor drift or offset; a level three (L3) of L2 data that have been analyzed to answer specific research questions. Typically, L3 data is used to create figures in a scientific publication. Other data organization strategies outlined by the rDMP are the definition of Datasets or Data Collections as data units that make sense from a disciplinary point of view, with unique names; propose a folder organization structure (based on the organizational structure of one of the subprojects, that we propose should be extended to the whole project); propose a file naming strategy for data files.

Metadata

A discipline specific metadata standard has been identified: Ecological Metadata Language. Tools to implement it have also been identified (Morpho, R), as well as a controlled vocabulary of keywords [2].

Sharing and preservation

Data will be published through Oregon State University's institutional repository, ScholarsArchive@OSU. The minimum requirements for depositing data in ScholarsArchive@OSU are outlined (documented, tidy data in open formats). L2 data will be

[2] See: (http://vocab.lternet.edu)

shared for all datasets, and sharing of L3 data is highly encouraged. Data will be made publicly available under a Creative Commons Universal (CC0) or a Creative Commons Attribution (CC-BY) license.

General considerations

Many sections of a rDMP are essentially the same as in a pDMP. The initial pDMP section of "Types of data produced" becomes a "Data inventory" section (discussed above) and a "Data organization" section, where the rDMP discusses strategies to organize and quality control the data. "Data and metadata standards" are discussed in an rDMP and a pDMP in a similar way. Data and metadata standards are identified, and if these do not exist or are not appropriate, the plan proposes solutions or remedies. Both rDMP and pDMP include a section about "Policies for access and sharing" where they discuss how to protect the privacy or confidentiality of the data, and identify any intellectual property or other rights issues. rDMPs and pDMPs also include a section on "Policies for re-use and distribution" that outlines how the data will be reused and distributed, and if there will be restrictions to access. Finally, "Plans for archiving and preservation" describe long term strategies to preserve the datasets.

Discussion

The conclusions outlined in this paper are reflections from the author as a result of writing the Data Management Plan for the Watershed Research Cooperative. The plan has not been implemented and, therefore, we cannot evaluate yet the success or failure of the strategy drafted in the rDMP. What this article hopes to demonstrate is that the structure of a DMP is a useful framework to manage datasets that are at different stages in the research life cycle than the planning stage.

This article defines rDMPs as Data Management Plans that are written at the final stages of research projects, after data has been collected, analyzed, and quality controlled – partially or fully– but before the project has finished and the data have been archived. This stage brings certain advantages: researchers are still highly involved in the data, and many important details are still remembered; the problems brought by bad data management practices are not a distant scenario, perceived as improbable or unimportant (as is sometimes the case with pDMPs), but a reality; the benefits that a rDMP can bring to the project will benefit not only the field of study or the community (as will happen when data is archived and made publicly available), but will, most likely, benefit also the individual researcher via increased efficiency. This stage also brings challenges: it is the stage that brings more conflict among researchers within the same group that are trying to share their data, and conflict is more likely to occur when data has been badly managed; the data is still changing and in some cases some of it may still need to be collected. Planning for past and future data at the same time can be challenging.

The fact that rDMPs can be useful in this particular stage of the research life cycle suggest that they may also be useful in other stages. Of particular interest would be the post-project state when the project has finished, the research has been published, and the data has been archived. The use of data management plans for legacy data, bringing in the role of archivists, is an interesting field to explore.

Hopefully the number of projects –especially large projects– executed without a

pDMP will keep diminishing, making rDMPs an unnecessary tool. The mandates from federal agencies in the US and Europe, and many other non-federal agencies, certainly suggest that they will. There is no data, known to the author, that approximates the number of projects being executed without pDMPs each year. There is also no data about the amount of legacy data that exists from projects that have been finalized. The existence of a rDMP framework such as the one described in this article could facilitate the task of recovering these kinds of datasets.

Acknowledgements

Thanks to Jon Souder and Laurie Bridges for their help during the project. This work was funded by the Fish and Wildlife Habitat in Managed Forests Research Program (FWHMF) in the College of Forestry at Oregon State University.

References

Adams, J. (2012). Collaborations: The rise of research networks. *Nature*, *490*(7420), 335–336. doi:10.1038/490335a

Souder, J., Hatten, J., Ganio, L. & Bladon, K. (2016). *From chaos to consistency: Moving towards data stewardship and sharing for the watershed research cooperative.* (Project proposal to the Fish and Wildlife Habitat in Managed Forests Research Program)

Wagner, C. S., Whetsell, T. A. & Leydesdorff, L. (2017, 1st Mar). Growth of international collaboration in science: revisiting six specialties. *Scientometrics*, *110*(3), 1633–1652. Retrieved from https://doi.org/10.1007/s11192-016-2230-9 doi:10.1007/s11192-016-2230-9