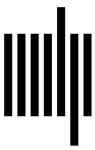
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022



The MIT Press

8 Developing a Data Management Plan

Susan Smythe Kung

1 Introduction to DMPs

A data management plan, commonly referred to as a DMP, is a written document that outlines a researcher's long-term and short-term plans for generating, handling, describing, organizing, processing, analyzing, preserving, and sharing the data resulting from a research project.[1] A DMP includes detailed procedures for data collection; all aspects of organization and processing before the data are shared or disseminated; and a plan for how the data will be released so that they can be found and accessed in perpetuity, with proper attention paid to relevant legal and ethical concerns. The DMP is the road map or guidebook for how the data will be handled during every phase of the research life cycle. The very act of writing a DMP can be enormously beneficial to researchers as it requires them to think through the logistics, ethics, and expenses associated with carrying out their proposed research. Though the purpose of writing a DMP is to establish a prescribed program for managing data, researchers should understand that they will need to review their DMPs periodically and revise them as necessary (this will be discussed in more detail in section 3).

While some researchers might never be required to write a DMP (NSF 2018), research data management (RDM)[2] is something that all researchers must do throughout their education and careers. RDM is not something that researchers simply write about and then forget, but rather they must practice it on a daily basis to do good research, keep their data orderly, and produce valid and reproducible results. Thus, for each research project that is undertaken, the researcher should create a comprehensive DMP that covers every aspect of data management during the data life cycle and that can be modified to satisfy the DMP requirements of any given research funder, publisher, or organization.

Creating a DMP will help you budget for all aspects of data management, including activities associated with generating, storing, analyzing, anonymizing, and archiving the data. Many researchers, when they plan their research budgets, include only the costs associated with the activities of data collection, storage, and analysis, and they do not think to include costs associated with things like anonymization or long-term digital preservation (archiving) of their data. You should learn what resources are available to you at your institution regarding RDM services such as storage, computational processing, preservation, archiving, and such.[3] Be practical and make sure that you do not include anything in your DMP that is beyond your resources; be realistic and plan accordingly with the resources that are available to you. It is especially important to get in touch with the archive or repository that you plan to use for the long-term preservation of your data to find out their requirements for format types, file names, documentation and metadata standards, deposit size limits, rights and restrictions, fees, time lines and such (for more information, see Andreassen, chapter 7, this volume). Build these requirements into your DMP from the beginning.

Make the first draft of your DMP as comprehensive as possible and do not worry about a page limit. Although the resulting DMP might be too long and overly detailed to submit to a particular funder or publisher, it will serve as a comprehensive plan for you (and your team) to follow, and it can be modified to make it applicable to any purpose, including tailoring it to the requirements of a particular funder (see section 3 for information on revising and repurposing your DMP).

Many university websites or librarians will point researchers to the DMPTool[4] or other templates that are specific to different funders. Because these tools and templates were designed to satisfy the DMP requirements of

various funders, they put more emphasis on the inclusion of funder requirements than on the inclusion of research activities themselves, and they put almost no emphasis on the resources needed to carry out the research activities (Williams, Bagwell, & Nahm Zozus 2017). Nevertheless, these tools are useful for tailoring a more comprehensive DMP into something shorter that will satisfy the requirements of a particular funder. While these tools can help you to create a DMP that meets the requirements of a particular funder, they cannot do the work of planning how you will manage your research data with the resources, budget, personnel, and time that you realistically have at your disposal. Only you can write an accurate and adequate DMP that will chart the course for how you will manage your research data.

Section 2 will guide you through the component topics that you should address in your DMP, including data generation, analysis, and handling (section 2.1); the legalities and ethics of data generation and use (section 2.2); data storage, backup, and security (section 2.3); data documentation and metadata (section 2.4); data dissemination, preservation, and sharing (section 2.5); and your responsibilities and the time line for carrying out your RDM (section 2.6). Section 3 provides guidance on when and how to revise your DMP, as well as how to adapt it to the requirements of funders. Section 4 summarizes the importance of a comprehensive yet flexible DMP and lists additional resources that you can turn to for help.

2 Components of DMPs

The aim of this section is to introduce you to various elements of a comprehensive DMP. Keep in mind that some components might not be relevant to your particular project. At the end of each subsection, there is a list of questions for you to consider with respect to your project; these questions are labeled as figures 8.1–8.6, whose numbers correspond to the associated subsection. You might find it helpful to read the list of questions for each subsection prior to reading the text of that subsection. As you read the text, answer the questions and take notes about the activities, resources, and issues that are relevant to your research circumstances. If you answer the questions as you go, you will have the necessary scaffolding on which you can then build your comprehensive DMP by the time you finish reading this chapter.

2.1 Data generation, analysis, and handling

A key concept that must first be established when writing a DMP is the definition of *data*. Because every subdiscipline of linguistics uses different types of data, and every project within that subdiscipline might use a subset of those types of data, there is no single generic definition that works for all DMPs. Rather, every researcher must explain what they consider to be data for their given research project (see Good, chapter 3, this volume). Furthermore, thinking about the *content* of the data that you plan to collect (e.g., paradigms, conversations, grammaticality judgements) will help you determine the *digital parameters* (file types and formats) of that data. For example, a cognitive linguist might make audio and video recordings of experimental protocols and later code those recorded experiments in spreadsheets; the digital data parameters might include .wav audio files, .mov video files, and .csv spreadsheets. When writing a DMP for a funder, the content of the data should be described in the proposal while the digital parameters should be made explicit in the DMP. However, a comprehensive DMP should include both.

The digital parameters of the data should also be described in terms of the best practices that are used for collecting particular data types. Best practices for data collection can vary depending on the intended use of the data, so they should be clearly articulated in the DMP. Furthermore, best practices are subject to change or revision over time, so it is important to document the best practices that are recognized by your field at the time that you collect your data.

Once you have established the digital parameters of the data you intend to collect or generate, you should next specify any equipment, software, and/or other tools that you will use to both generate and analyze the data (see Han, chapter 6, this volume). For example, you might use a Zoom H6 digital recorder to collect audio recordings of tokens in .wav format that you will later analyze with Praat software for phonetic analysis to create XML files in .TextGrid format, and you might code vowel formants using a spreadsheet in .csv format.

If you plan to create any analog data (non-digital data such as drawings, sketches, diagrams, notes, and so on), describe them, and explain how those data will be handled and analyzed. If they will eventually be digitized, list the resulting digital formats.

Next, consider any data you plan to use that you will obtain from other sources, such as archives, data

corpuses, collaborators, among others. What formats will those data be in and what formats will you create as you work with them? If you plan to use any existing data, you should explain the source of these data, as well as their types and formats and any changes that you will make to them. If the data are analog, explain how you will get them into a digital format that you can use for your project.

It is important to note whether any of the data must be kept private, confidential, or restricted in some way. Explain in detail the subset of the data that will be affected and what measures will be undertaken to maintain privacy, confidentiality, or restrictions. This must be done for all data types, no matter the source.

If you plan to create or generate any raw data (e.g., run experiments or make audio and video recordings), you will need to establish a plan for differentiating the raw data from your working files. And no matter what kind of data you are using—raw or working—you need to establish a system for version control. The need for quality assurance and/or quality control procedures varies greatly among (sub)disciplines, projects, and data types, so make sure to describe any procedures that you plan to implement for your project (e.g., training activities, computer visualization, computer or human review).

If you plan to use a research management tool, such as a database, name the tool and describe how you will use it. Research management tools can vary greatly between different disciplines, including different subdisciplines of linguistics, and they might come and go faster than you can plan and carry out a project.[5]

If you will have a team of people working on this project (even a team of two), you should establish the roles, responsibilities, and tasks of each team member with respect to data generation, analysis, handling, and quality assurance. Describe team members in terms of their project titles, responsibilities and tasks, qualifications or training to do these tasks, and allowed access to different data types, particularly with respect to confidential data.

Finally, do not put anything into your DMP that you do not understand. Do your homework and research things such as tools, software, or metadata schema to use and best practices in your field to follow. Do not put something into your DMP just because your colleague or friend did and especially if you do not fully understand what it is. Misunderstood tools, practices, and the like

After or as you read section 2.1, "Data generation, analysis, and handling," answer the following questions:

1. What kind of data will your project produce?
    a. What type, format, and amount of digital data will you produce?
    b. Will you be creating any analog data? If so, what kind?
    c. Will the analog data be digitized? When? In what format(s)?
    d. Will you reuse existing data? If so, describe those data and their source(s).
2. What best practices for data collection/generation (relevant to your subdiscipline and the data types) will you follow for collecting the data?
3. What equipment, software, and/or other tools will you use to generate and analyze the data?
4. Will any of the data need to be kept private, confidential, or restricted? If so, explain why and how that will be accomplished.
5. How will you differentiate raw data from working data files?
6. How will you manage file versioning?
7. Will you need to implement any kind of quality assurance or quality control protocols for generating or handling the data? If so, describe them.
8. Will you use any research management tool(s) or software?
9. Will anyone besides you be generating or handling the data? If so,
    a. Who are the team members?
    b. What roles will they play in the research?

Figure 8.1
Questions: Data generation, analysis, and handling.

will be glaringly obvious to the experts who will read your DMP, and those misunderstandings could work against you in the long run.

2.2 Legalities and ethics of data generation and use

Some of the legal and ethical issues associated with carrying out research projects include data ownership of newly generated data, the intellectual property associated with data generated during the project as well as existing data that might be used, and ethical considerations for some or all of the data. This section will not go into great detail regarding the intellectual property law or ethics associated with RDM (see Collister, chapter 9, this volume, and Mattern, chapter 5, this volume, respectively, for more detailed discussions of these topics); rather, it explains what needs to be covered in a DMP.

When planning a research project, many people do not stop to think about who will own the data, or its inherent intellectual property (IP),[6] that they collect or create. Some never even think about ownership, while others assume that they, as the project owner, will also own the data and any associated IP. However, that is not always the case. In actuality, the owner might be the research funder, the university or lab that sponsored the funded research, the researcher(s) who designed the project and got the funding, the researcher(s) who collected the data, the language consultants or project participants who provided the data by sharing their personal histories or cultural knowledge, the communities to which these language users belong, or someone else. In the United States, it is often the case that the research institution or university that sponsored the research owns the data and the inherent IP (Blum 2012, cited in Henderson 2016). This means that while the researcher is allowed to disseminate (publish, present) ideas, theories, and conclusions drawn from that data—and thus own the copyright to those publications—that researcher's university owns the actual data. If you are affiliated with a university, or some other research institution, regardless of your status (faculty or student), you should find out whether this is the case at your institution. If you are a staff member or someone who was hired specifically to work on a sponsored project, then the work you do is likely to be considered *work for hire*, depending on the terms of your employment; if it is, then you might not own the data or the IP, depending on the laws of the country where you work. Because every university or institution is different, it is in your best interest to investigate the policy of your institution and your country.

If your project is funded by a private funder, then that funder might own the data and the inherent IP. Read the fine print and ask questions so that you will understand your legal rights to use the data for your particular project both while it lasts and after it ends.

Moreover, the details of copyright laws vary from one country to another, so data created in one country might be subject to different IP laws than data created in another country. If you are working in multiple countries, you might want to consult a copyright professional at your institution or the WIPO Lex,[7] an IP database maintained by the World Intellectual Property Organization. Thus, it is important to explain in the DMP—according to country-specific IP law—who will own any data produced by the project.[8]

If you plan to reuse existing data, you must explain in the DMP your legal right to use that data (e.g., a nonexclusive license from the copyright holder, fair dealings or fair use, public domain)[9] or how you will get permission to reuse that data. Bear in mind that there might be costs associated with using certain data sets, such as fees to use a particular database or corpus or a license fee to reproduce a recording. Be sure to consider these fees as part of the costs of RDM. It is your responsibility to investigate the legalities of the ownership and use of the data that you plan to work with. Once you understand who owns the data and its IP, make the details of ownership explicit in your DMP.

Ethical issues can affect how the data are collected and stored, who can access or use the data, how long the data may be kept, and whether the data must be destroyed at the end of the project. Your DMP should address such ethical issues that are covered by institutional review boards (IRBs) or research ethics boards (REBs), such as informed consent, anonymity or deidentification, and privacy, as well as ethical issues that might not be covered, such as the need to protect Traditional Knowledge and long-term data archiving or preservation.

A key area in which the accepted ethics of working with human subjects for a linguistics project differ from other disciplines is in the anonymization of data. Rather than automatically anonymizing linguistic data, it is an accepted practice to give each project participant (i.e., language user) the opportunity to be named in the project metadata so that they will receive proper attribution

for their part in the research.[10] They must also give their consent for the project data to be put into a repository or archive without being anonymized. If they do not give their full consent, the data must be anonymized or excluded according to their wishes. This practice of proper and faithful attribution applies to all project members as well, including interviewers and interviewees, transcribers and translators, and annotators and coders of the data. While this practice is especially prevalent for language documentation projects, it is spreading to other subdisciplines in linguistics (Berez-Kroeker et al. 2018).

There are research environments in which you might be required to enter into a contract or memorandum of understanding (a non-legally binding agreement between two friendly parties) to gain access to a research location or population, such as an industry workplace or a tribal reservation. If this is the case, seek legal advice from your institution's legal department or somewhere else.

For some types of projects, such as language documentation or acquisition projects, you might have an ethical responsibility to digitally repatriate (return) some or all of the collected data (Kung 2021). If so, your DMP should explain how and when you will do repatriation activities, and you should take the associated costs into consideration when planning your project budget. It is quite simple to copy files onto external storage media such as a USB drive or share the files via file sharing or storage systems, but these costs need to be planned for and built into the project from the start.

2.3 Data storage, backup, and security

In the DMP, researchers must lay out a clear plan for storing and backing up data, migrating them as necessary, and keeping them secure. This plan should include an estimation of the approximate volume of data to be collected and stored; a plan for how the data will be securely stored and redundantly backed up during all phases of the research project; and an explanation of who will have access to the data, how they will be given access, and how personal information about research participants will be protected.

Most researchers have experienced some sort of data mishap in which they have lost access to an important file or to the latest version of that file. The loss might

After or as you read section 2.2, "Legalities and ethics of data generation and use," answer the following questions:

1. What parts of your data are subject to copyright?
2. Who will own the intellectual property (IP) rights to the data that you generate?
    a. Are you working under any contracts or terms and conditions? If so, what are the terms of "ownership"?
    b. Are you working in different countries that might have different IP laws? If so, investigate the IP issues for each country.
3. If you are reusing data, do you need to get permission or a license to do so?
    a. Are there any fees associated with reusing existing data?
4. What are the ethical considerations associated with the data?
    a. What are the requirements of your institution's REB or IRB?
    b. How will you obtain informed consent?
    c. Will you need to anonymize or de-identify the data?
    d. What country-specific privacy laws will apply to the data?
    e. Do the data include Traditional Knowledge that needs to be protected?
5. What ethical or copyright issues might be associated with archiving and sharing the data?
6. Will you have to enter into a contract or memorandum of understanding to engage in research in your field site?
7. Will you need to digitally repatriate the data to the research community or participants? If so, how will you do that?

Figure 8.2
Questions: Legalities and ethics of data generation and use.

have been due to any number of events both in their control (user errors such as failing to save work or accidental deletion) and out of their control (system errors such as a failed automatic backup, loss or theft of a hard drive, unexpected power outages or surges, and so on). Furthermore, different storage media types are subject to different types of issues. External media and laptops can be lost, stolen, dropped, erased, or overwritten. Servers can go down or be inaccessible during maintenance. Cloud storage might not be available or allowed. External storage media (hard drives, USB drives, DVDs) have limited life spans. Storage and backup failures will happen, so it is extremely important to have a backup plan in place at the outset of the project.

There are a few mnemonics in the literature on RDM that are designed to remind us of the importance of storage and backup. *The 3-2-1 rule* (Leopando 2013, cited in Briney 2015) says to keep three copies on (at least) two types of storage media in (at least) one off-site storage location. The 3-2-1 rule would be satisfied by keeping three copies of the data on your lab computer and in cloud storage (i.e., two media types, one off-site).[11] While *LOCKSS* is the name of a digital preservation solution used by some academic libraries, it also refers to a good practice for personal data preservation: lots of copies keep stuff safe.[12] In your DMP, you should lay out your plan for storage and backup: how many copies you will keep, what kind of media you will use, and where your storage media will be located.

If you plan to use information technology (IT)-managed or cloud-based storage,[13] you should name the service provider and discuss the associated backup and security benefits or issues. IT-managed storage should include regular backups or snapshots of the data on a revolving schedule, so note this schedule in the DMP. If your storage service does not include regular backups, explain how you will back up your data and at what intervals. What other storage media will you use, and how will you back them up?

You should always make a copy of your raw data as soon as possible after you collect or generate it, and then keep the raw data separate from the working data, that is, the data that you plan to process, manipulate, analyze, and/or anonymize. Describe how you will keep the raw data separate from the working data.

If you have ever left a file made with a proprietary program untouched for several years and then tried to open it using the latest version of that program, then you know that this can be difficult or problematic to do. Because operating systems and software are constantly updated (and sometimes discontinued), you need to be proactive about migrating your files into the updated versions, including both your working files and your raw data files. Be sure to cover periodic file migration in your DMP, especially if you plan for your project (and your raw data) to last for several years. A way to make file migration easier is to save and store your data in lossless, stable, open (non-proprietary) file formats.

If you plan to collect any analog data (e.g., notebooks, drawings or sketches, physical artifacts or samples), explain how you will store and back these data up. If they will be digitized, explain how and when, as well as where the original, physical artifact will be stored.

In cases in which the data are generated in a fieldwork location where there is no access to IT-managed or cloud-based storage, researchers should describe how the data will be transferred securely from the field to the lab or home office (see Robinson 2006).

One of the hardest aspects of planning for data management is estimating the approximate amount of data that will be compiled; in other words, approximately how much data in giga- or terabytes and how many files do you anticipate will result from this project? For projects that also require audio and video data, estimate the number of recording hours for each format. Estimating the total amount of data is a crucial step, as the number and size of the files that will have to be stored and archived have direct impacts on the cost. If your project will last for multiple years, at what rate will your storage capacity needs increase?

Topics that fall under the heading of data security include all laws and rules that regulate data confidentiality, privacy, and cultural sensitivity. Your IRB/REB will require you to have a plan in place to deal with all but the last of these whenever you do research with human subjects. If you are an independent researcher or affiliated with an organization that does not have an IRB, you must familiarize yourself with the relevant rules and laws that will be applicable to your research.[14]

The definition of confidential data can vary between countries and between universities in the same country. In the United States, all universities consider social security numbers and student grades to be confidential. Medical information in the United States is subject to the Health

Insurance Portability and Accountability Act of 1996.[15] In the European Union, the General Data Protection Regulation[16] that went into effect in May 2018 protects the personal information of all EU citizens; and a similar law, a Lei Geral de Proteção de Dados Pessoais, went into effect in Brazil in September 2020.[17] These laws are intended to protect personal data that are collected, processed, and used by businesses, and they protect data that are transported across international borders. At the time of writing, it is still not clear how these laws will affect research done with human subjects in the relevant countries or through international research collaborations.

Though few countries have national laws that protect Indigenous IP or Cultural or Traditional Knowledge, many tribes or groups have particular protocols (laws, rules, or belief systems) that classify information and regulate how and when it can be accessed and by whom. Make sure that you address these protocols in your DMP and explain how you will protect these sorts of sensitive data and who will have access to the data.

In general, if you are working with a team, explain which team members will have access to any sensitive data and their qualifications for accessing them. Explain the security of your storage media and whether different types of data will be stored in different places. If you use IT-managed storage, will anyone besides your team have access? Regardless of the storage media, how will you prevent data manipulation? Will any of your data have to be encrypted? While encryption might make data more secure, it can also make the data more difficult to backup and impossible to preserve. Consider carefully which, if any, data types need to be encrypted.

2.4 Documentation and metadata

Have you ever opened a folder on your computer, external hard drive, or cloud storage that you have not opened in a long time and realized that you cannot remember what any of the files or subfolders contain or even their relationships to each other? Or have you opened a spreadsheet and been completely baffled by the contents of the rows and columns or the relationship between them? The solution is documentation. *Documentation* explains the context of a research project and how that project is carried out (methodology,

After or as you read section 2.3, "Data storage, backup, and security," answer the following questions:

1. How will you store and back up your data during data collection and analysis?
    a. Will you be using a data storage service (e.g., IT-managed storage, cloud storage)?
    b. Who will be responsible for backing up the data? You? A data storage service?
    c. How often will the data be backed up?
2. How will you keep raw data separate from working data?
3. How/when will you migrate your data?
4. How will analog data be stored and backed up?
5. If you will be generating data at a field site, how will you safely and securely transfer it to your office/home?
6. How much data will you need to store?
    a. Estimate of the amount of anticipated data in gigabytes or terabytes.
    b. For projects that also require audio and video data, estimate the number of recording hours for each format.
7. What are the costs associated with storage and backup?
8. How will you keep your data secure?
    a. Who will have access to the data and how will they be given access?
    b. How will personal information about research participants be protected?

Figure 8.3
Questions: Data storage, backup, and security.

protocols, workflows, procedures, manuals, programs, equipment configurations, software settings, and such); how data are organized, managed, stored, and backed up; how data files, points, or sets are related; and how data quality is controlled or ensured (Michener 2015). Good documentation helps to prevent misunderstandings, and well-documented data are easier to find, understand, analyze, share, and reuse (Henderson 2016). Think about the documentation as being the instructions that a future researcher (or your future self) will need to understand your project to be able to reuse the data.

Documentation should be ongoing and updated regularly, and especially when there is a change of any kind to any aspect of the DMP. Two common methods of documentation are *readme* files and *data dictionaries*. *Readme* files (e.g., Readme.txt) are meant to be human-readable forms of documentation present in every digital directory (folder) that contains project data. The top-level directory (main folder) should include a project description that puts the entire project in context. Readme files can be used to document all terms, conventions, codes, abbreviations, units of measure, recording frequencies, software settings, and so on used in the project (also called a "data dictionary"); this document lists everything that someone new to the project will need to know. At lower directory levels, the readme file should explain what the individual files are and how they are related to each other. A data dictionary is a key to a database system; it lists all of the terms, definitions, conventions, codes, abbreviations, units of measure, and such that are used in a project database, and it explains how the different tables (or files) are related to each other (Briney 2015; Henderson 2016).

Documentation should include an explanation for file-naming and version control practices that are to be used for the project. Before picking a file-naming schema, check with the repository that will preserve your data to see whether it has a required file-naming schema or set of conventions that you should follow (Henderson 2016; Kung et al. 2018). Using the repository's file-naming schema or conventions from the start could save you a lot of time and effort later. If your intended archive does not require you to follow a particular file-naming schema or convention, you should nevertheless adhere to some best practices, such as those that follow.

Keep file names as short as possible (fewer than 25 characters, including the extension) and use only letters A to Z, numbers 0 to 9, the hyphen, and the underscore. Avoid special fonts, diacritics, spaces, periods (except to separate the file name from the format extension), and other special characters because these might be problematic for some operating systems or scripts (if not yours, then perhaps those of the repository that will provide your long-term digital preservation). In choosing a file-naming schema, pick two or three things that will help you distinguish or remember the files contents, such as date, location, protocol, or participant identifiers (Henderson 2016; Kung et al. 2018). Avoid using participants' names or initials in file names in case they decide that they want to remain anonymous; remember that they might make this decision years after your research is complete, which would be especially problematic if you (or someone else) have already published a data set containing their name or initials. Dates should be in the international archival standard (International Organization of Standardization's ISO 8601), with or without hyphens, YYYYMMDD or YYYY-MM-DD, but bear in mind that using hyphens will make your file names longer. Use leading zeros for any numbered file names or versions (Henderson 2016; Kung et al. 2018).

File versions should be indicated with either a version number or a date appended to the end of the file name, such as filename_v03.txt, filenameV04.txt, or filename20181025.txt. Special version control software such as Git saves the differences between files rather than duplicates of the entire file (Briney 2015; Kung et al. 2018). Whatever schema you decide to use for file-naming and version control, make sure to document them.

The term *metadata* refers to "structured information about an item" (Henderson 2016:72), and this structured documentation makes your data discoverable and machine readable in systems. Tracking metadata is a crucial component of research documentation. *Descriptive metadata* include information such as author, title, abstract, keywords, publication date, and so on. *Administrative metadata* include the technical information about a file, as well as the rights management (copyright, licenses) and preservation information. *Structural metadata* are information about the relationship between files or other objects in a data set (Henderson 2016; Riley 2017; Thieberger & Berez 2012). Metadata should be documented about each file at the time of creation or

as soon as possible afterward. The more comprehensive the metadata, the more useful the data (Michener 2015).

All research sponsors require you to name in your DMP the metadata schema that you will use. There are many different metadata schemas, including Dublin Core, Metadata Object Description Schema, Metadata Authority Description Schema, Schema.org, Web Ontology Language, Data Documentation Initiative, and Preservation Metadata: Implementation Strategies, to name just a few. The metadata that should be collected varies from (sub)discipline to (sub)discipline, so how do you decide on an appropriate metadata schema to use for your project? If you have identified the data repository that you plan to use, you should adopt the metadata schema that is used in that repository. Many humanities and social science repositories (including many language archives) use either Dublin Core or Metadata Object Description Schema, depending on the repository software, and some even use both. However, be aware that many general data repositories do not cater their metadata elements to any specific discipline; thus, it is a good idea to find out the types of metadata fields that are commonly used in your field or discipline and collect those to make your data useful for your discipline.

For research to be reproducible, researchers must be transparent about their methodology for data collection, handling, and analysis, as well as about the sources of their data (Berez-Kroeker et al. 2018). This requires thorough documentation of the methodology and tracking of the metadata, both of which are crucial to properly describe the nature of the data, as well as the context under which the data were generated. The DMP should include explanation of the processes by which documentation and metadata will be captured or created and the standards that will be followed. If you are working with a team, assign documentation tasks to particular project members to increase the chances that the documentation is done consistently. Good documentation is crucial when/if project personnel change.

2.5 Data dissemination, preservation, and sharing

Dissemination of research findings has traditionally included scholarly publications and conference presentations, but now scholars increasingly are expected to disseminate the research data on which their findings are based by means of data preservation, data sharing, and data publication. Original linguistic research data and data sets must be accessible for the research to be reproducible, and reproducibility is a necessary component of verification and accountability of published findings (Berez-Kroeker et al. 2018). Thus, more and more publishers, research funders, universities, and departments require research data to be shared through data archiving and/or publication. Furthermore, many government-sponsored funders maintain the position that if the research data were collected with public funds, then those data should be made available to the public whenever possible (Stebbins 2013; Horizon 2020 Programme 2017). Additional reasons for data preservation and sharing include the following. Archived data sets can be used for new research. Data sets from different sources can be used to create new data sets. Data sets of similar data from different time periods can be longitudinally compared for new research findings. Published or archived data sets can be considering for hiring, tenure, and promotion decisions. Published data sets can be used for public outreach and classroom teaching (kindergarten to twelfth grade and higher education).

After or as you read section 2.4, "Documentation and metadata," answer the following questions:

1. How will you document any relationships between digital files? Between digital files and analog data?
2. What file-naming schema will you use?
3. How will you control versions?
4. What metadata schema will you use?
5. How will you track your metadata?
6. Who will be responsible for maintaining documentation for the duration of the project?

Figure 8.4
Questions: Documentation and metadata.

Archived files are automatically migrated to new formats as technology changes, and individual researchers no longer have to be responsible for this cumbersome task (Henderson 2016; Kung et al. 2018).

While all research data must be stored for the duration of the project, only a subset of the data should be preserved (Henderson 2016; Kung et al. 2018). Most researchers do not realize that there is a difference between *storing* data and *preserving* data. *Data storage* refers to the location where you keep your files, for example, cloud storage, IT-managed storage, or a hard drive, so that you or your team can access them. *Data preservation* goes beyond simply *storing* data to include management and production of all of the activities that must be done to digital files and their metadata to ensure that they can be accessed into the future as software and hardware change (Beagrie & Jones 2008). When you put your data files in cloud storage you are simply storing them; when you deposit data files into a digital repository, you are entrusting them to an organization that is committed to digitally preserving them for an agreed duration of time. Once data files are preserved in a digital repository, they are considered published, and they are both discoverable and accessible online for reuse (provided you have not placed embargoes or other restrictions on them).

However, not all data should be preserved (in their original state). Some data must be deleted or destroyed at the end of a research project if required by the IRB or REB protocol; some data sets that include private or confidential information might need to be anonymized or deidentified; and some data types might need to be restricted or embargoed in some way (Briney 2015; Henderson 2016; Kung et al. 2018). The DMP should include a detailed description of the future of the data to be generated by the project and an explanation of how, when, and where the data will be archived and made available for reuse (i.e., shared). Researchers should explain any modifications to the data that will be needed before they can be submitted to the repository, including anonymization or format conversion. Data should be as open as possible, but as closed as necessary (Horizon 2020 Programme 2017), so you should discuss any access and reuse limitations that will be placed on archived data.

Your choice of data repository might be influenced by the funding source, the publication journal, your home institution, or your discipline. If you need help finding a data repository for your data, consult one or all of the following lists: the Registry of Research Data Repositories (https://www.re3data.org/), the Digital Endangered Languages and Musics Archives Network (http://www.delaman.org/), or the Open Access Directory's Data Repositories list (http://oad.simmons.edu/oadwiki/Data_repositories). Software should be deposited in a software repository like GitHub (https://github.com/) or GitLab (https://about.gitlab.com/). For more tips on how to find an appropriate repository, see Andreassen (chapter 7, this volume).

Before you name a digital repository in your DMP, you should first contact that repository to make sure that it is able to accept your data. You should also familiarize yourself with its policies regarding metadata schema, file format types, file size limits, deposit or collection limits, file names, data delivery procedures, access policies, deposit schedules, fees,[18] and restrictions or embargoes.[19] If you know that you have data that must be restricted or embargoed, search for a repository that allows restricted data sets. If you plan to restrict data indefinitely or control access to the data, you should have an *inheritance plan* in place for what will happen to those data when you are no longer available to control them; this is essentially a will for the data because archived data will likely outlive the data collector (Kung et al. 2018).

Many repositories set limits on the amount of data they will accept from a given researcher or a particular project, so you might have to carefully select the most important data from your project to archive. Moreover, most data repositories and archives require data depositors to curate their own files into organized collections of data sets, so you should include this work in your time line (see section 2.6). It can be challenging to decide exactly which data should go into the repository, and data curation is time-consuming and tedious work that, if left until the end of the project, can prove difficult and overwhelming (Kung et al. 2018). Many researchers who leave their data selection and curation until the very end of their projects fail to budget sufficient time, resulting in poorly organized data sets and collections, insufficient metadata or documentation, missed deadlines for final project reports, and sometimes even rejected final reports. Data curation that is done on a regular basis results in well-organized, well-described, and well-documented data collections that can be easily discovered, accessed, and reused. For instructions on

how to appraise, select, and prepare your own data for deposit in a digital repository, see Andreassen (chapter 7, this volume). For information on how to prepare language documentation data for archiving in a language archive, see Kung et al. (2018); although this work is intended as a resource for language documentation collections, much of the content is applicable for all types of linguistic data. Williams, Bagwell, and Nahm Zozus (2017) suggest that the DMP should be archived along with the data that resulted from the research project given that it is an important and comprehensive part of the documentation of the project.

Files submitted to archives or repositories for long-term preservation should be in lossless, standard, open (non-proprietary) formats. Many repositories limit the format types that they will accept to make the work of digital preservation more sustainable. While some repositories will allow you to upload any file format you want (including proprietary formats), this does *not* mean that the repository is promising to migrate that file format as technology and formats change; read the policy pages carefully, ask questions, and then plan to deposit only standard, open formats. Never put encrypted files into a repository because they cannot be migrated to new formats; and use standard character encoding such as eight- or sixteen-bit Unicode transformation format (check with the repository to see what they support).

You must explain in your DMP how you will license your research data when you archive them. Licensing determines how data files can be shared and reused, and there are several different licensing systems, including traditional copyright (all rights reserved), Creative Commons licenses,[20] Open Data Commons licenses,[21] Local Contexts' Traditional Knowledge licenses,[22] and GNU licenses.[23] Before you pick a type of license to use, determine which licenses are used by the repository where you plan to deposit the research data and make sure that you understand the differences between the different licenses and their uses. If you used existing data from some other source in your project, it is likely already licensed; discuss with the repository manager what you should do with those data, then document that discussion and your decision in your DMP.

If you collected data of any kind in a community (Indigenous or not) that does not have access to your chosen data repository (e.g., access is restricted to the affiliates of the university where the repository is located) or to the Internet in general, establish a plan in your DMP for returning a copy of the research data to the community in a form that will be useful for the community members (Kung 2021). Moreover, if any of the data contains Traditional Knowledge that needs to be restricted for cultural heritage reasons, explain these reasons in your DMP, verify that your chosen archive will accept these materials,

After or as you read section 2.5, "Data dissemination, preservation, and sharing," answer the following questions:

1. How will you disseminate and share data from your project?
2. What portion of the data must be archived?
3. Will a portion of the data have to be deleted or destroyed?
4. Will a portion of the data need to be anonymized or deidentified?
5. Where will the data (or code) be archived?
    a. Name the digital repository where you plan to archive your data.
    b. Who has access to this digital repository?
6. What are the potential costs associated with the archiving and long-term preservation of your data?
7. Will a portion of the data need to be embargoed or restricted?
8. How will you license the data for reuse?
9. If relevant, how will the data be shared with or repatriated to the speech community?
10. What, if any, special protocols or rules (Indigenous or Traditional Knowledge) will apply to the data?
    a. How will those protocols be implemented in the chosen archive?

Figure 8.5
Questions: Data dissemination, preservation, and sharing.

and establish a plan—both with the Indigenous Community and the repository—for who may access the data and how. If the Indigenous Community has its own protocols or rules of access for particular types of data, note these protocols in your DMP, and determine whether your chosen archive has a way to enforce the community's protocols. There is always the possibility that the data might have to be controlled by a gatekeeper (you or a community member) who is familiar with the protocols and can enforce them. Discuss the possibilities, including an inheritance plan, with the repository and the community before you deposit the data.

2.6 Time line and responsibilities

Every DMP must include a time line for its implementation from start to finish. When writing your DMP, you must explain who is to be responsible for implementing the DMP and for ensuring it is followed, reviewed, and revised according to the time line (even if that person is you). Name every person, department, or organization that will be responsible for carrying out some aspect of your data management; what they/it will be responsible for doing (e.g., data collection, data entry, transcription, translation, annotation, coding, quality assurance or control, metadata creation and documentation, lab or methodology documentation, storage, backup or snapshots, systems administration, data curation, repository submission); and when they will do it. Consider what level of expertise is needed for each role and assign the role appropriately. Be explicit about any resources (including hardware, software, technology, skills, and such) that will be required to carry out the assigned tasks.

The time line should indicate when data will be collected and analyzed, as well as when data will be submitted to the repository, when the repository will process and ingest the data, when the data will be available for public access, and when any access embargos will expire. Be aware that while many of the tasks detailed in your DMP will be in your control to schedule and carry out, this is not necessarily the case when it comes to archiving your data. Be sure to consult the repository that you plan to use to establish a time line for deposits and archiving that will work for you both; this will ensure that the repository will be expecting your data according to your prearranged time line. While you might hope to be able to deposit all of the research data during the last funded month of your project, that most likely will not be possible for the repository. Keep in mind that many university-based repositories have very few full-time, non-student staff, so there might not be sufficient technical staff available to help you with your deposit during the summer and winter breaks.

3 Revising and adapting a DMP

Once the DMP is written, it provides a guide to or road map for the steps that will be followed while carrying out the research. Miksa et al. (2019) describe DMPs as "living documents" and go on to explain that "the amount and granularity of information contained within them evolves over time—from high-level estimates and expectations down to precise descriptions of actions that have actually been taken" (10). Thus, the DMP should be revised any time a change of any sort is made to the project, including the type or amount of data to be collected or the protocols for how the data are collected, analyzed, stored, preserved, and so on. As the DMP is revised, check whether any costs (e.g., storage, archiving) need to be adjusted accordingly. Document the changes to your DMP using version control and noting all changes that were made, who made them, and when (Miksa et al. 2019).

After or as you read section 2.6, "Time line and responsibilities," answer the following questions:

1. Who will be responsible for implementing and overseeing the DMP?
2. What is your time line for this project (from data generation to final archiving and dissemination)?
    a. When will each major task take place on this time line?
    b. Who will be in charge of implementing the tasks according to this time line?

Figure 8.6
Questions: Time line and responsibilities.

Once you have written your comprehensive DMP that covers every aspect of data management relevant to your project and circumstances, you can then edit it to address (only) the specific requirements of particular funders. Williams, Bagwell, and Nahm Zozus (2017) examined DMP requirements of different funders and identified 43 different required DMP topics; however, they found very little overlap in these required topics between funders. While the DMP requirements of many funders focus almost entirely on postcollection or postpublication data management, some funders also require that the DMP cover data management during the data collection and analysis phases of research as well. While there are online tools designed to aid researchers in writing a DMP, in particular the DMPTool[24] and DPMonline,[25] these tools have different DMP templates for different funders. Thus, there is no single template for a DMP that can satisfy the requirements and simultaneously meet the page limits of every possible funding agency. Moreover, Williams, Bagwell, and Nahm Zozus found that most funders that require a DMP put more emphasis on sharing data sets on which publications are based, and less emphasis on research activities and resources that actually "impact data quality, provide traceability or support reproducibility" (130).

Some funders now require researchers to submit both human-readable and machine-actionable DMPs (maDMPs; see Miksa et al. 2019; NSF 2019).[26] Online tools such as the DMPTool and ezDMP[27] can be used to create maDMPs once the researcher has all the necessary information to plug into the template. A comprehensive DMP can easily be used for this purpose, as well.

4 Summary

By now you should have a better understanding of the importance of a comprehensive DMP for your own RDM. It is true that, at the outset of a research project, it is simply not possible to know the exact details of all the possible variables, such as the amount of data that will be generated, all of the possible files types that will be created, all of the software that will be used, or the exact number of collaborators who might contribute to the project. Nevertheless, you need to make a rough plan that estimates these details to help you think through your project and plan your budget. As your project evolves, you should update and revise your DMP to reflect the changes, and as you revise the DMP, make sure to version it according to your plan for versioning the rest of your data. Your goal should be to write a comprehensive—but flexible—DMP that will aid you in planning your research project from start to finish and that can be easily modified for submission to research sponsors, publishers, and repositories. "When conceptualized and operationalized as comprehensive documentation of the data lifecycle for a study, a data management plan is a powerful tool and an integral component of the data management quality system" (Williams, Bagwell, & Nahm Zozus 2017:135).

If you still need guidance in drafting your DMP, your first stop should be your institution's data (management) services department or unit, which is usually affiliated with the university library. If your institution does not provide data management services or you are not affiliated with a university, consult some of the resources that are included in the references such as Berez-Kroeker, Collister, & Kung (2017), Digital Curation Centre (2013), Inter-university Consortium for Political and Social Research (2012), Kung et al. (2018), and Penn State (2019). Finally, sample DMPs can be found in Kung (2019), a data set containing supplementary materials that accompany this chapter. If you use the samples as templates for your own DMP, make sure that you customize the information so that it is relevant to your research project.

Even though writing a comprehensive DMP seems like a lot of work at or prior to the outset of a project, a well-organized plan for data management will pay off in the long run. Burnette, Williams, and Imker (2016) worked with a team of researchers at the University of Illinois to write and implement a DMP. At the end of the project, the principal investigators reported reductions in lost data and time, as well as stress and anxiety levels. Burnette et al. quote an unidentified principal investigator as saying, "It's not good science unless the data is managed well since 'you are only as good as your data'" (8). All investigators involved agreed that though creating and initially implementing the DMP took a lot of work at the outset, having a plan to follow saved them a great deal of time and effort and gave them peace of mind as the project advanced. Thus, writing a DMP is well worth the effort. Not only will you save yourself time and energy later, but you will produce orderly data that is suitable for analyzing, archiving, sharing, and reusing.

Notes

1. This chapter is based on Berez-Kroeker, Collister, and Kung (2017). I would like to acknowledge my two coauthors of that work, Andrea Berez-Kroeker and Lauren Collister, and thank them for trusting me to write this chapter on my own. Any mistakes in this current work are entirely my own. This material is based on work supported by the National Science Foundation under grant numbers SMA-1447886 and BCS-1653380.

2. See Mattern, chapter 5, this volume.

3. Note that these services might change from one year to the next; the services that were available for your last research project might have changed.

4. The DMPTool (https://dmptool.org/) is free for anyone regardless of university affiliation; see the "Quick Start Guide" at https://dmptool.org/help.

5. Two popular tools at the time of writing include the Open Science Framework (OSF, https://osf.io/) and AirTable (https://airtable.com/).

6. *Intellectual property* (IP) is a term that covers copyright, patents, trademarks, and trade secrets. The two types of IP that are the most relevant to research data are copyrights and patents (see Collister, chapter 9, this volume, and Alperin et al., chapter 13, this volume).

7. https://wipolex.wipo.int/en/main/legislation.

8. This is especially important if you expect any patents to result from your research project. See Alperin et al. (chapter 13, this volume) for some discussion of patents resulting from research.

9. See Collister (chapter 9, this volume) for more information about these terms.

10. See Berez-Kroeker et al. (2018) for a discussion for the need for attribution in linguistics.

11. However, note that if the cloud storage syncs to a folder on your hard drive (like Dropbox and Box can do), this counts as only one copy because if you delete a file in one place, it is also deleted from the other (Briney 2015).

12. LOCKSS is an open-source solution for peer-to-peer (distributed) digital preservation and integrity assurance that was founded at Stanford Library (https://www.lockss.org/).

13. IT-managed storage is a storage environment that is managed by a service provider. Most universities or institutions have some sort of storage that is managed by their IT department.

14. There are several online training programs for research ethics and compliance with both national and international foci; the CITI Program (Collaborative Institutional Training Initiative; https://about.citiprogram.org/en/homepage/) is one that is frequently used by organizations and individuals in the United States. Whatever training program you use, make sure to include the costs associated with research ethics training in your project budget.

15. https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html.

16. https://gdpr.eu/.

17. http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm.

18. Long-term digital preservation can be very expensive, so many repositories charge fees. Plan your budget accordingly.

19. An *embargo* is a restriction that is applied for a limited time, for example, five years to finish a degree, two years while research results are published, and so on.

20. Creative Commons licenses (https://creativecommons.org/) are widely used in the humanities and social sciences and by institutional data repositories.

21. Open Data Commons licenses (https://opendatacommons.org/) are frequently used for databases and code.

22. Traditional Knowledge licenses (http://localcontexts.org/tk-licenses/) are intended to be used by Indigenous Peoples to protect and share their Traditional Knowledge.

23. GNU licenses (https://www.gnu.org/licenses/) are for licensing software.

24. For the DMPTool, see note 4.

25. As of February 2018, the US-based DMPTool and the UK-based DMPonline have merged into a single tool (see DMPTool 2018); however the DMPonline tool is still available at https://dmponline.dcc.ac.uk/.

26. Thus far the requirement for maDMPs is limited to certain STEM programs and has not spread to linguistics programs. It will likely take time for all disciplines to catch up to STEM, but it is inevitable that maDMPs will soon become the norm.

27. The ezDMP (https://ezdmp.org/index) tool creates maDMPs specifically for National Science Foundation grants; the user must log in with either a Google or an ORCID (Open Researcher and Contributor ID; https://orcid.org/) account.

References

Beagrie, Neil, and Maggie Jones. 2008. *Preservation Management of Digital Materials: The Handbook*. Glasgow: Digital Preservation Coalition (DPC). https://www.dpconline.org/docs/digital-preservation-handbook/299-digital-preservation-handbook/file.

Berez-Kroeker, Andrea, Lauren Collister, and Susan Smythe Kung. 2017. Workshop on data management plans for linguistic research. LSA Summer Institute, University of Kentucky, July 29–30, 2017. *Archive of the Indigenous Languages of Latin*

America. Access: Open. PID: ailla:254604. https://www.ailla.utexas.org/islandora/object/ailla%3A254604.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56 (1): 1–18. https://doi.org/10.1515/ling-2017-0032.

Blum, Carol. 2012. *Access to, Sharing and Retention of Research Data: Rights and Responsibilities*. Washington, DC: Council on Governmental Relations. https://www.cogr.edu/sites/default/files/access_to_sharing_and_retention_of_research_data-_rights_&_responsibilities.pdf.

Briney, Kristin. 2015. *Data Management for Researchers*. Exeter, UK: Pelagic Publishing.

Burnette, Margaret H., Sarah C. Williams, and Heidi J. Imker. 2016. From plan to action: Successful data management plan implementation in a multidisciplinary project. *Journal of eScience Librarianship* 5 (1): e1101. https://doi.org/10.7191/jeslib.2016.1101.

Digital Curation Centre (DCC). 2013. *Checklist for a Data Management Plan, v.4.0*. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/data-management-plans.

DMPTool. 2018. "New DMPTool launched today!" *DMPTool Blog*. February 27, 2018. https://blog.dmptool.org/tag/enhancements/.

Henderson, Margaret E. 2016. *Data Management: A Practical Guide for Librarians*. Lanham, MD: Rowman and Littlefield, ProQuest Ebook Central.

Horizon 2020 Programme. 2017. Guidelines to the rules on open access to scientific publications and open access to research data in Horizon 2020, version 3.2. European Commission Directorate-General for Research and Innovation, March 21, 2017. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

Inter-university Consortium for Political and Social Research (ICPSR). 2012. *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle*, 6th ed. Ann Arbor, MI: ICPSR. https://www.icpsr.umich.edu/files/deposit/dataprep.pdf.

Kung, Susan Smythe. 2019. Data management plans for linguistic research. *Texas Data Repository Dataverse*. https://doi.org/10.18738/T8/538EEN.

Kung, Susan Smythe. 2021. Data archiving, access, and repatriation. In *The International Encyclopedia of Linguistic Anthropology*, ed. James Stanlaw. Hoboken, NJ: Wiley Publishers.

Kung, Susan Smythe, J. Ryan Sullivant, Vera Ferreira, and Alicia Niwagaba. 2018. How to organize your materials and data for a language archive (CoLang2018_Curation_Workshop_Slides.pdf). Linguistic Data Curation Tutorials. *The Archive of the Indigenous Languages of Latin America*. Access: Open. PID ailla:257452. https://www.ailla.utexas.org/islandora/object/ailla:257452.

Leopando, Jonathan. 2013. World backup day: The 3-2-1 rule. *Trend Micro*. April 2, 2013. https://blog.trendmicro.com/trendlabs-security-intelligence/world-backup-day-the-3-2-1-rule/.

Michener, William K. 2015. Ten simple rules for creating a good data management plan. *PLoS Computational Biology* 11 (10): e1004525. https://doi.org/10.1371/journal.pcbi.1004525.

Miksa, Tomasz, Stephanie Simms, Daniel Mietchen, and Sarah Jones. 2019. Ten principles for machine-actionable data management plans. *PLOS Computational Biology* 15 (3): e1006750. https://doi.org/10.1371/journal.pcbi.1006750.

National Science Foundation. 2018. Data management for NSF SBE directorate proposals and awards. May 15, 2018. https://www.nsf.gov/news/news_summ.jsp?cntn_id=118038.

National Science Foundation. 2019. NSF 19–069 Dear colleague letter: Effective practices for data. *National Science Foundation* (website). May 20, 2019. https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp?WT.mc_id=USNSF_25&WT.mc_ev=click.

Penn State University Libraries. 2019. *Data Management Toolkit*. https://guides.libraries.psu.edu/dmptoolkit.

Riley, Jenn. 2017. *Understanding Metadata: What Is Metadata, and What Is It For? A Primer*. Baltimore: National International Standards Organization (NISO). https://www.niso.org/publications/understanding-metadata-2017.

Robinson, Laura C. 2006. Archiving directly from the field. In *Sustainable Data from Digital Fieldwork*, ed. Linda Barwick and Nicholas Thieberger, 23–32. Sydney: Sydney University Press. http://hdl.handle.net/2123/1291.

Stebbins, Michael. 2013. Expanding public access to the results of federally funded research. *The White House Blog*. February 22, 2013. https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research.

Thieberger, Nicholas, and Andrea Berez. 2012. Linguistic data management. In *The Oxford Handbook of Linguistic Fieldwork*, ed. Nicholas Thieberger, 90–118. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571888.013.0005.

Williams, Mary, Jacqueline Bagwell, and Meredith Nahm Zozus. 2017. Data management plans: The missing perspective. *Journal of Biomedical Informatics* 71 (July): 130–142. https://doi.org/10.1016/j.jbi.2017.05.004.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363