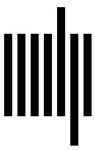
This is a section of doi:10.7551/mitpress/12200.001.0001

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

5 The Linguistic Data Life Cycle, Sustainability of Data, and Principles of Solid Data Management

Eleanor Mattern

1 Introduction

With the growth of data management requirements from funding agencies and a recognition of the value of reproducibility, replication, and data reuse, there have been efforts by disciplinary communities, administrators of data repositories, and libraries to develop guidance and services to support researchers as they care for and share data. This chapter is not written by a linguist. Instead, it is one library and archives professional's effort to connect discussion, guidance, and research on the management of digital data to the linguistics discipline. Because there are disciplinary differences in the types of data collected and used, varying expectations from funders and journals for data preservation and sharing, and distinct traditions for open research, this chapter takes a high-level view. As a starting point, this chapter considers the data life cycle model as a means for perceiving the persistent and ongoing nature of data management. It reviews guidance and best practices for sustaining data, emphasizing the value of consistency and a future-minded orientation as the core principles that should underlie this work.

2 Life cycle of research data

Archivists and records managers have long employed the metaphor of a records life cycle as a means for conceptualizing distinct stages of an information object: its creation, a period of active use, an inactive phase in which its long-term value is assessed; and the destruction or the long-term preservation of the record in an archival repository (Bantin 1998). Through the records life cycle model, archivists and records managers identified actions (e.g., selection for archiving or destruction) and infrastructure (e.g., an archival repository for a selected record) required to manage the records and support the longevity of selected records.

With research data, we have seen similar efforts, both from the library and archival communities and from domain-based researchers, to conceptualize the life cycle of digital data. Data service providers such as libraries have employed these visual representations of research and data workflows as a means to communicate the data curation activities that facilitate the sustainability of research data and to identify support services in place to assist researchers with data management. Alex Ball, data librarian at the University of Bath, writes that "the importance of lifecycle models is that they provide a structure for considering the many operations that will need to be performed on a data record throughout its life" (2012:3). For a researcher, mapping their research workflows against a life cycle model can help to encourage data management practices that can facilitate data integrity, value, and persistence (Poole 2016:963).

Life cycle models differ in subject and scope; some visualize the life cycle of research and others provide a more granular representation of the life cycle of research data. They take different shapes, with some representing cyclical processes and others linear sequences of steps. There are life cycle models that depict a multidirectional, recursive process, and others a unidirectional, forward moving one (Cox & Tam 2018). The language of "life cycle" and the cyclical nature of many of the representations, however, suggests an aspiration for data reuse. Cox and Tam write "Circular lifecycles can also be seen as having a strength in improving on the visualisation of research as a chain, by expressing the desire for data reuse, stressing that in some sense the process is to be repeated" (151).

For all their differences, there are common, high-level stages that we find in life cycle models, with the UK Data

Service's model serving as a simple, domain-neutral illustration of these general stages:

- Planning research
- Collecting data
- Processing and analyzing data
- Publishing and sharing data
- Preserving data
- Reusing data[1]

In this chapter, we will look at some good practices for data management that are associated with these stages and that cut across a life cycle model. Figure 5.1 adapts the US Geological Survey Science Data Lifecycle Model. Through the three bottom arrows, this model illustrates the cross-cutting data curation activities that are not confined to one stage. In this particular model, describing data sets, managing the quality of the data, and backing up data are ongoing, continuous data management actions that cut through a research project (Faundeen et al. 2013).

In the UK Data Service's cyclical model and in the linear US Geological Survey model, we see a neat, sequential representation of a data workflow. Research and data life cycle models are simplified representations of researchers' workflows and absent of some of the messier realities that characterize research and data workflows. In the Mattern et al. (2015) study of humanities and social scientists' research workflows, participants sketched their research workflows and annotated their sketches to indicate where they encounter data-related challenges. Their life cycle sketches included research and data stages that are not depicted in the neater models that libraries and research organizations tend to publish. Two participants in this small study, for example, described "confusion" as a phase in their work. Like the participants in the Mattern et al. study, linguists are unlikely to view their workflows as simple, sequential stages. However, the value of a life cycle model is akin, as Cox and Tam (2018) argue, to the value of a methods textbook: they represent, at a high level, the movement and stages inherent to research and data workflows.

Thieberger and Berez-Kroeker (2012) conceptualize a "workflow for well-formed data" for linguistic fieldwork, which is a step toward a domain and community-specific data life cycle. They write,

> The workflow begins with project planning, which for our purposes includes preparing to use technology in the field and deciding on which file naming and metadata conventions you will use before you make your first recording. . . . After recordings are made and metadata are collected, data must be transcribed and annotated with various software tools, and then used for analysis and in representation of the language via print or multimedia. Note that depositing materials in an archive is carried out at every phase of the procedure. (96–97)

We can draw connections between this discussion of a workflow for linguistics fieldwork data and the stages in the UK Data Service and the US Geological Survey Science Data Lifecycle Model. Thieberger and Berez's model similarly begins with a planning stage. All three describe a data collection and a data analysis stage. Like the US Geological Survey's life cycle, Thieberger and Berez describe data management activities that cut across all phases in a workflow. Unlike most life cycle models that represent data archiving as a final stage, Thieberger and Berez characterize this sharing and preservation work as ongoing during the life of a research project. This is a notable departure but suggestive of a research reality: few projects may have one distinct end point at which

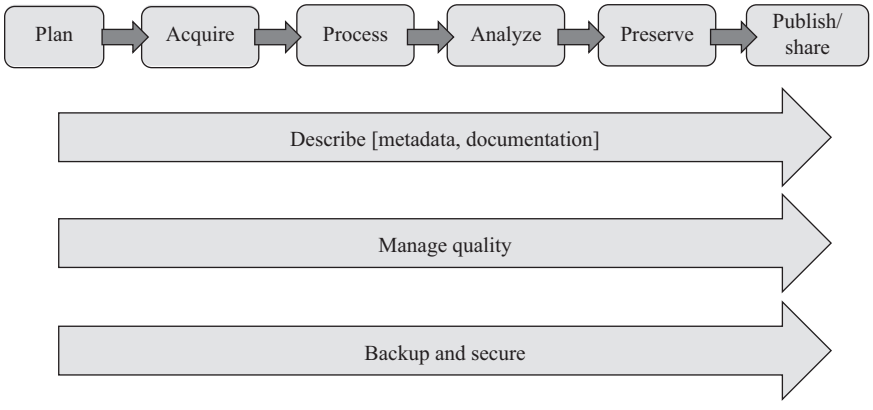


Figure 5.1
Representation of the US Geological Survey Science Data Lifecycle Model.
Source: Faundeen et al. (2013).

time data becomes archival. Indeed, because of the fragility of digital media, archiving and preservation are best addressed throughout research.

Linguists may collect more data, perhaps in the form of large numbers of audio and video recordings, than they ultimately transcribe, annotate, and analyze. This means that different data in a single research project may move through a data life cycle in different ways. Drawing on the UK Data Archives' model, this means that some data, for example, would not reach the processing and analyzing phase or the preserving phase. A researcher may build more robust metadata records for annotated and analyzed data and may ultimately choose to archive a discreet subset of the larger data corpus. For all data, however, a linguist should ensure that a plan is in place.

3 Sustainability of research data

As Lavoie (2012) reminds readers, the term "sustainability" has multiple connotations generally and in the context of research data. He introduces three meanings of sustainability pertinent to digital records such as digital research data sets: (1) economic sustainability, or the resources involved in digital stewardship; (2) social sustainability, or a "shared commitment to preservation among groups of stakeholders with a common interest in long-term access to a particular set of digital materials"; and (3) "sustainability from a technical perspective, in the sense of developing repository architectures, workflows, tools and preservation techniques that are robust, flexible and scalable" (68). These three meanings are worthy of consideration.

Lavoie's chapter focuses primarily on economic sustainability and provides a valuable discussion of the issues, challenges, and approaches of resourcing research data management that is relevant for all disciplines. He argues for the importance of selecting research data for long-term preservation. This is an understood reality in libraries and archives, with archivists referring to this selection process as "appraisal." Because resources are finite, selection of specific materials with evidential or research value (in this case, selected research data sets) for long-term preservation and access is aligned with economic feasibility.

Fran Berman, professor of computer science at Rensselaer Polytechnic Institute and former director of the San Diego Supercomputer Center, offers a second perspective on sustainability, arguing that there is a role for disciplinary communities to set criteria and methods for selection of research data for long-term preservation. Describing a model of social sustainability, Berman (2008:52) writes, "the need for community appraisal will push academic disciplines beyond individual stewardship, where project leaders decide which data is valuable, which should be preserved, and how long it should be preserved (except where regulation, policy, and/or publication protocols mandate specific stewardship and preservation timeframes)." With this, she calls on linguists and other domain-based researchers to develop selection frameworks. Indeed, we see examples of community-developed appraisal criteria that guide inclusion of data sets in data repositories; the Inter-university Consortium for Political and Social Research (ICPSR), for example, has a published and clearly defined set of collecting priorities (ICPSR 2012). Berman's call for community-based selection criteria is resonant with the second meaning of sustainability that Lavoie introduces: social sustainability, or a community's commitment to research data management. There are a number of efforts in the linguistics scholarly community that point to this commitment, with this volume, the Linguistics Data Consortium,[2] the *Austin Principles of Data Citation in Linguistics* (Berez-Kroeker et al. 2018), and the Research Data Alliance Linguistics Data Interest Group,[3] among them.

Lavoie's third meaning of sustainability—the workflows, infrastructure, and practices that support the longevity of research data—overlies the entirety of the research data life cycle. He writes, "when we speak of 'sustainable research,' it is perhaps more accurate to say we are speaking of sustainable data curation activities" (2012:81). This is the meaning of sustainability that forms the focus of section 4: sustainable practice that supports sustainable research data.

4 Data management practices

This chapter references the data life cycle as a framework for examining key practices for responsible and consistent data management, including actions that cut across all stages of the life cycle. In recent years, there have been a host of useful resources created and published that aim to assist researchers in sustaining their research data.[4] This section draws on these to provide a broad discussion on good habits, frames of mind, and actions.

4.1 Planning

In Theieberger and Berez's (2012) workflow, they explain that they decide on conventions for their file naming and metadata before collecting data in the field and, with this, point to a key practice in responsible data management: planning. Funding agencies, as described in Kung (chapter 8, this volume), are increasingly requiring data management plans (DMPs) as part of a grant application, but even where there is no requirement, investing time in crafting a plan can help to refine existing data practices (Mannheimer 2018:15) and encourage efficiency (Corti et al. 2014; Kung, chapter 8, this volume). DMPs generally include a description of the data that will be collected, the metadata and documentation that will be produced, the ways the data will be stored and backed up, security and privacy protections for relevant data, data access policies during a project, and a long-term plan for data sharing and preservation (Digital Curation Centre 2013; Burnette, Williams, & Imker 2016:2). With this broad coverage, DMPs "typically cover all or portions of the data life cycle" (Michener 2015). An effective plan for a collaborative project will additionally describe the responsibilities for all research partners, to help ensure that all involved understand what their roles are in managing data (Corti et al. 2014:29).

A good practice around DMPs is to treat them as "living documents," reviewing and editing them to reflect changes in data management practices and to address emergent data management needs and challenges (Michener 2015). In a data life cycle model, then, good practice would have researchers returning to the planning stage regularly. While DMPs are often characterized as mechanisms that save researchers time in the long-run, linguists Gawne and Berez-Kroeker (2018) realistically acknowledge that "management and curation of data for archiving is a time-consuming process, even when the documentation workflow is set up to optimize the process" (25).

4.2 Sustainable file formats

The selection of file formats that linguists use to store and preserve their data is a fundamental data management practice that can help to ensure sustainability and reuse of these research outputs. The evaluation and selection of data formats should ideally occur before data collection begins, making it a decision that occurs at the planning stage of a life cycle and that is then implemented throughout later stages.[5]

Much of guidance on data management best practices addresses the distinction between open versus proprietary file formats and the relevance of this distinction to sustainability of research data. *Open file formats* are formats that can be accessed using more than one software program and that are supported by more than one developer; *proprietary formats*, conversely, are supported by one developer and may be dependent on only one software application for use. While there are proprietary commercial software and corresponding file formats that are ubiquitous (for example, Adobe Photoshop and .psd files), where there is dependency on a single software application, there are vulnerability and limitations to access. As Trevor Owens (2018) of the Library of Congress explains, "the more a format depends on a particular piece of hardware, operating system, or software to render, the more susceptible it is to being unrenderable if one of those dependencies fail" (121). If saving data in open formats would result in a loss of functionality or information, retaining data in the proprietary format and making a copy in an open format is a recommended practice, particularly at the preservation and archiving stage of the data life cycle (Van den Eynden et al. 2011:13; Stanford Libraries, n.d.).

There are more immediate implications that the selection of a proprietary or dependent format introduces. If the researchers' goal is to support wide use of data sets that they make available, making the data available in formats that would not require the purchase of commercial software removes a barrier for reuse.

When assessing the sustainability of a digital file format, there is less probable risk associated with widely used file formats. Owens (2018:121) writes, "If you have PDFs, MP3s, JPEGs, or GIFs, you've got every reason to believe that people will be able to open those files. If those formats become risky, their wide use makes it likely that tools and services will be created to help migrate them." Selecting file formats that are widely adopted, in other words, provides some security for the future accessibility of the data records.

There are a number of resources available to help researchers select sustainable file formats for the types of data that they are creating.[6] The UK Data Service's recommended file formats table is a useful starting point. For a tabular data file such as a spreadsheet, the UK Data

Service's recommended formats are comma-separated values (.csv) and tab-delimited file (.tab). The organization also identifies acceptable formats, including the widely used Microsoft Excel formats (.xls/.xlsx), reflecting Owens's commentary on the relationship between file format adoption and sustainability.[7]

For researchers seeking a more detailed assessment of the preservation-friendliness of selected file formats, the Library of Congress's "Sustainability of Digital Formats" web resource provides detailed descriptions for a range of file formats; in addition to information about the degree of adoption and dependencies associated with the format, the Library of Congress considers the level of documentation that exists for the format and whether there is metadata embedded in the file, additional factors supporting format sustainability.[8] The Library of Congress's guidance on sustainable formats additionally points to the importance of selecting 'lossless' formats, or formats that do not lose information when compressed or made smaller; for images, for example, a TIFF file is a lossless format and a JPEG is a 'lossy' one.[9]

Table 5.1 depicts selected data types in linguistics research (Himmelmann 2012; Language Archive 2019) and, using the aforementioned resources and considerations, associated sustainable file format types.

The recommendations in table 5.1 reflect the Library of Congress's overview of formats and sustainability factors, as well as guidance from the US National Archives and Records Administration.[11] However, the table also accounts for the functionalities of linguistics-specific software programs, namely Praat and ELAN. A researcher working with audio files and Praat, for example, should be aware that Praat is unable to open the common audio file type .mp3 files and instead supports the .wav format (Styler 2017). Guidance on Praat, moreover, stresses the lossy nature of the .mp3, stating "friends don't let friends save phonetic data in lossy formats (e.g., .mp3, AAC, .wmv)" (63). For researchers working with spectrograms in Praat, the software allows for export of "Praat Pictures" as PDFs (48). This means the researcher will lose the interactivity with the spectrogram that the software provides, but will be able to generate a static file for preservation and access purposes.

4.3 File names and organizational structures

Anyone who uses a digital camera or a cell phone camera has likely encountered the systematic, yet inscrutable, file naming scheme associated with their images. Load the images onto a computer and one encounters file names that mean little to the creator, a string that begins with IMG and is followed by a number. When there are hundreds or thousands of these similarly named files, locating a desired image can be a challenging task indeed. A solid file-naming convention is one

Table 5.1
Recommended file formats for sustainability

Data type	Recommended format
Audio recordings	• Free Lossless Audio Codec (.flac) • Waveform Audio File Format (.wav)
ELAN files (.eaf EUDICO Annotation Format)[10]	• Tab Delimited Text (.txt). • Timed Text Markup Language (TTML) (for annotations)
Photographs, spectrograms and images (e.g., functional magnetic resonance imaging)	• TIFF (.tif) • PDF/a
Spreadsheets and databases	• XML-based formats • Comma-separated values files (.csv) • SQLite • SIARD (Software Independent Archiving of Relational Databases)
Text files (e.g., transcripts and observational notes, translations with interlinear glossing); annotations	• eXtensible Mark-up Language (.xml) • Plain text format (.txt) (encoding: ASCII, UTI-8, UTF-16) • Rich Text Format (.rtf) • Portable Document Format/Archival (PDF/A)
Tabular data and databases	• XML-based formats • Comma-separated values files (.csv) • SQLite • SIARD (Software Independent Archiving of Relational Databases)

that is meaningful to the researcher and consistently employed. Again, the determination of this convention ideally happens at the start of the research workflow, at the planning stage in our data life cycle and should be documented by the research team. This section will briefly review existing guidance for developing a convention and approaches for managing versioning.

There is no one way to name files and good practice is a memorable and sustainable one for the individual researcher and team. Libraries and the UK Data Service, again useful starting points for guidance, identify a number of elements that a researcher might choose to include in a file name (Corti et al. 2014:67):

- Date: Using a consistently structured date at the beginning of a file name can be helpful for sorting files chronologically, if that is relevant to the nature of the research. The International Standards Organization[12] format for a date, YYYY-MM-DD, facilitates this chronological sorting and will be widely understood as a date by other users (Witmer 2017).
- Project name or acronym
- Researcher name or initials
- Version number
- Ordinal numbering system: Using leading zeros (001, 002, 003, etc.) will assist with sorting. (Smithsonian Library, n.d.)

General guidance on filing naming suggests that brevity, rather than cumbersome and lengthy conventions, will be most sustainable in practice. Moreover, some software applications have difficulty with spaces and many special characters in file names; avoiding spaces and instead separating elements with an underscore, hyphen, or camel case is advisable (Witmer 2017).

With a new project, it is also necessary to develop a strategy for organizing files that is documented and simple enough to consistently follow throughout the life cycle of a project. A hierarchical organization, with all relevant files grouped under a common top-level project directory, is a common and advisable approach (Noble 2009). As with file-naming conventions, the best organizational strategy for subfolders is one that is logical and easy to employ. A simple text file that lives in the top-level folder and that overviews the organizational approach for the project files can function as a useful memory tool for a researcher and a valuable guide for someone approaching the project data for the first time.[13]

4.4 Data storage

This section considers decision making around active storage and the value of a distributed approach to active data storage. In Hart et al. (2016) "Ten Simple Rules for Digital Data Storage," we again encounter the importance of systematizing a data management practice and developing this system early. As strategies and required resources are dependent on the volume of project data, Hart et al.'s guidance for large data sets (terabytes to petabytes) is particularly valuable, offering insight into time-saving solutions for requesting project data stored on commercial cloud solutions. For all researchers, the authors soundly emphasize the necessity of a storage backup scheme and the importance of regularly evaluating whether the scheme is functioning; backups may fail, the authors acknowledge, even when a solid procedure is in place. Specifying at the planning stage how often the backups will be assessed and by whom—and then implementing that plan—can mitigate the risk of loss of the backup.

The National Digital Stewardship Alliance (NDSA), a consortium focused on building digital preservation capacity, has published a set of recommendations for sustaining digital records that, while geared toward organizations, are pertinent to the individual researcher. The NDSA (2013) advises keeping three copies in at least two geographic locations, a standard to which Hart et al. (2016) also subscribe. Hart et al. advise, "Ideally you should have two on-site copies (such as on a computer, an external hard drive, or a tape) and one off-site copy (e.g., cloud storage), with care taken to ensure that the off-site copy is as secure as the on-site copies. Keeping backups in multiple locations additionally protects against data loss due to theft or natural disasters." We regularly encounter this recommendation in library guidance, framed as the 3-2-1 rule (three copies; in two different storage media; with at least one off-site copy).[14] In the linguistics community, Thieberger and Berez (2012) address the importance of backups in research, offering different strategies for creating copies. For linguists doing fieldwork, they suggest that external hard drives, cloud storage, and USB "on-the-go" devices can all be good backup strategies, depending on the circumstances in the field (99–100).

4.5 Metadata and data documentation

This section provides an introduction to the purpose and importance of metadata and data documentation in

supporting the sustainability of research data. Funders commonly request that researchers indicate in a DMP what documentation and metadata will describe their data and, as the US Geological Survey's data life cycle (figure 5.1) depicts, the researchers will carry out the associated work throughout a research project.

We encounter metadata regularly in our lives: in a library catalog, in the product descriptions on Amazon, and on the title page and front matter of a book. *Metadata* are information about an object that helps us to understand, find, and use that object. As Miller (2004) explains, metadata help an individual other than the original creator

> to decide whether or not [an information object] is of value to them; to discover where, when and by whom it was created, as well as for what purpose; to know what tools will be needed to manipulate the resource; to determine whether or not they will actually be allowed access to the resource itself and how much this will cost them. Metadata is, in short, a means by which largely meaningless data may be transformed into information, interpretable and reusable by those other than the creator of the data resource. (4)

Thoughtfully created metadata that provide context into the creation and scope of a research data set are not only good practice but also essential to supporting data reuse. As the ICPSR (2012) explains, metadata are "often the only form of communication between the secondary analyst and the data producer, so they must be comprehensive and provide all of the needed information for accurate analysis."

There are a number of forms that metadata take in relation to research data. Structured information, often encoded as XML, is one. To share data in a disciplinary or institutional repository, a researcher will often be expected to provide information about the data set in a specified format, or metadata schema.

When developing a DMP and if planning on archiving their data in a repository, researchers are well advised to look at the metadata requirements in a data repository and to plan accordingly. When a linguist chooses to archive and share their research data, as discussed in Andreassen (chapter 7, this volume), they may encounter either a metadata schema that is disciplinary-agnostic or one that has been built with linguistics research in mind. In the case of a university institutional repository, the linguist would be more likely to encounter a domain-neutral schema. A linguist depositing in the Data Archive at the Max Planck Institute for Psycholinguistics, on the other hand, will be invited to provide more robust metadata to describe session recordings and annotations; this repository employs the ISLE Meta Data Initiative, "a metadata standard to describe multi-media and multi-modal language resources" (Geerts 2018).

In addition to International Standard for Language Engineering, there are a number of metadata schemas that linguists may adopt for describing their research data sets or that they may encounter when archiving data in a repository. Dublin Core, likely the most well-known metadata schema, is one of these. Made up of fifteen elements (e.g., creator, date, title), Dublin Core is a simple, all-purpose scheme and "has the most mapped element sets among and across domain-specific and community-oriented metadata standards" (Zeng & Qin 2008:16). Among these community-specific metadata schemas built from Dublin Core is the Open Language Archives Community standard. Designed to facilitate sharing of linguistics data, Open Language Archives Community metadata include all fifteen Dublin Core elements and elements that would make "it possible to describe language resources with greater precision" (e.g., discourse type, linguistic field) (Simons & Bird 2008).

The DataCite metadata schema is a domain-neutral set of elements, or fields, developed for structuring information about a data set (DataCite Metadata Working Group 2017, 2018). The standard includes a small number of required metadata elements (e.g., creator, title, resource type) and additional recommended and optional elements (e.g., description, rights, version). Numerous disciplinary metadata schemas have also been developed and the Digital Curation Centre (n.d.) has produced a valuable catalog of these standards. In the social sciences, the Data Documentation Initiative standard "an international XML-based standard for the content, presentation, transport, and preservation of documentation (i.e., metadata)" for data sets is widely accepted and employed.[15] ICPSR (n.d.), for example, uses Data Documentation Initiative as the repository metadata standard.

In addition to structured information that accompanies a data set, metadata may take the form of a readme file or data dictionary. A readme file is a simple text file that helps other users to understand and contextualize the data and to discern the interconnections among project records and data sets. The readme file additionally

highlights the methods for data collection and the ways in which the data were processed.[16] Cornell University Libraries offers a comprehensive overview of the information that should be included in a readme file (table 5.2).

Data dictionaries share the purpose of a readme file in that they provide necessary contextual information to support the understandability and clearness of the data set. However, data dictionaries generally focus on content included under the "Data-specific information" in Cornell's readme file template. Often created to support a database or spreadsheet, data dictionaries serve as variable glossary and key and are generally structured in a tabular format.[18] DataONE, a scientific community-led project focused on building researcher capacity for data management and sharing, explains that "a data dictionary provides a detailed description for each element or variable in your data set and data model. Data dictionaries are used to document important and useful information such as a descriptive name, the data type, allowed values, units, and text description."[19] For linguists building databases—for example, of endangered language metadata or for a crosslinguistic typological study—a data dictionary can help ensure data quality, interpretation, and reuse.[20]

Table 5.2
Cornell University: Recommended content for a readme file

General information
- Provide a title for the data set
- Name/institution/address/e-mail information for
  - Principal investigator (or person responsible for collecting the data)
  - Associate or coinvestigators
  - Contact person for questions
- Date of data collection (can be a single date, or a range)
- Information about geographic location of data collection
- Keywords used to describe the data topic
- Language information
- Information about funding sources that supported the collection of the data

Data and file overview
- Short description of what data it contains
- Format of the file if not obvious from the file name
- If the data set includes multiple files that relate to one another, the relationship between the files or a description of the file structure that holds them (possible terminology might include "data set" or "study" or "data package")
- Date that the file was created
- Date(s) that the file(s) was updated (versioned) and the nature of the update(s), if applicable
- Information about related data collected but that is not in the described data set

Sharing and access information
- Licenses or restrictions placed on the data
- Links to publications that cite or use the data
- Links to other publicly accessible locations of the data
- Recommended citation for the data

Methodological information
- Description of methods for data collection or generation (include links or references to publications or other documentation containing experimental design or protocols used)
- Description of methods used for data processing (describe how the data were generated from the raw or collected data)
- Any instrument-specific information needed to understand or interpret the data
- Standards and calibration information, if appropriate
- Describe any quality-assurance procedures performed on the data
- Definitions of codes or symbols used to note or characterize low quality/questionable/outliers that people should be aware of
- People involved with sample collection, processing, analysis and/or submission

Data-specific information
- Count of number of variables, and number of cases or rows
- Variable list, including full names and definitions (spell out abbreviated words) of column headings for tabular data
- Units of measurement
- Definitions for codes or symbols used to record missing data
- Specialized formats or other abbreviations used

Source: Cornell University's "Guide to Writing 'readme' Style Metadata."[17]

5 Conclusion

Section 4 looked at basic practices to support sustainable research data. This concluding section considers broader principles. Lavoie (2012) argues that sustainable research data can be equated to sustainable data curation practices. Sustainable practices are ones that are most compatible to researchers' existing workflows.

Core to effective data management is the creation of a strategy that is compatible to a researcher's existing workflows. If a researcher regularly uses software or applications that can be adopted to strengthen data management practices, this may be the more effective approach than learning an entirely new software system to assist with data management; for example, if a linguist is comfortable with Excel or a relational database system, they might consider using this familiar tool for metadata creation, rather than a distinct metadata creation tool.[21] This is because the best data management strategy is one that the researcher is able to consistently employ throughout the research life cycle.

Perhaps the most important principle to data management is that a future-minded orientation is essential. A consistent, effective data management approach ensures that the data creator is able to make sense of their own data two weeks, three months, or four years from when it was collected. A researcher should assess what would be valuable for their own memory, their own continued access, and their own future use of the data.

Moreover, funders and open data advocates, both major players in advancing the development of data management policy and approaches, have emphasized the value of data for future reuse. In linguistics, it is not difficult to recognize why reusable data is so essential. In the subfield of language documentation, for example, it is critical to future research that there is a reusable record of a language community that has no remaining fluent users or that is endangered. In linguistics subfields more broadly—whether experimental research in phonetics or psycholinguistic studies—there are benefits that come from having sustainable data. Linguists can use data sets for replication and for expanding on previous lessons drawn from the data.

The linguistics research community has made concerted efforts to consider long-term data stewardship, as evidenced by the establishment of language-focused data archives and metadata schemas (see Andreassen, chapter 7, this volume, and Buszard-Welcher, chapter 10, this volume). Ultimately, however, the stewardship of data through its life cycle and oversight of research data sustainability fall primarily to individual linguists. While supporting the longevity of information was once the role of librarians, repository managers, and archives, inaction on the part of researchers will place data at risk for loss and impede the possibility of reuse. Part II of this volume offers specific case studies of how linguists in different subfields and on different projects have approached data management in practice.

Acknowledgments

I am grateful for the thoughtful feedback and valuable recommendations from the reviewers assigned to this chapter and the editors. I am particularly grateful for their assistance in situating this chapter more deeply in the linguistics context.

Notes

1. UK Data Service, "Research data lifecycle," https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx.

2. Linguistics Data Consortium, https://www.ldc.upenn.edu/.

3. Research Data Alliance, "Linguistics Data IG," https://www.rd-alliance.org/groups/linguistics-data-ig.

4. As a starting point, see Corti et al. (2014). Guides by academic libraries are valuable sources of information about data management practices: University of Minnesota Libraries, "Research Data Services," https://www.lib.umn.edu/datamanagement.

5. Australian National Data Service, "File formats," https://www.ands.org.au/guides/file-formats.

6. Academic library guides are useful sources for a high-level discussion of sustainable formats. See, for example, Stanford Libraries, "Best practices for file formats"; University of Pennsylvania Libraries, "Data management best practices: Sustainable file types," https://guides.library.upenn.edu/datamgmt/fileformats.

7. UK Data Service, "Recommended formats," https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats.aspx.

8. Library of Congress, "The sustainability of digital formats," last updated March 25, 2019, https://www.loc.gov/preservation/digital/formats/index.html.

9. Cornell University Research Data Management Service Group, "File formats," https://data.research.cornell.edu/content/file-formats.

10. Language Archive (2019).

11. National Archives and Records Administration, "Appendix A: Tables of file formats," last updated September 2019, https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html.

12. International Standards Organization, "Date and time format—ISO 8601," https://www.iso.org/iso-8601-date-and-time-format.html.

13. Cornell University Research Data Management Service Group, "File management," https://data.research.cornell.edu/content/file-management.

14. University of Virginia Library Research Data Services+Sciences, "Data storage and backups," https://data.library.virginia.edu/data-management/plan/storage/.

15. Stanford Libraries, "Advanced metadata," https://library.stanford.edu/research/data-management-services/data-best-practices/creating-metadata/advanced-metadata.

16. University of Pittsburgh Library System, "Research data management @ Pitt," https://pitt.libguides.com/managedata/.

17. https://data.research.cornell.edu/content/readme; Cornell's guidance is licensed under a Creative Commons Attribution 4.0 International License.

18. Several data dictionary templates exist that can be adapted for a linguistics project. The open government data community has provided templates to guide the creation of data documentation for shared data sets, but these have value beyond this sector. See, for example, NYC OpenData's data dictionary template (accessible under "Resources and guidelines," https://opendata.cityofnewyork.us/open-data-coordinators/) and the US Department of Agriculture's template (https://data.nal.usda.gov/data-dictionary-blank-template).

19. DataONE, "Create a data dictionary," accessed October 2, 2019, https://www.dataone.org/best-practices/create-data-dictionary.

20. Linguistics examples were offered by volume peer reviewer.

21. Stanford Libraries, "Metadata tools," https://library.stanford.edu/research/data-management-services/data-best-practices/creating-metadata/metadata-tools.

References

Ball, Alex. 2012. *Review of Data Management Life cycle Models*. REDm-MED project document. Bath, UK: University of Bath. https://purehost.bath.ac.uk/ws/portalfiles/portal/206543/redm1rep120110ab10.pdf.

Bantin, Philip C. 1998. Strategies for managing electronic records: A new archival paradigm? An affirmation of our archival traditions? *Archival Issues* 23 (1): 17–34.

Berez-Kroeker, Andrea L., Helene N. Andreassen, Lauren Gawne, Gary Holton, Susan Smythe Kung, Peter Pulsifer, Lauren B. Collister, The Data Citation and Attribution in Linguistics Group, and the Linguistics Data Interest Group. 2018. *The Austin Principles of Data Citation in Linguistics*, version 1.0. http://site.uit.no/linguisticsdatacitation/austinprinciples/. Accessed October 24, 2019.

Berman, Francine. 2008. Got data? A guide to data preservation in the information age. *Communications of the ACM* 51 (12): 50–56.

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): 1–12. https://doi.org/10.7191/jeslib.2016.1101.

Corti, Louise, Veerle Van den Eynden, Libby Bishop, and Matthew Woollard. 2014. *Managing and Sharing Research Data: A Guide to Good Practice*. London: Sage Publications Inc.

Cox, Andrew Martin, and Winnie Wan Ting Tam. 2018. A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management* 70 (2): 142–157. https://doi.org/10.1108/AJIM-11-2017-0251.

DataCite Metadata Working Group. 2017. *DataCite Metadata Schema for the Publication and Citation of Research Data*, version 4.2. http://doi.org/10.5438/rv0g-av03.

DataCite Metadata Working Group. 2018. *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data*, version 4.2. https://doi.org/10.5438/bmjt-bx77.

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

Digital Curation Centre. n.d. *Disciplinary Metadata*. http://www.dcc.ac.uk/resources/metadata-standards.

Faundeen, John L., Thomas E. Burley, Jennifer A. Carlino, David L. Govoni, Heather S. Henkel, Sally L. Holl, Vivian B. Hutchison, et al. 2013. *The United States Geological Survey Science Data Lifecycle Model*. Reston, VA: US Geological Survey. http://dx.doi.org/10.3133/ofr20131265.

Gawne, Lauren, and Andrea L. Berez-Kroeker. 2018. Reflections on reproducible research. In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 22–32. Honolulu: University of Hawai'i Press. https://scholarspace.manoa.hawaii.edu/bitstream/10125/24805/ldc-sp15-gawne.pdf.

Geerts, Jeroen. 2018. *Deposit Manual*, version 1.1. Nijmegen, the Netherlands: The Language Archive, MPI for Psycholinguistics. https://archive.mpi.nl/deposit-manual.

Hart, Edmund, Pauline Barmby, David LeBauer, François Michonneau, Sarah Mount, Patrick Mulrooney, Timothée Poisot,

et al. 2016. Ten simple rules for digital data storage. *PLoS Computational Biology* 12 (10): e1005097. https://doi.org/10.1371/journal.pcbi.1005097.

Himmelmann, Nikolaus P. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation and Conservation* 6:87–207. http://hdl.handle.net/10125/4503.

Inter-university Consortium for Political and Social Research (ICPSR). 2012. *Guidelines for Effective Data Management Plans*. Ann Arbor, MI: ICPSR. https://www.icpsr.umich.edu/files/datamanagement/DataManagementPlans-All.pdf.

Inter-university Consortium for Political and Social Research (ICPSR). n.d. *Metadata*. https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/metadata.html. Accessed October 24, 2019.

Language Archive. *ELAN—Linguistic Annotator*, version 5.8. Last updated October 8, 2019. https://tla.mpi.nl/tools/tla-tools/elan/.

Lavoie, Brian F. 2012. Sustainable research data. In *Managing Research Data*, ed. Graham Pryor, 67–82. London: Facet Publishing.

Mannheimer, Sara. 2018. Toward a better data management plan: The impact of DMPs on grant funded research practices. *Journal of EScience Librarianship* 7 (3): 1–18. https://doi.org/10.7191/jeslib.2018.1155.

Mattern, Eleanor, Wei Jeng, Liz Lyon, Daqing He, and Aaron Brenner. 2015. Using participatory design and visual narrative inquiry to investigate researchers' data challenges and recommendations for library research data services. *Program: Electronic Library and Information Systems* 49 (4): 408–423. https://doi.org/10.1108/PROG-01-2015-0012.

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.

Miller, Paul. 2004. Metadata: What it means for memory institutions. In *Metadata Applications and Management*, ed. G. E. Gorman and Daniel G. Dorner, 4–16. Lanham, MD: Scarecrow Press.

National Digital Stewardship Alliance (NDSA). 2013. *Levels of Digital Preservation,* version 1. https://ndsa.org/activities/levels-of-digital-preservation/.

Noble, William Stafford. 2009. A quick guide to organizing computational biology projects. *PLOS Computational Biology* 5 (7): e1000424. https://doi.org/10.1371/journal.pcbi.1000424.

Owens, Trevor. 2018. *The Theory and Craft of Digital Preservation*. Baltimore: Johns Hopkins University Press.

Poole, Alex H. 2016. The conceptual landscape of digital curation. *Journal of Documentation* 72 (5): 961–986.

Simons, Gary, and Steven Bird, eds. 2008. *Recommended Metadata Extensions*. Open Language Archives Community. http://www.language-archives.org/REC/olac-extensions.html.

Smithsonian Library. n.d. *Smithsonian Data Management Best Practices: Naming and Organizing Files*. https://library.si.edu/sites/default/files/tutorial/pdf/filenamingorganizing20180227.pdf.

Stanford Libraries. n.d. Best practices for file formats. https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats.

Styler, Will. 2017. *Using Praat for Linguistic Research*, version 1.8. Last updated December 25, 2017. http://wstyler.ucsd.edu/praat/.

Thieberger, Nicholas, and Andrea L. Berez. 2012. Linguistic data management. In *The Oxford Handbook of Linguistic Fieldwork*, ed. Nicholas Thieberger, 90–118. Oxford: Oxford University Press.

Van den Eynden, Veerle, Louise Corti, Matthew Woollard, Libby Bishop, and Laurence Horton. 2011. *Managing and Sharing Data: Best Practices for Researchers*, 3rd ed. Colchester, UK: UK Data Archive. https://data-archive.ac.uk/media/2894/managingsharing.pdf.

Witmer, Scott David. 2017. Personal digital archiving guide part 1: Preservation planning. *Bits and Pieces* (University of Michigan Libraries blog). April 26, 2017. https://www.lib.umich.edu/blogs/bits-and-pieces/personal-digital-archiving-guide-part-1-preservation-planning.

Zeng, Marcia Lei, and Jian Qin. 2008. *Metadata*, 3rd ed. New York: Neal-Schulman Publishers.

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