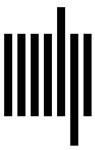
This is a section of doi:10.7551/mitpress/12200.001.0001

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

7 Archiving Research Data

Helene N. Andreassen

1 Introduction

Imagine you are a senior researcher in neurolinguistics, and you have recently submitted a paper to an international journal that requires data sets underpinning the analyses to be made openly available to the readers.[1] You get very positive feedback overall from the reviewers, but they ask you to provide more details on some aspects of the data collection methods, which you agree to do. After revision, your paper gets published and your results provoke quite some discussion among the researchers in your field. A fellow neurolinguist, whose work you've read but who you've never met in person, performs a replication study, using the original data set, to test the validity of your findings, and this fortunately proves successful. Moreover, she has ideas about how to take this research further and invites you to collaborate with her team on a future project.

Now imagine you are a PhD research fellow in phonology, and you are deep into the analysis chapter of your dissertation. You read a peer-reviewed paper about relevant phonological processes in a related language, where the authors claim the existence of a given pattern. This is stellar evidence to include in your own argument, but there are no examples in the paper, nor any link to the data. You are unsure whether you can trust this claim without seeing any empirical evidence, but you don't know the authors personally and fear it would be perceived as impolite to contact them and ask for data. You thus end up not incorporating the alleged finding into your chapter.

Finally, imagine you are a language teacher at your tribe's immersion school, and you come across a collection of newly digitized language recordings from a previous generation of speakers. You are able to delve into these recordings and enhance your lesson plans on particular grammatical constructions by providing new examples of them in use. Or imagine you are a postdoctoral researcher in morphosyntax with funding from the regional center for indigenous research, education, and knowledge production. Because this data collection is published with open access, you are able to compare recordings across generations and get one step further in identifying the context for an ongoing morphological change.

These scenarios together underline the importance of being thorough and transparent in terms of research data management (RDM) and of working toward a culture of sharing knowledge.[2] This is unfortunately not always the case. While Thomason (1994) observes too many instances of sloppy and unreflective practices in authors' treatment and analysis of research data, over twenty years later, Berez-Kroeker et al. (2017) and Gawne et al. (2017) demonstrate that there is still much room for improvement regarding the transparency of methods and location of research data in linguistic publications.[3]

To approach an ideal state of research transparency, Thomason (1994) suggests advice that places responsibility on both the authors and reusers.[4] Authors should provide sufficient details about the sources of data and methodology of data collection to allow for a deeper understanding and evaluation of their research. Reusers, on their part, should consult and cite primary sources when available; *primary source*, as Thomason uses it, refers to the fieldworker linguist's publication, which for practical purposes is the nearest one can get to the language user. It is worth noting that Thomason wrote her editorial note in the pre-Internet era, when there was no technical infrastructure linking publications and supporting research data, nor even any data repositories.[5] Publications were the primary source and the main window to empirical evidence. Today, with the infrastructure in place and an ever-increasing number of high-quality repositories (see Whyte 2015), we should conceive of the

primary source as the publication *in combination with the supporting research data*.

In this chapter, I discuss archiving research data: why and how to do it.[6] Although the focus is on archiving open data, the large majority of advice also holds for data with restricted access. In section 2, I present potential benefits of archiving, and in section 3, I turn to some key barriers to data sharing. In section 4, I focus on how to archive research data, including how to select a data repository, and in section 5, I turn to the challenge of archiving data with personal information. Section 6 concludes the chapter.

2 Potential benefits of archiving research data

Data collection is time-consuming, labor-intensive, and costly; sharing these valuable resources with others might not be the first thought that comes to mind when your data set is finalized.[7] Recently, however, there has been increasing focus on open science in academia, where two key arguments are knowledge sharing and research transparency. In what follows, I detail a set of potential benefits of opening up your research via data archiving (see also Tenopir et al. 2011).

First, archiving might be advantageous for the authors of research data. For instance, if reviewers of your manuscript can also access the data, they may be able to provide more complete feedback on your work, which might improve its quality. It is true that the review of data for peer-reviewed publications is currently not standard procedure, but initiatives such as the Peer Reviewers' Openness Initiative (Morey et al. 2016) indicate an increased focus on data access. Moreover, access to well-described data improves the replicability of your research,[8] and replication studies, in turn, may strengthen the credibility of your arguments (Peels 2019), or, conversely, speed up the correction or retraction of your publication (see, e.g., Ijzerman et al. 2015). Furthermore, because data sets can be conceived of as scholarly products in their own right, as suggested, for example, by the San Francisco Declaration on Research Assessment (DORA, n.d.), they deserve intellectual recognition and should be cited with as much care and detail as other sources (see Conzett & De Smedt, chapter 11, this volume). If others adhere to this practice when they consult or reuse your data, you get the credit you deserve. Another potential benefit of archiving your data is the enhanced visibility of your research, which might boost the chances of your work having a greater impact on future research and innovation. More concretely, visibility may lead to new collaborations or to more people using your research and citing you. Publications containing a link to the data have been found to have increased citation rates compared with publications with no such link (Drachen et al. 2016; Leitner et al. 2016). Finally, an increasing number of funding agencies and publishers require or highly encourage that the data underpinning publications be made openly available, providing there is no legal, ethical, security, or commercial reason not to do so (see, e.g., European Commission 2016). Willingness to share data might thus improve your future chances of getting funding and of being evaluated for publication in desired channels.

Creating knowledge is a joint project (Bender & Good 2010), and archiving data has great potential benefits to the research community. For instance, when you deposit your data in a high-quality repository, you ensure that peers can find and reuse them.[9] This might reduce the risk of duplication of effort and thereby positively affect cost efficiency. The time you allot to data archiving could be evaluated in light of the probability of the reuse of the data, to obtain a certain efficiency gain (Pronk 2019). Archiving linguistic data may be particularly important regardless, because many studies in the field focus on observation of individuals and are therefore not directly replicable (Berez-Kroeker et al. 2018). In addition, the fact that 47.1% of the world's living languages are to some degree endangered (Belew & Simpson 2018:28) underlines the importance of securing future access to existing data. Another potential positive side effect of data sharing is the improvement of scientific methods. Many repositories recommend that authors include—or at least link to—information about the methods used in the data collection and in the data analysis (Pedersen 2008; Wieling, Rawee, & van Noord 2018). With access to both data and methods, your research community is better equipped to discuss and evaluate the methods, which in turn may guide the design of future research. Finally, data archiving may facilitate comparison or aggregation of data sets from different studies. It is true that comparable data are not easily achieved, given that most individual data collections and treatments are carried out with specific research questions in mind. However, if data sets are uploaded to a repository with a generous reuse license, other researchers may recode and reanalyze the data and make them

fit into larger data sets (see, e.g., Kendall 2015). In addition, the ongoing discussion of strategies for facilitating the comparison and aggregation of data sets may in time reduce the expected workload (Cieri 2014; see also Bhattacharya et al. 2018).

Even though most linguistic data are collected for research purposes, they may also serve a role in education. This section therefore ends with some thoughts on how data archiving might benefit teachers and students. While instruction in higher education has traditionally focused on transmitting information to students, current awareness of the positive effect of student-centered teaching (Freeman et al. 2014) should encourage teachers of language and linguistics to actively involve students in the learning process. Also trending is the closely related research-based teaching approach, where learning activities center around the research process and research skills, "such as the ability to . . . gather and analyse data" (Huet 2018:728). Many teachers see the learning benefit of students collecting data themselves, but this may be too time-consuming or otherwise impractical, especially if the desired language variety is geographically inaccessible. Data collection might also be too challenging for students if the language users belong to a marginalized group, because this might require a high level of expertise on the part of the data collector (von Benzon & van Blerk 2017). With the increasing availability of open data, teachers can carry out research-based learning activities using real research data, without having to leave the classroom (Atenas, Havemann, & Priego 2015). The exploratory approach of the "three Is"—that is, *illustration* (looking at data), *interaction* (discussing the data), and *induction* (discovering rules)—is one example of data-driven learning that fits nicely with an active learning pedagogy (Johns 1991; McEnery & Xiao 2010), and which is feasible thanks to reusable research data.

This section has presented a number of potential benefits of data archiving, which in principle should encourage researchers to devote time to this in the research process. The next section reflects on a set of barriers that might explain why many researchers nevertheless still refrain from archiving.

3 Barriers to archiving research data

Although there are many potential benefits from archiving research data in a repository, less than 50% of researchers actually practice it (Berghmans et al. 2017; Stuart et al. 2018). At the same time, a higher number of researchers view data sharing favorably, or at least see that they could benefit from having access to other people's data (Tenopir et al. 2011; Berghmans et al. 2017). Why this mismatch between attitude and practice?

Archiving the research data that underpin text publications is fairly new, and many scholars have not acquired the skills needed to organize and document their data according to best practices. Nor do they know which repository to use. While these are legitimate barriers, current measures are in place to support researchers in this work. For instance, doctoral education in many countries is increasingly including RDM training, and there is also a plethora of free, online courses on basic RDM, such as, the FOSTER Open Science Training Courses (FOSTER, n.d.) and the Technische Universiteit (TU) Delft Open Science MOOC (TU Delft, n.d.).[10] Finally, an increasing number of institutions have their own RDM teams, typically located in the library, which offer courses and guidance on how to structure, document, and archive data. Nevertheless, preparing data for archiving takes time, regardless of whether you implement good practices from the beginning or make an all-out effort, and isolating hours to do this amid other commitments is not an easy task. In addition, because the publication of research data currently does not reward any publication points, is generally associated with little prestige, and because there are few immediate consequences if authors don't publish their data, the relatively low percentage of researchers who archive is, to a certain extent, understandable. If it is your intention to archive your data, I suggest writing a data management plan early in the project, which provides a rough overview of what needs to be done to get the data ready for archiving (see Kung, chapter 8, this volume).

A second barrier to data archiving might be the fear of what happens when others get access to the data. Some might be afraid that others will carry out and publish research that the authors of the data could have done themselves. Others might be afraid of having their data scrutinized and of possible criticism from peer reviewers or colleagues. While critique might require that they work on the data more—or, in the worst-case scenario, lead to their paper getting retracted—it might also provoke a fear of not having done the data collection well enough. Again, these are legitimate fears that should be taken seriously by

advocates of open data. However, given the current focus on the transparency of science in many fields and institutions, we can hope that RDM gains more focus and prestige in the research community in the foreseeable future, and that data sharing is not seen as a response to institutional or journal requirements, but more as an integrated part of the publication process where authors and peers can constructively communicate with reference to both text and data material.

One final barrier, which pertains to all research involving human beings, is ethical and legal concerns. As considered in sections 4 and 5, there are many complex issues related to data protection, and many researchers are unsure about what they can and cannot archive openly, how to archive protected material, and whom to ask for advice. I haven't come across any data-sharing regulations requiring researchers to openly archive sensitive material, which clearly indicates that, for all stakeholders, data protection trumps data sharing. However, the issue is not black and white. Even research with protected data should be transparent; this can be achieved rather easily with open metadata.[11] We return to this in section 5.

The barriers highlighted here are genuine and do not always come with quick-fix solutions. For this reason, I strongly suggest that you introduce these barriers as topics of discussion in your research group. This way, with many individual researchers expressing the same concerns, the barriers might be put on the agenda in relevant forums and ultimately contribute to positive change. I also would suggest that you seek advice from relevant entities outside your research group that might help you overcome some of these barriers, whether an institutional RDM team, legal team, or data protection officer.

In the next section, I focus on how to archive research data. First, I offer some advice on how to select a repository, before turning to the key actions of the archiving process itself.

4 How to archive research data

If you plan to deposit data in a repository, you must first become familiar with the basic aspects of the RDM life cycle (see Mattern, chapter 5, this volume).[12] The reason is simple: how you plan your project and how you manage your data throughout the project period will determine which data you can archive, and how efficiently you can do it. In this section, which focuses on the publication phase of the RDM life cycle, I use information from two different repositories for support and illustration: while the Archive of the Indigenous Languages of Latin America (AILLA, n.d.; see deposit guide in Kung & Sullivant 2018) focuses on data from a geographically restricted region, and typically contains large collections with restricted access, the Tromsø Repository of Language and Linguistics (TROLLing, n.d.; see deposit guide on DataverseNO, n.d.b) focuses on open, processed data sets, typically replication data for published research papers.

4.1 Selection of repository

When we write research papers, we do not haphazardly select the journal to which we submit. Rather, we examine whether the journal has a peer-review system, whether its readership corresponds to our intended one, and—more and more often, as requirements on open science become stricter—whether it allows open access publication or self-archiving of manuscripts. You should be equally critical when selecting a data repository, in particular if you are preoccupied with the visibility and safeguarding of your research data.

4.1.1 Repository search Repositories come in different types:[13]

- *Domain-specific* repositories, because they are closely linked to a research community, typically focus on a certain type of data and normally provide extensive curation services.[14] Examples: CHILDES (Child Language Data Exchange System, n.d.) and PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures, n.d.).
- *General* repositories typically serve a broad range of disciplines, but generally provide few curation services. Examples: Figshare (n.d.) and Zenodo (n.d.).
- *Institutional* repositories, built to document and preserve research produced at an institution, are typically run by libraries. Curation services may vary with resources. Example: DataverseNO (n.d.a).
- *National* repositories may complement institutional repositories, for example, by offering services for large data sets or data with sensitive information. Example: CESSDA Data Catalogue (Consortium of European Social Science Data Archives, n.d.).
- *Project-specific* repositories are tailored to specific data sets and contributor groups. Examples: ESLO (Enquêtes sociolinguistiques à Orléans, n.d.), TGDP (Texas

German Dialect Project, n.d.), and Oahpa (n.d.; online language-learning resources for the Sami languages).

Before embarking on data collection, you should have an idea about which repository to use, as repositories may vary with regard to requirements on content and metadata. It may happen that you discover, later in the research process, that the planned repository is not optimal after all, given the nature of the data you ended up collecting. I therefore suggest that you continually evaluate the match between the data and the preselected repository to make the deposit process the least cumbersome possible.

To find a suitable repository, I encourage you to discuss possibilities with your research group or others with expertise on your research topic and methodology. Another strategy is to examine the repository information in data citations in the bibliographic reference lists of relevant scientific literature. You could also use the Registry of Research Data Repositories (re3data, n.d.), a registry where you can filter for subject (e.g., linguistics), and where you can find information about, for example, the topic and terms of use of each repository. You could further consult the repositories listed in the Open Language Archives Community (OLAC, n.d.) or browse the Common Language Resources and Technology Infrastructure Virtual Language Observatory (CLARIN, n.d.b). In most cases, the repository homepage provides information on whether it is open for deposit by external scholars.

4.1.2 Repository checklist Alter and Gonzalez (2018) recommend using domain-specific repositories, as these are most likely to have domain-specific expertise and curation services that will enhance the value of the data. They further recommend "trusted" repositories, which support archival standards for discovery, documentation, and preservation. However, you should also consider the audience(s) you want to reach with your data, and how you can achieve this. In this section, I detail five aspects to consider when selecting a repository, taken from Whyte (2015):

1. Is the repository reputable?
2. Will the repository take the data you want to deposit?
3. Will the data be safe in legal terms?
4. Will the repository sustain the data value?
5. Will the repository support analysis and track data usage?

Information that may help you answer these questions can normally be found in the repository mission statement, deposit guidelines, or curator guidelines. If you find the terminology to be too cryptic or the information too vague, I recommend asking colleagues who have archiving experience or your local RDM team or contacting the repository directly.

Is the repository reputable? The repository should at least be listed in a repository registry, for example, TROLLing is listed in the Registry of Research Data Repositories and AILLA is listed in OLAC, or be broadly recognized in the research community. You might also want to investigate whether the repository has been endorsed by relevant funders, publishers, or societies. Some repositories are awarded a certificate stating their compliance with specific international standards, such as, CoreTrustSeal (n.d.). However, certification is rather new, and there are many repositories recommended in the research and publisher communities that are not (yet) certified (Husen et al. 2017); therefore, certification should not necessarily be used as the element that tips the scales in one direction or the other.

Will the repository take the data you want to deposit? Different discipline-specific repositories may accept different types of data. You might want to choose a repository that focuses on a specific type of data, such as, TROLLing, which primarily contains processed, open linguistic data and code, or a repository that focuses on linguistic data in general, for example, one of the CLARIN centers (CLARIN, n.d.a). You could search for a thematic repository with a solid international reputation in the domain and which publishes data similar to those you deposit, for example, CHILDES if you work with child language data, or PARADISEC if you work with endangered languages in the Pacific region around Australia. In brief, domain-specific repositories may have more or less strict requirements on data sets pertaining to research topic, methods, degree of processing, and technical specifications that you should investigate before starting to prepare your data for deposit.

Will the data be safe in legal terms? Depending on your type of data, you will need to consider the relevance of various legal terms and conditions. You should ensure that the ownership of the data is not transferred to a third party, but that it remains with the original owner, which is typically your institution, for example, from AILLA (n.d.): "This agreement does not take away

any rights from the depositor or any other creator of these materials; all parties to the creation of the materials retain all of their original rights." In addition, you should be able to determine which license to apply to the data[15] and to explicitly confirm that the data were created in accordance with legal and ethical criteria. This is typically done by signing the repository's terms and agreements before depositing data, for example, from TROLLing (n.d.):

> In order to submit a Dataset, you represent that . . . nothing in the Dataset, to the best of your knowledge, infringes on anyone's copyright or other intellectual property rights . . . nothing in the Dataset violates any contract terms (e.g., Nondisclosure Agreement, Material Transfer Agreement, Terms of Use, etc.) . . . nothing in the Dataset contains any private information, confidential information, proprietary information of others, export controlled information, or otherwise protected data or information that should not be publicly shared.

If you have data with personal information, these must be handled with particular care, as these are typically subject to data protection legislation. If your research is subject to the General Data Protection Regulation (GDPR, n.d.), the following requirements are in force. In all cases, you need to make sure that your institution has a data processing agreement with the desired repository and that the repository fulfills any security requirements issued by your institution. You must also check that you retain the right to control access to the data after depositing them. Finally, if you plan to archive the data outside your jurisdiction, keep in mind that the rules protecting them at home still apply, and that there are mechanisms for ensuring legal transfer to a so-called third country (see GDPR, n.d.:articles 44–50).

Will the repository sustain the data value? The FAIR data principles (see Janda, this volume; Wilkinson et al. 2016) constitute a set of guidelines to ensure that research data are *findable*, *accessible*, *interoperable*, and *reusable*. An increasing number of repositories support these, but unfortunately, many still don't.[16] If you want your data to be FAIR, you need to pay attention to aspects such as metadata, persistent identifiers, file format requirements, version control, and the possibility of linking to related materials. First, data sets should be discoverable at least through metadata of the title, author/creator, and date for deposit. Most repositories provide metadata fields where other information can be entered, such as domain, topic, language, type of data, data collection methods, and collection date. Metadata enhance the visibility of your data set and the ease of reuse, which is further facilitated if the repository uses metadata that are compliant with metadata standards in the field, such as, the Dublin Core Metadata Initiative (n.d.) for the humanities—and the OLAC Metadata (2008) extension for language resources specifically—and the Data Documentation Initiative (DDI, n.d.) for the social sciences. These ensure standardized descriptions of the data and facilitate comparison and aggregation of data sets (Cieri 2014). The repository should also provide a digital object identifier (DOI) or another persistent identifier for the data set landing page, which provides a persistent link to its location on the Internet. Some repositories issue persistent identifiers on subset- or data file–level, which can be useful functionality in the case of large data sets. I further recommend that you examine the repository's requirements on file formats, and whether it has a system for detecting non-persistent ones in the deposited data sets. The repository should also offer version control to ensure that all changes made to the data set after publication are tracked and explicitly detailed. Finally, check that there are metadata fields for related material. If the data are replication data for a text publication, the metadata should include a reference to the publication. If the data set is a subset of a larger collection archived elsewhere (or not archived), this should be specified in the metadata. In general, I recommend that you develop an understanding of how the desired repository is run, whether it provides curator services, and if there are long-term preservation strategies.

Will the repository support analysis and track data usage? When you archive your data, you might be interested in making them maximally visible, reused, and cited. If so, you should investigate whether the repository supports the harvesting of metadata by search engines or library discovery services. Moreover, in particular for larger repositories, you should check whether it is possible to retrieve your data set within the repository using keyword filters. Finally, you might want to know whether you can monitor the activity on your data set, for example, the number of views and downloads.

4.1.3 Requirements Funders, institutions, and journals may have requirements regarding the type of repository their researchers use. The US National Science Foundation,

for instance, which already requires that research data resulting from their funding must be shared, has expressed ambitions to investigate the repository landscape and to develop repository standards (see National Science Foundation 2015:7). Turning to requirements at the institutional level, these largely vary when it comes to levels of specification. At the more explicit end of the scale, for instance, is the TU Delft Research Data Framework Policy, which requires that institutional data that can be shared are archived in a repository that adheres to the FAIR principles and preserves the data for at least ten years. Their policy also requires that restricted data have archived metadata, and that any publication based on these data state why access is restricted and who can access them (TU Delft 2018:7). For an example of a journal publisher's requirements, I will refer to the research data policy of Springer Nature (n.d.) and in particular the policy of their journal *Natural Language and Linguistic Theory* (n.d.). This journal requires authors to provide sufficient information about the data collection in their manuscript, as well as repository information if the data have been archived. If such information is not given, the manuscript will be returned to the author, prior to review.

4.2 Key actions of data archiving

It is the researcher's responsibility to determine which data can be archived and which data should be archived. When making this decision, keep in mind that you cannot predict the future use of your data (Lindsay 2015), and that you need to think beyond the current specialist research community. Also keep in mind that future reuse of your data might necessitate access to one or more of the following:

- raw data
- processed data
- pilot data
- incomplete data sets
- notes
- negative results
- source code, statistical code
- experimental material

If you are in doubt about what to archive, consult the standard procedures in your subfield or discuss the potential value of the data with your research group (see also Digital Curation Centre 2014). When you have decided which repository to use, and you have finalized your data files, you can start working on your deposit. You can speed up this process by logging your RDM during the collection and treatment of the data, as much of this information is relevant for the metadata fields and the readme file.[17] Tenopir et al. (2011) observed a lack of awareness among researchers regarding the importance of metadata, and I therefore encourage you to become familiar with types of metadata and develop a strategy for recording them early in the RDM. The repository metadata templates typically contain both required and optional fields, and I recommend that you enter as much information as possible, in particular if one main purpose of archiving your data is reuse by others: keep in mind that the metadata you enter constitute what will be searchable by others. It may also be helpful to imagine yourself in the position of the reuser and reflect on which metadata you would need to trust the data set.

As mentioned in section 3, data archiving is a time-consuming process, and it might be tempting to simply enter a reference to your text publication in the data set metadata, directing the reuser there to find all of the necessary information on the research question and data collection methods. However, it is important to remember that in many cases there are still paywalls preventing researchers from accessing desired text publications.

You also need to evaluate the format of your files to secure future access. Both AILLA and TROLLing require that files come in a persistent format,[18] which means that, for example, the much-used Microsoft Excel and Word files must be converted prior to deposit. If you don't know how to convert your files, and if there is nothing in the deposit guide to help you, seek help online or from your local RDM or information technology team. Furthermore, most repositories come with recommendations about which license to apply to the data, that is, which type of reuse you will allow for your data after publication. There are many types of licenses, for example, Creative Commons (n.d.), and if this is unknown terrain, I encourage you to investigate what the license recommended by the repository actually implies.

A common desire among researchers is to keep their data locked down until publication of their research paper. However, you may want to cite your data in the text manuscript. Because one main component of a data citation is the location of the data (e.g., a DOI), you would

need to at least create the landing page of your data set prior to the submission of your manuscript. Repositories vary when it comes to flexibility in this regard. Both AILLA and TROLLing allow a temporary embargo on the data files, meaning that you can create and publish your data set, but only the metadata will be publicly available until the embargo period is over, typically when the paper is published.[19] You might also want the peer reviewers to access and provide feedback on your data set while evaluating your paper. Again, repositories vary, but in TROLLing, for instance, the system can create a private URL to the unpublished and still-modifiable data set, which you can send to the editor alongside the manuscript.

Finally, you may want to modify your data set after its initial publication; a typical change to the metadata would be adding the reference information for the corresponding research paper. More substantial changes may also occur, such as adding new files, either to complement the existing data set or to replace a file containing errors or personal information. Note that many repositories do not allow the deletion of files, but these often have version control, which automatically creates a new version number for the revised data set, where you can enter an explanation of the changes.

More information on how to prepare your data for archiving may be found in Mattern (chapter 5, this volume). In the next section, I turn to challenges associated with archiving data containing personal information.

5 How to archive data with personal information

As mentioned in section 3, an oft-cited concern among researchers is the protection of the language users who have volunteered to contribute valuable empirical material to our research (see Holton, Leonard, & Pulsifer, chapter 4, this volume). With the application of the GDPR on May 25, 2018, affecting all researchers in EU member states, stricter rules are now imposed on projects that involve data with personal information. Among other things, the regulation introduces the principle of accountability, whereby the data controller, the person who "determines the purposes and means of the processing of personal data" (article 4 no. 7), must demonstrate compliance with all aspects of the GDPR, including what happens with the data in the archiving process.[20] In general, institutional and journal policies clearly state that data should not be shared openly if there are ethical or legal reasons not to do so. These data nevertheless should be safeguarded in a suitable repository, except if they are subject to destruction for some reason. Furthermore, the metadata that can be shared should be archived openly to make the data set discoverable; see, for example, Meyerhoff and Schleef (2015). This way, even though access to the data is restricted, the research stays transparent in that the data set is discoverable and in principle accessible (Meyer 2018).

Repositories vary in what they support when it comes to file protection. For instance, while TROLLing requires all data files to be open, possibly after an embargo period, AILLA offers four different access levels. Given the totality of your data set, you need to determine what you can and cannot do. There are at least four possible alternatives:

1. Open: Raw data, processed data, metadata. Restricted: Nothing.
2. Open: Processed data, metadata. Restricted: Raw data.
3. Open: Metadata. Restricted: Raw data, processed data.
4. Open: Nothing. Restricted: Raw data, processed data, metadata.

Many projects will have some data that cannot be openly shared, for example, interviews with sensitive content, and some data that can, for example, annotated intonation curves extracted from interviews. For such projects, you might want to keep all files together in a repository with access control, such as AILLA, or select two different repositories for the different types of data, for example, AILLA for the interview files and TROLLing for the annotated intonation curves, and then link the data sets via metadata. Needless to say, the actual data landscape is more nuanced than what is spelled out here; one example that illustrates this is a project documenting the linguistic and musical diversity of the Warruwi community in Australia (O'Keeffe et al. 2018). The project had a dual purpose—to make the data available for research and make them reusable by the language community. The data collection contained narratives that shouldn't be heard by male language users, and while some female informants felt assured that a label such as "women and girls only" would be respected, others required that the data be archived with access control. We take from this that by collaborating closely with the language community, and by including the language

users as informed decision makers in different stages of the research process, researchers can manage the data in a way that also respects cultural differences (see Kirilova & Karcher 2017 for a recent reflection on archiving qualitative research data).

If the golden standard is for data to be "as open as possible, as closed as necessary," to cite the European Commission (2016:4), researchers constantly need to balance the trade-offs between sharing and risk and between ease of access and data protection. Again, writing a data management plan early in the project might help you identify possible challenges that could be overcome by interacting with the language users. When it comes to selecting a repository for data with personal information, Kirilova and Karcher (2017), from the Qualitative Data Repository (n.d.), suggest that you target one with personnel deeply familiar with your scientific domain and methodology, who can guide you on deposit-related aspects even during the research process.

6 Conclusion

This chapter has focused on archiving research data. I have presented some of the potential benefits of archiving as well as some perceived barriers, and I have given advice on the archiving process. I would like to end the chapter with a strong encouragement to not think of archiving as extra work on top of your research, but rather as an integral part of it. Archiving according to best practices is not only meant to respond to requirements and facilitate your research data management, but also, and more importantly, to make your research better and more transparent. If you consider it too time-consuming, with little reward, consider the fact that you as a researcher can contribute to the cultural change that is needed to make sharing worth it. If you're a junior researcher, prepare for the future by already building archiving into your routines. If you're a senior researcher, support and advocate the junior scholars who do a good job of archiving, and otherwise use your experienced voice to highlight the importance of research data and good research data management. Ultimately, if more people archive and share data, the potential benefits presented in this chapter may become more common, which in turn may improve the overall quality of the scientific enterprise.

Notes

1. I would like to thank Per Pippin Aspaas, Laura A. Janda, Ingvild Stock-Jørgensen (University of Tromsø – The Arctic University of Norway [UiT]), Andrea Berez-Kroeker (University of Hawai'i at Mānoa), and two anonymous reviewers for valuable comments on a previous version of this chapter. I would also like to thank the RDM team at the UiT University Library and the RDA Linguistics Data Interest Group for fruitful discussions on this topic over the years.

2. By *transparent*, I mean being explicit about the evidence supporting scientific claims, i.e., the application and implementation of methodology, the collection and analysis of data, and the interpretation of outcomes (Munafò et al. 2017).

3. By *publication*, I mean a scientific text publication.

4. By *reusers*, I mean people who read and cite publications or data sets in their own work, or people who use already published research data for different purposes.

5. By *repository*, I mean "a database or a virtual archive established to collect, disseminate and preserve scientific output . . . [where] the action of depositing material . . . is (self)archiving" (OpenAIRE 2018).

6. By *archiving*, I mean transferring data to a resource provider, e.g., a repository or a data center, all while complying with any documented guidance, policies, or legal requirements.

7. By *data set*, I mean data with content of a particular kind, that are related and treated collectively, and which have a shared and distinctive intended application (Renear, Sacchi, & Wickett 2010).

8. By *replicability*, I mean a "study . . . having certain features such that a replication study of it could be carried out" (Peels 2019:4).

9. See Vines et al. (2014) for a thought-provoking example from biology.

10. MOOC: Massive Online Open Courses. Links to web resources are included in the References.

11. By *metadata*, I mean "descriptive or contextual information which refers to or is associated with another object or resource . . . [which] usually takes the form of a structured set of elements which describe the information resource and assists in the identification, location and retrieval of it by users, while facilitating content and access management" (Higgins 2007; see also Mattern, chapter 5, this volume).

12. The RDM life cycle can be divided into three broad phases: the planning phase, where you prepare your project; the active phase, where you collect data and perform analyses of results; and the publication phase, where you publish your paper and archive your research data.

13. The description of domain-specific, general, and institutional repositories is taken from Alter and Gonzalez (2018).

14. By *curation*, I mean "maintaining, preserving and adding value to digital research data throughout its lifecycle" (Digital Curation Centre n.d.).

15. I refer to Collister (chapter 9, this volume) for details on copyright and licenses.

16. See Abu-Alam (2019) for a thought-provoking example from polar research.

17. By *readme file*, I mean a document that provides an overview and a short description of the data set; see Andreassen and Lyche (2017) and Arkhangelskiy (2019) for examples. This is often required by repositories.

18. See an example of file format guidelines here: https://site.uit.no/dataverseno/deposit/prepare/.

19. By *embargo*, I mean that access to the data files is restricted for a given time period.

20. For more information about the GDPR and language resources, see Kamocki, Ketzan, and Wildgans (2018).

References

Abu-Alam, Tamer S. 2019. Open Arctic Research Index: Final report and recommendations. https://doi.org/10.7557/7.4682.

AILLA (Archive of the Indigenous Languages of Latin America). n.d. https://ailla.utexas.org/. Accessed June 10, 2019.

Alter, George, and Richard Gonzalez. 2018. Responsible practices for data sharing. *American Psychologist* 73 (2): 146–156. https://doi.org/10.1037/amp0000258.

Andreassen, Helene N., and Chantal Lyche. 2017. Readme_schwa.pdf [data file]. In *Replication Data for: Le rôle de la variation dans le développement phonologique: Acquisition du schwa illustrée par deux corpus d'apprenants norvégiens*. DataverseNO, V1. https://doi.org/10.18710/QULOBC/TAWHYH.

Arkhangelskiy, Timofey. 2019. 00_ReadMe.txt [data file]. In *Replication Data for: Russian verbal borrowings in Udmurt*. DataverseNO, V1. https://doi.org/10.18710/5N34CG/RS5CIZ.

Atenas, Javiera, Leo Havemann, and Ernesto Priego. 2015. Open data as open educational resources: Towards transversal skills and global citizenship. *Open Praxis* 7 (4): 377–389. https://doi.org/10.5944/openpraxis.7.4.233.

Belew, Anna, and Sean Simpson. 2018. The status of the world's endangered languages. In *The Oxford Handbook of Endangered Languages*, ed. Kenneth L. Rehg and Lyle Campbell, 1–36. Oxford: Oxford University Press.

Bender, Emily M., and Jeff Good. 2010. A grand challenge for linguistics: Scaling up and integrating models. https://faculty.washington.edu/ebender/papers/GrandChallenge.pdf.

Berez-Kroeker, Andrea L., Lauren Gawne, Barbara F. Kelly, and Tyler Heston. 2017. A survey of current reproducibility practices in linguistics journals, 2003–2012. https://sites.google.com/a/hawaii.edu/data-citation/survey.

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.

Berghmans, Stephane, Helena Cousijn, Gemma Deakin, Ingeborg Meijer, Adrian Mulligan, Andrew Plume, Alex Rushforth, et al. 2017. Open data: The researcher perspective—Survey and case studies. Mendeley Data, version 1. https://doi.org/10.17632/bwrnfb4bvh.1.

Bhattacharya, Tanmoy, Nancy Retzlaff, Damián E. Blasi, William Croft, Michael Cysouw, Daniel Hruschka, Ian Maddieson, et al. 2018. Studying language evolution in the age of big data. *Journal of Language Evolution* 3 (2): 94–129. https://doi.org/10.1093/jole/lzy004.

Child Language Data Exchange System (CHILDES). n.d. https://childes.talkbank.org/. Accessed June 10, 2019.

Cieri, Christopher. 2014. Challenges and opportunities in sociolinguistic data and metadata sharing. *Language and Linguistics Compass* 8 (11): 472–485. https://doi.org/10.1111/lnc3.12112.

CLARIN (Common Language Resources and Technology Infrastructure). n.d.a. Depositing Services. https://www.clarin.eu/content/depositing-services. Accessed June 10, 2019.

CLARIN (Common Language Resources and Technology Infrastructure). n.d.b. Virtual Language Observatory. https://www.clarin.eu/content/virtual-language-observatory-vlo. Accessed June 10, 2019.

Consortium of European Social Science Data Archives (CESSDA). n.d. CESSDA Data Catalogue. https://datacatalogue.cessda.eu/. Accessed June 10, 2019.

CoreTrustSeal. n.d. https://www.coretrustseal.org/. Accessed June 10, 2019.

Creative Commons. n.d. About CC Licenses. https://creativecommons.org/about/cclicenses/. Accessed June 10, 2019.

DataverseNO. n.d.a. https://dataverse.no/. Accessed June 10, 2019.

DataverseNO. n.d.b. Deposit Guide. https://info.dataverse.no/. Accessed June 10, 2019.

DDI (Data Documentation Initiative). n.d. http://www.ddialliance.org/. Accessed June 10, 2019.

Digital Curation Centre. 2014. Five steps to decide what data to keep: DDC checklist for appraising research data, version 1. Edinburgh: Digital Curation Centre. https://www.dcc.ac.uk/guidance/how-guides/five-steps-decide-what-data-keep.

Digital Curation Centre. n.d. What is digital curation. https://www.dcc.ac.uk/about/digital-curation. Accessed June 10, 2019.

DORA (Declaration on Research Assessment). n.d. https://sfdora.org/. Accessed June 8, 2019.

Drachen, Thea Marie, Ole Ellegaard, Asger Væring Larsen, and Søren Bertil Fabricius Dorch. 2016. Sharing data increases citations. *Liber Quarterly* 26 (2): 67–82. https://doi.org/10.18352/lq.10149.

Dublin Core Metadata Initiative. n.d. http://www.dublincore.org/. Accessed June 10, 2019.

Enquêtes sociolinguistiques à Orléans (ESLO). n.d. http://eslo.huma-num.fr/. Accessed June 10, 2019.

European Commission. 2016. H2020 programme: Guidelines on FAIR data management in Horizon 2020, version 3.0. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

Figshare. n.d. https://figshare.com/. Accessed June 10, 2019.

FOSTER (Facilitate Open Science Training for European Research). n.d. Open Science training courses. https://www.fosteropenscience.eu/toolkit. Accessed June 10, 2019.

Freeman, Scott, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences* 111 (23): 8410–8415. https://doi.org/10.1073/pnas.1319030111.

Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker, and Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation and Conservation* 11:157–189. http://hdl.handle.net/10125/24731.

GDPR (General Data Protection Regulation). n.d. https://gdpr-info.eu/. Accessed June 10, 2019.

Higgins, Sarah. 2007. What are metadata standards. Digital Curation Centre. http://www.dcc.ac.uk/. Accessed June 10, 2019.

Huet, Isabel. 2018. Research-based education as a model to change the teaching and learning environment in STEM disciplines. *European Journal of Engineering Education* 43 (5): 725–740. https://doi.org/10.1080/03043797.2017.1415299.

Husen, Sean Edward, Zoë G. de Wilde, Anita de Waard, and Helena Cousijn. 2017. Recommended versus certified repositories: Mind the gap. *Data Science Journal* 16 (42): 1–10. https://doi.org/10.5334/dsj-2017-042.

Ijzerman, Hans, Nina F. E. Regenberg, Justin Saddlemyer, and Sander L. Koole. 2015. Perceptual effects of linguistic category priming: The Stapel and Semin (2007) paradigm revisited in twelve experiments. *Acta Psychologica* 157:23–29. https://doi.org/10.1016/j.actpsy.2015.01.008.

Johns, Tim. 1991. Should you be persuaded—two samples of data-driven learning materials. *English Language Research Journal* 4:1–16.

Kamocki, Pawel, Erik Ketzan, and Julia Wildgans. 2018. Language resources and research under the General Data Protection Regulation. In *CLARIN Legal Issues Committee (CLIC) White Papers Series*. https://www.clarin.eu/.

Kendall, Tyler. 2015. Making old data sources into new data sources: On the aggregation of sociolinguistic data sets and the future of real-time and cross-study analysis. *From Data to Evidence*, Helsinki, Finland, October 19–22. https://www.helsinki.fi/en/researchgroups/varieng/d2e-from-data-to-evidence.

Kirilova, Dessi, and Sebastian Karcher. 2017. Rethinking data sharing and human participant protection in social science research: Applications from the qualitative realm. *Data Science Journal* 16 (43): 1–7. https://doi.org/10.5334/dsj-2017-043.

Kung, Susan, and Ryan Sullivant. 2018. AILLA self-deposit tool training. Archive of the Indigenous Languages of Latin America. https://ailla.utexas.org/.

Leitner, Florian, Concha Bielza, Sean L. Hill, and Pedro Larrañaga. 2016. Data publications correlate with citation impact. *Frontiers in Neuroscience* 10:419. https://doi.org/10.3389/fnins.2016.00419.

Lindsay, Greg. 2015. The latest medical breakthrough in spinal cord injuries was made by a computer program. *Fast Company*, October 14, 2015. https://www.fastcompany.com/3052282/the-latest-medical-breakthrough-in-spinal-cord-injuries-was-made-by-a-computer-program.

McEnery, Tony, and Richard Xiao. 2010. What corpora can offer in language teaching and learning. In *Handbook of Research in Second Language Teaching and Learning*, ed. Eli Hinkel, 364–380. London: Routledge.

Meyer, Michelle N. 2018. Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science* 1 (1): 131–144. https://doi.org/10.1177/2515245917747656.

Meyerhoff, Miriam, and Erik Schleef. 2015. *Sociolinguistics and Immigration: Linguistic Variation among Adolescents in London and Edinburgh* (data set). UK Data Service. https://doi.org/10.5255/UKDA-SN-851797.

Morey, Richard D., Christopher D. Chambers, Peter J. Etchells, Christine R. Harris, Rink Hoekstra, Daniël Lakens, Stephan Lewandowsky, et al. 2016. The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science* 3:150547. https://doi.org/10.1098/rsos.150547.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, et al. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1:0021. https://doi.org/10.1038/s41562-016-0021.

National Science Foundation. 2015. Today's data, tomorrow's discoveries: Increasing access to the results of research funded by the National Science Foundation. https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf.

Natural Language and Linguistic Theory. n.d. Instructions for authors: Data management. https://www.springer.com/journal/11049/submission-guidelines#Instructions%20for%20Authors. Accessed June 10, 2019.

Oahpa. n.d. http://oahpa.no/index.eng.html. Accessed June 10, 2019.

O'Keeffe, Isabel, Linda Barwick, Carolyn Coleman, David Manmurulu, Jenny Manmurulu, Janet Gardjilart Bumarda Mardbinda, Paul Naragoidj, and Ruth Singer. 2018. Multiple uses for old and new recordings: Perspectives from the multilingual community of Warruwi. In *Communities in Control: Learning Tools and Strategies for Multilingual Endangered Language Communities. Proceedings of FEL XXI Alcanena 2017*, ed. Nicholas Ostler, Vera Ferreira, and Chris Moseley, 140–147. Hungerford, UK: Foundation for Endangered Languages.

OLAC (Open Language Archives Community). n.d. http://www.language-archives.org/archives. Accessed June 10, 2019.

OLAC Metadata. 2008. http://www.language-archives.org/OLAC/metadata.html. Accessed January 15, 2020.

OpenAIRE. 2018. What are repositories? Modified October 11, 2018. https://www.openaire.eu/. Accessed June 7, 2019.

Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). n.d. http://www.paradisec.org.au/. Accessed June 10, 2019.

Pedersen, Ted. 2008. Empiricism is not a matter of faith. *Computational Linguistics* 34 (3): 465–470. https://doi.org/10.1162/coli.2008.34.3.465.

Peels, Rik. 2019. Replicability and replication in the humanities. *Research Integrity and Peer Review* 4 (2): 1–12. https://doi.org/10.1186/s41073-018-0060-4.

Pronk, Tessa E. 2019. The time efficiency gain in sharing and reuse of research data. *Data Science Journal* 18 (10): 1–8. http://doi.org/10.5334/dsj-2019-010.

Qualitative Data Repository. n.d. https://qdr.syr.edu/. Accessed June 10, 2019.

re3data (Registry of Research Data Repositories). n.d. https://www.re3data.org/. Accessed June 10, 2019.

Renear, Allen H., Simone Sacchi, and Karen M. Wickett. 2010. Definitions of *dataset* in the scientific and technical literature. In *ASIS&T '10 Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem*, ed. Cathy Marshall, Elaine Toms, and Andrew Grove, 1–4. Silver Springs, MD: American Society for Information Science.

Springer Nature. n.d. Research data policies. https://www.springernature.com/gp/authors/research-data-policy/data-policy-types/12327096. Accessed June 10, 2019.

Stuart, David, Grace Baynes, Iain Hrynaszkiewicz, Katie Allin, Dan Penny, Mithu Lucraft, and Mathias Astell. 2018. Practical challenges for researchers in data sharing (whitepaper). Springer Nature. March 21, 2018. https://doi.org/10.6084/m9.figshare.5975011.

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. Data sharing by scientists: Practices and perceptions. *PLOS One* 6 (6): e21101. https://doi.org/10.1371/journal.pone.0021101.

Texas German Dialect Project (TGDP). n.d. https://tgdp.org/. Accessed June 10, 2019.

Thomason, Sarah G. 1994. The editor's department. *Language* 70 (2): 409–413. http://www.jstor.org/stable/415877.

TROLLing (Tromsø Repository of Language and Linguistics). n.d. https://trolling.uit.no. Accessed June 10, 2019.

TU Delft. 2018. TU Delft Research Data Framework Policy. https://doi.org/10.5281/zenodo.2573160.

TU Delft. n.d. Open Science: Sharing your research with the world (MOOC). https://online-learning.tudelft.nl/courses/open-science-sharing-your-research-with-the-world/.

Vines, Timothy H., Arianne Y. K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, et al. 2014. The availability of research data declines rapidly with article age. *Current Biology* 24 (1): 94–97. https://doi.org/10.1016/j.cub.2013.11.014.

von Benzon, Nadia, and Lorraine van Blerk. 2017. Research relationships and responsibilities: 'Doing' research with 'vulnerable' participants: Introduction to the special edition. *Social & Cultural Geography* 18 (7): 895–905. https://doi.org/10.1080/14649365.2017.1346199.

Whyte, Angus. 2015. Where to keep research data: DCC checklist for evaluating data repositories, version 1. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/.

Wieling, Martijn, Josine Rawee, and Gertjan van Noord. 2018. Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics* 44 (4): 641–649. https://doi.org/10.1162/coli_a_00330.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018. https://doi.org/10.1038/sdata.2016.18.

Zenodo. n.d. https://zenodo.org/. Accessed June 10, 2019.

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