9 Copyright and Sharing Linguistic Data

Lauren B. Collister

1 Introduction

A key component underlying many aspects of a data management process is intellectual property rights, specifically copyright. Intellectual property rights are based on the question: who owns the data that are being collected? The follow-up question that scholars need to ask is: how does that ownership impact what a researcher can do with the data when it comes time to publish and share?

Many scholars operate under the assumption that because they are doing the work of collecting and managing the data that means that the data belong to the scholar. This is not always the case, however, both ethically due to cultural principles of ownership of language (see Holton, Leonard, & Pulsifer, chapter 4, this volume) and legally due to the particularities of copyright law. An understanding of copyright and its intersection with the ownership of data can save a headache later in the project; as Newman (2007:29) writes, "the failure of scholars to pay attention to such [copyright] matters has had serious negative consequences." Many scholars have had an experience with an unanticipated copyright question, such as having to prove that they have permission to include an image or figure in a published journal article. These unanticipated questions can be particularly troublesome when they potentially impact an entire data set on which a research project is founded. Many scholars discover the complex intellectual property questions about their data far too late in the process to easily deal with any concerns or complications and find themselves looking for work-arounds or last-minute solutions. These situations often result in the inability of a scholar to share the data that he or she so painstakingly collected. This chapter is intended to help readers get ahead of these questions by providing an overview of intellectual property, specifically copyright, and how these laws apply to linguistic data and how they can enable the sharing of linguistic data.

With the focus on linguistic data, I must mention that the world of intellectual property and scholarship is much bigger than this chapter can cover; where I can, I provide pointers to helpful references and tools to pursue more information. However, this chapter necessarily has some limitations. First, this discussion of copyright and data is situated in a broader context because copyright also applies to other scholarly products, such as journal articles, dissertations, and teaching materials. For a good grounding in copyright issues beyond data, Newman (2007) is essential reading. Second, because linguistic data is so diverse (see Good, chapter 3, this volume), the overview provided by this chapter cannot cover every possibility for all the types of linguistic data that currently exist or that will exist. Finally, this chapter will contain some information about ethics, especially when ethical considerations intersect with copyright, but will not contain an overview of ethics for all of data. Readers are strongly encouraged to review chapter 5 by Holton, Leonard, and Pulsifer in this volume for more information on ethical considerations for data as well as laws, principles, and frameworks that may apply, such the OCAP (Ownership, Control, Access, and Possession) principles from the First Nations Information Governance Centre in Canada (2014) and other principles and guidance resulting from the Indigenous Data Sovereignty Movement.

With the above-mentioned limitations in mind, the intent behind this chapter is to provide foundational knowledge to enable a linguist to ask the right questions about intellectual property with the goal of sharing linguistic data. *Share* is an intentionally broad term that encompasses a wide range of activities from publishing data alongside an article or book to depositing data in a repository to posting a data set on a website; what all of

these activities have in common is making data accessible and findable on the internet. While *sharing* data may include person-to-person data exchanges over e-mail or the direct transferring of files, this chapter will be most relevant to those who want to put their data set online in some way.

Sharing data is essential to the goal of reproducible research to avoid the "file drawer problem" discussed by Gawne and Styles (chapter 2, this volume). Houtkoop et al. (2018) have shown that the primary barriers to sharing of data are cultural issues in academic research—namely that it is not the regular practice of people in the field (yet). When scholars *are* interested in sharing their data, they express concern and confusion about intellectual property (especially when it comes to open data), and the lack of established practice in the field means that they do not have examples to look to for guidance. Chapter 8 of this volume covered how to share data as part of a data management plan, including where one might ultimately archive the work (see Andreassen, chapter 7, this volume). The goal of this chapter is to enable scholars, first, to understand how intellectual property affects their work and, second, to ultimately ensure open access (free of barriers to access, re-use, and distribute) to their linguistic data when ethically appropriate. This work to enable access to data can facilitate easier discovery and citation of linguistic data (see Conzett & De Smedt, chapter 11, this volume), which will lead to metrics and tracking of re-use of one's data set (see Champieux & Coates, chapter 12, this volume) and ultimately an essential addition to a research portfolio (e.g., Alperin et al., chapter 13, this volume).

To begin this section of the data journey, I will start with a definition and explanation of copyright, including how and when it applies to data. With this definition in hand, I will next address exceptions to copyright, followed by intersecting concepts that can impact copyright and data. Having identified whether copyright applies to data and how, along with other considerations for determining the copyright status of data, this chapter closes by addressing intellectual property rights and responsibilities when sharing data.

2 What is copyright?

Copyright law is intended to give authors of original works certain rights to those works, including the right to reproduce, distribute, publicly display and perform, and make adaptations of the work in question. While this definition seems simple enough, what counts as an *author* and what counts as an *original work* have important consequences for scholarly work, especially data. When starting a data collection project, it is pertinent to ask whether the data being collected or used are covered by copyright. In this section, I will describe what kinds of data might be covered by copyright, followed by the time limitations and scope of copyright; this section will help linguists understand when and how copyright might apply to their data.

Is copyright the only intellectual property that linguists need to worry about? Copyright is just one type of intellectual property, and other types of intellectual property in the United States include *patents* (a grant of a property right by the government to an inventor to exclude others from making, using or selling an invention) and *trademarks* (a name, symbol, or phrase used in interstate commerce to identify the source of a product or service) (Barnett, Collister, & McAllister-Erickson 2019). It is unlikely that trademarks will intersect with data, and if a data set is a component of a patent then consultation with a lawyer or legal counsel is recommended; both of these are outside of the scope of this discussion. Another type of intellectual property, *sui generis* rights in the European Union and South Korea, may also apply to some data sets, and these will also be covered in section 2.1.

2.1 What copyright covers

Copyright laws typically cover *original* works created by an *author*. In the Copyright Law of the United States, "Copyright protection subsists, in accordance with this title, in original works of authorship" (US Copyright Office 2016:section 102(a)), and in the United Kingdom, "a work should be regarded as original, and exhibit a degree of labour, skill or judgement" (UK Copyright Service 2017:section 4).

An initial question to answer is who counts as an *author* when it comes to the "original works of authorship" covered by copyright. In identifying legal authorship for copyright, an author is typically a person who "makes creative or editorial decisions about how ideas and facts are expressed" (Carroll 2015:4). This legal definition of authorship is not the same as contributions to a scholarly work, which are addressed by initiatives such as the

Contributor Roles Taxonomy (CRediT) from the Consortia Advancing Standards in Research Administration (CASRAI),[1] in which people have assisted a work of scholarship beyond the writing of the text (e.g., data curation, software development, conducting experiments) may be listed either as authors of a work or in an acknowledgment section (Brand et al. 2015). The legal definition of authorship may also not match, or be in direct conflict with, community and cultural ideas about ownership, especially of language (see Holton, Leonard, & Pulsifer, chapter 4, this volume). The author as the legal entity who owns the rights to the work is the person or entity that makes decisions about the work. Typically with academic works, the copyright holders and authors are those listed on the bylines of journal articles and books; with data sets that have many contributors, copyright should be negotiated among the contributors (contracts can be an essential part of this process, and are discussed in section 2.4). More than one person can hold copyright to a work, and each author has the full rights of copyright to the work and can legally (although perhaps not ethically) exercise those rights independently without the permission of the other copyright holder(s).

If a work must be original to be covered by copyright, then it follows that non-original work is not subject to copyright, which applies quite often to data used in scholarly work. Michael Carroll (2015) described copyright's relationship to scholarship well when he wrote that "copyright law is founded on certain science-friendly policies. Copyright imposes no restrictions on the sharing of the basic building blocks of knowledge—facts and ideas—which are part of the public domain. Researchers routinely rely on this freedom to copy in their daily practice" (3). Measurements of and facts about the world do not fall under the protection of copyright. The US Copyright Office explicitly defines some of these exclusions by stating, "In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work" (2016:section 102(b)). Work that is not protected by copyright or any other intellectual property laws (patents and trademarks) is said to be in the public domain, which means that "the public owns these works, not an individual author or artist. Anyone can use a public domain work without obtaining permission, but no one can ever own it" (Stim 2013). The public domain, in the words of Duke University's Center for the Study of the Public Domain (2011), is "'free' as in 'free speech,' not 'free' as in 'free beer'—because it is unprotected by intellectual property rights, it is free of centralized control as a legal matter, and you can use it without having to get permission." It is important to remember that *public domain* is a legal term with a specific definition—creative work that is not protected by copyright—and does not refer to anything that is freely available to view. Sometimes people use the phrase *public domain* inaccurately to refer to material that is free to view and download online—even though some material may be free to access and view, copyright still applies and re-use, translation, adaptation, or selling of the material would require permission from the copyright holder. It is important not to confuse free to view with legally free to use.

Further complicating matters is the distinction between *data* and a *database*. While the facts and measurements may not be subject to copyright because they are not original, the arrangement or compilation of these facts potentially could be if that arrangement or compilation is sufficiently creative (Sims 2012). Additionally, in the European Union and South Korea, databases created entirely within the borders of these countries that require "substantial investment" to assemble or maintain are protected by a specific set of laws referring to *sui generis* rights. These rights protect against "extraction or reutilization of substantial parts of a protected database as well as frequent extraction of insubstantial parts of a protected database" with exceptions given for non-commercial research (Carroll 2015:5–6). Database rights, including *sui generis* rights and copyright, may impact corpora, lexicons, or other grammars that linguists may use or create as data sets.

It is therefore the case that copyright may not apply to data sets if they are measurements of or facts about the world, but copyright may apply to analyses and representations of those data sets; it may be the case that the researcher may own her written observations about an object, but she may not own the object itself (Borgman 2015:178). This has some simple examples that are often used in the physical or natural sciences: measurements of rainfall, coordinates of locations, recipes, and formulas. In these cases, the researcher would own any text that she wrote about those measurements or the

creative visualizations that she created to display those measurements, but the actual measurements themselves would be in the public domain and therefore usable by anybody without permission needed.

In linguistics, because of the nature of the field, the situation can become complicated quickly. Some linguistic data may be subject to copyright because linguists deal in words, phrases, and sentences that may be in themselves creative expression, not measurements or facts about the world. Take, for example, a situation in which a linguist wants to compare the difference between [a] in two language varieties and uses as data recordings of radio interviews done with speakers of the two varieties. The linguist may excerpt all examples of [a] and analyze the formants and frequencies. In this situation, copyright would not apply to those vowel measurements or the method of doing those measurements, but copyright would apply to the recording that was the source material for the measurements. This is because the recording itself almost certainly contains material that could be classified as an *original* work—unless the recording were very dry indeed, such as a speaker reading nothing but a list of measurements or telephone numbers. The question for the linguist becomes then, who owns the copyright to the recordings, and can permission be gained to use the data? Before asking for permission to use the data, two more pieces of knowledge are needed: understanding whether copyright may have expired (and therefore the work is in the public domain), and whether Fair Use (or Fair Dealing) exceptions to copyright may apply and be sufficient for the project at hand.

2.2 When copyright applies

The *when* question of copyright asks both when copyright comes into effect and when it expires. In the United States, copyright is granted to an author of an original work automatically when that work is "fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device" (US Copyright Office 2016:section 102). In other words, when the work becomes perceivable by another human being, whether or not another human being has actually perceived the work, the author is automatically granted copyright. The author does not need to apply for formal copyright protection or fill out any forms to have copyright over the work. This situation applies to most of the countries of the world; text in the Berne Convention for the Protection of Literary and Artistic Works, an international treaty signed by 176 countries as of the time of this publication,[2] states that material is protected when it is "fixed in some material form" and requires that authors must not have to comply with any formalities to be granted the rights of copyright (World Intellectual Property Organization 1979:article 2(2)). The "material" or "tangible" form referred to in these laws is intended to provide some proof of the existence of the material that is sufficiently permanent to allow it to be perceived by another person after its creation. Words spoken out loud dissipate and, although they may be heard by another person immediately, merely speaking words is not sufficiently permanent to qualify as "fixed in some material form"; however, these words may be fixed through audio or video recording or by writing the words down on paper or computer. Finally, Tribal lands have their own intellectual property laws that may differ from the countries they border, and for linguists working with Indigenous languages (whether doing new data collection or working with legacy data from an archive), it is important to consult the Tribal laws before making any decisions about copyright and its applicability (see Reed, forthcoming, for a thorough discussion).

If copyright applies when the work is fixed in a material form, when does it expire? This is a much more complicated question and varies not just by country, but by when the work was created and the laws that were in effect at that time. The Berne Convention grants protection for the life of the author plus fifty years, but allows each signatory country to set longer term limits (World Intellectual Property Organization 1979:article 7). In the United States, for example, for new works or those which have been created since 1978, copyright is in effect for the life of the author plus seventy years (US Copyright Office 2016:section 302). In other situations (such as in the United States pre-1978), whether a work was published or unpublished impacts the duration of copyright, and in some cases, the material had to be accompanied by a set copyright statement.

Thinking back to the example of the linguist analyzing [a] in recordings of radio broadcasts, she would be dealing with copyright because the broadcasts were recorded and therefore fixed in a material form. Because copyright status differs according to a number of considerations

such as the year of publication, the registration status, and the law at that time, there are a number of tools that have been developed to identify whether an item is covered by copyright or not. Wikimedia Commons has a helpful guide to copyright rules by country that includes length of copyright.[3] Peter Hirtle of Cornell University has developed an extensive chart showing dates and parameters for copyright status in the United States.[4] For Canadian copyright status, the University of Alberta's Copyright Office has an excellent flowchart to determine whether an item is in the public domain.[5] In the European Union, public domain calculators are available for several countries via the Out of Copyright website.[6] Depending on when and where the radio broadcasts occurred for her project, the linguist for this example data set should check the specifics for the recordings in question and whether they might be in the public domain: what year were they made? Where were they done? These tools, or perhaps a local copyright librarian, could help her find out the status of the recordings and whether they are in the public domain or not.

2.3 Exceptions to copyright: Fair Use and Fair Dealing

If copyright still applies to the data set (that is, if the data are not in the public domain), then it still may be used for research purposes without obtaining explicit permission. Copyright law sometimes contains features that allow people to use copyrighted works under certain conditions. Fair Use in the United States is one example of these features, which allows people to use portions of copyrighted material for purposes such as commentary, criticism, scholarship, or parody, as long as the use does not "interfere with the copyright holder's legitimate economic interests" (Newman 2007:35). To make a Fair Use assessment, there are four considerations: the purpose and character of the use, the nature of the work being used, the amount of the original work being used, and the effect of the use on the potential market of the original. There is another consideration that often comes into play, which is whether the use transformed the copyrighted material "by using it for a different purpose than that of the original, rather than just repeating the work for the same intent and value as the original" (International Communication Association 2010:6). To help scholars make these assessments, a number of checklists exist to help users make a Fair Use evaluation; two examples of helpful checklists are the Thinking Through Fair Use tool from the University of Minnesota Libraries and the Fair Use Checklist from the Columbia University Libraries.[7] Additionally, when dealing with Indigenous cultural materials, linguists are recommended to consult the discussions of Fair Use as cultural appropriation by Trevor Reed (2020, forthcoming).

In the United Kingdom, Canada, Australia, and elsewhere, Fair Dealing is a user's right to use copyrighted works without permission or payment of royalties. Fair Dealing and Fair Use are not the same in all countries; for example, in Canada, "fair dealing for the purpose of research, private study, education, parody or satire does not infringe copyright" and specific requirements for mentioning the source are defined for criticism, review, and news reporting. Canada also has exceptions to copyright for non-commercial user-generated content, reproductions for private purposes, and recording broadcasted programs for later use (Canada, Minister of Justice 1985:C–42, section 29). However, Fair Dealing uses are subject to Moral Rights, which allow for the preservation the integrity of a work or performer (Canada, Minister of Justice 1985:C–42, section 28). For Fair Dealing assessments, the University of Ottawa provides a Fair Dealing decision tree.[8]

Fair Use and Fair Dealing intersect with linguistic data when a researcher wants to use copyrighted works as a source of data. One increasingly common example of this situation is text and data mining of copyrighted material such as books. Fair Use or Fair Dealing may apply to these research works and allow for them, but these provisions do not necessarily allow for the re-sharing of the source data when publishing the work. It might be Fair Use to compile a corpus all of the books by Stephen King to perform text analysis on them, but sharing that corpus openly online would interfere with the economic interests of the copyright holder. Under the "transformative use" component of Fair Use, a data set that contains word frequency counts derived from the corpus could be shared as long as it was not directly and extensively quoting the books in a way that could be a substitute for reading or purchasing the books.[9] For our example linguist with her radio interviews, she might be able to use the recordings as data even though they are under copyright using a Fair Use argument; she may then share vowel measurement data and potentially audio file snippets of individual vowels depending on how extensive the quotation of the original source is. To direct people

to the original source for the data, she could choose to share a link to the source recordings if they are available online or a citation where others could find the source data without her re-sharing or re-publishing the original recordings herself. When choosing data sources for projects, it is important to consider both whether copyright and Fair Use/Fair Dealing may allow the source to be used for research as well as whether the final data set will be shareable when the research project is complete.

2.4 Intersections with copyright

In addition to copyright, when working with data specifically, two other important constructs exist that may intersect with copyright questions and need to be considered: contracts and ethics.

Contracts are agreements made between two (or more) parties that address the rights and responsibilities of each party. While copyright is the default status typically assigned automatically when the work is fixed in a tangible form, a contract is an active agreement that can alter or override copyright. Scholars most often encounter contracts when publishing papers or books—these contracts are between a publisher and the copyright holder (the scholar/author) and lay out the rights that the publisher has over the copyrighted material (the article). Sometimes, these contracts are called *copyright transfer agreements*, and in them a scholar signs over all rights under copyright to the publisher to publish the work; in return, the scholar may get royalties, limited re-use rights, or the right to make derivative works. Other times, the contract is a license that states that the scholar keeps copyright and assigns to the publisher a license that permits the publisher to do certain things on behalf of the author, such as the right to be the outlet of first publication and the right to make and distribute copies. Our example linguist using radio interviews for data may encounter these contracts because the radio station or media entity will most likely be the owner of the content, not the speakers in the actual interview, and a contract between the speakers and the media outlet may determine both who owns the rights and what can be done with the content.

For an author, these copyright transfer agreements are important to read, especially when research data are a part of a publication; be careful when transferring the rights to a research publication to a third party, especially when the word *exclusive* is used. When a publisher is the exclusive holder of all rights associated with copyright, a scholar such as our example linguist may find herself in the unenviable position of having to request permission from a publishing company to re-use her own data set if she published it with a journal or other publisher and signed a copyright transfer agreement. In an even worse scenario that came across my desk in 2017, a graduate student was asked by a publisher to pay a licensing fee to use material from their own published article in their dissertation. Publishing contracts are an important agreement that can have long-lasting effects, including on research data, and therefore it is extremely important to pay attention to the agreements and ask for clarity from the publisher when there are any questions. It is also good practice to enlist a librarian to help with this conversation, as publisher ownership of research is of great interest and importance to the work of librarianship.

Another common scenario that scholars may encounter is contracts spelling out the requirements of grant funding. In many countries, government grants may come with a requirement to publish all work done from grant funding in an open access publishing outlet, and many private foundations are enacting similar policies. For example, in the United States, any article that results from a grant that comes from the National Institutes of Health (NIH) must be deposited in PubMed Central, a repository created to facilitate the open sharing of the outcomes of federally funded research. Many EU funders have signed on to Plan S, an agreement to require all publications that result from research funded by public grants to be published in compliant Open Access journals or platforms.[10] Policies on data sharing are expanding from funders as well; the National Science Foundation (NSF) in the United States has a policy that states that grantees are "expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants."[11] When applying for a grant from any organization, governmental or otherwise, look for their requirements for data sharing and ownership of data to fully understand the scope and implications for data created or gathered during the project.

Contracts also appear in the forms of Terms of Service or other license agreements, such as those for a database, website, or media service. This situation may be particularly relevant to scholars working with language corpora,

which are often subject to licensing and Terms of Service, and the publishers of the corpora may have specific rules about how to excerpt and cite material. While many people do not read these contracts before clicking "I agree" (and analysis of the language of these contracts has shown that they are "far beyond what a functionally literate adult could be expected to understand" [Luger, Moran, & Rodden 2013:2687] so even those who do read them can hardly be expected to comprehend the terms), it is important when using services such as these for research to read the license agreement. These contracts may stipulate what a user can and cannot do with the material provided by the service, and this can mean that a scholar cannot do some things that they might usually expect to be able to in a research environment (such as sharing their data, or even publishing excerpts of the data in a journal article). These Terms of Service can also change over time, especially when using social media or other online corpora that are subject to rules that change often and supersede previous agreements (Wheeler 2018). When publishing or sharing research data online, Terms of Service may also violate ethical considerations regarding ownership of data or sharing of sensitive or private data (see Holton, Leonard, & Pulsifer, chapter 4, this volume, for some examples).

Contracts also come into play between scholars and both their research subjects and their research assistants. When the data being generated by a subject or consultant of a research project is sufficiently creative—for example when creating a narrative, telling a story, or performing a song or poem—copyright can apply to the material in the scholar's data set. To be able to quote or excerpt the data, the scholar needs a good contract in which the interviewees or research subjects agree to allow the researcher to use their material, and in which the scholar discloses her plan for sharing or disseminating the data. This is typically part of the informed consent process and required for most work with human subjects; it is very important to include the sharing of research data in this consent process and to explain clearly to participants how their data will be re-used and shared, and to make certain that they agree to allow their material to be distributed. A researcher may also employ a translator to work with data collected, make translations or glosses of texts, or to be an intermediary between the researcher and community members. Students may be employed to help annotate, clean up data, or develop visualizations based on data sets. Because some of this work may qualify those people as authors of material and therefore owners of copyright, it is important to have a good contract setting out who owns the work being produced during the project. Typically these contracts fall under the concept of Work for Hire, in which the employee agrees that all material created in the course of her employment is the property of the research project or the researcher. The scholar typically offers compensation (monetary, course credit, or stipend) in return for the employee's work, and additionally may credit their colleagues for their contribution. Contracts are a helpful tool for understanding and communicating these situations and can spell out all of these responsibilities and rights in a clear way so that all parties know who owns what, as well as what will be done with the material being created. It is essential to work directly with the community and within established framework and guidance to create these contracts and to update them when necessary (Holton, Leonard, & Pulsifer, chapter 4, this volume); this is vital work for a scholar to infuse ethical scholarship into the legal aspects of academic work.

Another important consideration is that even though something might be technically legal under copyright law, permissible with a Fair Use argument, or allowed under a contract agreement, this does not mean that the act is ethical. *Ethics* refers to "norms for conduct that distinguish between acceptable and unacceptable behavior" and there are many ethical norms in research and data collection and sharing that are important to consider in conjunction with legal and contractual rights (Resnik 2015). The sharing of personal data is a major component in ethical considerations. In 2016, a group of researchers released to the public a data set of the personal profiles of around 70,000 users obtained from the online dating website OkCupid. The researchers argued that the data were publicly available, although their methodology section does not discuss privacy settings, and that all they were doing was presenting the data "in a more useful form" (Zimmer 2016). Whether or not these profiles were legal and accessible, ethical guidelines about the release of personal data should have been considered in this case. The General Data Protection Regulation in the European Union is one example of law governing the sharing of personal data and it impacts how researchers should process and anonymize personal data (Klavan, Tavast, & Kelli 2018).

Ethics also intersects with culturally sensitive or protected material, which is codified in documents such as the UN Declaration on the Rights of Indigenous People that states that Indigenous people "have the right to maintain, control, protect and develop their intellectual property over such cultural heritage, traditional knowledge, and traditional cultural expressions" (2008:article 31). Principles and practices of cooperative fieldwork (e.g., Dwyer 2006) can help linguists collect data in ways that address and respect ethical concerns. For guidance on how to approach scholarship and data in a people-centered, ethical way, see Holton, Leonard, and Pulsifer (chapter 4, this volume).

3 Copyright and sharing data

Gawne and Styles (chapter 2, this volume) set out an argument for making linguistic data available to facilitate reproducibility and verifiability of research in our field. Once a research project is complete—and sometimes even before it is complete—scholars are able to make their data sets available for others to use (subject to the above-mentioned ownership, ethical, and privacy considerations). This shared data, when done without barriers to access or re-use, is called open data (Dietrich et al. 2009). Copyright status, as well as ethics and contract situations, can impact how data sets can be shared and in what form. In section 2, I set out ways to identify whether copyright applies to data sets during the collection phase. In this section, I will cover the impact of copyright on the act of sharing data.

3.1 Data in the public domain

If copyright does not apply, the data can be legally considered to be in the public domain. This means that, taking into account ethical considerations such as anonymization of personally identifying information and cultural considerations of ownership and access to language, researchers are free to share their data sets in the most open way possible. To be completely open, a researcher may explicitly designate their data set as in the public domain, including the arrangement and description of the data (e.g., a readme file). Typical scholarly practice is to still cite the source of data, and a data set should come with a suggested citation, whether it is provided by a repository or archive or the suggested citation is created by the researcher. Conzett and De Smedt (chapter 11, this volume) provide an overview of data citation guidance that will be helpful in doing this work.

Even if it is unclear whether data are in the public domain or not, scholars who may have ownership of the data can remove all doubt by dedicating the data set to the public domain. This can be done with a statement such as "this dataset is dedicated to the public domain" (Stim 2013) or with a Creative Commons zero (CC0) license, which is a legal tool for waiving copyright (Creative Commons, n.d.). These are illocutionary acts, specifically a declaration (Searle 1975:366), and by making the statement on the document, the owner changes the status of those documents and makes them available for others to use and re-use freely.

3.2 Data owned by the scholar

If copyright does apply and the scholar or data collector is the author who owns copyright (whether through being the original author creating the material or having rights assigned to the scholar through contracts), then the author can choose what to do with the data set. Because copyright is automatic and defaults to all rights reserved, without any act by the author, the data are not free to be re-used by others unless the author acts to make it so. The author can share the data set without any additional copyright information, but if any other scholar wants to re-use the data, that scholar will have to ask for permission from the data set's author.

To facilitate open data, the author can apply an open license that allows the author to retain their copyright but allows others to re-use the data set with certain conditions. A license is a contract between the owner of the data and the users of data that allows use of the data in certain ways. If a data set is subject to copyright in any way, a license can help others know how to re-use it and how to attribute it properly, and they save the author the time and hassle of granting permission to individual requests.

Creative Commons (CC) licenses are an example of open licenses that can be used by the author of content. These licenses are legal documents that a copyright holder can apply to their work, with the basic stipulation being that if someone re-uses the data set, she is required to attribute the source with a citation of the original data set (this is the BY clause in a CC-BY license). Other parameters of CC licenses include a non-commercial use restriction (NC), a prohibition on changing the content

(no derivatives, or ND), and a requirement that all works based on the original must be also openly licensed (share alike, or SA).

Another license that can appear on data sets is the GNU General Public License (GNU GPL, sometimes including the version number and appearing as GPLv3) (Smith 2014). The GNU GPL is a free software license that allows the creator to retain copyright but has very permissive re-use rights, with the only restriction being that any derivative or improved version of the work must also be released under a free software license.

Regardless of which particular license is chosen for a data set, the Research Data Alliance recommends that "access to and re-use of research data should be open and unrestricted as a default rule, or otherwise be granted to users with the fewest limitations possible" (RDA-CODATA Legal Interoperability Interest Group 2016:3). Using the most open possible license encourages the open and easy re-use of data for future projects. A restrictive license imposes conditions on re-use that may make a data set incompatible with another data set, thus limiting a future researcher's ability to combine data sets for a single project.

Many tools exist to help scholars choose a license for their data. If the data come with software or other code, the Choose an Open Source License tool will be helpful.[12] The Public License Selector tool guides scholars through a series of questions that will help determine which license to use.[13] Creative Commons also operates their own license selector specifically for CC licenses.[14]

3.3 Data owned by another party

When using material owned by another party as data for a research project, it is sometimes possible to share the data as part of a research project, but this may require an extra step on the part of the researcher. The question to ask in this case is whether the material contains a permissive license for sharing, and if not, what is the process for obtaining permission to share?

The material may be openly licensed using a CC or other license as described in section 3.2. In this case, sharing, including re-distributing the material online and potentially publishing it in a repository or other outlet, is allowed under certain parameters; as long as the researcher follows those parameters, the data set can be shared. Some Terms of Use or other contracts may set out conditions for sharing of data sets. In these cases, it is important to follow the requirements closely in accordance with the contract's terms. For scholars who are working with multiple data sets with different licenses, those licenses may contradict each other and reduce interoperability; in this situation, consult the RDA-CODATA *Legal Interoperability of Research Data Principles and Implementation Guidelines* (2016).

When data are not licensed in any way or they contain a license that prohibits re-use or sharing, permission must be granted for the data to be shared. This may happen to the example linguist with the radio interviews that may be owned by a media or broadcast entity. If the linguist wants to share those original interviews, she can write to the data owner (in this case, the media company or radio station) and inform the owner about the research project and her wish to create an open data set including the materials owned by the company. The researcher should disclose where and how the data set will be shared (e.g., in a data repository) and what license she wishes to apply to the data. In some cases, the owner of the material will consent to this open sharing as long as attribution is retained (e.g., the linguist credits the media company with a full citation of the interview recording, air date, and program name). Other times, the owner of the data would not allow for an open license to be placed on their material, but may consent to having it included as part of the materials for the study but retaining their copyright—if this is the case for our example linguist's recordings, it will be her responsibility to label her data set clearly and appropriately, stating that the recordings belong to the media company, and (if she wants to be helpful) including contact information for others to use to obtain permission. This latter case would mean that if another scholar wished to re-use these recordings found in the data set created by our example linguist, that scholar would have to obtain her own permission to use the recordings. In both of these cases, the owner of the content may require a payment for the re-use of their content.

If sharing permission is not granted by the owner, this does not preclude a Fair Use of the data for the research project. While Fair Use or Fair Dealing may allow for the use of copyrighted material owned by another person or entity than the scholar, these do not allow for the sharing of the data set containing that copyrighted material, which may negatively impact future research built on the project as well as impede reproducibility. For the

example linguist, she has a potential solution where she can share the transformed data set of her vowel measurements openly and include citation information and a link to the original broadcast recordings for other scholars to find. This allows her to share her analysis based on the facts of the data set and direct people to the source location that exists elsewhere; it requires an extra step for those who are looking to re-use the data or reproduce the study, but still makes clear the data's provenance. For other linguists who have compiled a corpus or other data set that reproduces the original source material wholly, the sharing options may be much more limited, and therefore re-use or replicability may be difficult or impossible. When possible, especially for research use, it is recommended to get permission to share the data early in the project to avoid a situation where sharing of research data becomes impossible.

4 Conclusion

This chapter provided an overview of copyright and intellectual property considerations for data, and how those considerations can impact the open sharing of data sets. Attention to intellectual property questions is important from the outset of a project involving data to facilitate the sharing of data associated with the research project and to avoid difficult situations at the end of a process. With the ultimate goal being sharing data as openly as possible, asking intellectual property questions can facilitate sharing and make the process much easier for the scholar.

Because of the difference in copyright law in different areas of the world and the variety of linguistic data, not every situation can be covered in this short introduction, but this chapter should be a good start. For more help with copyright, scholars can consult with a librarian. Many academic libraries have a staff member dedicated to copyright or intellectual property. This person will likely have a title like Copyright Librarian; additionally, Scholarly Communication Librarians can provide guidance when it comes to intellectual property. In large cities, public libraries may also have a copyright expert who can assist with questions about these issues. Because they are librarians and (usually) not lawyers, while they can help with resources and information, they cannot offer legal advice. When drawing up contracts for participants in a research study, an Institutional Review Board (IRB) should be able to offer guidance as part of an informed consent process. It is important to inform IRB staff of intent to openly share data so that they can advise on any ethical or legal issues in the data collection period that may need attention specifically for the end goal of sharing data. The General Counsel at a college or university should be the resource for Work for Hire contracts when working with translators, data collectors, or research assistants.

The most important message about copyright and the sharing of data is that these should not be left for the end of a project. Attending to copyright before data collection will put a researcher on solid footing when proceeding to writing and analysis, and sharing of a data set can not only benefit a scholar with more attention to her work, but also benefits the field by making linguistic work more reproducible.

Notes

1. https://www.casrai.org/credit.html.

2. https://www.wipo.int/treaties/en/ShowResults.jsp?&treaty_id=15.

3. https://commons.wikimedia.org/wiki/Commons:Copyright_rules_by_territory.

4. https://copyright.cornell.edu/publicdomain.

5. https://www.ualberta.ca/copyright/resources/tools.

6. http://outofcopyright.eu/.

7. https://www.lib.umn.edu/copyright/fairthoughts; https://copyright.columbia.edu/basics/fair-use/fair-use-checklist.html.

8. https://copyright.uottawa.ca/what-is-copyright/exceptions-copyright/fair-dealing-decision-tree.

9. For an interesting example of this, see the 2008 case *Warner Bros. Entertainment, Inc. v. RDR Books*, 575 Federal Supplement 2d 513 (S.D.N.Y. 2008) where an "unauthorized" lexicon of terms from the Harry Potter book series was found to not be Fair Use because, although the court found that the lexicon itself was transformative, the text of the lexicon quoted extensively from the novels and movies, which outweighed the transformative aspect of the lexicon. https://www.copyright.gov/fair-use/summaries/warnerbros-rdrbooks-sdny2008.pdf.

10. https://www.coalition-s.org/.

11. https://www.nsf.gov/bfa/dias/policy/dmp.jsp.

12. https://choosealicense.com/.

13. https://ufal.github.io/public-license-selector/.

14. https://creativecommons.org/share-your-work/.

References

Barnett, John, Lauren Collister, and Jonah McAllister-Erickson. 2019. Copyright and intellectual property toolkit. *LibGuides, University of Pittsburgh*. https://pitt.libguides.com/copyright.

Borgman, Christine L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge: MIT Press.

Brand, Amy, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. 2015. Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing* 28 (2): 151–155. https://doi.org/10.1087/20150211.

Canada, Minister of Justice. 1985. *Copyright Act. Consolidated Federal Laws of Canada*. C. C–42. https://laws-lois.justice.gc.ca/eng/acts/C-42/page-9.html#h-26.

Carroll, Michael W. 2015. Sharing research data and intellectual property law: A primer. *PLOS Biology* 13 (8): e1002235. https://doi.org/10.1371/journal.pbio.1002235.

Center for the Study of the Public Domain. 2011. Public domain frequently asked questions. https://law.duke.edu/cspd/publicdomainday/2011/pddfaq.

Creative Commons. n.d. CC0. *Creative Commons*. https://creativecommons.org/share-your-work/public-domain/cc0. Accessed March 27, 2019.

Dietrich, Daniel, Jonathan Gray, Tim McNamara, Antti Poikola, Rufus Pollock, Julian Tait, and Ton Zijlstra. 2009. What is open data? In *The Open Data Handbook*. London: Open Knowledge Foundation. http://opendatahandbook.org/guide/en/what-is-open-data.

Dwyer, Arienne M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In *Fundamentals of Language Documentation: A Handbook*, ed. Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel, 31–66. Berlin: Mouton de Gruyter. https://kuscholarworks.ku.edu/handle/1808/7058.

First Nations Information Governance Centre. 2014. *Ownership, Control, Access and Possession (OCAP™): The Path to First Nations Information Governance*. Ottawa: First Nations Information Governance Centre.

Houtkoop, Bobby Lee, Chris Chambers, Malcolm Macleod, Dorothy V. M. Bishop, Thomas E. Nichols, and Eric-Jan Wagenmakers. 2018. Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science* 1 (1): 70–85. https://doi.org/10.1177/2515245917751886.

International Communication Association. 2010. Code of best practices in Fair Use for scholarly research in communication. *The International Communication Association*. http://cmsimpact.org/code/code-best-practices-fair-use-scholarly-research-communication.

Klavan, Jane, Arvi Tavast, and Aleksei Kelli. 2018. The legal aspects of using data from linguistic experiments for creating language resources. In *Human Language Technologies—The Baltic Perspective*, ed. Kadri Muischnek and Kaili Müürisep, 71–78. Frontiers in Artificial Intelligence and Applications 307. Amsterdam, IOS Press. https://doi.org/10.3233/978-1-61499-912-6-71.

Luger, Ewa, Stuart Moran, and Tom Rodden. 2013. Consent for all: Revealing the hidden complexity of terms and conditions. In *CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2687–2696. New York: ACM. https://doi.org/10.1145/2470654.2481371.

Newman, Paul. 2007. Copyright essentials for linguists. *Language Documentation* 1 (1): 28–43. http://hdl.handle.net/10125/1724.

RDA-CODATA Legal Interoperability Interest Group. 2016. *Legal Interoperability of Research Data: Principles and Implementation Guidelines*. *Research Data Alliance*. https://doi.org/10.5281/zenodo.162241.

Reed, Trevor. 2020. Fair Use as cultural appropriation: Why the "forgotten factor" matters. American Library Association Copyright, Legislation, Education, and Advocacy Network. Video, 1:01:40. Copytalk Webinar Archive. http://www.ala.org/advocacy/copyright/copytalk.

Reed, Trevor. Forthcoming. Creative sovereignties: Should copyright apply on Tribal lands? *Journal for the Copyright Society USA*, Available at SSRN: https://ssrn.com/abstract=3736137.

Reed, Trevor. Forthcoming. Fair use as cultural appropriation. *California Law Review*, Vol. 109, 2021, Available at SSRN: https://ssrn.com/abstract=3456164.

Resnik, David B. 2015. What is ethics in research and why is it important? *National Institute of Environmental Health Sciences* (blog). December 1, 2015. https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm.

Searle, John R. 1975. A taxonomy of illocutionary acts. *Language, Mind, and Knowledge*. *Minnesota Studies in the Philosophy of Science* 7:344–369. http://conservancy.umn.edu/handle/11299/185220.

Sims, Nancy. 2012. Friday fun: Facts, expression, and illustrations. *Copyright Librarian* (blog). August 3, 2012. http://simsjd.com/copyrightlibn/2012/08/03/facts-and-expression.

Smith, Brett. 2014. A quick guide to GPLv3. *Free Software Foundation*. https://www.gnu.org/licenses/quick-guide-gplv3.html.

Stim, Rich. 2013. Welcome to the public domain. *Stanford Copyright and Fair Use Center* (website). April 3, 2013. https://fairuse.stanford.edu/overview/public-domain/welcome.

UK Copyright Service. 2017. P-01: UK copyright law fact sheet. *The UK Copyright Service*. September 27, 2017. https://www.copyrightservice.co.uk/copyright/p01_uk_copyright_law.

United Nations. 2008. United Nations Declaration on the Rights of Indigenous Peoples. *United Nations*. https://www.un.org/development/desa/indigenouspeoples/declaration-on-the-rights-of-indigenous-peoples.html.

US Copyright Office. 2016. Copyright law of the United States and related laws contained in Title 17 of the United States code. Circular 92. *United States Copyright Office*. https://www.copyright.gov/title17.

Wheeler, Jonathan. 2018. Mining the first 100 days: Human and data ethics in Twitter research. *Journal of Librarianship and Scholarly Communication* 6 (2): eP2235. https://doi.org/10.7710/2162-3309.2235.

World Intellectual Property Organization. 1979. *Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979) (Authentic text)*. *WIPO Lex* (database). https://wipolex.wipo.int/en/text/283698.

Zimmer, Michael. 2016. OkCupid study reveals the perils of big-data science. *Wired*. May 14, 2016. https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science.

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

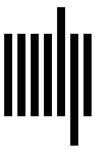
The Open Handbook of Linguistic Data Management

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 open access edition of this book was made possible by generous funding and support from the authors



The MIT Press

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