4 Indigenous Peoples, Ethics, and Linguistic Data

Gary Holton, Wesley Y. Leonard, and Peter L. Pulsifer

1 Introduction: Linguistic data and Indigenous peoples

The world is dominated by just a few large languages that mediate mass communication, social media, education, politics, and many other domains. A study by Kornai (2013) found just sixteen of the world's nearly seven thousand languages to be "digitally thriving," with a firmly established online presence and the tools necessary to live and interact in an increasingly digitally connected world. These sixteen languages are spoken natively by some 2.8 billion people, or nearly 40% of the world's population. These are the languages of Big Data, machine translation, automated speech recognition—the ones that technology companies care most about. For these languages, ethical protocols are largely driven by commercial interests and entail regional legal structures pertaining to data governance.[1] But the vast majority of the world's linguistic diversity is found elsewhere: namely, within the thousands of minority languages, many of which belong to small, often politically and economically marginalized Indigenous groups.[2] Data from these small and often critically endangered languages are key for understanding linguistic diversity—a major focus of linguistic science—but also for maintaining that diversity through language maintenance and reclamation efforts. Linguistic research on Indigenous minority languages takes place against a backdrop of increasing threats to Indigenous language vitality and pressures to shift away from Indigenous languages toward languages of wider communication—often colonial languages (e.g., English, Spanish, Mandarin). We emphasize that language endangerment, along with the response by various stakeholders such as linguists, archivists, and especially the communities these languages come from, is central to the discussion of Indigenous peoples, ethics, and linguistic data. While the causes of language endangerment are many and complex, social and cultural dislocation due to unequal power relations between minority communities and majority populations have played major roles in facilitating language shift (Grenoble 2011). Traditional models of linguistic research often mirror these unequal power relationships (Leonard 2018), with the result that linguists researching Indigenous languages may be seen as agents of social and cultural dislocation as well. Moreover, Indigenous communities may view linguistic research as out of step with the impending threat of language loss. In particular, communities experiencing rapid language shift and consequent language endangerment may take a more holistic view of language research as being embedded within a process of language reclamation (cf. Leonard 2017) and psychological healing (cf. Meek 2010; Jacob 2013) against a backdrop of numerous ethical violations that underlie language shift. Hence, any discussion of ethics in linguistic data requires a discussion of Indigenous data and must adhere to protocols for working with Indigenous data, as well as to the broader sociopolitical contexts in which language work takes place. However, where formal, legal frameworks do exist governing Indigenous and minority language data, these frameworks tend to be modeled on those developed for large languages rather than the cultural values or political concerns of Indigenous populations. We thus focus in this chapter on ethical issues in relation to Indigenous languages and the communities they come from, for it is in this context that the intersection of people, ethics, and data has been least formalized, despite its significant implications.

1.1 Who defines "linguistic data"?

Some of the limitations in theorizing this intersection reflect that in most scholarly literature the notion of

"linguistic data" is not explicitly defined. A working definition might be "data used in the study of language" (Good, chapter 3, this volume), but even this seemingly broad definition assumes a particularly narrow and decontextualized view of the relationship between people and data. Often, what counts as data—and by extension what counts as an ethical response to data collection and management—is in the eye of the beholder, thus opening the door to several possible ethical breaches. For example, a particular string of speech may be viewed as data by a researcher but as a sacred incantation by language users. A more general issue is the tendency for language researchers to equate language with data, and by extension to view language as a mere data point. This reductionist view groups everything produced through research as "data" and thus serves to dehumanize and decontextualize language. Especially through any belief system in which language is defined in relation to its users, the assumption that linguistic data could exist in isolation becomes odd, and such an approach is especially problematic for any discussion of ethics because ethics emerge from people and particular contexts. Hence, any useful definition of linguistic data must avoid divorcing data from their sources, with the particular understanding of what constitutes "data" in a given context clarified.

Although the concept of linguistic data itself may be difficult to define, it is nevertheless useful to distinguish among different types of data. Himmelmann (2012) distinguishes among raw data, primary data, and structural data, based on a cline of decreasing language user involvement. *Raw data* consist of original, unannotated recordings and (non-standardized) writings. *Primary data* consist of the annotations, especially transcriptions and translations, applied to the raw data. *Structural data* consist of structural and typological inferences—the "facts" of language. Inherent in this typology is the notion that at least some types of linguistic data are "manufactured" or "produced" rather than collected. Through this lens, primary and structural data may be construed by some as research products and thus creations of the researcher rather than the language community.

While this typology has some utility within the field of linguistics with respect to theorizing language documentation and language archiving, among other areas, it is insufficient for understanding ethics in relation to linguistic data. In particular, the distinction between raw data (produced by speakers and signers) and primary/structural data (produced by researchers) reflects a Western epistemology of data that potentially disenfranchises language users by removing their agency, while concurrently absolving researchers from acknowledging that they always play roles in representing languages because even basic annotations emerge through particular cultural lenses and conventions. At the extreme, this leads to objectification that obscures the fact that language as a social practice is "embedded in a broader cultural matrix, and it depends critically on that matrix for the activity to be meaningful" (Whaley 2011:344). In contrast, Indigenous approaches to linguistic data tend to reflect a "holistic understanding of language as contextualized language" (Fitzgerald 2017:e291). A linguistic message may be encoded as a string of phonemes built into morphemes and clauses, but this string itself also has meaning, expressing information that may have unique cultural significance attached to the people involved in creating it (which may go beyond actual speakers and signers), the place where it was created, and other factors. From this broader perspective, there is not one set of ethical principles for raw data and another for primary/structural data. Given that all data types ultimately derive from speakers and signers in language communities, all data types must engage equally with ethics.

1.2 Who "owns" linguistic data?

Another complexity to the intersection of people, ethics, and data is the notion of language ownership—by whom and to what extent—and the related notion of whether it is ownership, as opposed to other types of relationships such as connection, kinship, or stewardship, that should guide policies and practices surrounding linguistic data. We adopt *ownership* as a working term, recognizing that this word is used in many existing discussions and policies involving language ethics in Indigenous communities (cf. Guerrettaz 2015). Furthermore, the grammars of most languages permit their users to assert ownership over languages and even individual speech forms. Thus, one can speak of "my language" or "my words" using a possessive form.[3] Nevertheless, we emphasize that understanding particular ethical contexts entails engagement with local understandings of the relationships between languages and communities. Central to this exercise is elucidating local meanings of *language* itself, as the type of relationship and associated

nuances of ownership often emerge from this definition. Notably, it is common in Indigenous definitions to link language and peoplehood (Leonard 2017), and some such as the following center the relationships between people, ethics, and language highlighted in this chapter:

> Language is
> our unique relationship to the Creator,
> our attitudes, beliefs, values, and
> fundamental notions of what is truth.
> Our languages are the cornerstone of
> who we are as a People.
> Without our languages,
> our cultures cannot survive.
>
> (quoted in Shaw 2001:39, from *Principles for Revitalization of First Nations Languages, Towards Linguistic Justice for First Nations,* Assembly of First Nations, Education Secretariat, 1990)

Emerging from examples such as this one, but also common in non-Indigenous communities, is recognition that languages are social constructs, codes shared by communities of language users. In this sense, linguistic data are very different from many other forms of data because the knowledge exists at the community level even though discrete productions of language occur by individuals. Linguistic data are also not completely public (as with, e.g., meteorological measurements), but nor are they completely private (as with, e.g., medical or genetic records). Moreover, given the social-intersectional nature of language as a communicative medium, privacy concerns are not always addressed at the time of data collection. Many of the questions of ethics and linguistic data center on issues of ownership and consequent rights of access. In considering these questions, it is important to bear in mind the special and unique place of linguistic data as simultaneously public and private, and how existing legal structures may fail to adequately recognize ownership of linguistic data. For example, within the legal structures of countries such as the United States, creative forms of language are often given legal protections (copyright), while everyday utilitarian language is considered to be in the public domain (Collister, chapter 9, this volume). Indigenous communities may, however, feel that all forms of language—whether deemed "creative" or not—should be legally protected and formally placed under community ownership.

The principle that linguistic data are imbued with ownership, which comes with rights and responsibilities, is embedded within current best practice standards in linguistic research, particularly those that provide guidelines for data citation. The *Austin Principles of Data Citation in Linguistics* note that "citations should facilitate readers retrieving information about who contributed to the data, and how they contributed" (Berez-Kroeker, Andreassen, et al. 2018). Similarly, Bird and Simons (2003:571–572) assert citation as one of seven key values that underlie language documentation efforts: "We value the ability of users of a resource to give credit to its creators, as well as to learn the provenance of the sources on which it is based." However, data citation standards remain in their infancy within linguistics (cf. Berez-Kroeker, Gawne, et al. 2018), and even where such standards have been adopted, there is little consistency as to who should be credited (though see Conzett & De Smedt, chapter 11, this volume, on emerging standards in this area). Is the creator or contributor the person who produced the language artifact, the linguist who recorded it, or both? Omitting these details from source citations has the effect of divorcing linguistic data from their sources and, by extension, from the larger sociopolitical contexts in which ethical concerns must be addressed.

One concept that is useful for the purposes of exploring the complexity of ethical approaches to linguistic data is the notion of "legacy" data, that is, those data "which were created when concerns surrounding intellectual property were less sensitive than they are today" (O'Meara & Good 2010:162), and for which usage restrictions often need to be newly considered or reevaluated. Many communities are confronting situations in which their ancestral languages have not been actively used for years or are remembered by just a few Elders.

Often this shift is a result of a history of colonization that has resulted in a transfer of language knowledge from communities to data repositories housed in non-Indigenous archives, where the materials have been deposited by non-Indigenous researchers (often linguists) and curated around Western norms. For example, language materials may be housed separately from ethnographic materials that from the community of origin's perspective should not be separated from language (but see O'Neal 2015 for a summary of efforts to decolonize archives; see also Linn 2014; Shepard 2016; and Wasson, Holton, & Roth 2016 regarding efforts specific to language archives). Archives and data repositories must recognize and respond to the diverse histories and agendas surrounding legacy data. As noted by Christen

(2011:209), "there is neither a singular call, nor a one-size-fits-all answer to the archival questions indigenous peoples bring to bear on the institutions that hold much of their cultural heritage."

We concur with Christen and others who emphasize the inappropriateness of a one-size-fits-all approach, but also observe the following common themes for language work based on legacy data. First is that the stakes tend to be very high when archival legacy data play a crucial role in language reclamation—and some reclamation efforts begin entirely with archival materials (cf. Spence 2018; Lukaniec, chapter 25, this volume). Second is that the original collection and curation of these data are removed in time and social context from these contemporary language reclamation practices, which are increasingly intertwined with broader decolonial efforts by Indigenous communities. As the ethics of legacy data may be different from the ethics of data being actively created, people working with legacy materials must take special care to ensure that ethical concerns are adequately addressed. Failure to do so can result in unintended consequences such as over- or undersharing of materials: Sensitive materials may (and often do) end up in the public domain, or communities may be barred from accessing materials recorded by ancestral community members. Even where consent was obtained, legacy research protocols may be out of step with modern practices and hence warrant reexamination. For example, some legacy materials may be openly available to researchers without legal restrictions; however, if those materials were originally gathered without explicit, documented consent, then the source community may desire to have a voice in determining access conditions. This may even be the case in situations where consent was given but where cultural norms and expectations have shifted since the time of original consent.

2 Indigenous communities, languages, and ethics in linguistics

The mismatch between the contexts of original creation and contemporary use in the case of legacy data aligns with the evolution of ethical practices in linguistics over the past half century, which we summarize in some detail in this section in recognition of how disciplinary histories and norms inform ideas about ethics and vice versa. Particularly important for the current discussion is linguistic fieldwork, which is the context wherein Indigenous language data are often collected. The traditional "ethical" fieldwork model, dubbed the "linguist-focused" model by Czaykowska-Higgins (2009:22), is concerned with minimizing ill effects to the speakers and signers—often stylized as "informants"—involved in the language work. This is very much in line with ethical concerns as expressed by institutional review boards and other bodies concerned with preventing harm to individual research "subjects." As Hale (2001:76) notes, "linguists are inevitably responsible to the larger human community which its [research] results could affect." But in spite of this reference to community, the traditional model reflects a view in which the individual is the natural unit of analysis for determining ethical issues (Leonard & Haynes 2010). Beyond obtaining consent from individual language users and ensuring that they are protected from physical harm, the traditional model is essentially extractive: "'Good' speakers, whose legitimacy is determined by the researcher . . . produce language that is transformed into 'data', which is conceptualised through a 'language as object' metaphor . . . that tends to emphasise structural properties at the expense of social practices" (Leonard 2017:18). In this model, which remains highly valued in the field, data are manufactured as part of the research process and then explicitly decontextualized.

The renewed focus on language documentation and conservation that has emerged over the past two decades has led to a major reexamination of ethical practices in linguistic research (cf. Rice 2006, 2010; Czaykowska-Higgins 2009; Innes & Debenport 2010; Dobrin & Berson 2011). Rice (2006) observes a transition toward a more empowering and participatory model of linguistic research. The most notable outcome of these discussions has been the emergence of a more community-based notion of ethics that emphasizes the responsibilities of researchers not just to individuals but also to communities. This change is reflected in the *Ethics Statement* adopted by the Linguistic Society of America (2009), which makes explicit reference to community:

> While acknowledging that what constitutes the relevant community is a complex issue, we urge linguists to consider how their research affects not only individual research participants, but also the wider community. In general, linguists should strive to determine what will be constructive for all those involved in a research encounter, taking into account the community's cultural norms and values.[4]

What counts as "constructive" in relation to the Linguistic Society of America *Ethics Statement* may vary from one community to another, but it is likely to include some form of what has come to be known as community-based research, that is, research not only *for*, *on*, and *with* language communities, but also *by* communities (Czaykowska-Higgins 2009:24). This typically involves a training component that develops research capacity within a community (Genetti & Siemens 2013).

At the same time, the trend toward more community-based models of language research has been accompanied by the emergence of two countervailing trends. First, what might be termed the endangered languages movement (cf. Krauss 1992)—which in many ways has spurred on this new discussion of ethics—has at the same time led to objectification and commodification of languages and their users (Hill 2002; Dobrin 2008; Dobrin, Austin, & Nathan 2009; Whaley 2011). This reductionist view envisions a kind of triage in which language research is prioritized based on typological characteristics and vitality assessments. Languages with rare or unusual sound systems or grammatical structures are viewed as important objects of study. Funding applications are justified based on perceived threats to the language, and by extension draw attention to linguistic data as products that must be collected and archived. Language communities are reduced to numbers of users and ranked according to position on scales of language vitality.

Second, the open data movement has led to a more empirical approach to linguistics, in which claims about language are grounded in data, and research results are expected to be both verifiable and reproducible (Berez-Kroeker, Gawne, et al. 2018). This trend has placed an emphasis on long-term archiving of linguistic data and led to the development of best-practice standards and archiving mandates from both funding agencies and academic institutions (Henke & Berez-Kroeker 2016). Moreover, there is also an increasing expectation that these archives be publicly accessible with few restrictions (Seyfeddinipur et al. 2019). These developments have resulted in an exponential increase in both the number of dedicated language archives and the volume of archival material. However, these repositories are most often located outside the control of the communities from which the archival deposits have been extracted (Shepard 2014), thus running counter to the spirit of community-based language work unless special provisions are made.

Underlying both the endangered languages and the open data movements is the notion of language as an object of study. This notion is in some ways fundamental to the traditional conceptualization of linguistic science, in which "language data are extracted from context of usage, and linguistic experiments are replicable" (Grenoble & Whitecloud 2014:344). Yet, this view is in direct conflict with the idea that these data never exist in isolation. Resolving this tension is central to linguistic ethics.

Countering the objectified view of language research is an emerging collaborative model that engages language communities as full partners and thus helps to highlight and validate Indigenous perspectives on language. For example, building on the notion of community-based language research, Leonard and Haynes (2010) propose a process of *collaborative consultation*, which involves continuous reflection and sharing throughout the research process. Referencing two North American Indigenous communities, Leonard and Haynes illustrate the collaborative consultation approach with the notion of speakerhood. They recognize that within any speech community, language research is inextricably tied with the notion of speakerhood; however, this issue is particularly challenging within endangered language communities, where knowledge of language is by definition in decline. A linguist-focused view of ethics might treat speakerhood, similar to the notion of what counts as data, as an objective "fact" that can be measured and assessed without community input. In contrast, a collaborative consultation model frames such issues not as unilateral decisions but rather as negotiable determinations developed through consultation among the stakeholders in a research project. In other words, this model assumes that language communities are not mere sources of data but instead become research partners from the outset (Leonard & Haynes 2010; Rice 2018). By extension, language communities' ethical norms and concerns guide the research development and implementation at all stages.

We observe that the community-based and collaborative approaches discussed are not only emerging as best practices in linguistic research, but are also increasingly a requirement for language researchers. For example, the Council of Athabascan Tribal Governments' Indigenous Knowledge Policy explicitly requires that researchers engage in collaborative research methodologies (Council of Athabascan Tribal Governments 2018). However, while such collaborative approaches

are becoming increasingly common in the academy, they remain heavily marked in the sense that they have to be explained and justified and are rarely the default in academic ethical protocols. For example, university ethical oversight structures continue to largely focus on ethics with respect to protecting *individual* research participants, thus easily overlooking or deemphasizing concerns that occur at the community level. This is insufficient when language ownership occurs at the community level and becomes completely deficient when a proposed research project does not technically meet the criteria for oversight by ethical review boards and thus gets no such review. The latter issue is of special concern for projects that involve legacy data that are legally deemed to already be within the public domain.

The response to such situations often involves incorporating Indigenous knowledges and protocols into the existing (largely Western) models, for example, by adding in some sort of required consultation with community leaders about a research project. We recognize beneficial outcomes to such efforts but suggest that this approach is inadequate because it largely maintains the power structures and research models that have facilitated exploitation of Indigenous peoples and languages. Language researchers are increasingly thinking about future uses of materials, but what about the deeper question of how current power relationships will change? Ethical approaches to linguistic data must consider not only the ethical uses of those data but also the ethics of the relationships underlying the data. Thus, we shift the question to one of how data protocols can begin from Indigenous knowledges and protocols, structurally embedding ethical community-centered concerns into all aspects of data collection, management, and use.

3 Indigenous research methods and data sovereignty

To address this question, we highlight important themes that emerge from Indigenous research methods, which collectively privilege Indigenous knowledge systems and protocols, and often critique the assumptions and ethics of dominant research practices (e.g., Wilson 2008; Kovach 2009; Chilisa 2011; Smith 2012; Lambert 2014; Tuck & Yang 2014; Snow et al. 2016). As noted by the Intercontinental American Indigenous Research Association, central to Indigenous research methods is the notion that knowledge is produced through relationships—"with people in a specific Place, with the culture of Place as understood through [specific Indigenous] cultures, with the source of the research data, and with the person who knows or tells the story that provides information."[5] By extension, and in strong contrast to the language-as-object approach, linguistic data become meaningful and interpretable through awareness of the social contexts in which they are produced and of the people who produce (and reproduce) them.[6] *People* in this case goes beyond the individuals who originally produced the data, such as individual speakers and signers, to also consider others such as Elders and other community leaders, the researcher's professional networks, and so on—stakeholders whose relationships with each other inform the context. Anchored in the strong focus on relationships and the associated accountability, it is common in discussions of Indigenous research methods to recognize several "*R*-words" that should guide research—and by extension, inform data ethics. Beyond *relationship*, we expand on the following four *R*-words outlined by Snow et al. (2016:360): *responsibility*, *respect*, *reciprocity*, and (conceived of as a single principle) *rights and regulations*.[7] *Responsibility* goes beyond accountability to individual research participants to include communities and their ways of knowing, with an eye toward the sociopolitical contexts in which research occurs and to the power structures that it reflects and affects. *Respect* to communities and to their knowledge systems entails centering both in the collection, management, and use of linguistic data. *Reciprocity* includes what many linguists working in community research contexts describe as "giving back" (e.g., by creating language pedagogical materials), but also entails reciprocal relations with respect to the construction of knowledge in areas such as data interpretation. The principle of *rights and regulations*, which attends to the protocols of participation and ownership that ensue from Indigenous self-determination, is central to the Indigenous Data Sovereignty (IDS) movement, which advocates for the direct involvement of Indigenous stakeholders in the collection, management, and use of data about Indigenous peoples. While the term data sovereignty is used primarily within Indigenous contexts, the movement is part of a broader dialogue and set of policies/laws within broader society that are focused on maintaining control over data about individuals and organizations. IDS includes the recognition of the right of Indigenous peoples and nations to govern collection,

ownership, and application of their own data, which are widely recognized as cultural, strategic, and economic assets.[8] Moreover, recent developments in the IDS movement are calling for the observance of core protocol and practices when working with Indigenous peoples and knowledge. These include, but are not limited to:

- recognition that tribes must exercise sovereignty when conducting research and managing data;
- following cultural protocols;
- being flexible;
- extending hospitality;
- ensuring appropriate compensation for expertise;
- understanding that access to knowledge is not a universal right;
- recognizing that responsible stewardship includes the task of learning how to interpret and understand data and research;
- accepting that research must benefit Native people.

(School for Advanced Research 2018; National Congress of American Indians 2018, as cited in Carroll, Rodriguez-Lonebear, & Martinez 2019)

In practice, IDS is being realized at national and regional scales through the development of principles and practices by Indigenous peoples, their representative organizations, and non-Indigenous stakeholders. For example, the First Nations Information Governance Centre (2014) in Canada developed and asserted the "OCAP Principles," which highlight both the relationships (Ownership, Control, Access) between Indigenous peoples and the data and information created by or about them, and the more concrete aspect of physical possession of data and information (Possession). The US Indigenous Data Sovereignty Network has proposed guidelines to facilitate harnessing Indigenous ways of knowing and doing and applying them to the "management and control of a Native nation's data ecosystem" (Rainie, Rodriguez-Lonebear, & Martinez 2017). Inuit Tapiriit Kanatami, the National Inuit Organization in Canada, established the National Inuit Strategy on Research, which has a priority area focused on ensuring Inuit access, ownership, and control over data and information. Although linguistic data are not explicitly addressed, the scope of National Inuit Strategy on Research spans all research activities involving or about Inuit, and an associated implementation plan (Inuit Tapiriit Kanatami 2018) seeks to eliminate exploitative and colonial approaches to research, as noted by Inuit Tapiriit Kanatami President Natan Obed:

> For far too long, researchers have enjoyed great privilege as they have passed through our communities and homeland, using public or academic funding to answer their own questions about our environment, wildlife, and people. Many of these same researchers then ignore Inuit in creating the outcomes of their work for the advancement of their careers, their research institutions, or their governments. This type of exploitative relationship must end. (3)

Additionally, the Inuit Circumpolar Council-Alaska, through its Alaskan Inuit Food Security Framework aims to establish a model where Indigenous knowledge is considered as part of environmental management and all other relevant activities from the outset. These integrative approaches establish Indigenous knowledge as an essential part of the research process (Inuit Circumpolar Council-Alaska 2015). Ultimately, this involves respect—not only for the data but also for the underlying knowledge systems. To ensure that community needs and knowledge frame all stages of such research and its applications, partnerships and relationships are at the heart of these and other Indigenous research protocols. For example, this is made explicit in the University of Hawai'i *Kūlana Noi'i* (research standards): "The Kūlana Noi'i provide guidance for building and sustaining not just working partnerships but long-term relationships between communities and researchers" (University of Hawai'i Sea Grant 2019). Similarly, in reference to research involving Pacific peoples, the University of Otago's *Pacific Research Protocols* specifically address the issue of balance with research relationships and partnerships. Moreover, the protocols recognize the need to acknowledge the appropriate function of shared knowledge and that the ownership of primary data lies with the people who contribute that knowledge (University of Otago 2011:14).

On an international scale, the International Indigenous Data Sovereignty Interest Group of the Research Data Alliance and the Global Indigenous Data Alliance put forward the CARE Principles for Indigenous Data Governance: **C**ollective benefit, **A**uthority to control, **R**esponsibility, and **E**thics.[9] These CARE principles, which are focused on people and purpose, intersect with the more data-oriented, broadly cited FAIR principles (**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable).[10] While the

CARE principles are arguably the most prominent of the emerging general protocols, a similar focus is found in a number of community-based protocols, such as the San Code of Research Ethics.[11] Centering principles, aspirations, and goals related to IDS, these examples provide a sound foundation for guiding and informing data-related activities, including reflexive and principle-oriented linguistic research and data collection. These laws and policies are partly a response to perceived and identified privacy issues and breaches on major social media platforms and institutional infrastructures (e.g., banks, credit agencies, and insurance companies). Thus, there is a broader societal concern about ethical data management and use that is translating into new social structures (e.g., norms and laws) that have normative and legal implications for linguists. In this way, linguistic data sovereignty is just one part of a larger approach in which Indigenous communities control not just access to linguistic data, but also the production, interpretation, and dissemination of those data.

4 Intersections of linguistic research with the open data movement

As previously indicated in this chapter and elsewhere in this volume, citable, open data is currently the dominant movement in the domain of research data management (Kitchin 2014:49; Nosek et al. 2015; Collister, chapter 9, this volume).[12] As a result, researchers are now much more likely to deposit linguistic data in archives, ensuring that these data are accessible both to the source communities and the broader public. However, if applied universally, open data principles can contradict IDS principles (e.g., understanding that access to knowledge is not a universal right; cf. Rainie et al. 2019). More nuanced, recent statements and declarations on open data management are identifying the need to include exceptions to fully open data. For example, the International Arctic Science Committee Statement of Principles and Practices for Arctic Data Management (International Arctic Science Committee 2013) uses the term "ethically open access" that identifies the following exceptions to this requirement of full, free, and open access:

- where human subjects are involved, confidentiality shall be protected as appropriate and guided by the principles of informed consent;
- where local and traditional knowledge is concerned, rights of the knowledge holders shall not be compromised;
- where data release may cause harm, specific aspects of the data may need to be kept protected (for example, locations of sensitive sites).

These exceptions are well recognized in current language documentation practices. For example, funding agencies do not require that sensitive or otherwise restricted documentation be archived (Seyfeddinipur et al. 2019). However, as linguists increasingly use cloud-based platforms, networked "apps," machine learning and artificial intelligence technologies, and other new tools (cf. Galla 2016), we further caution that there is a need to be aware of the potential ethical implications of using these tools. For example, if linguists use popular platforms such as YouTube or Google Drive in a research workflow that includes linguistic data, they may assign certain rights to the platform provider, whether intended or not. For example, the Google Terms of Service grant Google license to "publish, publicly perform, or publicly display your content, if you've made it visible to others" and to "modify and create derivative works based on your content."[13] And the Zoom video conferencing Terms of Service grant the company the right to store recordings of meetings on its servers.[14] By accepting such terms of service, a linguist may unintentionally contravene (Indigenous) data sovereignty and general ethical principles and policies by licensing content to a third party. The act of providing (uploading, storing) content (e.g., recordings, transcripts) on such a platform without prior, informed consent from the outset may result in an unwitting ethical breach. Thus, while linguistics moves toward open data culture and practice, the related ethical nuances and caveats must always be considered. Establishing protocols for proper consent related to data collection and use, and for data management of Indigenous languages is critically important and urgent. There is a need for research and data management planning to be driven by Indigenous peoples, communities, families, and organizations. There is a need for infrastructure and resources so this can be realized.

5 Conclusion: Linguistic data cannot be divorced from their sources

Ethics are often framed as problems to be worked around. However, that data are intimately tied to their sources

need not be seen as a hindrance to overcome, but instead can be a positive step, providing context necessary for interpreting those data. Anchoring data analysis in the relationships and contexts from which the data come can lead to outcomes that would otherwise be missed (cf. Mithun 2001; Rice 2001). Moreover, explicitly acknowledging the links between data and their sources facilitates both reproducible research and applied use.[15] This is true for archival documentation just as much as it is for new data collection: the current ethical context of the use of legacy materials may differ from the ethical context in which the materials were created, but the ties between data and sources remain nonetheless. In essence, the intersection of people, ethics, and data is about relationships—past, present, and future. Reflecting on ethics in language research, Czaykowska-Higgins (2018) characterizes this emphasis on relationships as "rehumanizing linguistics, acknowledging the centrality of relationships and difference in language documentation work, emphasizing accountability to those relationships, and grounding ethical research methodologies in social relations" (116). The actions that language researchers take now in terms of data management decisions will guide how the relationships will evolve.

To return to the principle from which we started this chapter: linguistic data cannot be divorced from their sources. While ethical practices may differ across different research contexts, this principle remains and lies at the heart of ethics in linguistic data. As the various chapters in this volume clearly attest, the field of linguistics is evolving quickly into a more data-driven science. There is ample evidence that recognizing—and indeed celebrating—the relationships between people, ethics, and data will ensure a more robust science of human language in which linguistic communities and associated knowledge systems play a more symmetrical role.

Notes

1. This chapter focuses on ethical issues in relation to Indigenous language data. The discussion of ethical issues in relation to world languages has taken place primarily within the subfields of applied linguistics and sociolinguistics. See Eckert (2013), De Costa (2015), and the various references therein.

2. While we will use the term Indigenous language throughout this chapter, it should be understood that many of the ethical concerns surrounding Indigenous peoples and languages apply equally to other minority language groups such as Cajun French, which are not typically referred to with the label Indigenous.

3. That said, Hinton (2002:151) questions whether a metaphor of ownership is actually evoked by the so-called possessive construction in English.

4. The emphasis on community is reiterated even more strongly in the 2019 Ethics Statement (Linguistic Society of America 2019).

5. https://www.americanindigenousresearchassociation.org/about-us/. Accessed May 11, 2020.

6. While Indigenous research methods are being increasingly adopted outside Indigenous contexts, these methods are more often construed as applying to the data collection process rather than the data themselves. Data are still typically seen as independent, divorced from their sources. As we argue in this chapter, this view is fundamentally flawed in the context of Indigenous and minority language research.

7. Other *R*-words include *relevance* to the community as a necessary goal for research, *reverence*, *reflexivity* as a necessary practice by researchers, and *relationality* (the concept that relationships and their associated complex interdependencies form a foundation to everything).

8. This is true not only in the realm of linguistic data but more broadly as well, as evidenced by recent trends toward the monetization of data, and resulting policy changes that grant rights of control over personal data (cf. Marelli & Testa 2018; Choi, Jeon, & Kim 2019). IDS itself, while increasingly well defined in its own specific context, similarly exists within a much broader context that includes discourse around open data, privacy, and the emerging ecosystem of platforms and methods (e.g., machine learning). Both IDS and the broader context have implications for each other—for example, the broader context of privacy control and intellectual property law has implications for IDS; similarly, IDS is having an impact on broader research data dialogues.

9. https://www.gida-global.org/care.

10. https://www.force11.org/group/fairgroup/fairprinciples.

11. https://www.globalcodeofconduct.org/affiliated-codes/.

12. For information about the broader open science movement, see the Organisation for Economic Co-operation and Development's overview: https://www.oecd.org/science/inno/open-science.htm; for an example specific to linguistic data, see https://linguistics.okfn.org.

13. https://policies.google.com/terms. Accessed May 11, 2020.

14. http://zoom.us/terms. Accessed May 17, 2020.

15. Data citation practices that explicitly acknowledge speakers and other contributors can help to maintain this connection between data and source (see Andreassen et al. 2019).

References

Andreassen, Helene N., Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell, and the Research Data Alliance Linguistic Data Interest Group. 2019. Tromsø recommendations for citation of research data in linguistics (Version 1). *Research Data Alliance*. https://doi.org/10.15497/rda00040.

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

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* 57 (1): 1–18. https://doi.org/10.1515/ling-2017-0032.

Bird, Steven, and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79 (3): 557–582.

Carroll, Stephanie Russo, Desi Rodriguez-Lonebear, and Andrew Martinez. 2019. Indigenous data governance: Strategies from United States Native Nations. *Data Science Journal* 18 (1): 31. https://doi.org/10.5334/dsj-2019-031.

Chilisa, Bagele. 2011. *Indigenous Research Methodologies*. London: SAGE Publications.

Choi, Jay Pil, Doh-Shin Jeon, and Byung-Cheol Kim. 2019. Privacy and personal data collection with information externalities. *Journal of Public Economics* 173:113–124. https://doi.org/10.1016/j.jpubeco.2019.02.001.

Christen, Kimberly. 2011. Opening archives: Respectful repatriation. *American Archivist* 74:185–210.

Council of Athabascan Tribal Governments. 2018. *Researching Gwich'in/Upper Koyukon Indigenous Knowledge in the CATG Region*. Fort Yukon, AK: Council of Athabascan Tribal Governments.

Czaykowska-Higgins, Ewa. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities. *Language Documentation and Conservation* 3 (1): 15–50. http://hdl.handle.net/10125/4423.

Czaykowska-Higgins, Ewa. 2018. Reflections on ethics: Rehumanizing linguistics, building relationships across difference. In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 110–121. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24813.

De Costa, Peter I., ed. 2015. *Ethics in Applied Linguistics Research: Language Researcher Narratives*. New York: Routledge.

Dobrin, Lise M. 2008. From linguistic elicitation to eliciting the linguist: Lessons in community empowerment from Melanesia. *Language* 84 (2): 300–324.

Dobrin, Lise M., Peter Austin, and David Nathan. 2009. Dying to be counted: The commodification of endangered languages in documentary linguistics. In *Language Documentation and Description*, vol. 6, ed. Peter K. Austin, 37–52. London: SOAS.

Dobrin, Lise M., and Josh Berson. 2011. Speakers and language documentation. In *The Cambridge Handbook of Endangered Languages*, ed. Peter K. Austin and Julia Sallabank, 187–211. Cambridge: Cambridge University Press.

Eckert, Penelope. 2013. Ethics in linguistics research. In *Research Methods in Linguistics*, ed. Robert J. Podesva and Devyani Sharma, 11–26. Cambridge: Cambridge University Press.

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

Fitzgerald, Colleen M. 2017. Understanding language vitality and reclamation as resilience: A framework for language endangerment and "loss" (Commentary on Mufwene). *Language* 93 (4): e280–e297.

Galla, Candace Kaleimamoowahinekapu. 2016. Indigenous language revitalization, promotion, and education: Function of digital technology. *Computer Assisted Language Learning* 29 (7): 1137–1151. https://doi.org/10.1080/09588221.2016.1166137.

Genetti, Carol, and Rebekka Siemens. 2013. Training as empowering social action: An ethical response to language endangerment. In *Responses to Language Endangerment: In Honor of Mickey Noonan*, ed. Elena Mihas, Bernard Perley, Gabriel Rei-Doval, and Kathleen Wheatley, 59–77. Amsterdam: John Benjamins.

Grenoble, Lenore A. 2011. Language ecology and endangerment. In *The Cambridge Handbook of Endangered Languages*, ed. Peter K. Austin and Julia Sallabank, 27–44. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511975981.002.

Grenoble, Lenore A., and Simone S. Whitecloud. 2014. Conflicting goals, ideologies, and beliefs in the field. In *Endangered Languages: Beliefs and Ideologies in Language Documentation and Revitalization*, ed. Peter K. Austin and Julia Sallabank, 337–354. London: British Academy. https://doi.org/10.5871/bacad/9780197265765.003.0016.

Guerrettaz, Anne Marie. 2015. Ownership of language in Yucatec Maya revitalization pedagogy. *Anthropology and Education Quarterly* 46 (2): 167–185. https://doi.org/10.1111/aeq.12097.

Hale, Ken. 2001. Ulwa (Southern Sumu): The beginnings of a language research project. In *Linguistic Fieldwork*, ed. Paul

Newman and Martha Ratliff, 76–101. Cambridge: Cambridge University Press.

Henke, Ryan, and Andrea L. Berez-Kroeker. 2016. A brief history of archiving in language documentation, with an annotated bibliography. *Language Documentation and Conservation* 10:411–457. http://hdl.handle.net/10125/24714.

Hill, Jane H. 2002. "Expert rhetorics" in advocacy for endangered languages: Who is listening, and what do they hear? *Journal of Linguistic Anthropology* 12 (2): 119–133. https://doi.org/10.1525/jlin.2002.12.2.119.

Himmelmann, Nikolaus P. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation and Conservation* 6:187–207. http://hdl.handle.net/10125/4503.

Hinton, Leanne. 2002. Commentary: Internal and external language advocacy. *Journal of Linguistic Anthropology* 12 (2): 150–156. https://doi.org/10.1525/jlin.2002.12.2.150.

Innes, Pamela, and Erin Debenport, eds. 2010. *Ethical Dimensions of Language Documentation*. Special issue, *Language and Communication* 30 (3).

International Arctic Science Committee. 2013. IASC Data Statement. https://iasc.info/data-observations/iasc-data-statement.

Inuit Circumpolar Council-Alaska. 2015. *Alaskan Inuit Food Security Conceptual Framework: How to Assess the Arctic from an Inuit Perspective*. Anchorage: Inuit Circumpolar Council-Alaska.

Inuit Tapiriit Kanatami. 2018. *National Inuit Strategy on Research*. https://www.itk.ca/wp-content/uploads/2018/04/ITK_NISR-Report_English_low_res.pdf.

Jacob, Michelle M. 2013. *Yakama Rising: Indigenous Cultural Revitalization, Activism, and Healing*. Tucson: University of Arizona Press.

Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE Publications.

Kornai, András. 2013. Digital language death. *PLoS ONE* 8 (10): e77056. https://doi.org/10.1371/journal.pone.0077056.

Kovach, Margaret. 2009. *Indigenous Methodologies: Characteristics, Conversations, and Contexts*. Toronto: University of Toronto Press.

Krauss, Michael E. 1992. The world's languages in crisis. *Language* 68 (1): 4–10.

Kukutai, Tahu, and John Taylor. 2016. *Indigenous Data Sovereignty: Toward an Agenda*. Canberra: Australian National University Press.

Lambert, Lori. 2014. *Research for Indigenous Survival: Indigenous Research Methodologies in the Behavioral Sciences*. Lincoln: University of Nebraska Press.

Leonard, Wesley Y. 2017. Producing language reclamation by decolonising "language." *Language Documentation and Description* 14:15–36.

Leonard, Wesley Y. 2018. Reflections on (de)colonialism in language documentation. In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 55–65. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24808.

Leonard, Wesley Y., and Erin Haynes. 2010. Making "collaboration" collaborative: An examination of perspectives that frame linguistic field research. *Language Documentation and Conservation* 4:269–293. http://hdl.handle.net/10125/4482.

Linguistic Society of America. 2009. *Ethics Statement*. https://www.linguisticsociety.org/sites/default/files/Ethics_Statement.pdf. Accessed May 17, 2020.

Linguistic Society of America. 2019. *LSA Revised Ethics Statement*, final version (approved July 2019). https://www.linguisticsociety.org/content/lsa-revised-ethics-statement-approved-july-2019. Accessed May 17, 2020.

Linn, Mary S. 2014. Living archives: A community-based language archive model. *Language Documentation and Description* 12:53–67.

Marelli, Luca, and Giuseppe Testa. 2018. Scrutinizing the EU General Data Protection Regulation. *Science* 360 (6388): 496–498. https://doi.org/10.1126/science.aar5419.

Meek, Barbra A. 2010. *We Are Our Language: An Ethnography of Language Revitalization in a Northern Athabaskan Community*. Tucson: University of Arizona Press.

Mithun, Marianne. 2001. Who shapes the record: The speaker and the linguist. In *Linguistic Fieldwork*, ed. Paul Newman and Martha Ratliff, 34–54. Cambridge: Cambridge University Press.

National Congress of American Indians. 2018. Resolution KAN-18–011: Support of US Indigenous Data Sovereignty and Inclusion of Tribes in the Development of Tribal Data Governance Principles. June 4, 2018. http://www.ncai.org/resources/resolutions/support-of-us-indigenous-data-sovereignty-and-inclusion-of-tribes-in-the-development-of-tribal-data.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, et al. 2015. Promoting an open research culture. *Science* 348 (6242): 1422–1425. https://doi.org/10.1126/science.aab2374.

O'Meara, Carolyn, and Jeff Good. 2010. Ethical issues in legacy language resources. *Language and Communication* 30 (3): 162–170. https://doi.org/10.1016/j.langcom.2009.11.008.

O'Neal, Jennifer R. 2015. "The right to know": Decolonizing Native American archives. *Journal of Western Archives* 6 (1): 1–17.

Rainie, Stephanie Carroll, Tahu Kukutai, Maggie Walter, Oscar Luis Figueroa-Rodríguez, Jennifer Walker, and Per Axelsson.

2019. Issues in open data: Indigenous data sovereignty. In *State of Open Data*, ed. Tim Davies, Stephen B. Walker, Mor Rubinstein, and Fernando Perini, 300–319. Cape Town: African Minds. https://doi.org/10.5281/zenodo.2677801.

Rainie, Stephanie Carroll, Desi Rodriguez-Lonebear, and Andrew Martinez. 2017. *Policy Brief: Indigenous Data Sovereignty in the United States*. Tucson: Native Nations Institute, University of Arizona.

Rice, Keren. 2001. Learning as one goes. In *Linguistic Fieldwork*, ed. Paul Newman and Martha Ratliff, 230–249. Cambridge: Cambridge University Press.

Rice, Keren. 2006. Ethical issues in linguistic fieldwork: An overview. *Journal of Academic Ethics* 4 (1): 123–155. https://doi.org/10.1007/s10805-006-9016-2.

Rice, Keren. 2010. The linguist's responsibilities to the community of speakers: Community-based research. In *Language Documentation: Practice and Values*, ed. Lenore A. Grenoble and N. Louanna Furbee, 25–36. Amsterdam: John Benjamins.

Rice, Keren. 2018. Collaborative research: Visions and realities. In *Insights from Practices in Community-based Research: From Theory to Practice around the Globe*, ed. Shannon T. Bischoff and Carmen Jany, 13–37. Berlin: Mouton de Gruyter.

School for Advanced Research. 2018. Community+Museum Guidelines for Collaboration. https://sarweb.org/guidelinesforcollaboration/.

Seyfeddinipur, Mandana, Felix Ameka, Lissant Bolton, Jonathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, et al. 2019. Public access to research data in language documentation: Challenges and possible strategies. *Language Documentation and Conservation* 13:545–563. http://hdl.handle.net/10125/24901.

Shaw, Patricia A. 2001. Language and identity, language and the land. *BC Studies: The British Columbian Quarterly* 131:39–55.

Shepard, Michael Andrew Alvarez. 2014. "The substance of self-determination": Language, culture, archives and sovereignty. PhD dissertation, University of British Columbia.

Shepard, Michael Alvarez. 2016. The value-added language archive: Increasing cultural compatibility for Native American communities. *Language Documentation and Conservation* 10:458–479. http://hdl.handle.net/10125/24715.

Smith, Linda Tuhiwai. 2012. *Decolonizing Methodologies: Research and Indigenous Peoples,* 2nd ed. London: Zed Books.

Snow, Kevin C., Danica G. Hays, Guia Caliwagan, David J. Ford Jr., Davide Mariotti, Joy Maweu Mwendwa, and Wendy E. Scott. 2016. Guiding principles for indigenous research practices. *Action Research* 14 (4): 357–375. https://doi.org/10.1177/1476750315622542.

Spence, Justin. 2018. Learning languages through archives. In *The Routledge Handbook of Language Revitalization*, ed. Leanne Hinton, Leena Huss, and Gerald Roche, 179–187. New York: Routledge.

Tuck, Eve, and K. Wayne Yang. 2014. R-words: Refusing research. In *Humanizing Research: Decolonizing Qualitative Inquiry with Youth and Communities*, ed. Django Paris and Maisha T. Winn, 223–247. Los Angeles: SAGE Publications.

University of Hawai'i Sea Grant. 2019. *Kūlana Noi'i*. Honolulu: University of Hawai'i Sea Grant College Program. http://seagrant.soest.hawaii.edu/kulana-noii/.

University of Otago. 2011. *Pacific Research Protocols*. Otago, New Zealand: University of Otago. https://www.otago.ac.nz/research/otago085503.pdf.

Wasson, Christina, Gary Holton, and Heather S. Roth. 2016. Bringing user-centered design to the field of language archives. *Language Documentation and Conservation* 10:641–681. http://hdl.handle.net/10125/24721.

Whaley, Lindsay J. 2011. Some ways to endanger an endangered language project. *Language and Education* 25 (4): 339–348. https://doi.org/10.1080/09500782.2011.577221.

Wilson, Shawn. 2008. *Research Is Ceremony: Indigenous Research Methods*. Black Point, Nova Scotia, Canada: Fernwood Publishing.

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